LAW AND ECONOMICS SEMINAR **Professor Polinsky**
Spring Term 2007

Thursday, March 1, 2007
4:00 - 5:30 p.m.
Stanford Law School
Room 280A

"Can You Believe Econometric Evaluations of Law, Policy, and Medicine?"

by

John J. Donohue

Yale Law School

<u>Note to Seminar Participants:</u> My apologies for the length of the paper, which is essentially a cut and paste job from a book I am writing, tentatively entitled, "Beautiful Models, and Other Threats to Life, Law and Policy." Readers who are not yet ready to make such an enduring commitment to reading my work will find that they will get the basic gist of the paper by reading the first two section (pages 1-15) and the Conclusion (pages 59-61). A quick look at Table 3 (on page 32) makes the point that minor tweaks in a model estimating the deterrent impact of the death penalty can lead to a wide range of results, from each execution saving many lives to each execution costing many lives (just cast your eyes across the estimates in the table to see the enormous range that can be generated almost without effort). Similarly, a glance at Table 9 (on page 53) can visually convey the message that different models lead to very different estimates of the impact of right to carry laws on crime -- the lines going up suggest that crime rises when such laws go into effect, and the lines going down suggest estimates of declining crime. Even more time restricted readers may find the abstract sufficient.

Abstract

The explosion in applied econometric work across all domains of business, the academy, law, policy, and medicine is one of the most remarkable intellectual achievements of the last thirty years. But the lowering of the barriers to entry into econometric analysis has raised issues both of quality control and bias as a mountain of studies pour forth, offering remarkable, but often wildly conflicting claims. The paper tries to convey a sense (for those who don't do this stuff) of how difficult it is to determine causal impacts of laws/policies/medical treatments, and why things often seem to go astray owing to the various problems of ideology, self-interest, bad data, the limits of the econometric models, and the flux in best practice as the science improves. As a result, some have invested too much faith in these studies, particularly when they support a preferred position. Decisionmakers, legislators, judges, and others should realize that there are so many ways that a regression study that is trying to do more than merely summarize data or find correlations without regard to establishing causation can go wrong that it is often more remarkable when one is right than when one proves inadequate. A host of studies evaluating the deterrent effect of the death penalty and the impact of right-to-carry concealed weapons laws is offered to illustrate the point that minor tweaking of standard econometric models can generate such a wide range of conflicting answers that supporters of any position can find supportive econometric evidence. Changes in the academy, the courts, and the process of peer review, as well as advances in the science of econometrics, are needed to improve the quality of decisions and realize the enormous potential of the continuing revolution in statistical theory and practice.

Can You Believe Econometric Evaluations of Law, Policy, and Medicine?

John J. Donohue
Yale Law School
February 22, 2007

I. Introduction

Business executives, policymakers, judges, doctors, academics, and, indeed, anyone tuned in to the print or electronic media are all aware of the explosive growth in the number of empirical studies that purport to answer every imaginable question. What are the most desirable child-rearing tactics, the most profitable investment and business strategies, the best ingredients to include in food to promote consumer health or business sales, the best ways to select applicants to colleges or universities, or the best ways to end a movie? Researchers using sophisticated statistical techniques have tried to provide the answer. How will stents or angioplasty impact the survival of heart attack patients, or vitamins influence health, or alcohol consumption, exercise or aspirin affect longevity or the incidence of Alzheimer's? The exploding field of evidence-based medicine is trying to find out. Will gun control laws or the death penalty or three-strikes laws or harsher degrees of incarceration influence crime?[2] Will damage caps affect the rate of medical malpractice,[3] does no-fault auto insurance increase car accidents, or do antidiscrimination laws affect the earnings of blacks or women? Do tradable emissions rights affect the quality of air, do laws permitting unilateral divorce harm children or reduce domestic

[1] My thanks for outstanding research assistance to Maile Tavepholjalern, Sascha Becker, and Tatiana Neumann.

[2] Thomas Marvel and Carlisle Moody, "The Lethal Effects of Three Strikes Laws," 30 Journal of Legal Studies 89 (2001)(concluding that these laws increase homicides by 10-29 percent, but have no deterrent or incapacitative benefit).

[3] As of June, 2002, "Georgia was identified by the American Medical Association as one of twelve states where medical liability insurance costs had increased to a level that was expected to result in significant numbers of physicians leaving or limiting clinical practice, retiring, or relocating to another state." The Georgia Board for Physician Workforce, *The Effect of the Medical Liability Insurance Crisis on Physician Supply and Access to Medical Care in Georgia*, Jan. 2003, p. 6, http://gbpw.georgia.gov/vgn/images/portal/cit_1210/23/14/54061804MASTER%20for%20PDF.pdf). A 2002 survey by the Georgia Board for Physician Workforce found that the largest effect of the medical liability insurance crisis on access to medical care was that physicians planned to stop providing high-risk procedures, such as delivering babies, reading mammography tests, and performing complicated surgical procedures. Moreover, more than 1,750 Georgia physicians reported that, in order to reduce their liability risk, they had or would stop providing coverage of emergency room services. Perhaps revealing a touch of gullibility (or disingenuousness), the Georgia legislature cited these findings as a basis for its decision to cap non-economic damages in medical malpractice cases. Ga. Code Ann. § 51-13-1 (2006).

violence? Does Israel's policy of assassinating Palestinian terrorists reduce terrorism?[4] The list of statistical evaluations in the realms of law, policy and medicine is seemingly endless.

The stunning successes of the new empiricism – from the identification of the problem of global warming to a richer understanding of an enormous array of aspects of our social and political world, to advances in medical treatment and business practices – are legion and constitute an incredible leap forward in the human capacity to ascertain important truths about our complex world. Indeed, one of the most important realizations in the last century in the legal, economic, and social realm – that relatively free markets could be a powerful tool for creating wealth – came not through any advance of theory but rather from the cold, hard demonstration that societies that unleashed the power of markets experienced far more economic growth and prosperity than the countries that stifled economic enterprise through governmental ownership of productive capital or over-regulation. While this demonstration was so dramatic that formal studies were not needed, the basic approach of comparing the economic performance of the United States, Western Europe, and Japan with that of Eastern Europe, coupled with the natural experiments afforded by East and West Germany, North and South Korea, Taiwan and mainland China before the introduction of market reforms, is emblematic of the type of causal analysis that has now become omnipresent in the American academy and professional schools.

This onslaught of empirical analysis has the potential to generate important ancillary social benefits beyond the many specific insights by nurturing a commitment to the concept of using empirical tools to discern causal impacts. The public's understanding of causal attribution cannot expand without commensurately dampening the widespread human tendencies to lapse into the clutches of potentially oppressive ideologies of nations, tribes, religions, and cultures. If the onslaught of econometrics serves to aid the development of a culture of truth, this will be an important contribution to social wellbeing.

One can hardly overstate the incredible benefits that would result if we could really answer all of limitless array of empirical questions with precision and confidence in a way that generated agreement – at least among the relatively small group of scholars in a position to evaluate the credibility of these studies. In the medical arena, it is likely that tens of thousands of lives could be saved in the U.S. each year, and, of course, these gains would be multiplied many times as the knowledge from better treatments and nutritional and health advice could be shared around the globe. Even greater benefits would result if precise estimates of the impact of the entire array of legal and policy decisions could be generated. If we really possessed this knowledge, the enormous set of policies that have no beneficial effect (yet impose nontrivial costs) could be eliminated, and the policies that have desirable effects could be implemented. This would revolutionize law and government -- heat would be dissipated, light would shine, and the supply of truth would explode.

---

[4] Asaf Zussman and Noah Zussman, "Assassinations: Evaluating the Effectiveness of an Israeli Counterterrorism Policy Using Stock Market Data," 20 Journal of Economic Perspectives (Spring 2006) 193-206 (concluding that it is an effective counter-terrorism strategy to kill senior Palestinian military leaders but a counter-productive strategy to kill senior political leaders).

But for all its immense promise, the empirical revolution faces major challenges. The explosion in the use of the highly complex tools of causal analysis has meant that there is substantial variance in the quality of the econometric products that are working their way into the public domain. Very smart researchers using very complex empirical approaches are spinning out flawed studies at a rapid rate, and referees and journal editors -- let alone student editors, policymakers, and the general public -- are not in a position fully to identify these flaws at a reasonable cost. Causal studies about complex social policies can be derailed in a thousand different ways, ranging from problems with the underlying data, coding problems that arise during the massive manipulations needed to organize the dataset, the research design, and the specification of the underlying econometric model, as well as the implementation of the model and the interpretation of the resulting statistical estimates. Researchers must make hundreds of potentially critical choices, each of which provides opportunities for error and manipulation (conscious or unconscious). Many of these decisions will never be recognized by journal editors or most readers, so the belief that only valid studies can satisfy the rigors of peer review is largely chimerical. Even those relatively few who are able to implement the best econometric practices will find that the evolving science leaves them with results that one day offer powerful support for a particular thesis, and the next day are recognized to be little more than the workings of chance. Moreover, even if one succeeds in securing the correct answer on some important causal question, the ability of others to raise ostensibly telling objections to the correct answer is substantial. If a good cross examiner can make an honest person appear deceptive, a talented econometrician can certainly make a sound study appear infirm. Wedding advocacy to flawed statistics creates a situation that can turn the promise of truth into an unholy mix of confusion, uncertainty, and falsehood masquerading as truth.

While these problems are endemic to causal studies across the board, in many ways the dangers are greatest in the legal academy. First, law professors have come out of a tradition of advocacy rather than from a culture based on a scientific search for truth, and the former can pose substantial difficulties in empirical research. While the new empiricism is sweeping the legal academy, as evidenced by the number of researchers, seminars, conferences, and journals, the transition to a culture of science and truth as an aspiration will only come slowly to a body of scholars who have traditionally sought to pursue desired political or legal objectives. Second, the skill set needed for successful empirical evaluation is very different from what law professors have been trained to do. Learning appropriate empirical methodologies is like learning a foreign language as an adult – it is not easy, it takes considerable time and commitment, and relatively few become highly proficient. It is no surprise then, even among Nobel economists, the most cited work from theorists comes at age 43 (on average), while for the empiricists, it comes at age 61.[5] Empirical researchers in law have the advantage that they can harness their econometric skills to answer questions that are specified more precisely in light of the underlying institutional details of law and policy, but they have an important disadvantage as well. The empirical revolution is represented by an explosion not only in studies but also by enormous flux and even contention about best econometric practices. Scholars who need to devote themselves first to substantive areas of law will have to run hard to keep up with the evolving statistical sciences.

[5] David Galenson

One of the goals of this book is to stress that there are two prevalent errors in the response to the flood of statistical studies. Some adopt a posture of uncritical acceptance, typically to studies that offer results they find appealing. Nothing is more common then to see such scholars fill in a critical gap in their proposed agendas with a statement to the effect that "the study by X has shown that Y is true." The message is that this critical issue has been put to rest. Others, some of whom may be troubled by the implicit detraction of their own skill set posed by the elevation of econometrics or who simply perceive that dubious claims are oversold on the basis of shaky statistical foundation, can push too far in the opposite direction by rejecting all such studies as inherently unreliable.

A more discerning middle ground must be reached. As Francis Galton stated over a century ago in his pioneering work *Natural Inheritance* (1889):

> **"**Some people hate the very name of statistics but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man."

The explosion in the number of empirical studies, by which I mean studies using sophisticated statistical or econometric approaches to ascertain true causal relationships, has generated many remarkable insights and intellectual advances but has not come without pain. Just as important innovations in the economic sphere create opportunities for those who are capable of exploiting them and impose burdens on the owners of outdated technologies, the rise of the new empiricism has created its own set of winners and losers. Those whose skills were more humanistic, philosophical, or rhetorical are relatively disadvantaged as the world turns to more technical modes of analysis. Those with more advanced technical statistical and econometric training find themselves in greater demand and benefit as the percentage of technical discourse grows in academic, business, legislative, medical, and regulatory fora.

But, as every gold rush attracts an array of sharp dealers and charlatans, the rise of empirical analysis has brought its share of unprincipled purveyors of econometric and statistical artifice. Many have learned that flawed studies can often confer tremendous advantages. Candidate George W. Bush buttressed his Presidential ambitions in 2000 with empirical studies designed to show how his state tort reforms promoted growth in the Texas economy. Bush's handlers felt that supportive econometric studies would assist his political aspirations, and the demand created the supply. Politicians and their advisors know that as long as the inadequacies do not become evident until at least one day after the election, a flawed study can prove as helpful as a legitimate study, and can often be obtained at far lower cost.

Rodney Paige, Bush's first Secretary of Education, rose to prominence as a result of cooked data that appeared to show his tough approach to education was generating large gains in graduation rates for minority students. In essence, he expelled a huge proportion of high-risk minority students, and then reported that after implementing his get-tough policies, the high school graduation rates for minorities soared (with the rates based on the remaining students and not including the expelled students). Similarly,

pharmaceutical companies, backed by their compliant analysts, have understood that artful suppression of bad news or creative presentation of unduly rosy drug evaluations can generate profits along the lines of tens or even hundreds of millions of dollars before the FDA process finally catches up with the facts. A small group of sophisticated researchers has begun to take advantage of their ability to sniff out bad empirical studies by selling short the stock of pharmaceutical companies that have trumpeted the benefits of a particular drug and then releasing their assessment of the flaws in the studies.[6] For products where regulatory oversight is less substantial, the capacity to deceive consumers through distortion by statistics is even greater, leading to more primitive, and more prevalent, evasions.

Regulators and even legislative staffs have responded to the growing reliance on statistical evidence by hiring legions of statistically trained personnel, and law firms have turned to litigation consulting firms to handle their need for sophisticated econometric evaluation. The laggard in this process of tooling up is, of course, the judiciary, which has largely remained unburdened by the human capital needed to evaluate this tsunami of econometric studies, except through the random chance of a technically trained law clerk. Indeed, like many members of the academy, judges find themselves in the midst of an enormous process of transition as they are increasingly asked to make sense of conflicting econometric studies offered by high-powered statistical experts while the skill set and the institutional practices that are needed to effectively achieve this goal have not yet been fully developed. The Federal Judicial Center has realized this problem and commissioned two top statisticians to write a 90-page manual for federal judges.[7] The manual is excellent, and any judge who took the time to slog through it -- yes, if you are learning statistics by reading on your own time, it is slogging -- would surely gain insight into basic statistical matters. But the manual necessarily had to be pitched at a level that is far below the heights at which statistical experts commonly duel, no doubt in part because one party realizes that judges and juries will not be able to officiate a dispute conducted at that level.

Of course, judges have the ability to order the parties to pay for a special master with statistical ability who will in essence work for the judge and assist in evaluating competing statistical claims, but this option is rarely used. One reason for this reluctance may be that judges are concerned that they would be ceding too much authority to such a master without fully knowing whether the master is also capable of exercising good judgment, devoid of excessive ideological baggage. This is hardly an unfounded concern. Courts might profit by having an on-staff statistical expert, but this raises its own

---

[6] This is not a game for the meek as targeted companies tend to come after such analysts with a vengeance. In one particularly heated litigation, Hemispherx Biopharma responded to claims that its highly touted drug Ampligen was worthless – causing the company's stock price to plummet and enriching the short selling counter-analyst Manuel Asensio – by suing Asensio for defamation, seeking the $80 million in the drop in value of the company as damages. Although Asensio prevailed in the first trial by a vote of 11-1, a new trial has been ordered. Since 1998, Asensio has attacked 28 companies for advancing poor drugs based on poor studies and some of the companies have been forced into bankruptcy or delisted from the stock exchanges following his charges. He has claimed his annual earnings are roughly $6 million per year. L. Stuart Ditzen, Philadelphia Inquirer Magazine, Sunday, July 14, 2002.

[7] Kaye, D.H. and D.A. Freedman (2000), "Reference Guide on Statistics," *Reference Manual on Scientific Evidence*, 2nd ed., pp. 83-178.

problems. A staff employee of this type would be able to assist the court in the low-level issues of statistics, but I suspect it would be difficult to hire statistical experts who can stay on top of the debates that must be officiated in high-stakes litigation. Clearly, the staff statistician could help the judge understand basic tools, such as ordinary least squares regression or panel data estimation, and define the opaque terms of the statistical lexicon such as multicollinearity and endogeneity bias. But like an alligator who drags his prey into deep water to facilitate the kill, sophisticated expert witnesses hired at a rate of $1000 an hour or more will often be able to ratchet up the complexity in ways that can generate appealing answers for their clients and that are beyond the capacity of a staff-level statistician. A staff statistician may be loyal, but when the stakes are high, the judge needs a statistical ally who can adequately referee the contentions of top-flight testifying experts.

During my time as a fellow at the Center for Advanced Studies in Behavioral Sciences, a very renowned statistician, Lincoln Moses, was available for consultation. Moses was a wonderful man, lucid thinker, person of exceptional character and good judgment, and eminent statistician in his heyday – vastly more talented than any staff statistician that a judge or court could expect to hire. But Moses quickly told those at the Center in the year 2000 who wanted help on the techniques of modern micro-econometric analysis of law and policy that he had not kept up with the explosion in the development and refinement of these tools. Moses was thoroughly conversant with advanced statistics circa 1975 and could provide penetrating insights on a vast array of complex, technical issues. Anything in the Federal Judicial Center's reference guide for judges was child's play for him, but when it came to questions about panel data analysis, two-stage least squares, standard error adjustments, endogeneity, instrumental variables, measurement error – all issues at the heart of current empirical evaluation of law, policy, and medicine – he stepped back and said that was beyond his ken. Those who today are at the top of the heap in terms of applied econometric work will be equally behind in ten years without major additional investments to stay on top of the rapidly evolving elements of best econometric practice. A tremendous process of tooling up would be needed to bring the modal empirical researcher of today up to the level of high-end high tech academics such as Jeff Strnad and Dan Ho. Even with this investment, a staff statistician would still be confronted with vexing questions about the validity of empirical studies. Without it, the likelihood a staff statistician would stumble into inappropriate causal inference will be unacceptably high.

Moreover, however unprepared most judges (and even staff statisticians!) confronted by dueling statistical experts will be to assess the terms of these debates, they clearly possess a far greater capacity for comprehension of these technical matters than typical jurors. Judges, of course, tend on average to be more educated than jurors, over time develop some modest statistical acumen from seeing expert witnesses in court (and perhaps by reviewing reference manuals), and can ask questions at trial if something is confusing them. Jurors come without these advantages and can only draw on what they knew about quantitative matters prior to a trial and what they can pick up from listening to the direct and cross-examinations of the competing experts.

An example from my own experience illustrates this scenario and illuminates that there is reason for concern. When I was teaching at Stanford Law School, I was hired by the state of Illinois to serve as a statistical expert in an employment discrimination case. I

was furnished the report of the plaintiffs' expert, which stated that the statistical evidence revealed that the promotion rate of black managers in the particular state office was statistically significantly lower than the promotion rate of white managers, thereby buttressing the plaintiffs' claim of intentional discrimination. The report contained the precise numbers of black and white managers who received promotions, so nothing could be simpler in statistics than calculating if the slightly lower promotion rate for blacks was statistically significantly different from the white promotion rate. I was perplexed, though, when I performed the test since it showed that there was no statistically significant difference. The report contained only the final conclusion that there was a statistically significant difference in the black and white promotion rates but gave no details of how the calculation was performed. I knew the plaintiffs' expert was a highly talented academic economist at an elite university, so I couldn't believe that he could have simply erred in making the calculation.

Perhaps, I speculated, the economist had used a more sophisticated approach rather than the simple test for a difference in two proportions. Conceivably, the expert had used a logit regression to assess the likelihood of promotion for the various state employees. The logit regression would enable another test of whether blacks were promoted at a statistically significantly lower rate than whites. I dutifully performed the test but again found no significant disparity. Might he have used a slightly different model for dealing with the probability of promotion – specifically, a probit model? Minutes later, I had my answer -- still no sign of any disparity that could be deemed statistically significant.

Perhaps, I further speculated, a simple linear probability model was used to assess these probabilities of promotion for the two racial groups. Such models are commonly used but are not preferred in that they have the disadvantage that predictions based on the linear probability model will often fall outside the range from 0 to 1, the range within which probabilities must fall.. Running my fourth statistical test, I again found no sign of any statistically significant difference. At this point, I gave up and just wrote my report saying that I had found no significant disparity using four different tests and noting that the expert had given no indication of how he had reached a contrary conclusion.

Later, it was revealed that the expert had in fact used the linear probability model, but when he did so, he had included a correction for the problem of heteroscedasticity. As he told the jury at trial, the variance of the error term in the linear probability model varies from one observation to another in violation of one of the assumptions of ordinary least squares regression. This necessitated the use of a correction for heteroscedasticity, and when that correction was made, the linear probability model revealed that blacks were promoted at a significantly lower rate. Needless to say, the jury was not exactly riveted by the discussion of these rather arcane econometric concepts.

Much of the time that I was on the stand at trial involved further discussion of this precise issue. After an hour or so, the judge turned to the jurors and told them it was time for the 15 minute afternoon break. He then added that he wanted to remind the jurors once again not to discuss the case during the break, and, in particular, that there should be no discussion of heteroscedasticity while they were in the bathroom. The judge's admonition produced the desired burst of laughter, but the story suggests a deep problem of statistical evidence as proof in a jury trial setting. Four statistical tests – a test for difference in proportions, logit and probit regressions, and the unadjusted linear

probability model -- had revealed no sign of a significant disparity in promotion rates. While I can make the case that, if one is going to rely on the linear probability model, the correction for heteroscedasticity is appropriate, I have never seen anyone use that approach to establish a finding that was contradicted by more standard statistical tests or models.

The example illustrates a number of important points. First, even in this extremely simple case, an econometrician can conduct a statistical analysis in a number of different ways. Presumably, the plaintiffs' lawyer came to the expert and asked if he could provide testimony that the black managers were being promoted at a rate that was statistically significantly lower than that of white managers. It is unlikely that one would begin this analysis using a linear probability model without any control variables but with a correction for heteroscedasticity. Instead, the expert probably tried the first four approaches that I employed and then settled on the fifth model when that gave a congenial answer. Of course, most econometric analyses are vastly more complicated than this simple comparison of promotion rates, and with this greater complexity the number of modeling and specification choices grows exponentially. The result is that the econometrician often has hundreds of estimates to choose from, which confers the ability to select the estimate that best supports the party's case.

Second, the plaintiffs' expert gave no indication in his expert report of how the estimate was calculated. Was this a sign that he did not want to draw attention to the approach he used or did not want to attach his name to a report that showed on its face that he had based his testimony on a rather unusual statistical model?

Third, what in the world was the jury supposed to do with the conflicting testimony of the experts? Were blacks promoted at a rate that was statistically significantly lower than the rate at which whites were promoted, as one of the five estimates suggested, or was this not the case, as the other four models indicated? The plaintiffs' lawyer tried to argue that the linear probability model was the appropriate one and that, if his model was used, it was appropriate to correct for heteroscedasticity. I replied that I wouldn't rely on that model if the conceptually superior logit and probit models gave a conflicting answer, which they did since they showed no evidence of a statistically significant difference. The jury was certainly in no position to make an independent choice on this matter. Did they just decide which expert they liked more or thought to sound more authoritative? Did they instead simply disregard all of the statistical evidence and reach a verdict based on the non-statistical evidence?

Fourth, and perhaps most controversially, the entire process of statistical testing may have been worthless anyway. All that the standard protocol really tells you is whether a disparity of a certain magnitude is unlikely to be generated by chance if the process generating managerial promotion is conducted randomly. Most employers, even in state government, typically do not select employees at random, so using the statistics to reject the null hypothesis of random selection is really not all that helpful to decisionmakers who are trying to decide whether discrimination has occurred against black managers.

Does this case suggest that, at least in the courtroom, econometrics is merely a tool to fool the jury into thinking that a party's case is stronger than it really is? Certainly judges – even equipped with the methodology of *Daubert* – are not in a position to serve as gatekeepers who ensure that only competent and useful econometric analyses come

before the courts. One advantage that exists in the litigation context if substantial sums are at stake is that an opposing party will invariably hire its own expert to point out the weaknesses and errors in any study introduced by the other side. A timely, critical review of an empirical study is often valuable, and this occurs far more often for a study introduced in litigation than it does for published studies in academic journals. In the latter context, certain topics will generate substantial, and occasionally, intense scrutiny, but as a general matter, studies introduced in trials will be scrutinized far more thoroughly and quickly than normal academic publications. But while the jury in the case above was presented with the competing statistical arguments, neither judges nor juries will typically be able to evaluate the shortcomings of any but the most blatantly incompetent studies.

Does the adversary system effectively insure that appropriate statistical tools are being employed? Contrast the treatment in the discrimination case with a hypothetical situation in which similar claims were submitted to a peer-reviewed journal. One possibility would be that the editor would simply refuse to publish the paper, which is roughly the equivalent of a judge throwing the expert testimony out after a Daubert hearing. Clearly, the editor of the journal, who typically possesses relatively strong empirical credentials and is backed by one or more referee reports, is in a better position to reject the manuscript that a lay judge would be. But the editor would just have the test based on the linear probability model with the heteroscedasticity correction and certainly wouldn't know that other tests would have led to a contrary conclusion about racial differences in promotion rates. My guess is that few judges would dismiss the evidence offered by a talented tenured professor from a top university, which is perhaps appropriate given the "more probable than not" standard of proof. This means, though, that it puts jurors (and judges) in a difficult position of having to evaluate statistical studies with very little reason to believe that they can make intelligent decisions with this choice.

Rather than quashing the piece (or at least relegating it to publication in a lesser journal), the editor would also have the opportunity – assuming that he or she noticed the issue -- to require the plaintiff to show the contrary findings under the other tests. In effect, this was the outcome in the litigation, where I was hired to investigate the possibility of these other outcomes. In this event, litigation had one advantage over peer review since I was presumably as capable as the journal editor, but, unlike the editor, I was paid to replicate the claims of the opposing expert and find out for myself whether different approaches would lead to different answers. An editor, backed by a referee, might ask an author to investigate different modeling approaches, but the author frequently reports back that "I tried the other specifications you suggested, but they had little effect on the estimates." An author who is up for tenure and needs to bolster a promotion file with one more published article is under great pressure to offer a self-serving interpretation of these results, and even if someone were to identify the problems years later, it is unlikely that it would ever be revealed that a dishonest answer had been given to the editor. In this way, the untenured academic may be in a similar position as the political candidate – a badly flawed study may still promote a private interest. This may sound cynical, but my efforts at replication of some published studies suggest that some authors are reluctant to reveal shortcomings in their work, even when specifically asked to do so by referees and editors. The incentives are powerfully arrayed for

securing tenure, and once years have been spent on a study, the prospect that publication (and a possible successful tenure bid) will be derailed because of some error or oversight in estimation is often something to be avoided at almost any cost.

II. The (as yet) Elusive Protocol for Truth

The tide of empirical studies has generated remarkable insights but has also unleashed a body of evidence that the public is largely incapable of evaluating. The Babel of conflicting studies has, at times, had a paralyzing effect on doctors, patients, and policymakers, who may come to feel either that all the studies are worthless, or that there is no way for them to sort out the worthless from the valuable. On the other hand, some have taken the opposite approach, investing inordinate faith in the findings of solitary studies – typically, those confirming a prior belief or conforming to some pre-existing ideology -- and embracing their conclusions with unquestioning alacrity.

In general, the explosion of studies has not been matched by a similar explosion in the number of individuals who are capable of conducting high-quality research, or in the number of individuals who are capable of evaluating this research. While the advances in statistics and econometrics over the last thirty years have been astonishing, the knowledge requirements for conducting state of the art research are rising rapidly. It is possible that the mean elite academic of thirty years ago, though admittedly painfully ignorant of econometrics, was closer to the frontier of applied practice of that era concerning the primary issues in empirical studies than is the mean elite academic of today. The explosion in technical statistics and econometrics over the last thirty years means that studies that are thought to be valid one day are shown to be utterly wrong the next when a new statistical or econometric paper is published.[8] While top researchers are learning much and getting better each year, the quality of the average empirical researcher is quite possibly falling as the advent of the personal computer, more user-friendly statistical packages, and more on line data has dramatically lowered the barriers to entry for those interested in empirical research.

Today, someone who can find online data on a few explanatory variables that might have a causal impact on a particular dependent variable can download the information to a personal computer, type in a few commands in a program such as Stata or SAS, and be "interpreting" regression output in a few minutes. I have even seen speakers in the midst of law school workshops take questions from the audience and, in response, run regressions on the spot, to the amazement and appreciation of all (or nearly all). A common refrain during policy discussions is that a certain claim about the world is not true because X did a study on that and found the claim to be false. But a single study is a very slender reed on which to rest a strong belief that an important issue of legal or medical policy has been resolved.

[8] A recent article called what's important in economics shows that a huge proportion of the top 100 most cited articles are technical econometrics articles. If the 1940s – 1960s was the time of the advance of economic theory, the last twenty-five years has largely been the time of the ascendancy of micro-econometrics, reflected in part with the Nobel Prize in the year 2000 going to James Heckman and Dan McFadden.

At present, far too many in the public, the policy/legislative world, the medical realm, and the academy respond to empirical papers with utter credulity if they happen to like the outcome or complete disbelief if they don't. Many who have been pummeled by those championing some erroneous, or even bizarre, study conclusion feel ill-equipped to respond to the econometric "evidence" presented, but intuitively appreciate the error of the conclusion and consequently develop a complete disdain for the entire enterprise of statistical analysis of law and policy. Let's begin with the first group, since they probably cause more problems than the second. A good illustration is provided by John McCall who emailed the Chicago economist Steven Levitt saying that he, McCall, had been trying to decide whether to buy Levitt's best seller (coauthored with Steven Dubner), *Freakonomics*. McCall stated that he happened to review Levitt's discussion of John Lott's work claiming that laws allowing citizens to carry concealed handguns greatly reduced crime. *Freakonomics* describes Lott's theory as intriguing – the idea is that the bad guys are deterred by the prospect of running into a gun-toting potential victim – but then notes that when scholars tried to "replicate" these claims, they found them not to be true – that is, they found that more guns did *not* lead to less crime as Lott had claimed. Referencing Lott's web page, McCall stated that a number of articles in the *Journal of Law and Economics* purported to support Lott's finding. While conceding that he had not read any of the articles, McCall stated that "*The Journal of Law and Economics* is not chopped liver," apparently suggesting that articles cited on Lott's web page in support of his work should be credited because they were published in a reputable journal. As the cognoscenti now know, Levitt replied to this email by stating that Lott had raised the funding for this issue of the journal, and thus the articles in the "Lott" symposium had not undergone the same selection process as would customary submissions to the journal. Levitt's precise language was a bit more colorful, telling McCall:

> It was not a peer-reviewed issue of the journal. For $15,000 he [Lott] was able to buy an issue and put in only work that supported him. My best friend was the editor and was outraged the press let Lott do this.

Amazingly, this single email to an apparent friend of Lott (who seems to have sent Levitt's email along to his buddy) was enough to fashion a defamation lawsuit against Levitt. (Lott also claimed that a critical comment in *Freakonomics* about what other scholars – including me – had concluded about his more guns, less crime hypothesis also was defamatory, but the court thankfully dismissed that charge.) Perhaps even more remarkably, on January 11, 2007, U.S. District Court Judge Ruben Castillo rejected Levitt's motion to dismiss this claim, stating: "this Court finds that Lott has demonstrated that the email statements qualify as defamatory *per se* because they impute a lack of ability in Lott's profession, and cannot reasonably be innocently construed." This case serves as only one illustration of the widening gap between those who can properly evaluate the validity of causal studies – such as those claiming that more guns leads to less crime – and those who cannot. Much mischief will be accomplished unless this gap is narrowed.

If Lott's defamation case against Levitt ends up going to trial, it will be interesting to see whether Levitt can show that his statements, although hyperbolically expressed, are in essence true. For example, is the fact that Lott raised the money for the

issue, without which it would not have been published, enough to justify the statement that Lott "bought" the issue? Most academics, I would think, would not understand Levitt's claim as one of bribery but rather of using resources to influence the character of a particular publication. For example, when an affluent candidate for public office spends millions of his own dollars to advance his electoral process is it defamatory to say that the candidate "bought" the election? Such statements are common in political discourse and presumably have a considerable degree of first amendment protection.

On the second point in Levitt's email, there is really no question that Lott selected five articles, including one of his own, that were more supportive of his pro-gun views than an even-handed review of the literature would have generated. As we will see, a National Academy of Sciences panel, which included Levitt and other top academics, concluded by a 15 to 1 margin that no credible evidence supported Lott's more guns, less crime claims. A review of the abstracts of the five papers in Lott's symposium underscores the point that they were far more positively disposed to Lott's views than the members of the National Academy Panel:

> 1) John R. Lott, Jr. and John E. Whitley. 2001. Safe-Storage Gun Laws: Accidental Deaths, Suicides, and Crime. The Journal of Law and Economics 44: 659 -689
> "It is frequently assumed that safe-storage gun laws reduce accidental gun deaths and total suicides, while the possible impact on crime rates is ignored. We find no support that safe-storage laws reduce either juvenile accidental gun deaths or suicides. Instead, these storage requirements appear to impair people's ability to use guns defensively. Because accidental shooters also tend to be the ones most likely to violate the new law, safe-storage laws increase violent and property crimes against law-abiding citizens with no observable offsetting benefit in terms of reduced accidents or suicides."
>
> 2) Jeffrey Miron. 2002. Violence, Guns, and Drugs. The Journal of Law and Economics 44: 615-633
> "Violence rates differ dramatically across countries. A widely held view is that these differences reflect differences in gun control and/or gun availability, and certain pieces of evidence appear consistent with this hypothesis. A more detailed examination of this evidence suggests that the role of gun control/availability is not compelling. This more detailed examination, however, does not provide an alternative explanation for cross-country differences in violence. This paper suggests that differences in the enforcement of drug prohibition are an important factor in explaining differences in violence rates across countries. To determine the validity of this hypothesis, the paper examines data on homicide rates, drug prohibition enforcement, and gun control policy for a broad range of countries. The results suggest a role for drug prohibition enforcement in explaining cross-country differences in violence, and they provide an alternative explanation for some of the apparent effects of gun control/availability on violence rates."
>
> 3) Olson/Maltz. 2001. Homicide in Large U.S. Counties. The Journal of Law and Economics 44: 747-770

“Recently, a number of states have enacted laws that allow citizens to carry concealed weapons. This "natural experiment" was analyzed by John Lott and David Mustard, who found that these right-to-carry laws reduced violent crime, with a substitution toward property crimes, in those jurisdictions that adopted this law. Of particular importance, they found that homicide was reduced significantly, with even greater declines in larger jurisdictions. Their findings came at the same time that major reductions in homicide were occurring in many cities and states that did not change their gun-carrying policies, which lead to questions of whether their finding was spurious, caused by problems with the data or methods. In this paper, we describe an analysis that looks at the effect of changing one aspect of their homicide analysis: disaggregating homicide data by weapon type, victim characteristics, and victim-offender relationships. The results show that the liberalized carrying laws are associated with a number of effects, some that are consistent with those found by Lott and Mustard and others that are not. It also illustrates the importance of being able to look beyond aggregate crime measures in this type of examination, which is currently possible on a national level only for the crime of homicide.”

4) Thomas Marvell. 2001. Juvenile Gun Possession. The Journal of Law and Economics 44: 691-713
“A 1994 federal law bans possession of handguns by persons under 18 years of age. Also in 1994, 11 states passed their own juvenile gun possession bans. Eighteen states had previously passed bans, 15 of them between 1975 and 1993. These laws were intended to reduce homicides, but arguments can be made that they have no effect on or that they even increase the homicide rate. This paper estimates the laws' impacts on various crime measures, primarily juvenile gun homicide victimizations and suicide, using a fixed-effects research design with state-level data for at least 19 years. The analysis compares impacts on gun versus nongun homicides and gun versus nongun suicides. Even with many different crime measures and regression specifications, there is scant evidence that the laws have the intended effect of reducing gun homicides.”

5) Jeffrey S. Parker. 2001. Guns, Crime, and Academics: Some Reflections on the Gun Control Debate. The Journal of Law and Economics 44: 715 -723
“This comment on Thomas Marvell's "The Impact of Banning Juvenile Gun Possession" analyzes Marvell's empirical findings and their policy implications for gun control legislation. While Marvell's article stresses the absence of any finding favorable to juvenile gun bans, this comment points out that the statistical results actually support the stronger finding that some of the juvenile gun bans are associated with a statistically significant increase in homicides nationwide. Under either finding, the juvenile gun bans are welfare reducing because of the inherently costly nature of conventional gun control legislation. The concluding discussion argues that the failure to draw appropriate policy conclusions from methodologically sound findings on controversial subjects such as gun control undercuts the value of academic research as compared with competing influences in the public debate.”

The 15-1 vote against Lott's views on guns in the National Academy Report at least suggests that Levitt was essentially correct in saying that Lott was not capturing a cross section of views expressed in top journals on the issue of the link between guns and crime. Again, Lott has the right to push the articles he thinks are best, but doesn't Levitt have the right to point out that he thinks Lott's selections are cherry-picked? Thankfully, it is unusual to see a public figure such as Lott suing a scholar of the first rank for criticizing Lott's research and efforts to promote that research in a private email reply. The lawsuit is even more troublesome given that McCall initiated the conversation by asking Levitt about the support for Lott's position on an issue of substantial legal, political, and social importance. But perhaps even odder and more dangerous was the view -- implicit in John McCall's email – that because Lott could cite some articles on his web page that were published in a reputable journal and that vaguely supported his position, one should give credence to Lott's conclusion. The lawsuit describes McCall as an economist from Texas. He should know better.

Perhaps it would help to recall Colin Powell's speech before the United Nations arguing that the evidence showed that Saddam Hussein possessed weapons of mass destruction or former CIA director George Tenet's statement that this was a "slam dunk" issue. Whether one is talking about a CIA report on a covert nuclear arsenal, a brain scan of a possible tumor, or an econometric analysis of a law, the evidence often must be evaluated by someone with tremendous skill, training, and objectivity if correct conclusions are to be reached. Anyone with a strong incentive to reach a particular conclusion will often see clarity amidst the ambiguous or even contradictory evidence. If there is a Presidential Medal of Honor in your future for overstating the conclusiveness of certain data, boy that data tends to look convincing. Similarly, if you love guns, work for the NRA, don't like those who favor gun control, or have established your career on claiming that more guns lead to less crime, a few supportive regressions in a sea of conflicting evidence can seem very powerful. The last of the five pro-gun articles from the Lott symposium – by Jeffrey S. Parker – concludes that "the failure to draw appropriate policy conclusions from methodologically sound findings on controversial subjects such as gun control undercuts the value of academic research as compared with competing influences in the public debate." Lott, who relied on methodologically unsound findings, of course, failed to draw the appropriate policy conclusion, and, as we will see, managed to persuade many legislators to vote for pro-gun laws on grounds lacking any credible statistical support. Years after these laws were passed, the process of academic discourse and verification revealed the flaws in Lott's work, but laws passed on the basis of flawed studies remain in effect. Thus, bad econometrics has the capacity to undermine democratic decisionmaking. Of course, there is no room for smugness here. Those who hate guns and are wary of America's extensive gun culture, or hope to pin an academic career on showing guns are bad, often set a very low threshold on the evidence marshaled against Lott's thesis.

This book will hopefully give a better sense of how empirical studies should be evaluated on issues of explosive normative significance or substantial legal or medical import. At present, the public is in a position of utter dependency on the good faith and competence of unknown researchers with whom one would have no particular reason to invest much confidence. The public naively enters into a trust relationship on very questionable grounds. "But it is written by a Harvard professor or published in the New

England Journal of Medicine, and it supports what I have always believed! Surely I can rely on such a study." Read on my pretties.

III. A Case Study – Does the Death Penalty Deter?

For those who wish to embrace the finding of an econometric study, it may be helpful to gain an understanding of the enormous numbers of ways in which an econometric study can go wrong in comparison to the far more limited ways in which it will go right. Even if a researcher finds one of the few correct paths to truth, lots of others will have an interest (corrupt, ideological, self-interested, mis-guided, wishful) in arguing that something else is the truth (and they will have all of those other ways of arguing that their position is right). The story of the econometrics of the death penalty may prove instructive.

Cass Sunstein and Adrian Vermeule wrote an article called "Is Capital Punishment Morally Required?" that was brought to my attention in the summer of 2005. In it, these two prominent law professors from the University of Chicago and Harvard Law Schools stated that "powerful," "impressive," "sophisticated multiple regression studies" show that "capital punishment powerfully deters killings." The article had a powerful influence on many, and I was encouraged by Carol Steiker of Harvard Law School to take a look at the various studies to which Sunstein and Vermeule had alluded. I was able to enlist my brilliant and energetic friend, Justin Wolfers, in this enterprise, and before long we were both consumed by the process of evaluating a host of studies that concluded that the death penalty strongly deterred murder.

Sunstein and Vermeule were in no way out on a limb without the support of top economists. Nobel Laureate Gary Becker, a University of Chicago colleague of Sunstein, was a strong proponent: "I support the use of capital punishment for persons convicted of murder because, and only because, I believe it deters murders…. I believe the preponderance of evidence does indicate that capital punishment deters."[9] Another brilliant University of Chicago colleague (and sitting federal appellate court judge) Richard Posner concurred: "Early empirical analysis by Isaac Ehrlich found a substantial incremental deterrent effect of capital punishment….[M]ore recent work by Dezhbakhsh, Rubin, and Shepherd provides strong support for Ehrlich's thesis; these authors found in a careful econometric analysis, that one execution deters 18 murders." Three months after Becker and Posner issued their statements, David Frum of the AEI jumped on the bandwagon: "Between 1995 and 2005, the number of murders in the United States dropped from nearly 25,000 a year to under 15,000. An American was less likely to be murdered in 2005 than in 1960…. Restore the death penalty, and you restore safety... Refuse the death penalty, and the job of reimposing legal order becomes much more difficult: citizens live in fear, trust in authority and law fade. There may be another way of protecting society. But why ignore success?"

Indeed, some argued that the econometric evidence that the death penalty deterred model was unequivocal. On April 21, 2004, economist and law professor Joanna Shepherd argued in a statement to the House Judiciary Committee that "13 [modern]

[9] Becker, Gary S. (2006) "On the Economics of Capital Punishment," The Economists' Voice. Vol. 3: No. 3, Article 4. at 1.

economic studies on capital punishment's deterrent effect have been conducted in the past decade. Most use new improved panel data and modern statistical techniques. They all use multivariate regression analysis to separate the effect on murder, of executions, demographics, economic factors, et cetera. The studies are unanimous. All 13 of them find a deterrent effect. …" Given evidence of this nature, and with such strong backing from top economists, Sunstein and Vermeule can hardly be faulted for acquiescing to the supposedly best evidence on this controversial policy question. However, as Justin Wolfers and I were to find over the next year, there are many slips between lip and cup when it comes to econometric studies of law and policy.

In one of my favorite sports, basketball, a constant cry from a defensive minded coach to his players is "Don't overcommit!" A player who "overcommits" by, say, charging at an opponent who is lining up an outside shot, is highly vulnerable to having his opponent scoot past him on the way to an easy layup. If you are moving too fast in one direction, it is often hard to put on the breaks, let alone reverse course.

One finds overcommitment in the world of ideas as well, and the consequences are often negative there as well if one is trying to properly evaluate empirical evidence. Becker and Posner provide a case in point. These two academic giants start from a strongly deductive view of the world: we know from price theory that demand curves slope downwards. The death penalty raises the price of murder, they argue, and therefore it must decrease the number of homicides. As deductive arguments go, this one has some plausibility, but, in fact, things are a bit more complicated.

The steps in the chain of logic are not as direct as might at first appear. First, is it clear that the death penalty raises the price of murder? The price of murder that is relevant to a rational criminal is the expected penalty that someone who goes ahead with a murder would expect to pay. It is not clear to me a priori, but instead is something that needs to be established, that potential murderers *at the time they are contemplating the commission of a capital offense* view a death sentence as worse than life imprisonment without the possibility of parole. The Becker-Posner offer of proof on this issue is again not bad: most on death row fight the imposition of the death penalty, thereby suggesting they would be happier spending the rest of their lives in prison. But even this is not dispositive. An impulsive gang member trying to secure a reputation as tough might well think of the prospect of execution differently when he is boldly strutting out on the street than when he is a broken convict sitting on death row. Becker and Posner point to evidence on the latter situation, but it is really the former that is relevant.

Second, the price theoretic argument captures an important influence on those criminals who are aware of the sanctions and weigh costs and benefits (assuming they view the death penalty as more severe than life without parole) but it is not the only influence. The European Union has decided that a humane nation must reject capital punishment, and one would assume that developing more humane attitudes and greater reverence for life could have benign consequences tending to lower rates of murder. Again, the issue requires empirical verification and cannot be answered by price theoretic claims.

Third, even if the net effect of capital punishment were to reduce murder, it is a bit odd for economists to be advocating it without at least considering whether the benefits of the system outweigh its costs. For example, for New York and New Jersey, two states that adopted death penalty statutes in the wake of the moratorium imposed by

the Supreme Court's 1972 decision in *Furman v. Georgia*, would be hard to argue that crime fighting dollars were being wisely expended. The two states literally spent hundreds of millions of dollars to operate death penalty regimes that have ended up executing no one. Ultimately, the New York Court of Appeals deemed its death penalty statute unconstitutional in 2004, and New Jersey has now imposed a moratorium on executions as the system is being re-evaluated. Presumably, the states could have spent that money on crime-fighting measures that are known to be effective, so the state by following the "tough on crime" death penalty strategy may have ended up with reduced funding for alternative expenditures, more murders or both.

Fourth, Becker's claim that the death penalty is a "sizeable deterrent" – a classic case of overcommitment -- can only be understood as an empirical claim. Nothing in theory would or could say anything about the added deterrent value of a death penalty sanction in addition to a life imprisonment without possibility of parole sanction. Becker and Posner are confident they know the sign of the death penalty effect, but knowing the sign is not enough. Determining good policy requires predictions not simply about direction but about magnitudes, and here theory can say virtually nothing.

Finally, while Becker's theory purports to give a strong answer to the deterrent question, an empiricist recognizes that the question "does the death penalty deter" is likely to be too broad and imprecise. Even within the United States, the death penalty has been applied in very different ways to different crimes, so one might need to narrow the inquiry to whether particular death penalty regimes, implemented in particular ways, deter murders. Thus, one might conjecture that the death penalty would be more effective prior to the 1970s than after the Supreme Court limited its use in a series of narrowing decisions. In the earlier period, executions were relatively more frequent, more vivid in that they involved shooting, hanging, electrocuting, or gassing the condemned, and the duration from time of commission to execution was relatively brief.

Over the last thirty years, the nature of capital punishment has changed substantially in the United States. The procedural changes mandated by the Supreme Court and other legislative modifications have altered the process in a number of ways that makes it considerably less likely to be a deterrent to murder. First, the death penalty has become so rare that potential criminals in many states may have no idea it exists or think their chance of getting executed is too remote to be considered. Second, the process of execution – generally by lethal injection – would seem much less vivid and terrifying than earlier modes of killing. In addition, the number of crimes that would subject one to the death penalty has been narrowed considerably. The most likely category appears to be extremely brutal, vicious murders, but these may be the ones committed by those who are least likely to be subject to deterrence.

Table 1: Preferred Estimates of the Impact of the Death Penalty on Homicides from 6 Panel Data Studies – Gross Lives Saved per Execution

Study	Geographical Unit of Analysis (Time Period of Analysis)	Methodology	Lives Saved per Execution
Dezbakhsh and Shepherd (Economic Inquiry, 2003)	State-Level (1960-2000)	OLS	8.8 lives saved
Katz, Levitt, Shustorovich (ALER, 2003)	State-Level (1950-1990)	OLS	0.2 lives saved (not statistically significant)
Mocan and Gittings (JLE, 2003)	State-Level (1977-1997)	OLS	5.3 lives saved
Shepherd (J. Legal Studies, 2004b)	State-Level (1977-1999)	OLS	2.8 lives saved
Zimmerman (J. Applied Economics, 2004b)	State-Level (1978-1997)	2SLS	2.8 lives saved
Dezbakhsh, Rubin, and Shepherd (ALER, 2003)	County-Level (1977-1996)	2SLS	19.5 lives saved

This brief discussion leads me to my first recommendation concerning empirical claims about law and policy. While having a strong theoretical foundation is essential for undertaking empirical analysis, one must be careful that the inevitable simplifications of theory do not eliminate the key real-world complexities that the empirical inquiry is designed to address. Strong priors can make one see weapons of mass destruction – and many other things – that simply don't exist.

With that caution in mind, let's turn to some of the most prominent recent

empirical studies on the death penalty. Table 1 sets forth six panel data studies that Shepherd included in her claim about the 13 major studies supporting the deterrence caused by executions. On the surface, the case looks quite strong. Top academics, offering strong theoretical arguments, are behind the claim that the death penalty deters, and many sophisticated studies support it (Congress is told "unanimously"). Intuition and simple causal stories seem to support the story, which also has backing from eminent think tanks – or, in some cases, "ideology tanks". Now the facts.

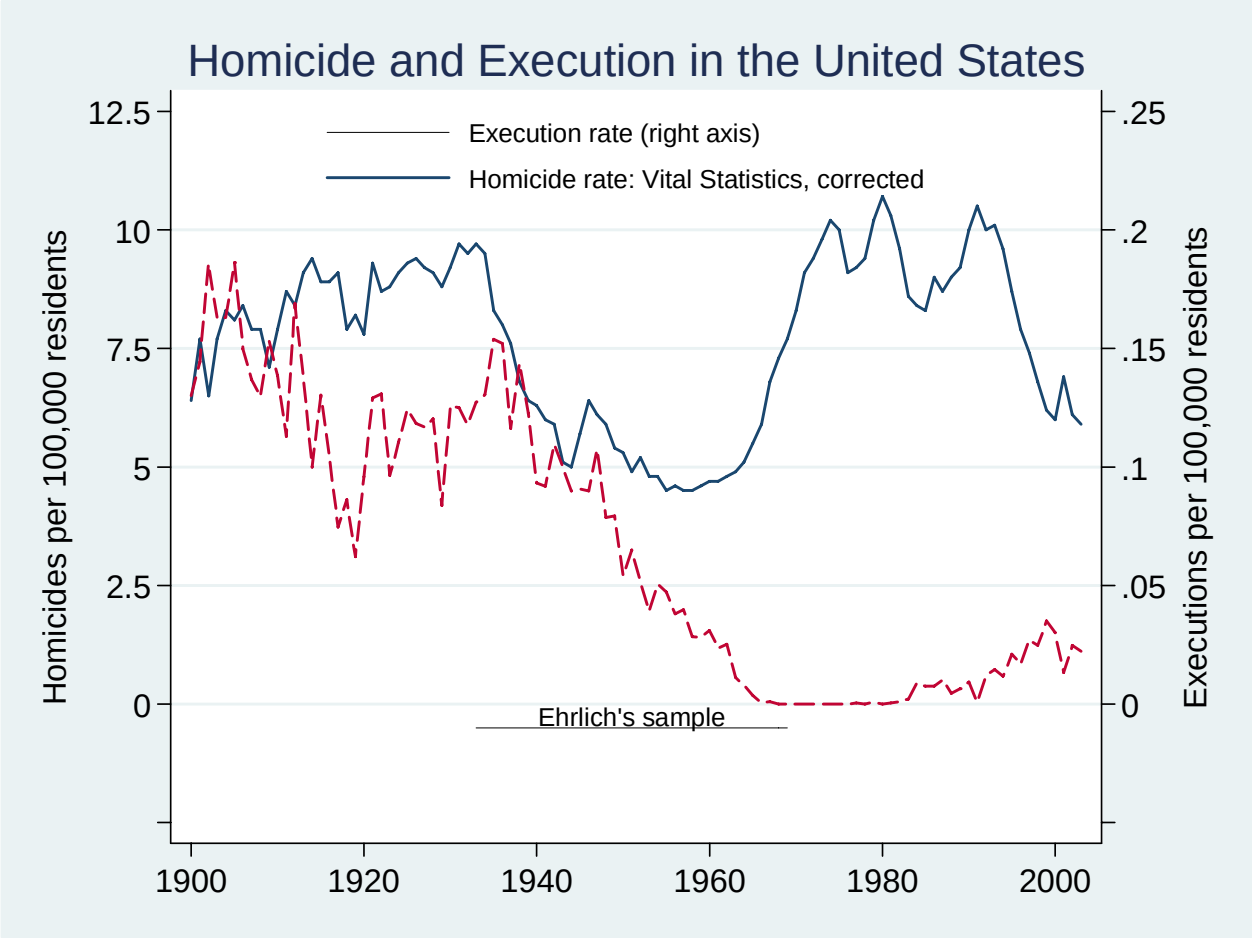
While Shepherd specifically included the paper by Katz, Levitt, and Shustorovich (KLS) as part of the "unanimous" evidence in support of deterrence, it bears mention, as noted in Table 1, that KLS's preferred estimate shows no significant deterrent effect. Moreover, KLS say: "there is little evidence in support of a deterrent effect of capital punishment as presently administered." In the above-referenced article in which Posner endorsed the deterrence of the death penalty, he only cited two studies – a 1975 paper by Isaac Ehrlich and the last paper listed in Table 1 by Dezhbakhsh, Rubin and Shepherd (DRS). Posner ignored Katz, Levitt, and Shustorovich, which appeared in the same issue of the journal and directly in front of the DRS paper. Katz (from Harvard) and Levitt (John Bates Clark Medal winner, on the University of Chicago economics faculty with Becker) are two of the most impressive economists in the country. I assume that Posner left that out study because it didn't comport with his priors. But as we will see, when the process of constructing an econometric estimate of the impact of a law or policy generates such wildly conflicting estimates, those with strong priors are naturally drawn to the estimates that support those priors. The result is that supportive regressions get published, disconfirming evidence is left on the floor, and those who share the same beliefs tout the econometric study as strong empirical support for their theories. This turns econometric analysis into little more than window dressing designed to provide apparent intellectual heft and justification to a purely theoretical belief. In this event, econometrics is not revealing truth but simply serving as a rhetorical device to rally supporters and try to persuade others that one's theoretical beliefs are true. No independent contribution is made when the empiricist can literally choose from among such widely varying estimates that make almost any conclusion possible.

An examination of the two studies that Posner cited -- the first national time series analysis of the death penalty by Isaac Ehrlich and the subsequent panel data analysis by DRS – and the KLS study that he ignored can tell us much about the advancing practice and current limitations of econometric evaluation of law and policy. Let's begin with Ehrlich's 1975 study, which was cited to the Supreme Court in a brief written by Frank Easterbrook when he was serving under Solicitor General Robert Bork. When the Supreme Court evaluated the constitutionality of the death penalty in 1976, Ehrlich's conclusion that every execution will save many lives was used to support the view that the death penalty was a substantial deterrent. The study, which came out when I was a law student, also influenced my decision to later pursue a Ph.D. in economics since I was intrigued by the prospect that questions that had vexed the public for centuries could now be resolved through this arcane and mysterious tool of econometrics.

Ehrlich used a national time series analysis, which is a relatively simple mode of analysis that essentially looks at the dependent variable each year (here the national murder rate) and sees how it changes with shifts in the relevant explanatory variable (here the national rate of execution). The obvious advantage of this approach is that national

data is often readily available, and the researcher needs to have only one set of observations per year (along with any "control" variables – such as demographics, state of the economy, or incarceration and police). Unfortunately, as is now widely understood, the disadvantages of national time series are profound, as a quick examination of Ehrlich's work should reveal.

Figure 1



Ehrlich's study was the first econometric study of the death penalty, and anyone who has had the misfortune of trying to read this impenetrable article has quickly perceived that it has all of the trappings of intellectual sophistication and econometric rigor. It embeds the empirical assessment into an ostensibly rigorous, theoretically based model, and purports to use the most sophisticated econometric techniques. The article has become a model not only for numerous other death penalty studies but also for other policies with possible effects on crime that were conducted over the next thirty years, such as John Lott's work evaluating right-to-carry laws. The paper, however, is nearly incomprehensible, which illustrates that the capacity of a bad econometric study to persuade increases with the opacity of the presentation.

Figure 1 essentially captures what drives Ehrlich's death penalty assessment. It shows the rate of executions and the murder rate for the United States for roughly the last century. The Figure also shows the time period for Ehrlich's analysis – from 1933-1969 – which turns out to be crucial to his findings. As the National Academy report on Ehrlich's study correctly observed:

> "the real contribution to the strength of Ehrlich's statistical findings lies in the simple graph of the upsurge of the homicide rate after 1962, coupled with the fall in the execution rate in the same period. His whole statistical story lies in this simple pairing of these observations and not in the theoretical utility model, the econometric type specification, or the use of best econometric method."

In essence what Figure 1 shows and what Ehrlich's analysis confirms is that the death penalty, which had dropped by 80 percent in the decades up to 1962 as the murder rate fell, itself declined to zero in the late 1960s as crime surged. Had he stopped his analysis in 1962, as Passell and Taylor showed, he would have found that his model predicted a *positive* relationship between executions and murder – that is, more hangings, more crime. Moreover, since Ehrlich concluded that the final drop in the death penalty in the late 1960s *caused* the sharp increase in murders at that time, he needed to employ an analysis that most heavily weighted that final drop (when crime jumped) rather than the large drop in executions over decades when murders were themselves falling. This was accomplished by using a log specification that in essence linked the percentage change in executions – which is obviously greatest when the execution rate is very low as it was in the late 1960s – with the change in the murder rate. A linear specification would have given greater emphasis to the 80 percent drop in the execution rate in the first three decades of his analysis, when the murder rate was falling.

A number of quick points should be noted. As Passell and Taylor promptly pointed out, a slight change in the dates of Ehrlich's analysis or in the particular specification, and the results went away.[10] The time series approach could only work if the model could predict changes in murders with great precision – which is not the case, since wide swings that are hard to explain with the standard quantifiable explanatory variables occur. In general, it is extremely hard to predict the effect of law or public policy with national time series data because all you have is a before and after comparison and it means you have to control for everything else that changed at that time to reliably predict the causal influence of the law. This is David Frum's problem, when he tried to attribute the crime drop of the 1990s to capital punishment. But the time series analysis has little capacity to choose among many possible explanations for the drop in murders, which starts around 1993: was it caused by the re-emergence of the death penalty or other factors that occurred at that same time period -- Bill Clinton's Presidency and the booming economy; the Brady Bill; 1994 federal assault weapon's ban; something else? (Yes, something else.)

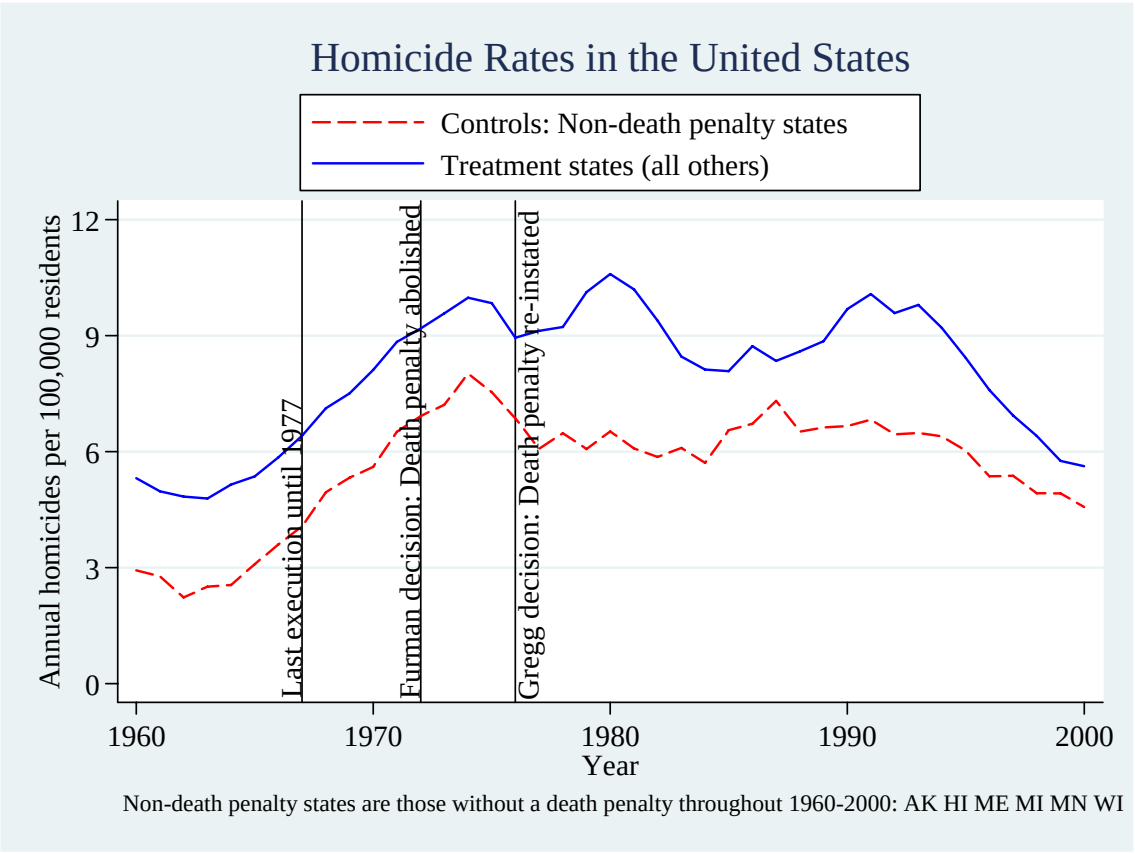
Moreover, since Ehrlich used national time series analysis there was no way to know whether increases in murders were occurring in states that were reducing executions. The curious will see that in Figure 2 below for a more nuanced way of presenting the data powerfully undercuts Ehrlich's causal story. This figure compares the time series for murder rates not for the nation but for two sets of states – those who never had the death penalty over the relevant time frame and all other states. The value of this comparison is that we know that a state that never had a death penalty law experiences no change in its penalty structure for murder when other states stop using their death penalty option or the Supreme Court forestalls this option. Thus, this comparison provides what

[10] Passel, Peter and John B. Taylor. "The Deterrent Effect of Capital Punishment: Another View." *The American Economic Review*. 67:3 (June 1977). 445-451.

Ehrlich's analysis lacks – a plausible treatment group (the states that stopped using their death penalty statutes in the late 1960s) versus a control group (those who never had a death penalty statute and thus experience no policy change during this period). As the table illustrates, the big run up in murders in the late 1960s is found in both states, thereby undermining the view that the drop in executions in the formerly executing states led to the jump.[11]

[11] One must hasten to add that the simple graphs don't control for other influences on crime. If other major influences on the murder rate were lurking in the background and were changing at the same time that the execution rates are changing, then the simple graphical depictions might lead to highly misleading conclusions. For this reason, one wants to conduct a full statistical analysis to ensure that the story that emerges from the graphical analysis is in fact true.

Figure 2


Homicide Rates in the United States
Controls: Non-death penalty states
Treatment states (all others)
Annual homicides per 100,000 residents
0
3
6
9
12
Last execution until 1977
Furman decision: Death penalty abolished
Gregg decision: Death penalty re-instated
1960
1970
1980
1990
2000
Year
Non-death penalty states are those without a death penalty throughout 1960-2000: AK HI ME MI MN WI

The comparison of Figures 1 and 2 illustrates the greater capacity for detecting causal relationships in moving from a single time series to a comparison of control and treatment states. Figure 2 provides a graphical depiction of the value of panel data, which combines both a time series as well as a cross sectional dimension. It is this treatment-control feature that better enables the panel data or difference-in-differences approach to identify true causal impacts of law and policy. By contrasting the two sets of states instead of looking at a national trend, one is essentially comparing the change in the murder rate of the treatment states after the death penalty is abolished with the change in the murder rate of the control group over the same period. If both groups of states experience identical changes in murder (that is, there is no difference in the differences), the argument that the death penalty influenced crime is undermined (even though in the simple time series analysis the murder rate rose at the same time the death penalty was eliminated). While an unscrupulous, or simply unsophisticated, researcher could rely on the national time series evidence to make the case for the death penalty (think of Isaac Ehrlich or David Frum), the more sophisticated researcher would reject that time series analysis after seeing Figure 2.

Some might conclude on the basis of Figure 2 that we can reject the conclusion that the death penalty is a deterrent. Some would go even farther and offer Figure 2 as proof that the lower murder rate of the non death penalty states is caused by their lack of capital punishment. But both of these inferences are impermissible. Some states, for a variety of reasons, simply have less crime, and this state specific trend is often enduring. Low crime states may simply feel less pressure to pass death penalty laws which would imply that the causal arrow runs from low crime to no death penalty rather than in the other direction.

While it would be nice if a simple graphical depiction could prove that the death penalty has no effect, life is not so simple. Indeed, graphs can often illuminate and frequently have useful rhetorical power, but complex causal questions can virtually never be resolved by a simple statistical or graphical comparison. What made the Ehrlich story compelling was that he was hooking up the elimination of the death penalty to a very substantial increase in crime and spuriously attributed the jump in murders to the decline in executions. Figure 2 suggests that any effect of the death penalty in the late 1960s is small, but this conclusion rests on there being no other influences on crime that differentially affected the two groups of states, thereby obscuring a deterrent effect. If that assumption is correct, all we can conclude is that the effect of the death penalty on murder – whether positive or negative – is small in the sense that it is below the threshold of perception from a graphical analysis. That does not mean that the deterrent effect is zero. First, it may simply be a small effect: while Figure 2 has the virtue of illustrating the prediction of a large impact as wrong, it might have the vice of inaccurately suggesting that there is no deterrent effect. Second, we have not controlled for any other influences on crime in these graphs. Perhaps most states restrained crime by using the death penalty, and the non-death penalty states controlled crime by relying on higher levels of police or incarceration. If both forces were weakened in the late 1960s, one

might observe similar murder rate increases and inaccurately conclude from the graphical depiction that the death penalty had no effect. To resolve these questions, we must move to a panel data analysis using the full set of controls.

The move from national time series analysis to panel data analysis is designed to improve the capacity of researchers to generate better causal estimates, and it undoubtedly does so in the hands of sophisticated users. But this important conceptual leap may also have some unfortunate unintended consequences. The more complex panel data analysis makes it harder for researchers and consumers to know what drives the estimates. Many more choices need to be made in constructing a panel data analysis, and more opportunities for error exist in more complex approaches. Moreover, the combination of greater inscrutability and more degrees of freedom enables researchers more easily - consciously or unconsciously - to pick and choose results that they find appealing.

To give a sense of how things can go astray, let's take a look at a paper by Dezhbakhsh and Shepherd that provides panel data estimates of the impact of the death penalty using data from 1960-2000. In this paper they assumed that the factor influencing criminals was the presence of a death penalty law, as opposed to the annual number of executions that Ehrlich had relied on in his national time series analysis. It should be obvious why Ehrlich had to rely on executions. He was conducting a single national analysis and therefore could not differentiate different death penalty policies across states in that analysis. But criminals could conceivably be influenced by the presence or absence of a capital punishment law in a particular state, and Table 2 depicts Dezhbakhsh and Shepherd's effort to test this hypothesis using panel data.

On its fact the first column of Table 2 suggests that presence of a death penalty statute dampens the murder rate by a substantial and statistically significant amount. The Becker-Posner thesis would appear to be confirmed. But wait. When we tried to replicate the Dezhbakhsh and Shepherd analysis we learned that using the latest corrections to the standard errors eliminated the statistical significance of the Dezhbakhsh and Shepherd estimates, as shown in the second column. One should hasten to add that Dezhbakhsh and Shepherd are not to be faulted by this over-estimate in the panel data standard errors. The science of statistics is developing at a rapid pace, and a 2004 paper revealed that the standard errors in panel data studies were inaccurate in a way that generally led to the inaccurate appearance of statistical significance.[12] The proposed adjustment yields vastly higher standard errors, as shown in column 2 of Table 2. The resulting confidence interval around Dezhbakhsh and Shepherd's preferred point estimate ranges from 119 lives saved per execution to 82 lives lost! I am prepared to believe that this interval captures the true effect of each execution, but this provides little guidance to policymakers.

[12] Betrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*. 119:1 (Feb. 2004) 249-275. While the precise reason for this finding is not essential to our discussion, the idea is that the observations on murder or even on the death penalty are not independent but tend to persist over time. Thus, the amount of independent data is actually less than the customary method of standard error calculation had assumed. Hence an adjustment was needed to make the standard errors somewhat larger to reflect this serial correlation in the panel data.

Table 2: Panel Data Estimates of the Effects of Death Penalty Laws on Murder Rates: 1960-2000

Dependent Variable: ***Annual Homicides Per 100,000 Residents*** $_{s,t}$				
	Dezhbakhsh and Shepherd (1)	Our Replication (2)	Controlling for Year Fixed Effects (3)	De Facto Versus De Jure Laws (4)
Death Penalty Law	-0.87*** (.21)	-0.95 (.57)	-0.47 (.74)	
Active Death Penalty Law (*≥ 1 Execution in Previous Decade*)				-0.57 (.63)
Inactive Death Penalty Law (*No Executions in Previous Decade*)				-0.45 (.77)
Decade Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	No	No	Yes	Yes
Adjusted R^2	.804	.791	.834	.834
Sample Size (Excludes DC, HI)	(unknown)	2009	2009	2009

One can almost hear the response of the pro-death penalty ideologue to this discussion: sure, column 2 shows that the results are not statistically significant, but the best evidence is still given by the point estimate, and it shows a sizable dampening effect on murder associated with the presence of a death penalty law – a drop of almost 1 per 100,000 in the murder rate over a period when the average rate was roughly 8. But column 3 shows that the column 2 estimate is roughly double the value that would be obtained using a more standard panel data approach. Recall the comparison of nondeath-penalty states versus all others in Figure 2, which we noted captures the essence of a panel data analysis. One has a treatment and a control group and just as one's eye confirmed that the pattern of murder rates was the same in the two groups of states depicted in Figure 2, the panel data estimates try to achieve the same analysis more formally using the statistical tool of regression. An almost universal approach to this form of panel data analysis would use a control for a year fixed effect, which simply means that the researcher assumes (as we did when eyeballing Figure 2) that there were uniform influences that might cause crime to jump up or drop down by a certain amount in each state each year. Dezhbakhsh and Shepherd presented their results without using

this standard year fixed effect control, instead simply controlling for a "time effect." We later learned that this was an effect that assumed that there was a constant effect across each state over each decade rather than each year. When we used the more customary treatment – the "year effect" -- the estimated effect in the death penalty was cut in half (moving from column 2 to column 3). We note in passing that other death penalty papers that have relied on the year fixed effect approach include:

- Dezhbakhsh, Rubin and Shepherd (2003)
- Shepherd (2005)
- Shepherd (2004b)
- Zimmerman (2004b)
- Mocan and Gittings (2003)
- Katz, Levitt and Shustorovich (2003) – using both the year effect and the decade effect.

One final tweak to Dezhbakhsh and Shepherd's analysis (column 4 of Table 2) reveals the sensitivity of the estimates. Here Justin Wolfers and I estimated the effects of the death penalty based not purely on the presence of the law but on whether the state has a law that has been used in the last decade. Again, we find that for both sets of states the standard errors exceed the estimated effects, thus leaving little statistical support for the claim that the death penalty deters.

The pro-death penalty forces might respond that the estimate is still negative, implying that we have at least some evidence of deterrence. While this is certainly a very weak foundation for the deterrence hypothesis, even this weak evidence exaggerates the support for the deterrence hypothesis. The reason is that the Dezhbakhsh and Shepherd paper is designed only to ascertain whether there is a correlation between executions and homicide rates while controlling for other factors. But death penalty laws may well be adopted as part of a "get tough on crime" attitude that might also lead to longer jail sentences, increased use of life without parole, harsher prison conditions or more aggressive policing. Since the existing estimates contain no (or inadequate) controls for these factors, we may have a classic case of omitted variable bias. The omitted factors are dampening crime and are correlated with an included factor (a death penalty law or the increased use of the death penalty), which would mean that the dampening effect of the omitted factors is incorrectly attributed to the included factor. Thus, omitted variable bias may be driving even the weak inverse correlation between homicides and death penalty law seen in Table 2.

This problem might conceivably be addressed using a two-staged least squares approach that essentially looks for quasi-experiments in which death penalty laws (or executions) are more likely to be present. The goal is to secure unbiased estimates of the impact of the law. In the medical world, and in some social experiments, randomization has been an immensely important tool used to generate an unbiased estimate of a particular treatment. The process is well known – a large group of individuals will be randomly assigned to either receive the treatment or be put into the control group and the two groups will be followed over some time and then compared. If the assumption that the two groups are comparable except for the fact that one group received the treatment and the other did not hold, the researchers conclude that the difference between the two groups represents the causal impact of the treatment. Note this is exactly what we were implicitly doing in our analysis of the nondeath penalty states versus all other states in

Figure 2, where we graphically tried to assess the difference in the performance of the two groups of states. As we have hinted, the Figure 2 analysis can break down because we did not randomly assign the treatment (a death penalty law) to the different sets of states. If, for example, many other get tough approaches are adopted when the Figure 2 states implemented their death penalty regimes, the two sets of states will differ in these other respects. Analogously, it would tough to assess the effects of vitamins in an experiment in which the treatment group not only got the nutritional supplement but were also put on an exercise program. In this case, it would be inaccurate to conclude that the different performance of the two groups was the result of the vitamins only.

The two-staged least squares technique is designed to more closely approximate that randomized experiment than is possible in the simple Figure 2 graphical analysis or the subsequently discussed panel data model of Dezhbakhsh and Shepherd. What is needed is something called an instrument, which can act somewhat in the fashion of the random assignment in the medical experiment. In fact, the random assignment is the perfect instrument because it is highly correlated with the treatment (hopefully perfectly correlated in the sense that everyone in the treatment group receives it and no one in the control group does), but it does not influence either group except through its assignment as treatment to one group. It would be terrific if we could find some instrument that makes it more (or less) likely that a state would adopt a death penalty law or execute more murderers but that did not influence the murder rate except through its influence on capital punishment. One can imagine somewhat fancifully a case where a group that is about to immigrate in large numbers to the United States and which is identical to the native population in every way but one – they are particularly enthusiastic about the use of the death penalty. As a result, when the wave of immigration hits, the states that get the largest numbers of these immigrants will experience a jump in the death penalty that is exogenous and in this way comparable to the imposition of the treatment by the researcher in the randomized medical experiment. We then look to see whether the murder rate bumps up or down in the states that received the pro-death penalty immigrants, and take that as our estimate of the impact of greater use of the death penalty.

This is somewhat similar to the process that DRS adopt in their two-stage least squares panel data model of the impact of the death penalty. The DRS equation looks imposing:

$$\frac{Murders_{c,s,t}}{(Population_{c,s,t}/100000)} = \beta_1 \frac{HomicideArrests_{c,s,t}}{Murders_{c,s,t}} + \beta_2 \frac{DeathSentences_{s,t}}{Arrests_{s,t-2}} + \beta_3 \frac{Executions_{s,t}}{DeathSentences_{s,t-6}}$$
$$+\gamma_1 \frac{Assaults_{c,s,t}}{Population_{c,s,t}} + \gamma_2 \frac{Robberies_{c,s,t}}{Population_{c,s,t}} + \gamma_3 CountyDemographics_{c,s,t} + \gamma_4 CountyEconomy_{c,s,t}$$
$$+\gamma_5 \frac{NRAmembers_{s,t}}{Population_{s,t}} + \sum_c CountyEffects_c + \sum_t TimeEffects_t + \eta_{s,t} + \varepsilon_{c,s,t}$$

In effect all it says is that it will try to explain the murder rate in every county for each year (the first term) by looking at county arrest rates, state death sentence rates given arrest, and state execution rates given death sentences while controlling for a bunch of factors. The other factors that they control for are two measures of the crime rate given

arrest, and state execution rates given death sentences while controlling for a bunch of factors. The other factors that they control for are two measures of the crime rate (assaults and robberies in each county), various aspects of the county's demographic and economic situation, and the rate of NRA membership in the state, plus controls for stable county traits that affect crime and for yearly national influences on crime (the year fixed effect that we spoke of earlier).

But we mentioned that this is a two-stage model, so what are the DRS instruments that are designed to serve the function of our pro-death immigrants in the fanciful example above? In fact, DRS use four different instruments that are designed to influence the arrest, receipt of a death sentence, and execution for murder but not influence crime except through the effect of increasing the rates of arrest, death sentencing and execution. The DRS instruments are:

- state-level nominal police payroll
- state-level nominal judicial expenditures
- state level Republican vote shares in presidential elections
- state-level prison admissions[13]

The parallel of Republicans to the hypothetical pro-death penalty immigrants may suggest that this is a good instrument, but the key to our example was that the fictitious immigrants were supposed to be otherwise identical to the native population, except for their pronounced predilection for the death penalty. If the pro-death penalty immigrants also had a host of other policy preferences (or behavioral differences) that influenced crime, then we wouldn't know whether it was these other factors or the increased use of the death penalty that was causing the impact on the murder rate. Thus, we can see the idea behind DRS's use of the Republican vote share instrument, but it still makes us uneasy.

How about the other instruments? They also are problematic. More spending on police might well influence murder arrests, say, so that sounds promising, but it also is thought to influence crime directly, in violation of the requirements of a valid instrument. It is also not clear how prison admissions influences murder arrests or executions, but it certainly could influence the murder rate directly by keeping violent criminals off the street. The prison admissions variable seems more like an explanatory variable for murders than an instrument for arrests, death sentences, or executions.

Let's see how things turn out in the DRS two-stage least squares estimation when these dubious instruments are used. Panel A of Table 3 depicts six different models that DRS offer as their estimates of the number of lives saved by each execution. Wow! Each execution is seen to save between 18 and 52 lives, with all the estimates being highly statistically significant. One can see why Posner, Sunstein, and Vermeule cited this study as supporting the conclusion that the death penalty was a major deterrent to murder. But one of the lessons of this book is that one cannot rely on the findings of any econometric study until someone gets the data, replicates the analysis, and really tries to

[13] Somewhat surprisingly the police, judicial, and prison variables are statewide aggregates, rather than per capita numbers, and the authors choose not to adjust either police payrolls or judicial expenditures to account for inflation.

figure out what is going on by identifying the ways in which the study is weak or the findings are fragile. When Justin Wolfers and I tried to replicate the findings in Table 3, we failed. Instead, we came up with the findings in Panel B, when we followed everything that DRS said they were doing. But wait. Panel B now gives six estimates that show that the death penalty is associated with *more* murders. Instead of deterrence, Panel B suggests anti-deterrence, with each execution costing between 1 and 54 lives. Now the capital punishment doesn't look so promising.

What explains this difference? It turns out that while DRS indicated in their paper that their instrumental variable was the state level Republican vote share in the most recent presidential election, they actually used a slightly different array of six variables based on the Republican vote share in each of the six separate presidential elections. This is all well and good, but note how sensitive the DRS results are to this one specification change. When they used separate variables for each presidential election to measure the Republican vote share, they found strong evidence of deterrence. When we followed their description of what they did, the opposite conclusion was reached. Which way is correct? That is a tricky question, but this example reveals why one cannot rely on regression results without extensive vetting. No referee could know that DRS had used a different variable definition than what they had described in their paper, nor would they know how critical this variable definition would be to the resulting estimates. In Panel A, DRS said their preferred model (model 4) revealed that each execution *saves* 18 lives, but Panel B shows that for this same preferred model using the variable definition described in their paper leads to an estimate that each executions *costs* 18 lives!

This is worrisome, but it gets worse. Panels C and D probe the sensitivity of the DRS estimates by dropping either Texas or California from their analysis and running the same regression that led to their pro-deterrence estimates in Panel A. In both cases, their "preferred model 4" generates huge anti-deterrence estimates – each execution leads to 29 to 42 *more* murders. But note with some judicious model selection, one could summon an estimate from Panels C or D using other models that would suggest that more than 30 lives would be saved by each execution. Note that when DRS presented their six Panel A models all showing statistically significant deterrent results, they were suggesting that any of their six plausible models were giving similar results. We now see that simply dropping either Texas or California from the analysis makes the numbers bounce wildly *using DRS's own six models*!

Table 3: Estimating the Effect of Executions on Murder Rates and Net Lives Saved: Testing the Sensitivity of the Dezhbakhsh, Rubin, and Shepherd (DRS) Estimates: 1977-1996

	Dependent Variable: *Annual Homicides per 100,000 Residents*$_{s,t}$					
	(1)	(2)	(3)	(4)	(5)	(6)
Implied Life-Life Tradeoff						
Panel A: Replication of DRS						
Net Lives Saved per Execution	**36.1***** (5.8)	**19.7***** (3.3)	**52.0***** (5.1)	**18.5***** (4.4)	**36.3***** (1.9)	**33.3***** (4.0)
Panel B: Allowing Only One Partisanship Variable						
Net Lives Saved per Execution	-24.5*** (8.0)	-53.8*** (6.0)	-43.3*** (8.2)	-17.7*** (6.0)	-0.9 (3.0)	-26.1*** (6.2)
Panel C: Dropping Texas						
Net Lives Saved per Execution	-21.5*** (7.6)	33.7*** (4.4)	6.5 (7.9)	-41.6*** (5.6)	32.5*** (2.1)	-11.3* (5.9)
Panel D: Dropping California						
Net Lives Saved per Execution	-26.1*** (7.0)	30.1*** (3.9)	33.3*** (6.5)	-28.7*** (4.9)	17.8*** (2.0)	9.6*** (4.8)

To generate an instrumental variable estimation, one needs at least one instrument for each variable about which one has concerns about endogeneity – in the DRS model this was the three variables of murder arrests, rate of death sentences, and rate of executions. We saw from our Panel B estimation that the particular way in which the Republican vote share instrumental variable was measured had an enormous impact on the estimated effect of the death penalty. When a single variable was used the results showed strong anti-deterrence but when they created six separate variables (one for each Presidential election), the results showed strong deterrence. Since DRS are creating six instrumental variables along with their Republican vote share measure and have three others (police expenditures, judicial expenditures, and prison admissions), we have more than the needed number of instruments to run the two-stage least squares regression. This means that we can probe how well these instrumental variables are doing in providing credible estimates of the impact of the executions, by selectively dropping some of the instruments. If the estimates are stable and robust, we would have greater confidence in the resulting predictions about the impact of the death penalty.

Table 4: Estimating Net Lives Saved per Execution: Exploring the Validity of the Dezhbakhsh, Rubin, and Shepherd (DRS) Instrumental Variables: 1977-1996

	Dependent Variable: *Annual Homicides per 100,000 Residents*$_{s,t}$					
	(1)	(2)	(3)	(4)	(5)	(6)
	Implied Life-Life Tradeoff					
	Panel A: Replication of DRS					
Net Lives Saved per Execution	36.1*** (5.8)	19.7*** (3.3)	52.0*** (5.1)	18.5*** (4.4)	36.3*** (1.9)	33.3*** (4.0)
	Panel B: Restricting the Instrumental Variables to Police Payrolls, Judicial Expenditure, and Prison Admission					
Net Lives Saved per Execution	**-85.6***** (13.7)	**-36.8** (28.3)	**-71.95***** (14.9)	**-52.3***** (9.2)	**-23.0***** (8.1)	**-85.7***** (13.6)
	Panel C: Restricting the Instruments to the Republican Vote Share					
Net Lives Saved per Execution	**429.4***** (21.2)	**82.0***** (4.6)	**286.5***** (11.1)	**288.8***** (15.7)	**53.1***** (2.2)	**242.3***** (9.3)

Panel A of Table 4 repeats the DRS estimates of the number of lives saved per execution. In Panel B, we make only one change to their model – we drop the Republican vote share instrument. Again, executions appear to be very dangerous for the public – each use of capital punishment leads to 23 to 86 more murders!

Supporters of the death penalty need not fear, however. Drop the other three instruments used in Panel B and rely solely on the Republican vote share and you are back in business, as Panel C dramatically reveals. Now each execution will save between 53 and a whopping 430 lives!

The lesson here is that instrumental variable estimation is a dangerous game if not used with extraordinary care. Tables 3 and 4 should reveal that in using the same data and essentially the identical statistical model, minor tweaks can make the resulting estimates bounce wildly. Of course, no one would publish an estimate that each execution saves 430 lives. It would suggest that if in the year 2004 we could have just gotten back to the muscular execution strategy of the hallowed days of 1999 when there were 98 executions instead of the anemic actual number of 59, we would have cut the number of murders by 16,770!! Since the actual number of murders that year was 16,137, the 430 lives saved estimate would essentially imply that returning to a policy in place only five years earlier would have eliminated murder from a major society for the first time in world history.

Similarly, the idea that each additional execution would lead to an extra 86 murders (as suggested by two models in Panel B) is equally fanciful. That would mean that the drop in executions from 1999 to 2004 would have saved 3,354 lives. Would that it were so. Hundreds of years of theorizing on what policies would reduce murder have given us only two usable bits of information: it helps to catch and punish murderers and it is not wise to punish small crimes as harshly as murder since, in so doing, you may incentivize criminals to kill witnesses or those who might capture them. If we are going to make advances beyond these rather obvious points, they will come through empirical evaluation of what helps to reduce murder. When researchers can generate such wildly varying estimates from the same basic model, the estimates that they choose to publish simply constitute a confirmation of their priors (or a reflection of their desired outcome). In this event, the empirical evaluation has contributed nothing, and we are left with our very meager theory-driven base of knowledge about what influences the rate of murder.

Given the beauty and the value of instrumental variable estimation and its promise as an important tool in helping to aid the process of teasing out the causal impact of law and policy, one doesn't want to end on such a nihilistic note. A few important lessons for the researcher should be emphasized. First, one must be extremely attentive to the requirements of a valid instrument if one hopes to usefully employ two-stage least squares estimation. Recall that the objective of the instrument is to provide a crude approximation of a randomized experiment, which requires two elements: first, one must identify a factor that stimulates (or depresses) the use of the death penalty, and, second, one needs to establish that the factor doesn't influence murder, except through its effect on the death penalty. Both of these issues must be addressed fully and carefully. It is not enough to conjecture, for example, that Republicans like the death penalty more than Democrats so their vote share should correlate with executions. This might establish an intuitive basis for believing the first element was present (although one should still establish this formally by showing that when Republican vote shares rose, the death penalty was used more commonly), but it doesn't even begin to address the second issue that Republican vote share must not influence murder except through its affect on the death penalty.

In response to our criticisms of the DRS instruments, Paul Rubin responded that "Most of our instrumental variables have been used in numerous empirical papers because previous researchers believed (often based on empirical testing) that the instruments were as uncorrelated with crime rates as one was likely to find."[14] Rubin is correct in the first part of his statement. These instruments have been used in many previous studies. Lott and Mustard (1997) and Rubin and Dezbakhsh (2003) used them to explain the impact of concealed gun laws, and Shepherd used them in three separate papers to assess the impact of truth-in-sentencing legislation, California's three strikes law, and sentencing guidelines. But of course the fact that these same instruments were used for these other studies should be the clue that they are problematic rather than the justification for their use. The second requirement is that the instruments do not affect murders, except through their influence on number of executions. But these other studies were investigating the impact of other laws that were deemed to have an impact on crime. Thus, the implicit experiment they create, at best, is to stimulate the use of the death

[14] *Rubin, Economists' Voice, April 2006*

penalty along with a host of other factors that influence crime. Instead, of getting a true estimate of the impact of executions, we get an impact of the influence of all of these other factors. No wonder that the numbers bump around wildly depending on small tweaks in the instruments. The effect of a huge array of factors is being lumped into this one estimate. A valid instrument can only have an impact on the variable of concern – here, murders – by operating through a single channel (here, the increase in executions). Invalid instruments, invalid results.

Oddly, all but one of the published studies that Shepherd testified about in her Congressional appearance concluded that the death penalty deterred murder, even though we have shown how fragile the estimates were. It is unlikely that the process of their model creation never generated statistically significant evidence of anti-deterrence or evidence of no effect, but the studies never mentioned such estimates. Cynics might suggest manipulation and cover-up, but I think a more likely explanation is that when one begins with a very strong theoretical belief that the outcome of a statistical analysis should come out a certain way, any opposing estimates are discarded as clearly the product of error. In fact, the one study, by Katz, Levitt, and Shustorovich, that did conclude that the death penalty had no effect was focused on exploring the impact of prison conditions on crime and was therefore not looking to draw a conclusion on the death penalty. Does this suggest that when the blinders of ideology or overcommitment to simple theory are lifted, the truth is more likely to be found? Let's take a look.

Table 5 sets forth details of the KLS model, which is essentially similar in basic structure to the panel data OLS model (rather than instrumental variable estimation) of Dezhbakhsh and Shepherd discussed above. While KLS estimated the Table 5 model over the period from 1950-1990, Justin Wolfers and I extended the KLS model through 2000 to have more complete post-moratorium data. We also extended the KLS model back in time to include the broadest period for which we have data --1934-2000.

KLS model the impact of the death penalty with the executions per prisoner variable, and we augment their analysis by using two other key explanatory variables: executions per capita and executions per lagged murder. Figure 4 provides estimates of the impact of executions using these three execution measures for each of the basic models employed by KLS. The figure contains not only point estimates of the number of lives saved (if positive) or lost (if negative), but also shows the 95 percent confidence interval around those estimates. The basic message is that the death penalty seems to cost lives rather than save them (once we subtract off the life of the one executed, 11 of the 12 estimates are negative), but there are many ancillary points that need to be made.

Table 5. Basic KLS Specification

Specification	$Murder_{st} = \beta_1 Death_{st} + \beta_2 Execute + \mathrm{X}_{st}\Gamma + \lambda_s + \delta_t + \varepsilon_{st}$
Complete set of RHS variables (included in X)	• Executions/1000 prisoners • Prison Death Rate/1000 prisoners • Prisoners per crime (lag 1) • Prisoners per 100,000 residents (lag 1) • Real per capita income • Insured unemployment rate • Fraction black • Fraction urban • Fraction 0 to 24 year-olds • Fraction 25 to 44 year-olds • Infant mortality rate • State fixed effects • Year fixed effects

Subscript *s* indexes states and *t* corresponds to time. *Murder* is the murder rate per 100,000 residents. *Death* and *Execute* are, respectively, the prison death rate (excluding executions) and the execution rate per 1,000 state prisoners. X is the matrix of criminal justice, economic, and demographic variables detailed below.[15] The indicator variables λ and δ represent state-fixed effects and time dummies. KLS also present three variants of this basic model: 1) replacing year fixed effects with region-year effects, which identifies the effect of executions only against other states within the same region (which may be important given the strong regional application of capital punishment); 2) replacing year fixed effects with decade fixed effects (the value of which is unclear); and 3) adding state time trends.

First, compared to many of the earlier estimates we have been discussing, these estimates are noticeably smaller in absolute value, which I think is a virtue. Whatever one thinks of the death penalty, its effect in the United States in the post-moratorium period is almost certainly very modest. Second, there is a great deal of imprecision in some of these estimates as suggested by the wide confidence intervals – primarily for the executions per capita variable. Third, Sunstein and Vermeule (2006) argued for the morality of the death penalty if each execution saved at least one murder victim. In Figures 4-6, the area above the horizontal line at zero would correspond to Sunstein and Vermeule's zone of moral executions (the deterrence zone). Becker (2006) would be willing to execute even if no net lives were saved, as long as the lives of the murderers were less socially valuable than the lives of their victims. Thus, Becker would potentially find executions acceptable for estimates in the band between 0 and -1 lives saved (the zone of incomplete deterrence). Finally, Becker stated that he would oppose executions

[15] The second two explanatory variables in Table 5 – prison death rate and prisoners per crime – might be thought to be somewhat non-standard controls in crime equations. We experimented by deleting the second two explanatory variables in Table 5 – prison death rate and prisoners per crime – from the various modified KLS regressions and found that they had relatively little impact on the estimated effect of capital punishment on murder.

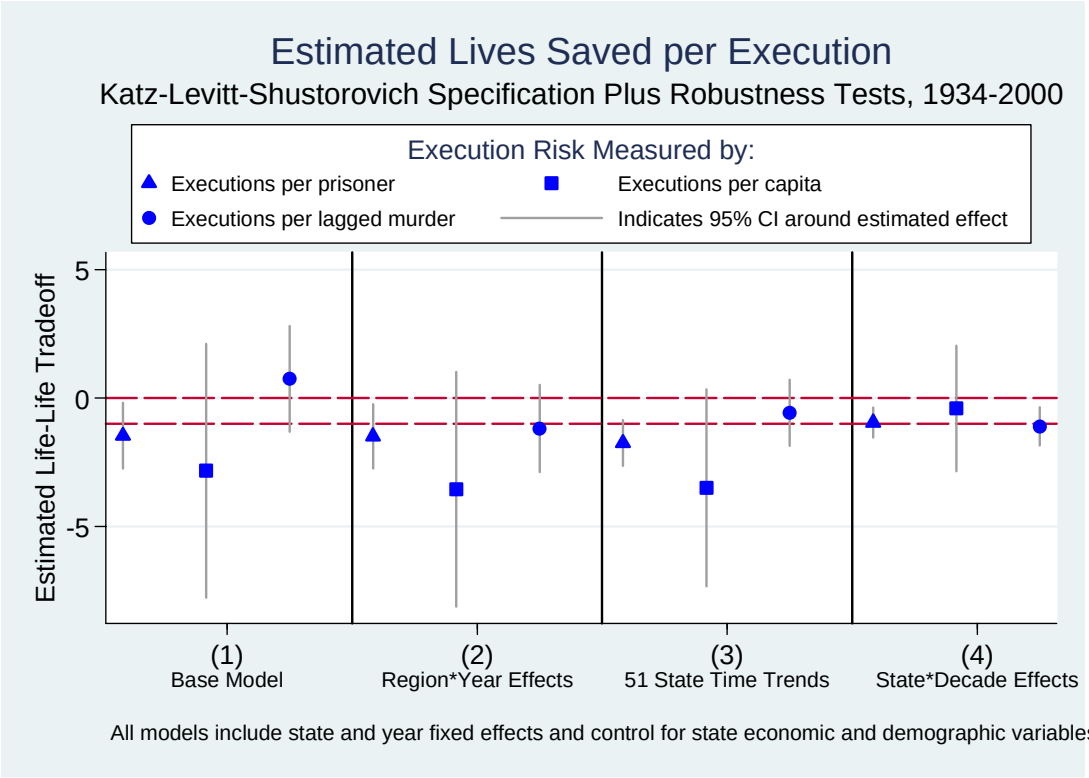
if they generated no deterrent benefits at all (the area below the horizontal line at -1, which is the anti-deterrent zone). Note that not a single confidence interval rests entirely in the area corresponding to Sunstein and Vermeule's notion of morally obligated executions, and only one point estimate falls in that category.

Fourth, the theme of this discussion has been that since researchers need to make lots of choices, they can essentially pick the answer they want under current econometric practice. My own judgment about the best variable to capture the influence of the death penalty is "executions per lagged murder." This means I could present the first panel regression to support a modest deterrent effect (if we don't care much about the life of the murderer, then the tradeoff looks better by one more life than the figure shows). The death penalty opponent could immediately counter that, even if one chooses the executions per lagged murder variable, the second panel is a more plausible regression (it probably is[16]), and that regression yields an estimate suggesting essentially no effect or even a modest anti-deterrence (albeit not statistically significantly). If one does choose that execution variable, then not much turns on whether one uses models 2 – 4 in this case. We noted earlier in our discussion of Dezhbakhsh and Shepherd, however, that they had chosen a model using state×decade effects (instead of the more standard year fixed effects), and in that case it made a big difference.

The bottom line from Figure 4 is that the confidence intervals around each point estimate encompass all three possible zones in 7 of the 12 estimates (the other five have anti-deterrence point estimates that straddle the anti-deterrence and incomplete deterrence zones). Overall, the results are somewhat more suggestive that the death penalty is useless or harmful rather than beneficial in saving lives, but this view must be tempered in light of the broad confidence intervals. One possible conclusion is that this evidence gives us no reason to reject the null hypothesis that the effect of executions on murder is zero: *All* the underlying coefficients on 66 years of state panel data were statistically insignificant. But, as we saw above in Figure 1, the simultaneous fall in executions and murders from the 1930s until the late 1950s raises a concern that the fall in murders is causing the fall in executions – a classic case of endogeneity which may bias the estimated effect of the death penalty in the pre-moratorium period.

[16] The first panel assumes that states that have executions are the treatment and the remaining states are the control. The second panel assumes that states like Vermont or New Hampshire are not good controls for death penalty states like Texas, and therefore it limits the control to other states within the same Census region.

Figure 3



Jeff Strnad has a wonderful new paper showing how Bayesian statistical techniques can offer more objective methods for selecting among competing models, and hopefully we will soon be at a point when many of these disputes can be resolved with the use of better statistical techniques.[17] Until that point is reached, though, we are in the position that I currently lament – the researcher has the capacity to pick and choose among a vast array of different models and specifications, and the hapless referees and reading public have little idea of what drives the results of most published papers. There is no doubt that tremendous intellectual energy is currently being invested in advancing the science of statistics right now, exactly because it is realized that the status quo is so unsatisfactory. Moreover, unless the new "best practice" protocols for selecting among the competing model and specification choices are of manageable complexity, this new frontier of statistical practice will represent a step back from the enormous lowering of the barriers to entry to conduct empirical research over the last thirty years.

While the Strnad article discussed the promised approaches that Bayesian statisticians have developed to negotiate between the choices of the varying statistical models, note that the KLS-based models depicted in Figures 3-5 are standard OLS panel data models while the previously discussed DRS models employed instrumental variables. If endogeneity problems are so severe that instrumental variables approaches are needed, the new Bayesian statistics will still not be of assistance if valid instruments are not to be

[17] Jeff Strnad, "Should Legal Empiricists go Bayesian?" (2006).

found. Again, the public is asked to accept different estimates of the effect of the death penalty where the differences are not easily resolved by any current statistical approach.

While an argument can be made that Figure 3 provides the most reasonable estimates of the impact of executions that now exist since they use the most plausible models in the current literature over the longest time span available. Using the longest possible time span will be beneficial if the effects of unusual omitted variables, such as the criminogenic influence of crack in the late 1980s and early 1990s (which makes the death penalty look good as crime fell and executions rose in the mid-1990s), and the crime-reducing effects of the ending of prohibition in 1933 (which makes the death penalty look bad as murder and executions both fall), tend to average out over the 66 year time span. But this is somewhat of a leap of faith.

It seems logical that if the death penalty had a deterrent effect, it would be most likely to be seen during a period when it was used more frequently, more rapidly, and with harsher and more vivid methods of execution. This is all true for the period prior to 1960, so we split the sample and estimated the effects for the period before that date (Figure 4) and from 1960-2000 (Figure 5). Interestingly, the estimates are the exact opposite of what we would predict – Figure 5 would lead one to believe the death penalty has an anti-deterrent effect when executions were common, and Figure 5 suggests that when they became rare, extensively delayed, and less harsh, they may have had a deterrent effect. My guess is that the two sets of results are driven more by simultaneity bias (Figure 4) and omitted variable bias (Figure 5), which again makes the hope that averaging over the entire period will generate valid estimates a bit heroic.

Perhaps one could counsel an interested legislature for or against the death penalty if one could make a case that one set of estimates clearly was more reliable, but in the end the amount of parsing would lose all but the most devoted adherents of econometric analysis – or the most intense ideologues. Of course, legislatures could borrow a page from the medical world and say that valid studies must be conducted before patients can be exposed to potentially harmful treatments. To really find out the effect of the death penalty, one could get a pact among all the current death penalty states, and then randomly assign states to a death penalty group or a non-death penalty group. After five or ten years, an econometric study would reveal the effect on murder. The pact might state that unless the law shows a benign effect of a pre-specified amount, the death penalty law will automatically sunset; conversely, non-death penalty states would automatically implement the law if it showed a benign effect of an appropriate magnitude. Death penalty champions might contend that this approach makes some states who want the death penalty forego it for a period of years, but similar issues arise in medical studies where the control group is taking a placebo and thus not getting a medical treatment that is promising enough to merit a very costly experiment. Since there is so little empirical evidence to support the death penalty, the states assigned into the no-death penalty group may be pleasantly surprised to learn that their controversial statute was unnecessary or even harmful. In any event, all the states should be pleased to move from a regime of policy by ideology and whim to policy by informed judgment.

Figure 4

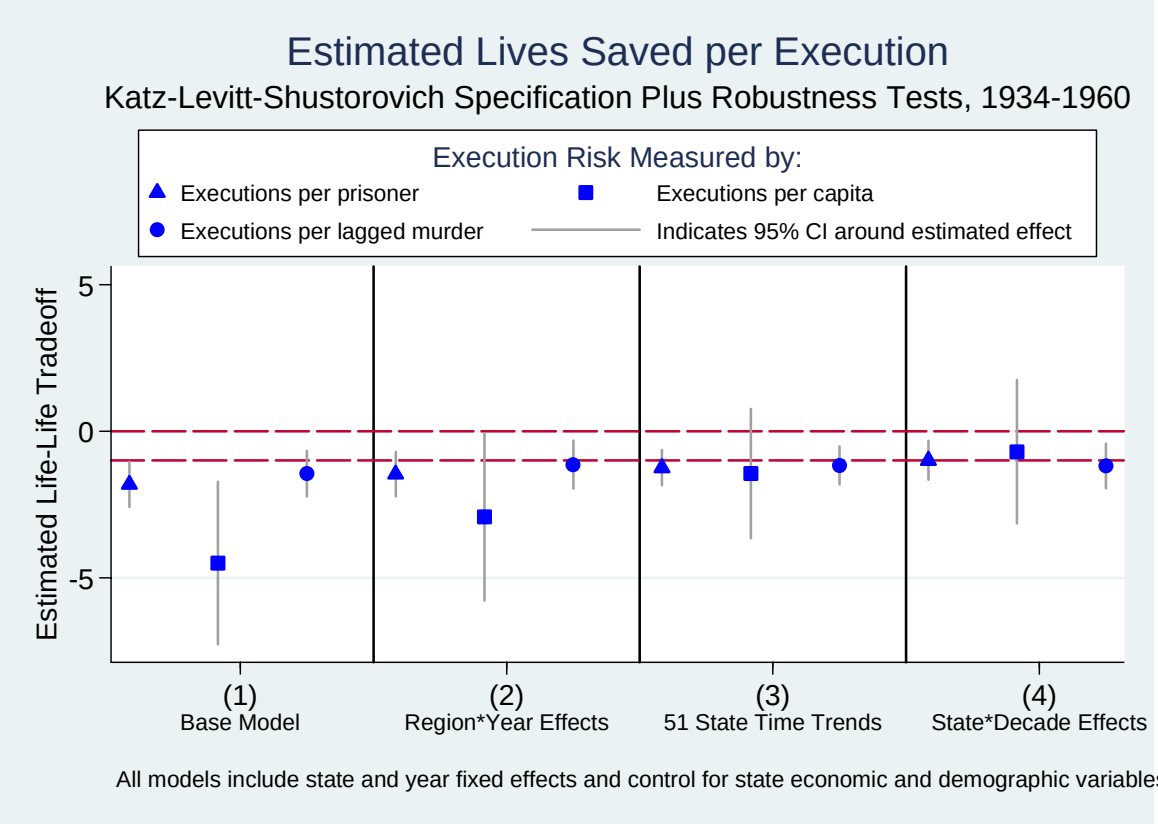

Estimated Lives Saved per Execution
Katz-Levitt-Shustorovich Specification Plus Robustness Tests, 1934-1960
Execution Risk Measured by:
Executions per prisoner
Executions per capita
Executions per lagged murder
Indicates 95% CI around estimated effect
Estimated Life-Life Tradeoff
5
0
-5
(1)
Base Model
(2)
Region*Year Effects
(3)
51 State Time Trends
(4)
State*Decade Effects
All models include state and year fixed effects and control for state economic and demographic variables

Figure 5


Estimated Lives Saved per Execution
Katz-Levitt-Shustorovich Specification Plus Robustness Tests, 1961-2000
Execution Risk Measured by:
Executions per prisoner
Executions per capita
Executions per lagged murder
Indicates 95% CI around estimated effect
Estimated Life-Life Tradeoff
10
5
0
-5
(1)
Base Model
(2)
Region*Year Effects
(3)
51 State Time Trends
(4)
State*Decade Effects
All models include state and year fixed effects and control for state economic and demographic variables

IV. The Strange Tale of Estimating the Impact of Right to Carry Laws

In 1997, John Lott and David Mustard launched what has come to be one of the most remarkable tales in the history of public policy evaluation when they announced that laws permitting citizens to carry concealed handguns – so-called right-to-carry (RTC) laws -- caused crime to fall. Hailed as heroes by the National Rifle Association (NRA) and its supporters, while derided as scoundrels by their staunchest critics, Lott and Mustard precipitated a scholarly and political odyssey that can teach us much about the techniques and limitations of sophisticated empirical research and the divergent norms of the scholarly and political realms.

The Theory Disproved

Lott and Mustard theorized that allowing non-felons who were not mentally ill and were willing to pay a permit fee to lawfully carry guns would serve to reduce crime because criminals would be more frequently thwarted by the armed resistance of potential victims or onlookers.[18] Skeptics immediately pointed out that the benefits of legitimate defensive use of guns by certain members of the public could well be outweighed by illegitimate uses, but ultimately the question was an empirical one – would RTC laws increase or decrease crime? Although Lott and Mustard's research was quickly denounced by some as unscientific nonsense, this criticism was unfair in that Lott and Mustard had taken a reasonable first step in trying to ascertain the effect of a RTC law by creating a panel data set of crime across all 50 states over the period from 1977-92 while using a fixed effects model to test whether RTC laws had any statistically significant effect on crime. Indeed, taken at face value (which we have now learned can be highly misleading), the first cut seemed to suggest exactly what Lott and Mustard said it did – that RTC laws reduced at least some categories of violent crime.[19]

With the benefit of hindsight (and much subsequent scholarly analysis on more complete and new data), one can now state rather confidently that "the first cut" was wrong -- RTC laws do not reduce crime. There is even statistical support for the view that RTC laws may actually increase crime, but further work is needed to sort out whether this evidence merely illustrates weaknesses in the panel data models of crime or in fact captures the true effect of the laws (more about this below). Donohue (2003) and

[18] Under the Texas RTC law, prior commitment to a psychiatric care facility or indeed any past psychiatric problem is not disqualifying as long as a licensed psychiatrist will state that the "condition is in remission and is not reasonably likely to develop at a future time." Texas Department of Public Safety, http://www.txdps.state.tx.us/administration/crime_records/chl/faq.htm.

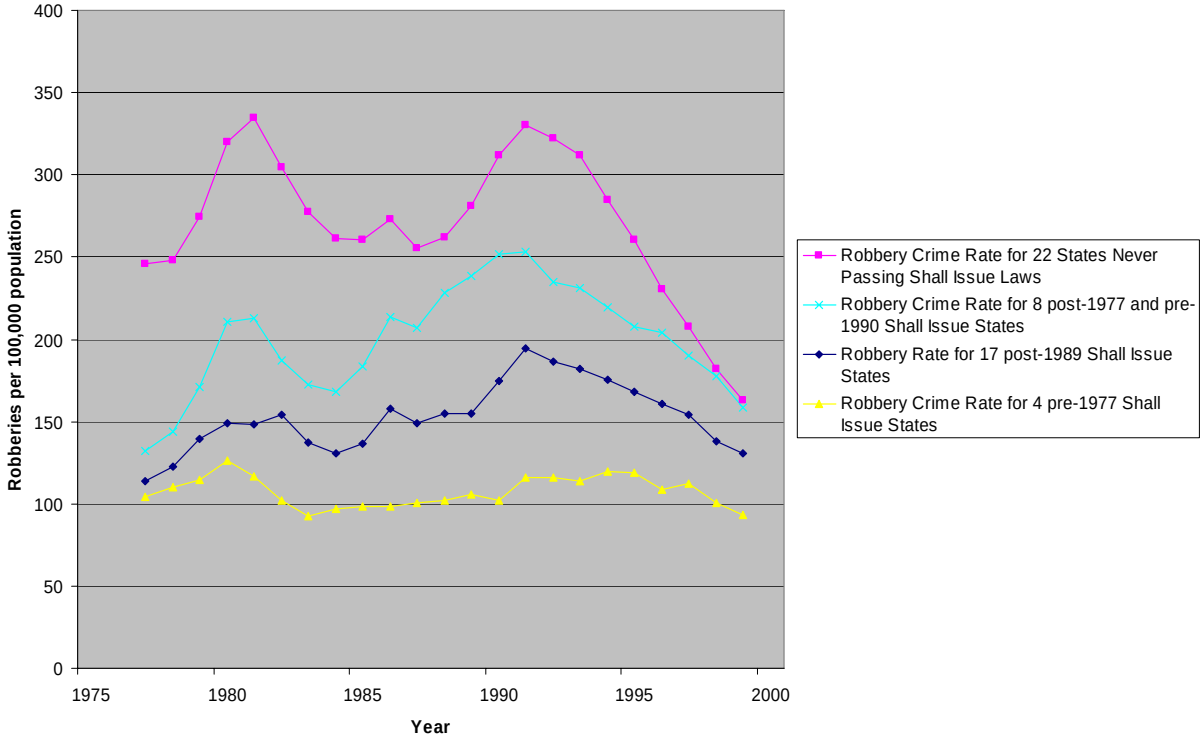
[19] Even apart from the usual disclaimers that association doesn't prove causation, Lott and Mustard's initial work contained important anomalies. There was no evidence that robbery declined, which cut powerfully against their thesis, since robbery is the crime most commonly committed outside, where a concealed gun would be expected to have its greatest potential benefit. Moreover, Lott and Mustard vacillated on whether RTC laws *increased* property crime, which they tried to explain as the result of criminals shifting away from robbery to crimes where they would not directly confront their victims (note again the centrality and anomaly of the robbery results, which conflict with the more guns, less crime thesis). Moreover, as historian Randolph Roth notes about Lott's initial study: Lott "biases his results by confining his analysis to the years between 1977 and 1992, when violent crime rates had peaked and varied little from year to year…. Had Lott extended his study to the 1930s, the correlation between guns laws and declining homicide rates that dominates his statistical analysis would have disappeared." Randolph Roth, "Counting Guns," 26 Social Science History 699, 700 (Winter 2002).

Ayres and Donohue (2003) extended the initial Lott and Mustard data set through 1999, probed the aggregated estimates that Lott and Mustard had previously championed, and revealed that their results were spurious.[20] The basic problem was that Lott and Mustard's data explored the impact of the adoption of RTC laws in 10 states starting in the second half of the 1980s, which was just about the time that the introduction of crack into certain urban areas led to a large increase in crime. States that adopted the laws tended to have less of a crack problem, and therefore, what Lott and Mustard thought to be the restraining effect on crime of RTC laws was more plausibly the product of omitted variable bias owing to the inability to control for the criminogenic influence of the introduction of crack cocaine. An early review of Lott's book "More Guns, Less Crime" by Ayres and Donohue noted that if the crack cocaine explanation for Lott and Mustard's initial results were correct, one would expect to see larger drops in crime in the 1990s in the states that had previously experienced the large crack-induced crime increases.[21] The subsequent work by Ayres and Donohue extending the Lott and Mustard data set through 1999 showed this to be the case (and the greater crime drop in the 1990s in non-adopting states holds true whether one looks at raw crime data or regression models controlling for the explanatory variables used by Lott and Mustard).

Specifically, while the 22 states that had not adopted RTC laws (through 1999) had substantially higher rates of violent crime, robbery, and aggravated assault in 1992 than the RTC states, this difference had been considerably narrowed by 1999. The pattern can be seen in Figure 6, which shows the robbery rates for the 22 non-adopting states, and for the three groupings of RTC states by date of adoption of their RTC laws – before 1977, between 1978 and 1989, and between 1989 and 1999. Note that of the four groups of states, only those without RTC laws experienced major declines in robbery over the 1977-1999 period. Indeed, the 22 non-adopting states had the *highest* rates of rape and property crime at the beginning of the 1977-99 period but the *lowest* rates by the end of that time period. While other factors undoubtedly explain why the non-RTC states had such large decreases in rape and property crime relative to RTC states, the relative crime patterns illustrate the uphill battle that exists for anyone trying to assert that RTC laws reduce crime.

[20] John J. Donohue, *The Impact of Concealed-Carry Laws, in* EVALUATING GUN POLICY: EFFECTS ON CRIME AND VIOLENCE 287 (Philip J. Cook & Jens Ludwig eds., 2003). Ian Ayres & John J. Donohue III, *Shooting Down the "More Guns, Less Crime" Hypothesis*, 55 STAN. L. REV. 1193 (2003).
[21] Ian Ayres & John J. Donohue III, *Nondiscretionary Concealed Weapons Laws: A Case Study of Statistics, Standards of Proof, and Public Policy*, 1 AM. L. & ECON. REV. 436 (1999).

Figure 6



Lott and Mustard as well as Ayres and Donohue primarily relied on panel data models across all 50 states with the date of adoption of the RTC law being the explanatory variable of interest. Kovandzic and Marvell tried a different approach that narrowed the focus by looking at actual concealed carry permits by county to identify the effect on crime of the RTC law adopted in 1987 in Florida, a state that Lott and Mustard considered to be highly supportive of their thesis.[22] Kovandzic and Marvell collected county data on crime and concealed handgun permits across Florida from 1980 – 2000 and concluded: "we find no credible statistical evidence that increases in permit rate growth (and presumably more lawful gun carrying) leads to substantial reductions in violent crime, especially homicide. Similar to Ayres and Donohue (2003), we find that our best, albeit admittedly imperfect, statistical evidence indicates that increases in permit rate growth may actually lead to slight increases in crime." I should note that while I liked the Kovandzic and Marvell paper and conclusions, it is not certain that if RTC laws had the protective umbrella that Lott contends, their methodology will reveal it. In essence, their approach assumes that criminals would recognize that more gun permits are being issued in some counties and would then shy away from crime in that county if the Lott story were true. They don't observe this – which is useful to know – but it still could be the case that criminals are deterred by the passage of the law but not

[22] Indeed, Ayres and Donohue (2003) estimated state specific effects, which suggested that Florida was one of the few states for which crime did drop significantly after the adoption of a RTC law – although there are reasons to believe that the crime drop was caused by other factors (such as the ultimate decline following the enormous crime run up induced by the Mariel boat lift of the early 1980s).

sophisticated enough to know of the gun permit disparities across counties, so that the crime drops would not be differentiated along those county lines. Again, just a caution that interpretation of econometric results is yet another area of contention.

Scholarly Triumph and Political Failure?

The weight of the evidence is now widely perceived, after the release of a National Academy of Sciences report so asserting, to have undermined Lott's claims that RTC laws have reduced the overall level of crime.[23] Thus, if the story were confined to the academic realm, it would seem to be a relatively happy one. In 1997, Lott and Mustard gave great prominence to an issue that was not widely known in the scholarly world – the NRA-led initiative to seek adoptions of state RTC laws – as they used a panel data set to provide the first step in analyzing the impact of these laws. Their strong conclusion that the laws reduced crime was counter-intuitive enough and soon seen to be politically salient enough that numerous other researchers stepped in to evaluate their data, which the two researchers creditably shared broadly. The first stage of this process of re-analysis proved to be inconclusive: one set of researchers pointed out potential problems in the analysis and interpretation of Lott and Mustard's 1977-92 data set while other researchers concluded that Lott and Mustard's initial results seemed robust to differences in specification and inclusion or exclusion of various explanatory variables.[24] In the second and perhaps final stage in the process of re-evaluation, researchers have been able to show, with the benefit of more complete data and/or superior econometric techniques, that the initial Lott and Mustard findings of crime reduction were largely spurious.[25]

Unfortunately, though, there is a dark side to this story. While it took about six years for the scholarly community to fully discredit the more guns, less crime hypothesis, the Lott and Mustard research had a major influence on public policy as a number of states adopted RTC laws recently with legislators touting the research of Lott as alleged proof that their action will cut violent crime.[26] Some may argue that Lott and Mustard's

---

[23] Committee on Law and Justice, National Research Council. Firearms and Violence: A Critical Review. National Academies Press. Washington D.C. December 2004.

[24] Among the articles not written by Lott and Mustard that support the more guns, less crime thesis are: Bruce Benson and Brent Mast, "Privately Produced General Deterrence" Journal of Law and Economics 44(2): 1-22, October 2001; and Carlisle E. Moody "Testing for the Effects of Concealed Weapons Laws: Specification Errors and Robustness" Presented at the Conference on Guns, Crime, and Safety, December 10-11, 1999 at the American Enterprise Institute, Pages 1-17, December 20, 2000.

Articles that disagreed with the Lott and Mustard findings include: Dan A. Black and Daniel S. Nagin, "Do Right-To-Carry Law Deter Violent Crime?" Journal of Legal Studies 27 (January 1998): 211; Jens Ludwig, "Concealed-Gun-Carrying Laws and Violent Crime: Evidence from State Panel Data," International Review of Law and Economics 18 (1998): 242; Zimring, Franklin and Gordon Hawkins. 1997. "Concealed Handguns: The Counterfeit Deterrent." The Responsive Community 7(2): 46-60; Dezhbakhsh, Hashem and Paul H. Rubin. 1998. "Lives Saved or Lives Lost? The Effects of Concealed-Handgun Laws on Crime." American Economic Review 88(2): 468-74; and Duggan, Mark. 2001. "More Guns, More Crime." Journal of Political Economy 109(5): 1086-1114.

[25] See also Willard Manning, *Comment* to John J. Donohue, *The Impact of Concealed-Carry Laws*, *in* EVALUATING GUN POLICY, *supra* note 2, at 331 (suggesting that correcting Lott and Mustard's results for autocorrelation would render all of their results statistically insignificant).

[26] From 1996-2000, there were no adoptions of RTC laws but since then Michigan adopted such a law in 2001, and Missouri, New Mexico, Minnesota, and Colorado all followed suit in 2003, with other states actively considering adoption. Alaska, which already has a RTC law that requires those wishing to carry

academic research was only window dressing and that it did not change any legislative outcomes, but this may be too optimistic a conclusion. At the very least, Lott's research and subsequent lobbying efforts gave cover to those who might have been reluctant to support RTC laws and emboldened their supporters to push harder for such laws. Indeed, Attorney General John Ashcroft asked the U.S. Supreme Court to adopt a more NRA-friendly interpretation of the 2d Amendment using Lott's research to argue that more guns would lead to less crime. In addition, Lott has inspired an entire cadre of gun-toters to believe that they are responsible for one of the most important benign trends in crime in American history (the large crime drops of the 1990s), even though the yet larger crime drops in the states that did not adopt the RTC laws should show the folly of that belief. As pro-gun groups like Guns Save Lives.com have sprouted across the country, these excitable and engaged (and armed) supporters of Lott's work have been highly resistant to the refutations of the more guns, less crime thesis, and have been energized to greater political activity on behalf of the NRA agenda. Of course, if RTC laws are harmful and Lott and Mustard's now discredited work has led to their greater adoption, then Lott and Mustard have imposed serious costs on the victims of the increased crime. Conversely, if the RTC laws have virtually no effect on crime but legislators voted for them thinking that they lowered crime, then at least there would be no blood on Lott and Mustard's hands but there would still be the harm to the democratic process of encouraging the adoption of laws on false pretenses (however innocent the erroneous findings originally were).

Thus, we have conflicting lessons from this episode. The benign lesson in the scholarly realm is that those who ask interesting and important questions may help stimulate the ultimate attainment of truth even if they themselves generate the wrong initial conclusion.[27] The lesson in the political realm, though, is far more malign. At least until the truth emerges – and perhaps even after it has become clear to open-minded scholars – those in the political realm who wish to push a particular agenda will do so as soon as a superficially supportive academic study hits the stands. Even after the study has been discredited, it may still have the capacity to provide cover for supporters of the erroneous conclusion, and it may still be cited and adamantly defended by fellow travelers.

Legislators and policymakers must keep in mind that it is very difficult to ascertain the effect of a law on a complex social phenomenon such as crime, and that regardless of the sophistication of the study, no one should have complete confidence in any study that has not been fully vetted by independent scholars preferably with both more complete data and better methods (as in the Ayres and Donohue paper) as well as using wholly different data and analytical approaches (as in the Kovandzic and Marvell study). Indeed, on at least three occasions I have found serious coding errors in the work

---

guns to secure a permit, has now adopted a law allowing anyone who can lawfully carry a firearm to do so without a permit.

[27] "Mr. Lott's 1997 paper on gun policy was, "to that point, the most important piece of empirical research that has ever been done in the social sciences," says Jeffrey S. Parker, a professor of law at George Mason University. "I doubt that even Ayres and Donohue would dispute that point."" David Glenn, "Scholarly Debate Over Guns and Crime Rekindles as States Debate Legalization," The Chronicle of Higher Education, April 30, 2003. I do think that Lott's work was important in that it stimulated the ultimate understanding that while RTC laws do not reduce crime, neither do they vastly increase it (since any overall crime increases are likely to be in the range of 1 to 2 percent).

of John Lott, which, when corrected, cut strongly (and in some case, overwhelmingly) against his thesis, which underscores the need for independent verification if the truth is to emerge.[28] Unfortunately, one cannot expect coding errors to be captured during the process of peer review, and perhaps still more unfortunately, John Lott has neither conceded the existence of these errors nor tried to correct them.[29]

It is also important for the political and scholarly audiences to be sensitive to signs of over-zealousness on the part of researchers as this may give clues that something more than the search for truth is motivating the research. Because it is easy to make mistakes that can undermine one's analysis – I assume the three sets of coding errors were honest mistakes -- it is important for scholars to quickly correct errors that they have introduced into political debates, and to be humble about pressing policy responses to their research until sufficient scholarly re-evaluation has been completed. This can take time – six years in this case. It also suggests the wisdom of having sunset provisions attach to legislation when some of the supporting legislators vote based on empirical studies that have not yet been fully vetted.[30] Since a high proportion of these studies turn out not to withstand scrutiny, it does seem to be a blight on our democracy to have erroneous studies saddle the electorate with laws that would not have been adopted had the truth been known at the time of passage. As the eminent sociologist Otis Dudley Duncan has stated: "The Lott episode is just one incident in a seemingly inexorable trend toward eliminating professionally competent research from discussions of social policy or overwhelming it with junk science. If that trend is not halted, the life blood of democracy itself will dry up. The people cannot make sensible choices without reliable information."[31]

---

[28] These coding errors are detailed in Ayres and Donohue (1999), n. 3 supra, and in Ian Ayres & John J. Donohue III, "The Latest Misfires in Support of the "More Guns, Less Crime" Hypothesis," 55 STAN. L. REV. 1371 (2003).

[29] One article noted that "Lott also points out that because the claim of coding errors appears in a law review, it has not been subject to review by third-party scholars, as would have been the case in a peer-reviewed economics journal." David Glenn, "Scholarly Debate Over Guns and Crime Rekindles as States Debate Legalization," The Chronicle of Higher Education, April 30, 2003. But Lott doesn't need anyone else to evaluate the claim. He can simply look at the Ayres and Donohue paper and concede (or refute) the claim of coding error, and concede (or refute) that its correction eliminates his more guns, less crime result. The Glenn article then goes on to note that: "Six tables that derive from the same allegedly miscoded data set appear in Mr. Lott's new book, The Bias Against Guns: Why Almost Everything You've Heard About Gun Control Is Wrong (Regnery, 2003). James Lindgren, a professor of law at Northwestern University, says, "If Donohue and Ayres's account is as it appears -- and I'm not in a position to judge that -- then Lott should withdraw the book for revision."" Indeed, Lindgren was the scholar who raised questions about misconduct on the part of Michael Bellisles, who resigned from Emory University amidst charges of academic fraud in his left-leaning gun research, as well as against John Lott, on claims that Lott may have made up a survey purporting to show that 98 percent of the time that guns are used defensively they are only brandished and not fired. As UCLA Professor Mark Kleiman has written: "If Lott were at a university, he would certainly be facing an inquiry into his professional ethics." http://www.markarkleiman.blogspot.com/2003_05_01_markarkleiman_archive.html#200233924

[30] Indeed, absent reenactment, the federal ban on assault weapons that was adopted in 1994 expired after 10 years in September of 2004, so there is clear precedent for including sunset provisions in federal weapons bans (even if there is no similar precedent for sunset provisions in state pro-gun laws, which may be even more important given the emerging technology issue discussed below).

[31] Otis Dudley Duncan, "John R. Lott, Jr. on Defensive Gun Use Statistics, " April, 11, 2003, http://www.cse.unsw.edu.au/~lambert/guns/duncan3.html.

But another element of the story – mimicking the discussion of the research on the deterrent effect of the death penalty – is how malleable the empirical research on the impact of RTC laws can be. Once again, ostensibly small changes in the panel data models can generate very different answers of the impact of RTC laws. To illustrate this fact, I present a number of different specifications in Figures 7 – 15 with each figure estimating the effect on each of the nine FBI Index I crime categories. Because of the powerful ideological motivations of many gun researchers, a legitimate fear is that an analyst trying to prove a certain point might choose among a vast array of possible statistical models simply to generate a desired result. To address this concern, I report not only a modified version of Lott's original model (called the "Modified Lott" set of explanatory variables),[32] but also the results of three other models that were developed by researchers to answer questions having nothing to do with RTC laws – one by Wentong Zheng (developed to look at the impact of lotteries on crime),[33] one by William Spelman (developed to look at the impact of incarceration on crime),[34] and one by John Donohue and Steve Levitt (developed to look at the impact of abortion legalization on crime).[35] These four different sets of explanatory variables are set forth in Table 6.[36] Whatever infirmities these last three models have, we know that they were created by serious academics without any attention to skewing the estimates of the impact of RTC laws. When we add a variable identifying the date of adoption of the RTC laws to these pre-existing statistical models, we can see if the results support – or refute – the more guns, less crime hypothesis.[37]

---

[32] The "Modified" Lott model starts with Lott's original set of explanatory variables and replaces one particularly questionable variable – the arrest rate – with a lagged incarceration rate. The problems with the arrest rate as Lott used it are numerous: the variable is badly measured, not exogenous (that is, it is not fair to assume that arrest rates influence crime since crime in year t will also influence the number of arrests in year t divided by the amount of crime in year t – thus, the contemporaneous crime rate ends up on both sides of the regression equation).

[33] Computer program supplied to author by Wentong Zheng, Graduate Student in Economics, Stanford University.

[34] Spelman, William (2000), "The Limited Importance of Prison Expansion," in *The Crime Drop in America*, edited by Alfred Blumstein and Joel Wallman (Cambridge, UK: Cambridge University Press).

[35] Donohue and Levitt, "The Impact of Legalized Abortion on Crime," Quarterly Journal of Economics (Vol. CXVI, Issue 2, May 2001), pp. 379-420.

[36] All of the estimates presented in this paper are based on state crime data, which is the approach used in the work by Zheng, Spelman, and Donohue & Levitt. While Lott does present estimates of the impact of RTC laws using state data, he prefers to use county data. I have been persuaded, though, that the county crime data is considerably less accurate than the state data, and since the intervention of interest – the adoption of a state RTC law – applies at the state level, I am more comfortable with the state data than with the county data.

[37] Interestingly, one might think that identifying the date of adoption of a RTC law would be easy, but even on this issue there is disagreement. It is not always straightforward to identify the precise date among a number of competing statutory enactments, when citizens of a state had a right to carry concealed handguns without demonstrating the need for a gun to a governmental official. Appendix Table A presents a number of different codings of RTC adoptions, and one can see that there are a number of differences across the various authors. This paper will rely on my latest coding scheme identified in the fourth column of Table A. Luckily, since most of the disputed dates are in small states whose impact in a population-weighted regression will necessarily be smaller, the coding issues do not seem to influence the results too strongly. Nonetheless, the coding issue illustrates how even simple issues become complex and disputed in empirical work.

Table 6.

	Stanford Law Review (Modified Lott)	Zheng	Donohue-Levitt	Spelman
Control Variables:				
Demographic:	36 race-age categories	%black %15-17 %18-24 %25-34		%black %0-14 %15-17 %18-24 %25-34
	Population size	Population size		
	Population density			
		%metro		%metro
Economic:	Personal income per capita	Income per capita	log(per capita income)	log(per capita income)
	Unemployment insurance per capita			log(unemployment rate)
	Per Capita income maintenance			
		Poverty rate	Poverty rate	
Criminal:	Lagged incarceration rate	Lagged prisoners per capita	log(lagged prisoners per capita)	log(lagged incarceration rate)
		Lagged police per capita	log(lagged police per capita)	log(police per capita)
Other:		Alcohol consumption per capita	Effective abortion rate	
		Governor party affiliation dummies		
State fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Regression Type:	OLS	OLS	OLS, adjusted for serial correlation with fixed effects (Bhargava)	OLS, standard errors adjusted for clustering at state-level
Additional notes:		No D.C.		

Let's start with the one crime – robbery in Figure 10 – for which we would expect to see a *reduction* in crime if the more guns, less crime hypothesis were true. The dummy model is the simplest statistical model, which simply tries to determine whether on average and controlling for the various explanatory variables of Table 1, crime was higher or lower after RTC adoption. These before and after estimates can be generated in

either of two ways – as a single estimate for the aggregate of all adopting states, or as multiple individual estimates for each adopting state (which can then also be averaged). Beginning with this dummy model and using the Modified Lott set of explanatory variables, we see four columns on the far left of Figure 10. The first thick column in the figure shows that four more states had positive estimates – implying that the RTC law *increased* crime – than had negative estimates (suggesting crime decreases).[38] One might argue that the first column numbers are less meaningful because they count positive and negative estimates that are not statistically significant in calculating the overall difference. The last of the four columns above "Dummy" in the "Modified Lott" portion of the figure limits the analysis to statistically significant estimates and finds that three more of the state estimates are positive than are negative. Column 3 takes a population weighted average of all the estimated state-specific effects and highlights whether the result is statistically significant by using a bright (dark) color. Column 2 avoids getting individual state estimates and simply generates an aggregated statistical estimate from the overall model. For this model (dummy using Modified Lott), neither of the estimates of the aggregated effects of the RTC laws is negative or statistically significant, and thus there is no support in that model for the more guns, less crime hypothesis with respect to robbery. Indeed the aggregate of the state specific effects (column 3) is close to 4 percent, which, if true, would indicate a rather substantial *increase* in robbery.

The same four column estimates are generated for the second model, which replicates the first, with one exception – rather than just looking at whether crime is lower or higher, the "dummy with state trends model" tries to see whether crime simply followed a pre-existing trend at the time of adoption. Here the results are terrible for the Lott model: columns 1 and 4 reveal that most states experienced crime jumps (whether one looks at all states or only those with statistically significant estimates) and columns 2 and 3 suggest that robbery increased by over 5 percent (with both estimates being statistically significant).

Looking to the third set of "Modified Lott" models – the spline model – we see that one estimate (that in column 2) suggests that crime goes down while the other estimates suggest the opposite. The spline model attempts to ascertain whether there is a change in the trend of crime after adoption.

These figures were constructed so that visually one would get a sense of whether the estimated crime effects were positive (in which case the RTC laws would increase crime and the columns would go upward) or negative (in which case the opposite would be true). The overwhelming impression from Figure 10 is that robbery *increases* crime. Yet, despite the fact that only 8 of the 48 columns are negative, one can immediately see the danger from someone who wishes to portray the data tendentiously. All one has to do is cherry-pick the estimates and present the aggregate spline model (as opposed to the weighted mean of the individual state estimates) using the Modified Lott or Zheng variables, which generates statistically significant negative estimates in the neighborhood of 1 to 2 percent. Of course, such conduct would violate statistical conventions, but this showing does reveal that even when the evidence points strongly in one direction, there are often scattered estimates that cut in the other direction against the weight of the evidence.

[38] The numbers associated with the first and fourth columns should be read from the index on the right hand side of the figure.

Figure 7

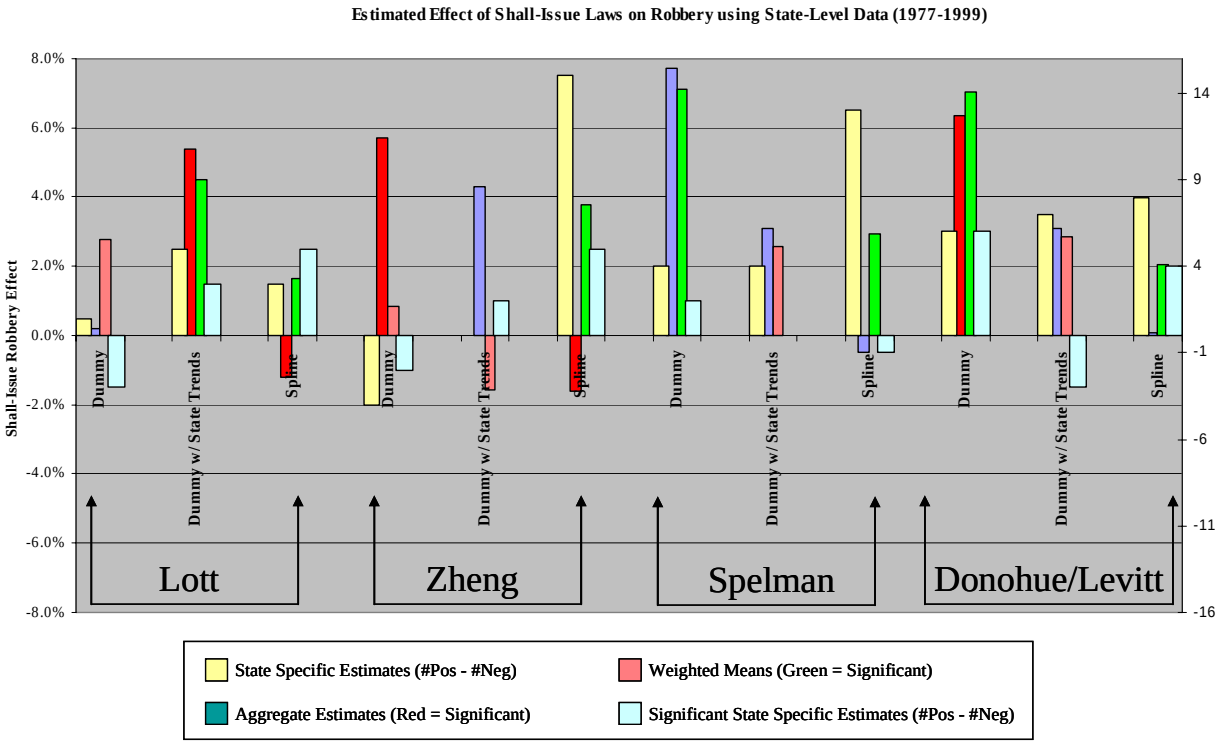


Figure 8

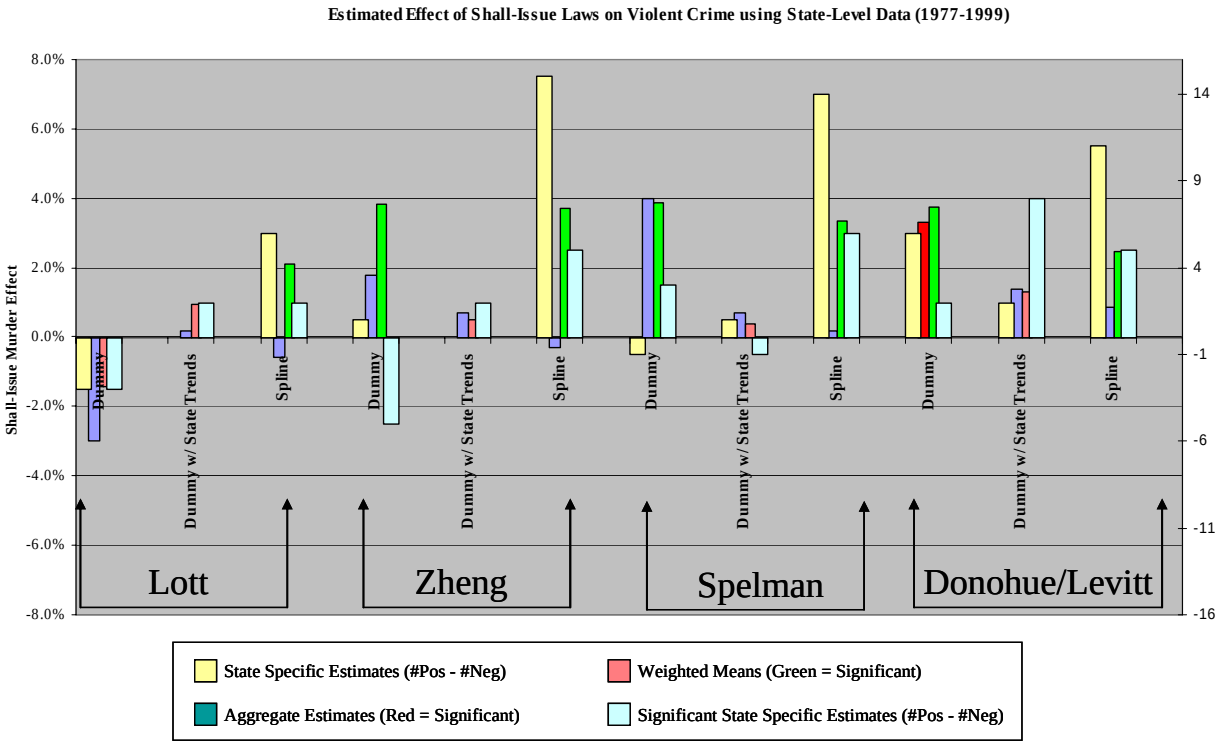


Figure 9

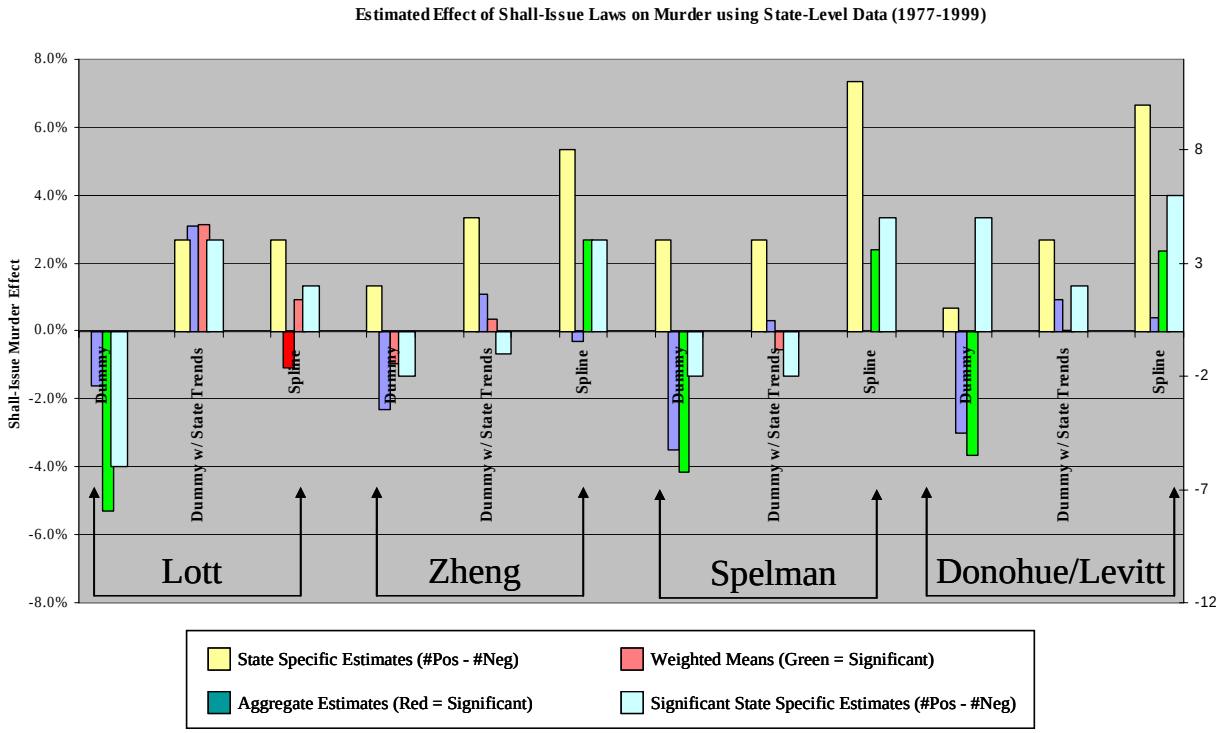


Figure 10

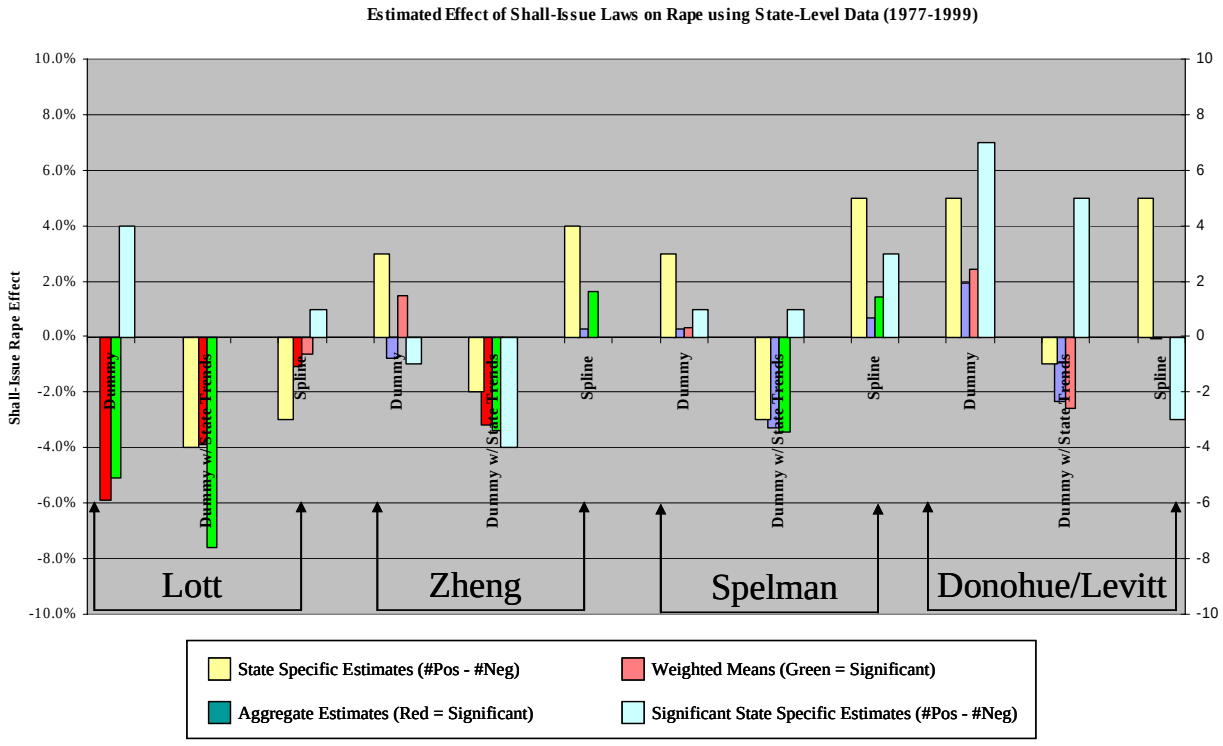
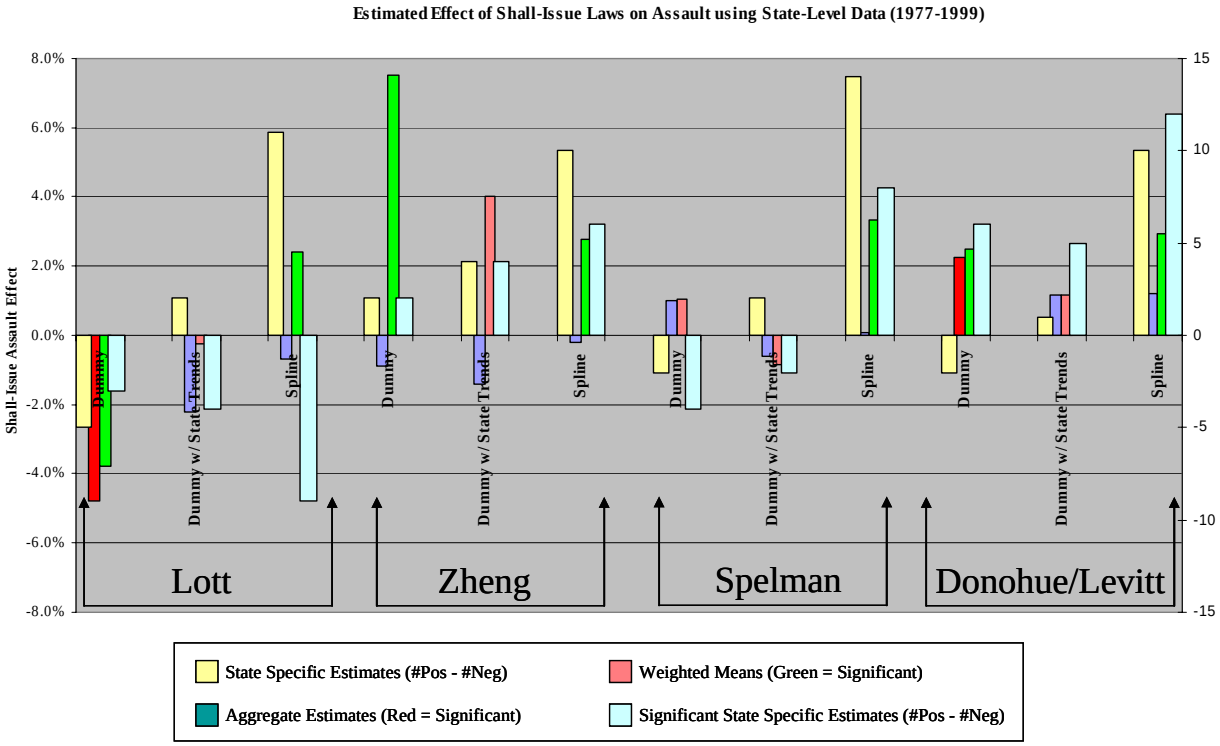


Figure 11



If one looks at the assault and even more powerfully at the violent crime figures, the evidence again supports – if anything – a positive relationship between RTC laws and crime. But again, a tendentious presentation of the data could find negative estimates, including a negative and statistically significant estimate of the impact of RTC laws on assault. On the four property crime categories, the estimates are overwhelmingly in the direction of suggesting *increases* in crime, but even there one could find two negative and statistically significant estimates for burglary.

In fact, the only two crimes for which the visual impact of the corresponding Figure is not overwhelmingly suggestive of a crime *increase* rather than a decrease are murder and rape. My basic read of the evidence on murder is that it provides no evidence of a drop in murder resulting from the adoption of the RTC law, despite the one statistically significant negative estimate. Given that there are 24 different estimates, the fact that only 1 of 24 is negative and significant, while 13 of the 24 estimates are *positive*, is more suggestive of a random influence rather than a true impact of the law. Indeed, looking at the state-specific estimates, one finds that 11 of the 12 sets of estimates show more states have crime *increases* (and 8 of 12 have more states with statistically significant positive estimates).

For the crime of rape, however, the visual impression is that RTC laws are associated with *lower* rates of crime. This is not to suggest that there are no estimates suggesting crime increases but the weight of the statistical evidence in Figure 10 seems to be suggestive of crime reduction.

For the four property crimes (although least powerfully for burglary), the data is suggestive only of crime increases associated with the adoption of the law. At this point, if one were to rely on the statistical evidence from these four sets of standard fixed effects panel data models – about which we will raise some concerns presently – one would probably reach the following conclusion: judging impressionistically, adoption of RTC laws seems to be associated with crime *increases* in all crime categories except murder, where the mixed evidence is probably most consistent with there being no impact, and rape, where the evidence is suggestive of crime decreases.

Figure 12

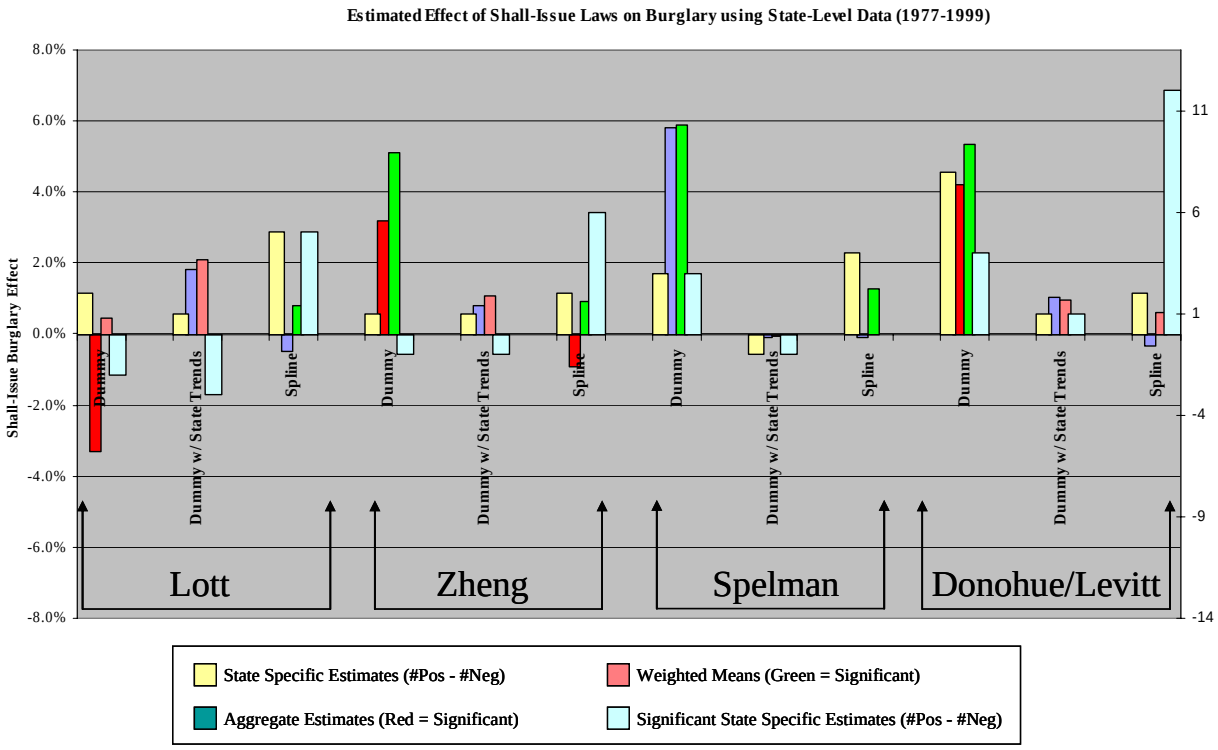


Figure 13

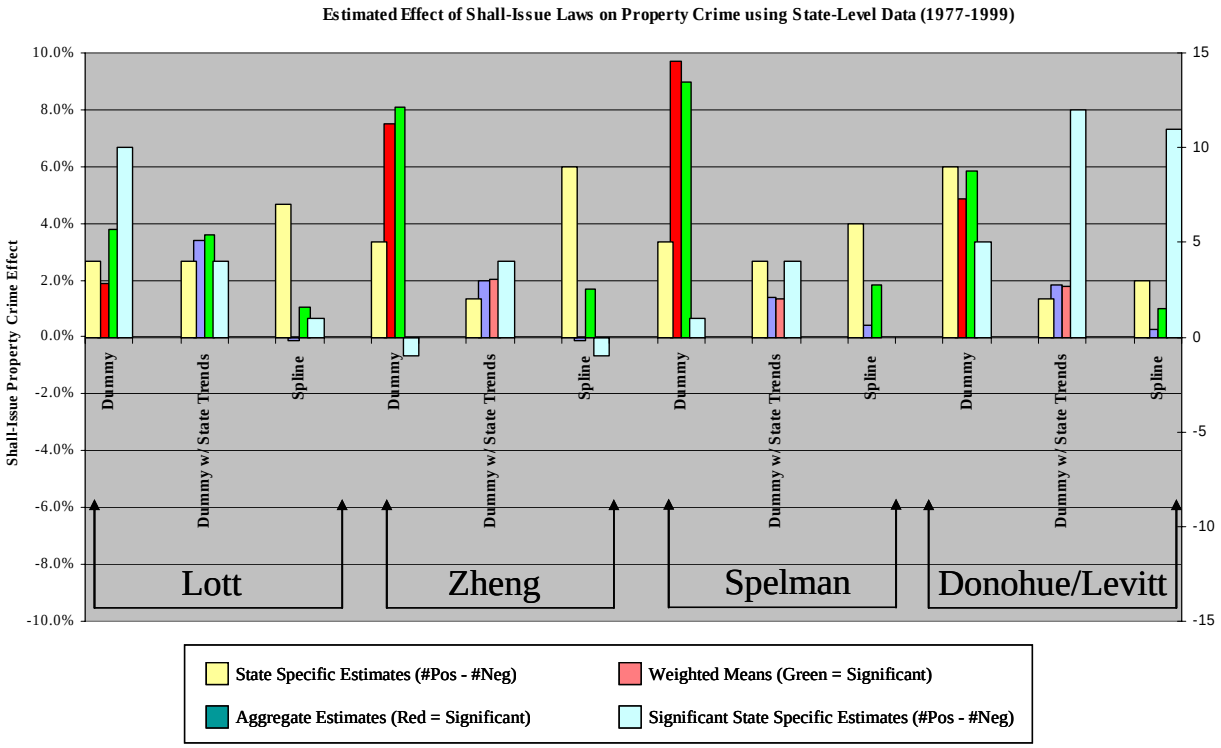


Figure 14

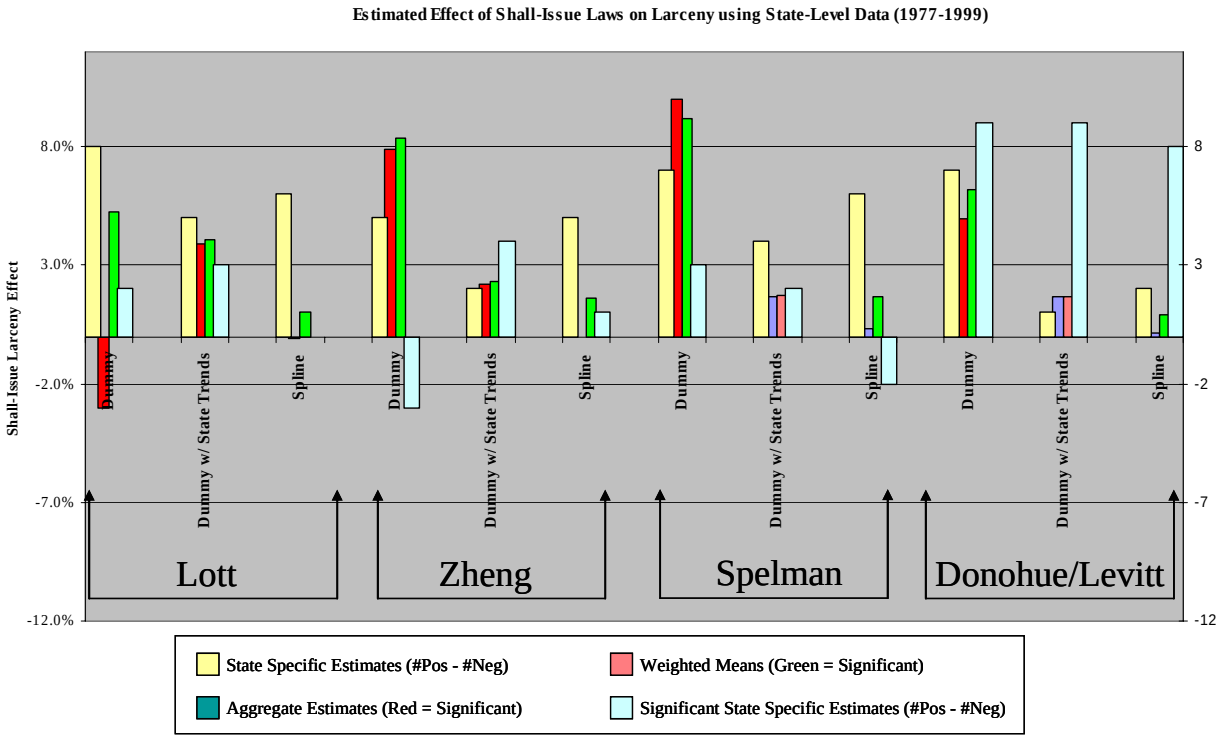
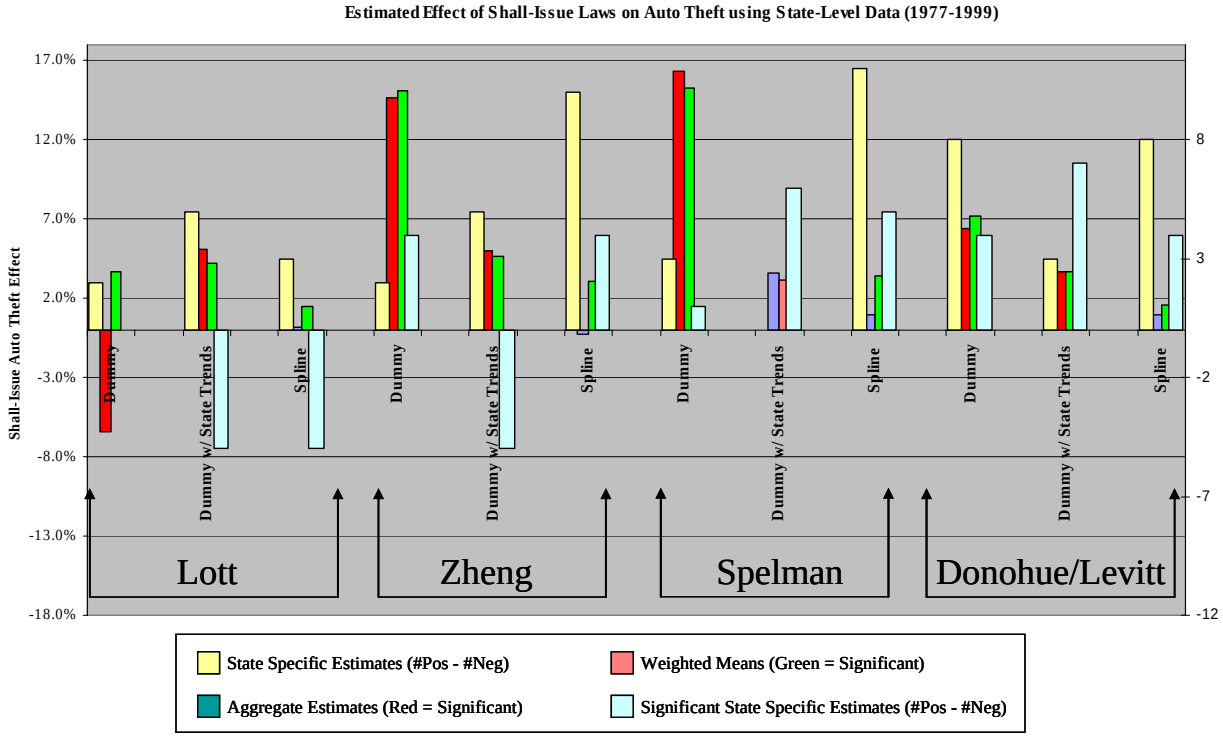


Figure 15



Thankfully, Jeff Strnad has tackled this morass of conflicting estimates by showing how Bayesian statistical techniques can provide a method for sorting among them.[39] But the current academic and political debate is being conducted at a far lower level of sophistication than that presented in his paper, and the journals are currently filled with articles in which researchers pick one of the hundreds of estimates that appear from the figures above, and argue that their preferred estimate is the one true effect of RTC (or other) laws. Moreover, while top level researchers like Strnad and Dan Ho will show the range of estimates that their models produce, current practice as judged by a review of hundreds of empirical papers shows that more commonly, one estimate is trumpeted followed by, in the stronger papers, a modest effort at establishing the robustness of the estimates. Unless some mechanism for establishing a brand name for the reliable researchers is found, most readers will simply be at a loss to know how much weight to attach to such studies. One institutional change which clearly represents an important step in the right direction is the decision by the American Economic Review to require all papers published there to provide data sets and program files that are then made available to the public. This enables the type of careful replication and probing of results that is essential to define the contours of reliability of modern econometric work. Already, this practice has uncovered many cases of data and computer coding errors that influence the outcome of important papers.

[39] Strnad, *supra* 14.

As the science of statistics advances in the next two decades, as it will very dramatically, we may get closer to the goal of having a protocol that can lead us to the correct statistical model for any question. At present, though, the component of empirical work that is art rather than science is still very high. The importance of functional form can be illustrated with the following simple example. Assume you had data on the areas of 30 circles, ranging from 1 to 30 in radius, and you wanted to use regression to estimate the area of the circle. This is a pretty simple case since we know there is only one explanatory variable that influences area – radius – and we can easily measure that variable. But if you simply regressed area on radius (rather than the theoretically appropriate regression of area on radius-squared), your estimated equation would be AREA = -519.4 + 97.4 * RADIUS. Even though this equation yields high t-statistics and an adjusted R-squared of .94, it is a rather poor equation. (Note that when the radius is small, this equation predicts a theoretically impossible negative area.) With the correct specification of the functional form, however, one would obtain the correct equation: AREA = 3.14159 * RADIUS-SQUARED. The lesson should be obvious: even when trying to estimate an equation where one knows that the dependent variable is fully explained by a single perfectly measured explanatory variable, one can get very disappointing regression results if the wrong functional form is employed. Since in most of the real-world empirical studies, very imperfect data is being used and we don't know and/or can't quantify all of the relevant explanatory variables AND we are uncertain about the proper functional form of the equation, one must entertain the prospect that whatever equation we come up with may deviate in unpleasantly large ways from the true but unknown equation.

At times, the debate over regression has an aspect reminiscent of the protracted vote count in the 2000 Presidential election. Those who wanted to limit the recount to re-running the punch cards through the vote counting machines perhaps realized that they were advocating a technology that was less precise than individual counting but preferring it because they feared that a hand recount would be more susceptible to partisan manipulation. Perhaps then regression might be analogized to the voting machines – imperfect, but approximately right, and preferable to other methods because it is less subject to manipulation. But the analogy is imperfect for two reasons. First, there is a reasonable alternative to use of the voting machines – one can count by hand. There is really no alternative to using regression for certain tasks. Even if regression can only imprecisely estimate, say, damages in an antitrust case, there is really no other tool that can do a better job. While regression is needed, the voting machines are not. In fact, if your life depended on an accurate vote count, hand counting would be preferable to the voting machines (albeit far more costly). Second, the voting machines are offered as a device to protect against manipulation, but this book tries to suggest that it is not only easy to get the correct answer through regression, but the tool is also subject to manipulation. The legal system seems to want to treat regression as a mystical tool, perhaps so that confidence in the outcome of trials will be greater.

V. Conclusion

The story of the debate over the death penalty and RTC laws suggests the need to re-evaluate how the public should respond to econometric studies. First, it must be underscored that a single study simply cannot resolve a public policy issue. An important study will initiate efforts at replication and extension and critical probing. It is the literature, not the first study, which will ultimately answer the public policy question. This realization will enable a more mature attitude to develop towards econometric research. If scholars begin to think of their work as an effort to shed light on the truth, they will be less defensive if subsequent work undermines their initial hypotheses. John Lott could be thought of as an important initiator of a literature, but he, and many commentators, seemed to think that if his study did not stand up to subsequent investigation, he was to be faulted. But the nature of econometric analysis is that even if you do everything perfectly, random influences will cause a factor that has no impact on some variable of interest to appear to be causally related 5 times out of 100. It seems a bit unfair to pillory someone who has done everything perfectly if they are a victim of chance events. If in fact Lott did falsify survey data to support his pro-gun views – as some prominent academics have suggested – then he is to be faulted for elevating ideology and/or ego over the scientific process. But the public needs to be clear that Lott's econometric work was not obviously flawed by the standards existing and the data available when he first conducted his initial analysis. His data ended in 1992 after a highly unusual crack-induced crime wave, which correlated with the adoption of right to carry laws, which led to a powerful but spurious correlation. Only when additional years of data were added was it easy to see that the initial estimates would not hold up.

Second, the need for humility in discussing results of econometric studies is particularly important. The 5 percent Type 1 error rate was just offered as a reason to be forgiving to those who have done fine work but have been victims of chance correlations. At the same time, it should encourage researchers to be humble in drawing conclusions, especially since the 5 percent error rate is only the tip of the iceberg. Problems with data can cause that error percentage to jump. Part of the reason for the explosion in econometric studies is the greater availability of data online, but dangers sometimes lurk in readily available datasets. For innumerable reasons, data that appears solid on its face can be quite severely flawed.

For example, substantial academic work has relied on the rich data sets that provide information concerning the frequency of abortion in the United States.[40] Researchers have tended to view this data as solid, and have frequently relied on it in an unquestioning manner. But as a report issued from one of the major sources of abortion data notes: "The collection of health and vital statistics in a country as vast and decentralized as the United States is a massive undertaking, fraught with problems of definition, compilation and verification. State reports, which form the basis of the collection effort in this instance, are often inadequate." The exact same is true for crime data and even more so for arrest data, where the FBI data that is relied upon so universally is full of holes. States not uncommonly fail to turn over complete data, and at times turn over no data at all. Yet the numbers get churned out despite the flaws, and researchers often use them with little appreciation of these weaknesses.

[40] AGI, "The Limitations of U.S. Statistics on Abortion," (January 1997), http://www.guttmacher.org/pubs/ib14.html: 01/1997

Indeed, there are times when the incentives to collect good data are actually undermined. The NRA constantly lobbies to prevent data collection that might impair their pro-gun agendas, and politicians frequently comply by imposing various direct bans on data collection or retention that might be used in studies that have conclusions that prove harmful to NRA interests. At other times, Congress creates inadvertent incentives to undermine data quality: for example, the welfare reform law -- the Personal Responsibility and Work Opportunity Act of 1996 – provides up to five "bonuses" of $20 million each to states that can show a reduction in the number of out-of-wedlock births in the 1998 fiscal year, along with a lowering of the state's abortion rate (with fiscal year 1995 as the comparison year). I wonder what the easiest way to show such a reduction in abortions might be when being a bit slack in the exhaustive process of collecting data from diverse sources may be rewarded highly?[41] As the Alan Guttmacher Institute has noted: "Given the chronic underreporting (or, in some cases, nonreporting) of abortion statistics by the states, the legislation—in essence—provides them with an incentive to be even less diligent in the future in producing accurate and complete data on abortions obtained by residents of their state." The problem is widespread. For example, the government-sponsored National Survey of Family Growth, which is a widely used source of data on pregnancy and contraceptive use in the United States, has been shown to dramatically undercount abortion at times. Women responding to the 1988 survey reported abortion frequency at a rate that was only 35 percent of the number of abortions known to have occurred at that time. Strenuous efforts were made in subsequent years to improve the reporting but the problem becomes clear: either the efforts succeeded, in which case, abortion would be deemed to have risen over the seven year period, or they failed or were only partially successful, in which case more inaccurate data was released.

Just recently, a CDC study of 14 states revealed that 16.8 of every 1000 eight-year old boys in New Jersey had some form of autism. The state with the second highest number, Utah, had only 12.7 cases per 1000 boys, and the last place state, Alabama, had only five. Immediately, empiricists started probing whether environmental insults might be greater in New Jersey, but my first suspicion would be that more cases were being counted as autism in New Jersey than in Alabama. The CDC researchers have now acknowledged that they were provided with 8 diagnostic records to make the judgment of autism in New Jersey, while other states only provided them with one or two.[42] Of course, if one's data is flawed, and data is always imperfect in some way, then the prospect for Type 1 error may rise far beyond the 5 percent level that one would expect when everything is done perfectly.

---

[41] The legislation acknowledges the danger in a section entitled "Disregard of Changes in Data Due to Changes in Reporting Methods," which states: "In making the determination required by subclause (I), the Secretary shall disregard … any differences between the rate of induced pregnancy terminations in a state for a fiscal year and such rate for Fiscal Year 1995 which is attributable to a change in State methods of reporting data used to calculate such rate." But this concern is only likely to ferret out the most crude and transparent attempts at data manipulation. Mayors who want crime to go down have not infrequently put pressure on the police not to count all crimes, and I suspect that when governors can similarly profit by showing a decline in the number of abortions, some will put pressure, or perhaps even simply cut budgets of the abortion counters, to make the numbers go down.

[42] Tina Kelley, "An Autism Anomaly, Partly Explained," The New York Times, Section 4, page 2 (February 18, 2007).

Given the fact that many observational studies claiming important effects of certain foods, drugs, or medical treatments have been undermined when randomized experiments were conducted, one has to be aware of the risk that similar problems lurk in the observational studies in the realm of social science and law.[43] This suggests the need for thinking more about using randomized experiments on such important issues. Once one adds on the problems of computer coding errors (which, given the enormous amount of data extraction and manipulation that these studies require, will be common even with small rates of error for individual maneuvers), limitations of uncertainty about model selection, imperfections in research design, endemic problems of uncertain magnitude of endogeneity in the adoption of laws and policies, and interpretive difficulties as the tools of statistics evolve in ways that researchers may not have anticipated or fully grasped, the need for humility in discussing empirical research should be strongly emphasized. Law and policies have important effects, but teasing out those impacts from observational data is one of the most difficult tasks being undertaken anywhere in the academic and policy world.

[43] The fantastically interesting results from the randomized Moving to Opportunity Program, analyzed by Jeff Kling, Larry Katz and other top researchers, shows that randomized results have in fact over-turned strong social science predictions from observational studies.