Evidence-Based Criminal Justice Policy: Summon the Randomistas?[*]

Andrew Leigh
Economics Program, Research School of Social Sciences
Australian National University

Speech to the NSW Bureau of Crime Statistics and Research
40th Anniversary Symposium

19 February 2009
Powerhouse Museum, Sydney

[*] This talk is given in my capacity as an Associate Professor at the Australian National University. The views expressed here are not those of the ANU or any other organisation. Email: andrew.leigh@anu.edu.au. Web: http://econrsss.anu.edu.au/~aleigh/.

Introduction

For the past ten years, I have taken a multivitamin pill once a day with my morning coffee. Like any good morning habit, it's a comfortable and familiar routine. But I've gotten a warm glow to know that each day begins with a simple act that helps take care of my body.

Last month, a friend suggested that a have a look at an article published in the *Journal of the American Medical Association*.[1] The authors set out to answer the question: do vitamin supplements make you live longer? To answer this, they drew together all the available A-grade evidence in the world; which in this case meant the randomised trials of vitamins A, C and E, beta carotene and selenium.

Just comparing those who choose to take vitamins with those who do not might give biased estimates, since those who take vitamin pills might also do other things to look after their bodies. But with randomised trials, you know that you are getting true causal effects. Putting together 68 randomised trials, the authors found no evidence that vitamins make you live longer. If anything, those who took vitamins seemed to live shorter lives.

Not wanting to send myself to an early grave, I stopped taking multivitamin pills.

What makes me so sure about this decision? First, because the evidence comes from a study published in one of the world's leading medical journals. Second, because medicine has a well-established hierarchy of evidence. Grade I evidence in that hierarchy is 'well conducted systematic reviews of randomised trials'. (Grade II are non-randomised controlled trials and uncontrolled experiments, while Grade III are descriptive studies and case reports.) There is a strong consensus in the medical profession that when it comes to questions like 'are vitamins good for you?', you cannot do better than a systematic review of randomised trials.

The Counterfactual Problem

Yet suppose I was a policymaker, charged with deciding whether to scrap or expand a criminal justice program. For many programs, it would probably turn out that the quality of the available evidence is not as good as the evidence on the efficacy of multivitamins. In the case of many criminal justice programs – old, new, or hypothetical – it is sadly the case that we cannot say with confidence whether or not they work.

Uncertainty over the effectiveness of our criminal justice policies is even more striking when you realise that – via our governments – every Australian spends an average of $50 annually

[1] Goran Bjelakovic, Dimitrinka Nikolova, Lise Lotte Gluud, Rosa G. Simonetti, and Christian Gluud. 2007. 'Mortality in Randomized Trials of Antioxidant Supplements for Primary and Secondary Prevention: Systematic Review and Meta-analysis'. *Journal of the American Medical Association*. 297(8):842-857. Available at http://jama.ama-assn.org/cgi/content/full/297/8/842. For an informal discussion of the issue, see Norman Swan's interview with one of the researchers on 5 March 2007, available at http://www.abc.net.au/rn/healthreport/stories/2007/1861068.htm. The researchers were at pains to point out that their findings should not be extrapolated to foods that are rich in vitamins, such as fresh fruit and vegetables.

on courts, $100 on jails, and $300 on policing.[2] Knowing more about which policies work and why could help us to cut crime, save money, or both.

Ask a handful of experts, and you will find no shortage of new ideas. Cognitive behavioural therapy, mentoring, neighbourhood policing, better rehabilitation, drug decriminalisation, 'three strikes' laws, boot camps for young offenders, longer sentences, shorter sentences, tougher jails, and milder jails are just a handful of the interventions that are perennially proposed.

Yet all too often, the evidence for or against a particular intervention is based upon anecdotes and case studies – what medical researchers would regard as the lowest grade of evidence in their hierarchy. While the evidence base in criminal justice policy is steadily advancing, a lack of data and an unwillingness to experiment are two major factors that hamper our understanding of what works and what does not.

In assessing any policy intervention, we need to know the counterfactual – what would have happened in the absence of the policy. If you are a farmer who is experimenting with a new fertilizer, this is pretty straightforward: put the fertilizer on every other plant, and the counterfactual is the unfertilized plants.

In the social sciences, this turns out to be a much tougher problem to address.[3] If we simply use time series variation, we may find it tricky to separate the policy change from secular changes in crime rates over time. For example, say that we wanted to estimate the impact of the 1996-97 National Firearms Agreement on gun homicides. The difficulty is that gun homicides in Australia were declining prior to the 1996-97 gun buyback, and continued to decline after the buyback. Teasing apart the effects of the buyback from these time trends is no easy task, and the conclusions can be quite sensitive to the particular assumptions that one makes about what the trends would have done in the absence of the NFA.[4]

Another approach is to construct a counterfactual by using those who choose not to participate in a program as the control group. For example, suppose that you wanted to see the impact of Neighbourhood Watch on crime rates. One evaluation strategy might compare crime rates in communities that chose to establish a Neighbourhood Watch group with those that did not. But if these two sets of communities are systematically different – say because the treatment group has more social capital than the control group – then such an approach might mis-estimate the impact of the intervention.

[2] See the Australian Bureau of Statistics (2008), 'Expenditure on public order and safety', *Year Book Australia, 2008*, Cat No. 1301.0. ABS: Canberra. Figures are for 2005-06.

[3] This is one of the reasons that I believe the social sciences should be known as the 'hard sciences' rather than the pejorative 'soft sciences'.

[4] For a more detailed discussion of this point, see Christine Neill and Andrew Leigh (2008) 'Do Gun Buybacks Save Lives? Evidence from Time Series Variation' *Current Issues in Criminal Justice*, 20(2): 145-162

The Strengths of Randomised Trials

One way of getting around these problems is to conduct a randomised policy trial, in which participants are allocated to the treatment or control group by the toss of a coin. The beauty of randomisation is that with a sufficiently large sample, you are statistically guaranteed to have two identical groups, both on observable characteristics and on unobservable characteristics. Just as in a medical randomised trial of vitamin supplements, the only difference between the treatment and control groups is the intervention itself. So if we observe statistically significant differences between the two groups, we can be sure that it is due to the treatment, and not to some other confounding factor.[5]

In Australian criminal justice, a canonical example of a randomised policy trial is the NSW Drug Court trial, conducted in 1999-2000. Since this audience is probably quite familiar with this trial, I will be brief in sketching the details. Offenders are referred to the Drug Court from local or districts courts, undergo a detoxification program, and are then dealt with by the Drug Court instead of a traditional judicial process. At the time it was established, the number of places in detoxification was limited, so participants in the evaluation were randomly assigned either to the treatment group (313 persons) or the control group (201 persons). They were then matched to court records in order to compare reoffending rates over the next year or more. The evaluation found that the Drug Court was effective in reducing the rate of recidivism, and that while it was more expensive, it more than paid for itself.[6] At a recent conference celebrating the tenth anniversary of the Drug Court, speakers broadly acknowledged the role that the Court has played in improving the lives of drug offenders and the general community.[7]

What is striking about the Drug Court trial is that it provides a ready answer to the shock jocks. Imagine the following exchange.

> Q: ‘Minister, is it true that your program spends more on drug offenders? Why should taxpayers fork out more money to put drug addicts through detox programs, when we could spend less and just throw them in jail?’
>
> A: ‘You bet we’re spending more on this program, and that’s because we have gold-standard evidence that it cuts crime. A year after release, those who went through the Drug Court were half as likely to have committed a drug offence, and less likely to have stolen. It’s probably the case that Drug Courts help addicts kick the habit. But

[5] On randomised policy trials, see for example Leigh, A. (2003) ‘Randomised Policy Trials’ *Agenda: A Journal of Policy Analysis and Reform* 10(4): 341-354; Farrelly, R. (2008), ‘Policy on Trial’, *Policy*, 24(3): 7-12; The Economist (2002), ‘Try It and See’, 2 March, pp.73-74.

[6] Lind, B, Weatherburn, D, Chen, S, Shanahan, M, Lancsar, E, Haas, M, De Abreu Lourenco, R 2002, *NSW Drug Court evaluation: cost-effectiveness*, NSW Bureau of Crime Statistics and Research, Sydney. Available at www.courtwise.nsw.gov.au/lawlink/bocsar/ll_bocsar.nsf/vwFiles/L15.pdf/$file/L15.pdf. A second study of the Drug Court has also been conducted, though this evaluation did not use random assignment. See Don Weatherburn, Craig Jones, Lucy Snowball and Jiuzhao Hua (2008), ‘The NSW Drug Court: A re-evaluation of its effectiveness’. *Crime and Justice Bulletin* No. 121.

[7] See for example Malcolm Knox (2009), ‘Applause for former drug users who turn their lives around’, *Sydney Morning Herald*, 7 February.

> even if you don't care a whit about their wellbeing, you should be in favour of Drug Courts because they keep you and your family safe.'

In the case of the Drug Court, many of us probably had an expectation that the policy would reduce crime. But high-quality evaluations do not always produce the expected result. Take the example of 'Scared Straight', a program in which delinquent youth visit jails and are lectured to by prison staff and prisoners about life behind bars. The idea of the program is to frighten young people into life on the straight and narrow.

Low-quality evaluations of Scared Straight, which simply compared participants with a non-random control group had concluded in the past that such programs worked, reducing crime by up to 50 percent. Yet after a while, some US states began carrying out rigorous randomised evaluations of Scared Straight. The startling finding was that Scared Straight actually increased crime, perhaps because youths discovered jail was actually not as bad as they had thought. It was not until policymakers moved from silver-standard evidence to gold-standard evidence that they learned the program was harming the very people it was intended to help.[8]

Another promising crime-fighting strategy is Neighbourhood Watch, which promotes citizen involvement in crime prevention through encouraging incident reporting, marking property, and conducting surveys on security. Anecdotal reports from Neighbourhood Watch organisers almost invariably report that the program has a substantial impact on reducing crime, while quasi-experimental evaluations also tend to suggest that Neighbourhood Watch worked. But evidence from randomised trials indicates no impact whatsoever. As criminologist Lawrence Sherman puts it: 'One of the most consistent findings in the literature is also the least well-known to policymakers and the public. The oldest and best-known community policing program, Neighborhood Watch, is ineffective at preventing crime.'[9]

Being surprised by policy findings is perfectly healthy. Indeed, we should be deeply suspicious of anyone who claims that they know what works based only on theory or small-scale observation. As economist John Maynard Keynes once put it in a different context, 'When the facts change, I change my mind. What do you do, sir?'[10]

A little modesty can go a long way. The great social policy reformer, US Senator Daniel Patrick Moynihan, was fond of quoting 'Rossi's Law' (named after sociologist Peter Rossi) that 'The expected value for any measured effect of a social program is zero.' Moynihan was

[8] For a review of the quasi-experimental and randomised evaluations of Scared Straight, see Petrosino A, Turpin-Petrosino C, Buehler, J. (2002) ''Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency' (Updated C2 Review). Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE), available at www.campbellcollaboration.org. See also Robert Boruch and Ning Rui (2008), 'From randomized controlled trials to evidence grading schemes: current state of evidence-based practice in social sciences' *Journal of Evidence-Based Medicine* 1(1): 41–49.

[9] Sherman, L.W (1997), 'Policing for Crime Prevention' In Sherman, L.W., Gottfredson, D.C., MacKenzie, D.L., Eck, J., Reuter, P. and Bushway, S. (eds) *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, D.C.: US Office of Justice Programs., Chapter 8.

[10] Reply to a criticism during the Great Depression of having changed his position on monetary policy, as quoted in Alfred L. Malabre (1994), *Lost Prophets: An Insider's History of the Modern Economists*, p. 220.

a great idealist, and Rossi’s Law does not mean that we should give up hope of changing the world for the better. What it means is that we should be sceptical about the capacity of any given program to achieve the goal. If you believe that relatively few social programs are effective, it commits you to more rigorous evaluation, and the process of patiently sifting through the evidence until you find a program that works.

But Rossi’s law need not involve putting aside passion about tackling the big challenges. Crime and its aftermath can have a horrendous impact on victims and their loved ones. Incarceration scars as often as it rehabilitates. Even those who only read about crime in the newspaper can be affected, as fear of crime – what English jurist Jeremy Bentham called ‘the secondary mischief’ can make people more stressed in their daily lives, more reluctant to leave their homes.[11]

Finding ways to reduce the impact of crime on Australian lives is a worthy goal to devote your life towards. But there is no contradiction between being passionate about ends, and scientific and critical about the means. A cancer researcher may expect that any given cancer cure will fail; yet still dedicate her career towards finding such a cure.

I have discussed the example of the randomised trial of the NSW Drug Court. Yet the pity is that this is the exception which proves the rule, since so few Australian social policy and criminal justice interventions are subject to rigorous randomised trials. Although it would be practically possible to randomly trial many more interventions, there is a reticence among policymakers to subject policy interventions to gold standard evaluation. In most developed nations, it is impossible to get a new pharmaceutical licensed without a randomised trial. Yet new social and criminal justice policies – often costing considerably more – require no such evaluation.

This is beginning to change in other parts of the world. In the United States, there is a move towards institutionalising high-quality evaluation in some areas of social policy.[12]

- The Second Chance Act, dealing with strategies to facilitate prisoner re-entry into the community, sets aside 2% of program funds for evaluations that ‘include, to the maximum extent possible, random assignment… and generate evidence on which reentry approaches and strategies are most effective’.

- The No Child Left Behind Act calls for evaluation ‘using rigorous methodological designs and techniques, including control groups and random assignment, to the extent feasible, to produce reliable evidence of effectiveness.’

[11] Bentham, J. 1781 [1996]. *An Introduction to the Principles of Morals and Legislation*. (edited by J.H. Burns and H.L.A. Hart), Clarendon Press, Oxford. See also Borooah, V.K. and Carcach, C.A. 1997. ‘Crime and Fear: Evidence from Australia’. *British Journal of Criminology* 37(4): 635-657

[12] These examples are drawn from the Coalition for Evidence-Based Policy, which is part of the Council for Excellence in Government (http://www.excelgov.org/), and a presentation by Adam Gamoran, titled ‘Measuring Impact in Science Education: Challenges and Possibilities of Experimental Design’, NYU Abu Dhabi Conference, January 2009.

- Legislation to improve child development via home visits directs the Department of Health and Human Services to 'ensure that States use the funds to support models that have been shown in well-designed randomized controlled trials, to produce sizeable, sustained effects on important child outcomes such as abuse and neglect'

The biggest challenge facing proponents of randomised policy trials is the question of ethics. When you have a program that you think is effective, how can you toss a coin to decide who receives it? The simplest answer to this is that the reason we are doing the trial is precisely because we do not know whether the program works. If you believe Rossi's Law, it mostly doesn't matter whether a given individual is allocated to the treatment group or the control group. Indeed, for some programs – such as Scared Straight – participants in the control group ended up better off than those in the treatment group.

Adam Gamoran, a professor at the University of Wisconsin-Madison, takes the ethical argument a little further. If you know for sure whether a program works, Gamoran argues, then it is unethical to conduct a randomised trial. But if you do not know whether the program works, then it is unethical *not* to conduct a randomised trial. Every dollar we spend on an ineffective program is a dollar that could have been directed to a better program or returned to taxpayers. The quicker we can find out what works and what does not, the sooner we can direct resources where they are needed most.

I do not mean to lightly dismiss ethical concerns about randomised trials; merely to suggest that in many cases, they are overplayed. Medical researchers, having used randomised trials for several decades longer than social scientists, have now grown relatively comfortable with the ethics of randomised trials. Certain medical protocols could be adapted by social scientists, such as the principle that a trial should be stopped early if there is clear evidence of harm, or the common practice of testing new drugs against the best available alternative.

One example, again from NSW, helps to illustrate how much further advanced medical researchers are when it comes to randomised trials. For the past three years, an NRMA CareFlight team, led by Alan Garner, has been running the Head Injury Retrieval Trial, which aims to answer two important questions: Are victims of serious head injuries more likely to recover if we can get a trauma physician onto the scene instead of a paramedic? And can society justify the extra expense of sending out a physician, or would the money be better spent in other parts of the health system?

To answer these questions, Garner's team is running a randomised trial. When a Sydney 000 operator receives a report of a serious head injury, a coin is tossed. Heads, you get an ambulance and a paramedic. Tails, you get a helicopter and a trauma physician. Once five hundred head injury patients have gone through the study, the experiment will cease and the results will be analysed.

When I was writing an article about the trial last year, I spoke with Alan Garner, who told me that although he has spent over a decade working on it, even he himself doesn't know what to expect from the results.[13] 'We think this will work', he told me a in a phone conversation, 'but so far, we've only got data from cohort studies'. Indeed, he even said 'like any medical

[13] Andrew Leigh, 'A Good Test of Public Policy', *Australian Financial Review*, 8 April 2008

intervention, there is even a possibility that sending a doctor will make things worse. I don't think that's the case, but [until HIRT ends] I don't have good evidence either way.'

For anyone who has heard policymakers confidently proclaim their favourite new idea, what is striking about Garner is his willingness to run a rigorous randomised trial, and listen to the evidence. Underlying the HIRT is a passionate desire to help head injury patients, a firm commitment to the data, and a modesty about the extent of our current knowledge.

The Limits of Randomised Trials

While randomisation is an underused tool in the policy drawer, it is not effective in all cases. Writing tongue-in-cheek in the *British Medical Journal* in 2003, Gordon Smith and Jill Pell argued that the quality of the evidence on parachute effectiveness was severely limited by the absence of randomised controlled trials.[14] They pointed out that the only evidence that parachutes prevent deaths when jumping out of a plane was based on observation and expert opinion, the lowest rank of evidence in the hierarchy. Their conclusion: we need randomised trials of parachutes!

In the sphere of criminal justice, there are other limits to randomisation. Some fundamental rights such as trial by jury are guaranteed by the Constitution, so a randomised trial would be unconstitutional. Other principles – such as the notion that two offenders should be sentenced according to the same criteria – are so central to our ideas of what is just and equitable that we would never think of randomising them. The *randomistas* – as economist Angus Deaton likes to call them – are not welcome everywhere.

Yet while it is unlikely that any judge would agree to randomised the sentence length of murderers, there is plenty that we could learn from more modest experiments. Does allowing more regular family visits reduce recidivism? Would more generous post-release payments help ex-prisoners falling back into bad habits? Do newsletters that inform people about local crime rates make communities safer? Perhaps some of these questions are already being answered by Australian randomised trials – if not, we should be open to subjecting our programs to more scientific scrutiny.

Another limit to what we can learn from randomised trials comes from scale effects. As anyone who has eaten cafeteria food knows, what works well on a small scale does not necessarily work on a large scale.[15] One problem is that small-scale programs are often 'boutique programs', which are resourced to a level that is not feasible if implemented across an entire system. Another risk is that small-scale programs might fail to measure spillover and displacement effects. In economic jargon, randomised trials are a very precise way of measuring partial equilibrium effects, but often do not allow us to get at the general

[14] Gordon C S Smith and Jill P Pell, 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *British Medical Journal*, 327:1459-1461.

[15] This analogy is shamelessly stolen from Janet Currie, 2001. 'Early Childhood Education Programs,' *Journal of Economic Perspectives*, 15(2): 213-238,

equilibrium effects.[16] Because of these factors, it is theoretically possible that we could one day end up doing too many randomised evaluations. In practice, I suspect this is unlikely to happen in my lifetime.

There are also limits to the power of good evidence to change the minds of policymakers. In a recent speech, Productivity Commission Chairman Gary Banks catalogued several instances – ranging from fertility policy to industry policy – in which the Productivity Commission had shown that policies had failed to achieve their goal.[17] Even in NSW, the government has been reluctant to expand the successful Drug Court model to population centres outside Sydney.[18] As these examples show, evidence-based policy is not only about good evidence – it is also about that evidence being accepted by the policy processes. But even here, raising the evidence bar matters. High-quality evaluations are harder for policymakers to ignore than low-quality evaluations.

Conclusion

Over the past forty years, the NSW Bureau of Crime Statistics and Research has played a critical role in improving criminal justice policy. Against the forces of prejudice and ignorance, sober facts and detailed analysis may seem like blunt weapons. But the Bureau has shown that over time, it is possible to shape even the most heated of debates.

Knowing the facts about crime rates is enormously important. We can scarcely begin to tackle crime unless we know which offences are rising, which are falling, who is committing the crime, and whether we are dealing with recidivism or one-offs. Seemingly simple questions often turn out to require painfully sophisticated analysis, and there is always a temptation to bury inconvenient facts. By maintaining a scrupulous commitment to truth-telling, the Bureau has kept its place as the honest broker in criminal justice policy debates.

Having built such a powerful reputation for marshalling the facts about crime, I believe that over the next 40 years, the Bureau should focus to a greater extent on understanding what works. This is not new territory, but it is where the greatest gains are to be made, by building on the Bureau's existing strengths in policy evaluation, pushing strongly for more rigorous evaluation, and insisting upon random assignment wherever possible.

The example of the NSW Drug Court evaluation is a powerful example of how a rigorous randomised evaluation can justify the expansion of a politically difficult intervention. But we should never forget the social benefit of evaluations whose result is to show us that a program

[16] For a thoughtful discussion of these issues in the context of international development, see the papers presented at a Brookings Global Economy and Development Conference entitled 'What Works in Development? Thinking Big and Thinking Small', held in Washington DC on 29-30 May 2008. Papers and presentations are available at http://www.brookings.edu/events/2008/0529_global_development.aspx. See also Angus S. Deaton (2009), 'Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development', NBER Working Paper No. 14690. NBER: Cambridge, MA.

[17] Gary Banks (2009) 'Evidence-based policy-making: What is it? How do we get it?', ANZSOG/ANU Public Lecture Series 2009, Canberra, 4 February. Available at http://www.pc.gov.au/__data/assets/pdf_file/0003/85836/cs20090204.pdf

[18] See eg. Veronica Apap (2008), 'No drug court for the Illawarra', *Illawarra Mercury*, 25 November.

does not work. Thanks to randomised evaluations of multivitamin tablets, my household now has $20 a year to spend on other things. The same is true of evaluations that find government programs to be ineffective. What Franklin Delano Roosevelt called ‘Bold, persistent experimentation’ should be more common in Australian criminal justice policy, and the Bureau should continue to lead the way.