Agenda, Volume 10, Number 4, 2003, pages 341-354

Randomised Policy Trials

Andrew Leigh

In the eighteenth century, the notion that new medical treatments should be evaluated through randomised trials began to gain acceptance. Despite criticism from doctors, who believed that their scientific expertise should be taken on faith, evidence-based medicine gradually gained adherents. Today, randomised trials are a necessary step in the licensing of drugs in Australia and throughout the developed world.

By contrast, in the Australian policy sphere, robust evaluation of the effectiveness of particular programs remains rare. While policy evaluations have become more common over recent years, the efficacy of most of these evaluations remains questionable. Because randomised trials are so rare in Australia, the political rhetoric is usually substituted for hard evidence. Yet in the United States (US), where randomised trials are most common, researchers in education, health and welfare have convincingly demonstrated that they are the most effective way of testing whether policies achieve their stated goals.

The discussion in this paper outlines why randomised trials tend to be superior to other forms of policy evaluation, and address six common objections to their use. The early evolution of randomised evaluation, first in medicine, and then in the social sciences is then discussed. This is followed by an analysis of several recent examples in which policy knowledge has been advanced through the use of randomised trials. The paper then concludes with some suggested areas in which evidence-based policy could be implemented in Australia.

Why Randomised Trials?

Randomised trials represent the most robust method of evaluation known to social science. Just as in medicine, when we want to know the impact of a policy intervention, randomised trials provide the most accurate answer — allowing policymakers to conduct a true 'policy experiment'.

In a randomised framework, the treatment and control groups are alike in all respects except the treatment itself. The alternative — commonly employed in non-randomised 'pilot programs' — is to make some heroic (and generally incorrect) assumptions about what would have happened in the absence of the intervention. Non-randomised evaluations generally take one of two forms. The first approach is to look at participants before and after the program, while the second is to compare participants with some other group of people who did not participate in the program. Both these alternatives are flawed. In the first case, it is impossible to know how the participants would have fared in the absence of the

---

Andrew Leigh is a PhD Student in the John F. Kennedy School of Government, Harvard University. Email: andrew_leigh@ksg02.harvard.edu.

program. In the second, there is a good chance that the decision to enrol is related to other factors that affect outcomes. For example, an evaluation of a quit smoking program might compare those who enrol with those who do not enrol. Yet we would probably think that those who sign up for a quit smoking course are more likely to quit, and hence that do not are a poor control group. This type of evaluation is inherently flawed, yet it is frequently used in government policy evaluations.

In effect, random assignment of participants to treatment and control groups provides us with the perfect counterfactual. It enables us to answer the question: what would have happened to these same participants if they did not enrol in the program? With random assignment, the outcomes for the control group (the non-participants) represent what would have happened to those in the treatment group in the absence of the program. With a sample size in the hundreds, we can expect that the two groups will be very close on any possible measures — not only in observable characteristics such as age, income, sex and race — but also in unobservable characteristics such as intelligence, motivation, family background, and prior knowledge of the program.

Objections to Randomised Trials

Since randomised trials are often criticised on a number of grounds, it is worth dealing with some of these criticisms. (For a more thorough rejoinder to the critics, see Cook and Payne, 2002.)

Objection 1: Randomised trials don't work because it is too difficult to define the goals of most policies

To successfully conduct a randomised trial, it is necessary to precisely define the outcomes that the policy aims to achieve. But can policy goals be measured? It is true that in some cases, policies have immeasurable elements. For example, the effect of school class sizes on students' test scores can be assessed, but the effect of class sizes on self-esteem is more difficult to gauge. Randomised trials would be likely to be more effective at gauging the impact of class size on test scores than on self-esteem. It is also the case that some policies do not need to be assessed because their impact is self-evident. For example, the provision of public pensions is a simple fiscal transfer, and is generally regarded as a success if the money reaches the intended recipient. So long as the payment process is in place, it is difficult to see why we would need to carry out a randomised trial of pension provision.

Yet between these two examples lie a vast array of policies whose outcomes are measurable, but whose effectiveness is questionable. In many cases, policy goals can be measured, yet in the absence of a randomised trial, it is not self-evident that the policy achieves its intended purpose. Industry assistance is generally aimed at creating extra jobs, and fostering research and development, yet governments frequently do not know what would have happened if they had not

provided assistance. Job search and job training programs are aimed at increasing employability and earnings, but are rarely subjected to proper assessment (for two exceptions, see Barrett and Cobb-Clark, 2001; Breunig *et al*, 2003). The Baby Bonus, introduced by the Howard Government in 2001-02 as a means of boosting fertility rates, was not subject to any rigorous trial to see whether it indeed achieved its goal. And at a state level, changes in policing policies often occur in a blitz of rhetoric, with minimal attention paid to results.

A subtler version of this objection is that randomised trials sometimes take place in an environment that creates unusual incentives (one such example is discussed below). In medicine, drug trials can be 'double blind', meaning that neither the subject nor the doctor know whether the patient is receiving the drug or a placebo. Because subjects and administrators cannot be 'blinded' in randomised policy trials, they may adjust their behaviour in ways that undermine the experiment. But this is not an argument against randomised policy trials *per se* — merely a warning that policy experiments should be conducted in a manner that does not create peculiar incentives for the participants.

Objection 2: Randomised trials involve denying treatment to worthy individuals

Because randomised trials involve withholding potentially beneficial treatments from some individuals, some critics have charged that they are unethical. Yet this ignores the fact that governments never provide assistance to all those who would benefit from it. With any rules-based system of administering welfare benefits results, bureaucrats will end up denying assistance to those who do not fully satisfy the set eligibility criteria but would otherwise benefit from the program.

Additionally, in the case of a pilot program, the objection that those in need will miss out is ameliorated by the fact that researchers genuinely do not know whether it is preferable to be assigned to the treatment or control group — otherwise they would not conduct the trial. Cook and Payne (2002) point out that a review of randomised medical trials shows that the treatment outperformed the control only about half the time. They quote Chalmers (1968), 'One has only to review the graveyard of discarded therapies to discover how many patients might have benefited from being randomly assigned to a control group.'

In the case of a program that is already in place, many would consider it unfair to deny it to some recipients. But an ethical way of testing the efficacy of existing programs is to provide financial compensation to ensure that no-one suffers as a result of participating in the trial. For example, in the RAND Health Insurance Experiment, researchers were able to persuade participants to volunteer by promising that those who were assigned to the 'minimal insurance' group would be given financial compensation to ensure they were still better off as a result of participating (though this proved a costly option).

Objection 3: There are already good alternatives to randomised trials

What about alternatives to randomised trials? According to some critics, there is no need to implement randomised trials, because the alternatives are just as good.

The proliferation of non-randomised 'pilot programs' is often cited as an effective way of discovering whether or not policies work. But unfortunately, too many pilot programs are methodologically suspect, and therefore probably a waste of public funds. Without randomisation, researchers lack an appropriate control group with which to determine what would have happened in the absence of the policy intervention. As has been discussed above, the two main alternatives to randomisation — looking at the participants before and after, or following a group of non-participants — produce results that are of questionable veracity. Policymakers should be suspicious of any pilot program relying on a control group that is not truly comparable to the treatment group.

Other critics of randomised trials claim that they are unnecessary because econometric advances now allow researchers to conduct 'quasi-experiments' (also known as natural experiments). It is true that advances over the past two decades have substantially improved researchers' ability to analyse policies in the absence of randomisation. Techniques such as differences-in-differences, regression discontinuity, instrumental variables, and propensity score matching, all help to simulate the conditions of an actual experiment (Angrist and Krueger, 1999). In certain instances, randomised trials will not be feasible for ethical reasons, because the expected effect is very small (necessitating an overly large study), because we are interested in general equilibrium effects, or because the policy is only thought to take effect with a long lag.

In the past decade, a number of seminal quasi-experimental studies have been produced.[1] Differences-in-differences techniques have allowed researchers to analyse the effect on employment of changes in state minimum wage rates, regression discontinuity has been used to study the effect of compulsory school attendance laws on lifetime earnings, and state policy changes have been used to analyse the effect of abortion legislation on crime rates. Quasi-experimental techniques can also be useful in dealing with problems that may occur in randomised trials — such as attrition bias or non-compliance with experimental protocols (Heckman *et al*, 2000). Yet many questions are simply not amenable to quasi-experimental studies, particularly where unobserved ability and selection into the program play a significant role. Where randomisation is feasible, it remains the gold standard in social research (Burtless, 1995).

Objection 4: Qualitative research is more important than quantitative research

In most cases, randomised evaluations focus on measurable outcomes, such as employment, earnings, crime rates, or test scores. While these measures are powerful ways of measuring what policymakers and citizens care about, they are generally incomplete. Qualitative outcomes, such as how policies affect

[1] A significant impediment to quasi-experimental studies in Australia is the difficulty in obtaining microdata. Although the Australian Bureau of Statistics makes some data available to Australian university researchers as Confidentialised Unit Record Files, significantly less information is available than in many other developed nations (Leigh and Wolfers, 2003).

participants' self-esteem, and what participants feel is the most important impact of a policy intervention, are also critical parts of any assessment.

As Weiss (1998:14) points out, a key result of the 'paradigm wars' of the late-1970s and 1980s, which pitted qualitative and quantitative researchers against one another, was the discovery that both methods could effectively complement one another. At their best, randomised trials can also include a qualitative component — for example, analysing the social context in which the policy is implemented. In the Moving to Opportunity trial (discussed in more detail below), researchers measured statistics on health, education and employment outcomes. They then carried out in-depth surveys with members of the treatment and control groups, asking them how they felt about their environment. Through this, they discovered that participants found the biggest impact of moving to a low-poverty neighbourhood was a substantial reduction in crime, and a lessening in participants' fear of crime (Kling, Liebman and Katz, 2001). Quantitative analysis allowed researchers to compare outcomes for the treatment and control groups, while qualitative research was necessary to find out what subjects saw as the most important aspect.

Objection 5: Politicians are interested in re-election, not results

Inevitably, the need for governments to win an election every three or four years shapes policies. All governments face the accusation from their opponents that they are 'poll-driven', and during election-time, the same charge is often levelled at oppositions. The growth of 'middle class welfare', in which taxation revenue is raised (with its associated deadweight loss) and then returned to the same taxpayers as non-means tested benefits, is frequently cited as an example of this form of politics.

Yet while the growth of telephone surveys in recent decades has facilitated poll-driven politics, it would be premature to conclude that Australian governments do not care about the results of their policies. Using data from Australian federal elections, Cameron and Crosby (2000) and Wolfers and Leigh (2002) show that governments' success in lowering unemployment is a significant predictor of whether or not they will be re-elected. Even governments that are only concerned with being re-elected should devote more energy to testing their policies. At the same time, political actors who are more concerned with good policy than election outcomes (the public service, think-tanks, academics, and perhaps the media) have an important role to play in holding governments to account. Calling for political claims to be tested via randomised evaluations is one way in which this can be done.

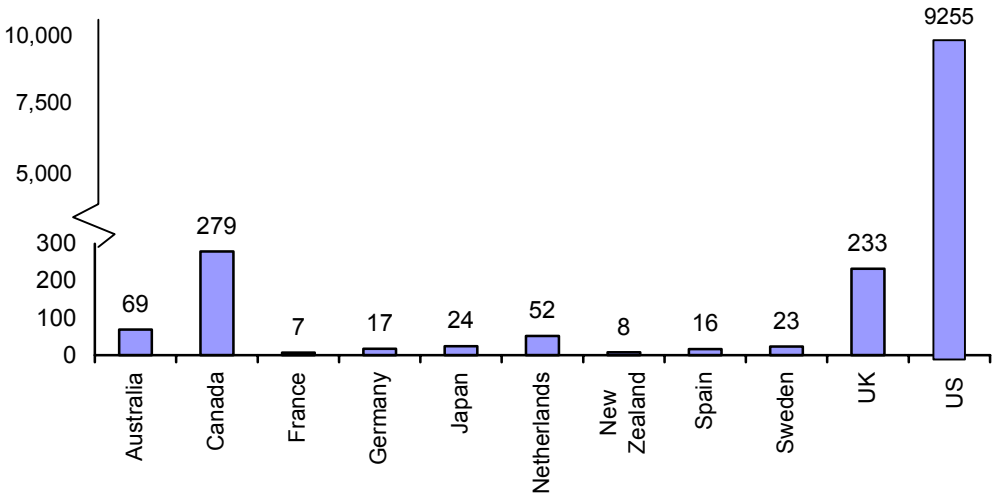
Objection 6: Only in America

While many of the examples in this paper are drawn from the US, it is not true to say that randomised trials are primarily confined to America. An international database of randomised trials in the social sciences (Campbell Collaboration, 2003), attempts to exhaustively catalogue all such trials. The database contains

more than 10,000 randomised policy experiments in the fields of social policy, psychology, education and criminology. Figure 1 shows the breakdown of these trials for selected countries. While the US clearly dominates, Canada and the UK have each carried out over 200 randomised policy trials. Sixty-nine Australian policy experiments appear in the Campbell Collaboration randomised trial database.

Further evidence of the international interest in policy trials is the international nature of the Campbell Collaboration itself. In February 2003, its conference in Stockholm, Sweden drew researchers from throughout Europe and the Americas. As will be discussed below, randomised trials of social programs may have had their genesis in the US, but today, policymakers and researchers in a variety of countries are looking at how they can be used to produce better policies.

Figure 1: Randomised Policy Trials by Country



Source: Campbell Collaboration (2003).

Note: Figures were obtained by searching all indexed and non-indexed fields of the C2-SPECTR database. To allow for variants on country names, the UK was searched as England or Great Britain or Scotland or Wales or United Kingdom or UK or U.K., while the US was searched as America or United States or U.S. or US.

The Evolution of Randomised Trials

In medicine, the theory underlying randomised trials has existed for at least three and a half centuries. Debus (1970:34) quotes John Baptista van Helmont, a physician writing in the mid-seventeenth century:

> Oh ye Schooles … Let us take out of the hospitals, out of the Camps, or from elsewhere, 200 or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them into halfes, let us cast lots, that one halfe of them may fall to my share, and the other to yours; … we shall see how many Funerals both of us shall have … Here your business is decided.

Yet it was not until the late-nineteenth century that these ideas began to gain wider acceptance. Clinical psychologists began using randomised trials around 1850-1880, while Denmark tested a diptheria serum in randomised trials in the late-nineteenth century. Following a formal demonstration of the statistical theory underlying randomisation (Fisher, 1935), the 1940s and 1950s saw a rapid growth in randomised trials for treatment of tuberculosis and poliomyelitis.

In social science, Lyndon Johnson's War on Poverty saw the evolution of randomised trials. During the late-1960s, the Head Start program, and the Ypsilanti pre-school program were among the first to be evaluated through randomised trials (Cicirelli and Associates, 1969). And commencing in 1968, the Mathematica Foundation conducted the 'New Jersey-Pennsylvania Income Maintenance Experiment' (Kershaw and Fair, 1976), a large-scale social experiment designed to test whether the payment of supplementary income to poor families acted as a work disincentive.

During the 1970s, randomised experiments began to flourish. A California jail used random assignment to test whether group counselling reduced recidivism (Ward and Kassebaum, 1972), while the Baltimore LIFE experiment (Rossi, Berk, and Lenihan, 1980) tested whether financial assistance to prisoners upon their release improved outcomes for former inmates. The Housing Allowance Design Experiment analysed the effect of various different types of rental subsidies on recipients' housing quality (Struyk and Bendick, 1981). The RAND Health Insurance Experiment analysed behavioural responses to a variety of different health care plans (Newhouse *et al*, 1981). And a randomised trial in Mexico tested whether watching Sesame Street on a daily basis improved educational outcomes (Diaz-Guerrero *et al*, 1976).

Boruch, De Moya and Snyder (2002), who analyse randomised trials across a number of spheres of social policy, including criminology, education, and welfare, concluded that the rapid growth in randomised experiments during the 1960s and 1970s may have been followed by a slight decline during the 1980s. Weiss (1998:13) speculates that this may have been due to a decline in federal funding after 1981, when Ronald Reagan took office, and notes that evaluation made a comeback under Bill Clinton's administration. In recent years, evaluation has also become substantially more popular at the state and local level (Weiss, 1998:14).[2] Yet randomised policy evaluations remain comparatively rare — for every

---

[2] There is one notable exception to this. During the 1980s and early-1990s, a number of states obtained waivers from the federal government to experiment with various alterations of their welfare programs. As a result of the 1996 welfare reform legislation, most of these waiver experiments were discontinued (Moffitt and Ver Ploeg, 2001:30).

randomised policy trial, 24 randomised medical trials take place (*The Economist*, 2002:74).

Policy Lessons From Randomised Trials

In the past two decades, randomised trials have become familiar to researchers in the US and elsewhere. But how have they contributed to our understanding of policy problems? Some of the most important randomised trials that have been conducted in recent years, and their implications for policy are now discussed.

Training for unemployed workers

The randomised evaluation of the Job Training Partnership Act (JTPA) by the US Department of Labor is one of the most important randomised trials to have been conducted in the job training sphere. One of the major challenges in evaluating job training programs is that such programs have a high attrition rate. This is equally true in Australia — a recent review of the Job Network found that its training programs had a 60 percent dropout rate (Productivity Commission, 2002:5.14, 5.17-5.18.). The dropout rate causes an 'attrition bias', since those who participate in job training are not a representative sample of the jobless. Participants are likely to be the most motivated of the unemployed, who would have been most likely to find a job even in the absence of the program. Comparing the outcomes for participants and non-participants is therefore likely to produce a biased estimate of the effect of job training.

By randomly assigning some participants to training and others to a control group, the JTPA evaluation found that job training for 16-22 year olds generally failed to boost participants' earnings (Orr *et al*, 1996). In one sense, this was a dispiriting finding, but in another, it focused energies into more effective ways of improving the employment prospects of young adults.

By contrast, Australian politicians of all persuasions have shied away from serious trials of job training programs. No randomised trial was conducted on Labor's Working Nation program, which cost over $1 billion per year. Under the Coalition, the Department of Family and Community Services has recently conducted two randomised trials on the Job Network (Barrett and Cobb-Clark, 2001; Breunig *et al*, 2003), but both have focused on the effectiveness of intensive caseworker interviews. Job training programs in Australia are yet to be subjected to a randomised trial.

A likely explanation for the failure to properly analyse labour market policies in Australia is that the available evidence (from non-randomised evaluations) suggests the returns to job training programs are limited (Productivity Commission, 2002; Chapman, 1999). Politicians may be fearful that the results from a randomised trial will indicate that their policies are having limited results.

Education

By comparison with labour market and welfare, randomised trials in education have been relatively rare. Burtless (2002) speculates that this may be due to the fact that social scientists have never exercised significant influence over education policy evaluation, coupled with the fact that educators and parents wield substantial political influence, and are reluctant to surrender control over any aspect of teaching or curriculum.

Nonetheless, it is clear that more educational randomised trials have been carried out in the US than in any other country. The most prominent of these is Tennessee's Project Star, an experiment in which students were randomly assigned to classes of varying sizes (Krueger, 1999). Here, randomisation helped avoid the problem that students assigned to smaller classes are often different from those who are assigned to regular classes. Because schools frequently place talented or struggling students in smaller classes, a non-randomised comparison can give a false picture of the effect of class sizes on student performance.

Unfortunately, Tennessee's Project Star suffered from the fact that teachers had been told prior to the experiment that if students in the smaller classes outperformed those in the larger classes, class sizes would be reduced in all schools. This created a false incentive for small-class teachers to work harder than large-class teachers — what is known as a 'Hawthorne Effect' (Hanushek, 1998).[3] While a randomised trial was a promising way of discovering the effect of smaller classes, the political context of Project Star undermined the reliability of its results.

Nonetheless, randomised trials have great potential to expand what we know about educational reforms. DARE, a school-based anti-drugs program, was revised in 2001 following randomised trials showing that the program did not deliver promised results (Boruch, De Moya and Snyder, 2002). Research on young driver education programs showed that rather than reducing road deaths, they actually increased the road toll — by encouraging high school students to drive at a younger age (Hatcher and Scarpa, 2001:55-56). In the debate over school vouchers, randomised trials have recently been implemented in Washington, DC, Dayton, OH and New York, NY. Their analysis and re-analysis has contributed substantially to policymakers' understanding of the effects of school choice on student performance (Peterson, Myers and Howell, 1998; Krueger and Zhu, 2003).

---

[3] The Hawthorne effect was named after a 1927-32 study conducted at the Hawthorne Works of the Western Electric Company, aimed at determining the effect of lighting conditions on worker performance. After finding that both increased and decreased lighting caused an increase in productivity, the researchers concluded that it was the act of being observed which caused the change (Roethlisberger and Dickson, 1939). For more discussion of the Hawthorne effect, see Rossi and Freeman (1993:236-238).

Neighbourhood Effects

Another field in which a randomised experiment has contributed substantially to the understanding of a social policy issue is the debate over neighbourhood effects and locational disadvantage. In the 1990s, a US program, known as Moving to Opportunity (MTO), analysed the effect of providing housing vouchers to enable poor families to move from high-poverty to low-poverty neighbourhoods. While social scientists had long speculated that the physical and social environment had a significant impact on individual outcomes, little supportive evidence existed. In the words of MTO researcher Professor Jeffrey Liebman, the problem arose because 'it was difficult to find truly comparable individuals living in different neighbourhoods' (quoted in Leigh and Wolfers, 2001:32).

Moving to Opportunity was conducted in five cities — Baltimore, Boston, Chicago, Los Angeles and New York. A four-year follow-up found that those who won the MTO lottery and moved to a low-poverty area reported substantially lower levels of exposure to violence — while those families who stayed in poor neighbourhoods reported continuing high levels of fear. Moving had benefits in a range of other dimensions, too. Mothers who moved reported being healthier, feeling calmer and less prone to episodes of depression. For younger children, moving boosted test scores for both reading and mathematics. Among older children, moving reduced absenteeism, lowered school dropout rates, and — at least among boys — lessened behaviour problems. Child health also improved, with asthma attacks declining markedly. While researchers were unable to discern any statistically significant difference in employment or earnings between the two groups, movers were healthier, safer and had better outcomes for their children than non-movers (Katz, Kling and Liebman, 2003). And as mentioned above, qualitative research found that families who moved nominated crime as the biggest difference MTO made to their lives (Kling, Liebman and Katz, 2001).

Treatment of Drug Offenders

While the previous three topics have drawn almost exclusively upon US randomised experiments, the treatment of drug offenders is an area in which a randomised evaluation was recently carried out in Australia — evaluation of the NSW Drug Court.

The background to the NSW Drug Court evaluation was conducted in 1999-2000. During the previous decade, several US states and European countries had experimented with drug courts, but evidence from randomised trials on the effectiveness of these courts was severely limited. In addition, no randomised evidence existed on the cost-effectiveness of drug courts (Lind *et al*, 2002).

The NSW Drug Court trial consisted of non-violent offenders who met a series of criteria, including dependence on illicit drugs, and willingness to plead guilty. Participants were then randomly assigned either to the Drug Court, or to a regular court. Over the two years in which the trial was in operation, 514 people participated in the trial. Because the Drug Court was in a pilot phase, its

detoxification program had only limited capacity — making the exclusion of some applicants politically palatable and acceptable by policymakers and administrators.

The evaluation found that individuals processed through the Drug Court had significantly lower rates of recidivism for drug offences than those processed through the normal criminal justice system. While the cost of the Drug Court exceeded the cost of regular courts, the cost *per offence averted* was substantially lower under the Drug Court (Lind *et al*, 2002). Using best practice social science techniques, the Drug Court evaluation demonstrated that innovative criminal justice methods could actually turn out to be cost-effective.

Conclusion — Try it and See

What is the future for randomised trials in Australia? As demonstrated by the Drug Court example and the two trials conducted by the Department of Family and Community Services, the door to randomised evaluation is not entirely closed. But in comparison to the US, where rigorous policy evaluation has become a part of the political landscape, evaluation in Australia has a long way to go. Some small part of this might be due to our three-year election terms, but a more likely factor is that Australia has few evidence-based think-tanks and lacks America's culture of policy contestation. In its absence, there is a risk that cost-benefit calculus will be carried out with electoral maps, rather than by policy analysts.

For an ambitious government, there is little limit to the policy questions that might be answered in Australia through well-designed randomised trials. Criminologists know little about the effects of different forms of incarceration, and different in-jail programs, on rates of recidivism and subsequent employment patterns. In education, more evidence is needed on how teacher merit pay affects student performance and whether after-school programs improve educational attainment. In welfare, job training, tax credits, and early childhood intervention programs could all usefully be tested. And in industry assistance, new industry assistance or trade promotion programs could be tested via a small-scale randomised experiment to assess their impact on wages, employment and productivity. In each case, the cost of undertaking randomised pilot programs would be cheap — particularly when compared with the cost of policy mistakes.

In some of these cases, Australian policy researchers can simply free ride off research conducted in the US and elsewhere. But in other instances, Australia's unique institutions should make us question whether policies that are successful in America are likely to be effective on the other side of the Pacific. In education, Australia has an unusually high rate of enrolment in non-government schools by comparison with most other developed countries. In the labour market, we have higher rates of long-term unemployment and part-time work than many other nations. In welfare, our unemployment benefits differ from those of most other countries in not being time-limited. And for Australian firms, our economy is substantially more dependent upon international trade and investment than the US economy. Because of these factors and many more, Australian policymakers

should consider carrying out research themselves, rather than merely relying on overseas evidence.

At the turn of the twentieth century, Australia was known as 'the social laboratory of the world'. The nation was known for its policy innovation — among the first to use the secret ballot, allow women to vote, provide workers with a minimum wage and unemployment benefits, and introduce an aged pension. Australian policymakers should summon up the vigour of their predecessors, and conduct randomised trials on a variety of current and proposed policies — providing evidence on what works, and what does not.

References

Angrist, J. and A. Krueger (1999), 'Empirical Strategies in Labour Economics' pp 1339-1344 in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol 3A, North-Holland.

Barrett, G. and D. Cobb-Clark (2001), 'The Labour Market Plans of Parenting Payment Recipients: Information from a Randomized Social Experiment', *Australian Journal of Labour Economics* 4(3):192-205.

Boruch, R., D. De Moya and B. Snyder (2002), 'The Importance of Randomized Field Trials in Education and Related Areas', pp 50-79 in F. Mosteller and R. Boruch (eds) *Evidence Matters: Randomized Trials in Education Research*, Brookings, Washington DC.

Burtless, G. (1995), 'The Case for Randomized Field Trials in Economic and Policy Research', *Journal of Economic Perspectives* 9(2):63-84.

Burtless, G. (2002), 'Randomized Field Trials for Policy Evaluation: Why not in Education?', pp 179-197 in F. Mosteller and R. Boruch (eds) *Evidence Matters: Randomized Trials in Education Research*, Brookings, Washington DC.

Breunig, R., D. Cobb-Clark, Y. Dunlop and M. Terill (2003), 'Assisting the Long-Term Unemployed: Results from a Randomised Trial', *Economic Record* 79(244):84-102.

Cameron, L and Crosby, M. (2000), 'It's the Economy Stupid: Macroeconomics and Federal Elections in Australia', *The Economic Record* 76(235):354-364.

Campbell Collaboration (2003), *Social, Psychological and Criminological Trials Registry (C2-SPECTR),* available at http://www.campbellcollaboration.org (checked 9 August 2003).

Chalmers, T (1968), 'Prophylactic Treatment of Wilson's Disease', *New England Journal of Medicine* 278:910-911.

Chapman, B (1999), 'Could Increasing the Skills of the Jobless be the Solution to Australian Unemployment?', pp 176-199 in S. Richardson (ed.) *Reshaping the Labour Market: Regulation, Efficiency and Equality in Australia,* Cambridge University Press, Melbourne.

Cicirelli, V. and Associates (1969), *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development,* Report to the

Office of Economic Opportunity, Ohio University and Westinghouse Learning Corporation, Cleveland OH.

Cook, T. and M. Payne (2002), 'Objecting to Objections to Random Assignment in Educational Research', pp 150-178 in F. Mosteller and R. Boruch (eds) *Evidence Matters: Randomized Trials in Education Research*, Brookings, Washington DC.

Debus, A. (1970), *Science and Education in the Seventeenth Century,* MacDonald, London.

Diaz-Guerrero, R., I. Reyes-Lagunes, D. Witzke and W. Holtzman (1976), 'Plaza Sésamo in Mexico: An Evaluation', *Journal of Communication* 26(2):145-154.

Fisher, R. (1935), *The Design of Experiments,* Oliver and Boyd, Edinburgh.

General Accounting Office (1995) *Multiple Employment and Training Programs,* GAO/T-HEHS-95-93, Washington DC.

Hanushek, E. (1998), 'The Evidence on Class Size', Occasional Paper Number 98-1, University of Rochester, Rochester NY, available at http://www.edexcellence.net.

Hatcher, J. and J. Scarpa (2001), *Background for Community-Level Work on Physical Health and Safety in Adolescence: Reviewing the Literature on Contributing Factors*, Child Trends, Washington DC.

Heckman, J., N. Hohmann, J. Smith and M. Khoo (2000), 'Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment', *Quarterly Journal of Economics* 115(2):651-694.

Katz, L., J. Kling and J. Liebman (2003), 'Early Impacts of Moving to Opportunity in Boston: Final Report to the US Department of Housing and Urban Development' in J. Goering and J. Feins (eds) *Choosing a Better Life? Evaluating the Moving to Opportunity Social Experiment,* Urban Institute Press, Washington, DC (forthcoming)

Kershaw, D and J. Fair (1976), *The New Jersey Income-Maintenance Experiment*, Academic Press, New York, NY, Vol 1.

Kling, J., J. Liebman and L. Katz (2001), 'Bullets Don't Got No Name: Consequences of Fear in the Ghetto', Joint Center for Poverty Research Working Paper 225, Chicago IL, available at http://www.jcpr.org.

Krueger, A. (1999), 'Experimental Estimates of Education Production Functions', *Quarterly Journal of Economics* 114(2):497-532.

Krueger, A. and P. Zhu (2003), 'Another Look at the New York City School Voucher Experiment', National Bureau of Economic Research Working Paper 9418, NBER, Cambridge MA.

Leigh, A. and J. Wolfers (2001), 'Moving to Opportunity' *AQ: Journal of Contemporary Analysis* 73(5):31-32.

Leigh, A. and Wolfers, J. (2003), 'Policy Improves by Putting Rhetoric on Trial', *Sydney Morning Herald*, 5 March:15.

Lind, B., D. Weatherburn, S. Chen, M. Shanahan, E. Lancsar, M. Haas and R. De Abreu Lourenco (2002), *New South Wales Drug Court Evaluation: Cost-Effectiveness,* NSW Bureau of Crime Statistics and Research, Sydney.

Moffitt, R. and M. Ver Ploeg (eds) (2001), *Evaluating Welfare Reform in an Era of Transition*, National Academy Press, Washington, DC.

Newhouse, J., W. Manning, C. Morris, L. Orr, N. Duan, E. Keeler, A. Liebowitz, K. Marquis, C. Phelps and R. Brook (1981), 'Some Interim Results from a Controlled Trial of Cost-sharing in Health Insurance', *New England Journal of Medicine* 305:1501-1507.

Orr, L., H. Bloom, S. Bell, F. Doolittle, W. Lin and G. Cave (1996), *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*, Urban Institute, Washington, DC.

Peterson, P., D. Myers and W. Howell (1998), 'An Evaluation of the New York City Scholarships Program: The First Year,' Mathematica, Washington DC.

Productivity Commission (2002), *Independent Review of Job Network,* Report No. 21, Ausinfo, Canberra.

Roethlisberger, F. and W. Dickson (1939), *Management and the Worker,* Harvard University Press, Cambridge MA.

Rossi, P., R. Berk and K. Lenihan (1980), *Money, Work, and Crime: Some Experimental Evidence*, Academic Press, New York NY.

Rossi, P. and H. Freeman (1993), *Evaluation: A Systematic Approach,* 5th ed, Sage, Newbury Park CA.

Struyk, R. and M. Bendick (eds) (1981), *Housing Vouchers for the Poor: Lessons from a Natural Experiment,* Urban Institute, Washington DC.

The Economist (2002), 'Try It and See' 2 March:73-74.

Ward, D. and G. Kassebaum (1972), 'On Biting the Hand that Feeds: Some Implications of Sociological Evaluations of Correctional Effectiveness', pp 300-310 in C. Weiss (ed.) *Evaluating Action Programs: Readings in Social Action and Education,* Allyn & Bacon, Boston MA.

Weiss, C. (1998), *Evaluation: Methods for Studying Programs and Policies,* 2nd ed., Prentice Hall, Upper Saddle River NJ.

Wolfers, J. and A. Leigh (2002), 'Three Tools for Forecasting Federal Elections: Lessons from 2001', *Australian Journal of Political Science* 37(2):223-240.

Thanks to Fred Argy, Louise Biggs, Richard Curtain, Glyn Davis, Barbara Leigh, Andrew Norton, Donald Speagle, two anonymous referees, participants at the Fulbright 2003 Symposium and participants in the Australian Public Policy Research Network online discussion for comments on an earlier draft.