Australian Government
Productivity Commission

Strengthening Evidence-based Policy in the Australian Federation

Roundtable Proceedings

Canberra, 17-18 August 2009
Volume 2: Background Paper

ISBN 978-1-74037-312-8

This publication is available in hard copy or PDF format from the Productivity Commission website at www.pc.gov.au. If you require part or all of this publication in a different format, please contact Media and Publications (see below).

Publications Inquiries:
Media and Publications
Productivity Commission
Locked Bag 2 Collins Street East
Melbourne VIC 8003

Tel: (03) 9653 2244
Fax: (03) 9653 2303
Email: maps@pc.gov.au

General Inquiries:
Tel: (03) 9653 2100 or (02) 6240 3200

An appropriate citation for this paper is:

Productivity Commission 2010, *Strengthening Evidence-based policy in the Australian Federation*, *Volume 2: Background Paper*, Productivity Commission, Canberra.

JEL code:

The Productivity Commission

The Productivity Commission is the Australian Government's independent research and advisory body on a range of economic, social and environmental issues affecting the welfare of Australians. Its role, expressed most simply, is to help governments make better policies, in the long term interest of the Australian community.

The Commission's independence is underpinned by an Act of Parliament. Its processes and outputs are open to public scrutiny and are driven by concern for the wellbeing of the community as a whole.

Further information on the Productivity Commission can be obtained from the Commission's website (www.pc.gov.au) or by contacting Media and Publications on (03) 9653 2244 or email: maps@pc.gov.au

Contents

Figures

Tables

Preface

The Productivity Commission's 2009 Roundtable aimed to promote discussion on how to strengthen the use of evidence to better inform policy decisions. This background paper was provided to participants to facilitate discussion by exploring what is meant by evidence-based policy, and examining how it might be implemented in practice. It was principally prepared by Terry O'Brien and Kristy Bogaards, from the Commission's Canberra office.

In 2008 and early 2009, the Commission's Chairman, Gary Banks, gave a number of presentations on evidence-based policy making, the most recent of these being an ANZSOG/ANU public lecture in Canberra titled *Evidence-based policy making: What is it? How do we get it?* (Banks, 2009). This paper complements and elaborates on that address. In particular, it expands on the principles sketched on pages 8 to 18 and the institutional ideas at pages 21 to 23.

The first section of the paper provides a definition of evidence-based policy and briefly examines some features of an evidence-based approach.

Section two sets out some high level principles for evidence-based policy making that identify the main recurrent issues in policy evaluation. Such principles could be useful to governments, officials, journalists and the general public in structuring their thinking about policy.

Section three examines the institutional arrangements that strengthen evidence-based policy by improving transparency, helping governments support each others' policy evaluation efforts and building evidence into the decision making process.

Key points

- Evidence-based policy in essence refers to a process that, to the fullest extent possible, transparently uses rigorous and tested evidence in the design, implementation and refinement of policy to meet designated policy objectives.
- This background paper explores how policy could be strengthened in practice, by collecting and using better evidence, and seeks to identify some general principles that would help.
- Good policy formulation methodologies have a number of features in common. They
  - carefully define the policy problem and establish clear objectives
  - develop a range of policy options drawing on a coherent framework of theory
  - use evidence to test those options, in a cost-benefit framework where possible
  - explicitly address what would have happened in the absence of the policy, and consider attribution issues and possible biases
  - examine direct and indirect effects on the economy and the community.
- The choice of evaluation methodology depends on the task at hand and the type of evidence available. Hierarchies of evidence can be a useful mechanism for sifting large volumes of evidence and focussing on the most robust. But governments face a wide range of policy problems, and there is no single 'gold standard' approach to evaluation that would work best in all circumstances.
- Evidence-based policy requires more than good policy formulation methodologies and data. It requires institutional frameworks that encourage, disseminate and defend good evaluation, and that make the most of opportunities to learn. Where evidence is incomplete or weak, good processes for learning, and for progressively improving policies, become even more important. Some of the institutional features that can assist include:
  - Improving transparency
  - Building in and financing evaluation from policy commencement
  - Using sequential roll-out, pilots and randomised trials where appropriate
  - Establishing channels to disseminate evaluations and share results across jurisdictions
  - Strengthening links between evidence and the decision making process.

1 Evidence-based policy

There is nothing new about the idea that the best available evidence should underpin policy decisions — it has been described as old as the state itself (Nutley et al. 2009, p. 3). But the term evidence-based policy is relatively new — popularised in the late 1990s when the Blair Labour government was elected on a platform of 'what counts is what works':

> We will improve our use of evidence and research so that we understand better the problems we are trying to address. We must make more use of pilot schemes to encourage innovations and test whether they work. We will ensure that all policies and programmes are clearly specified and evaluated, and the lessons of success and failure are communicated and acted upon. (Blair and Cunningham 1999)

In the ensuing decade, the ideals of evidence-based policy became a global movement. In Australia, the Prime Minister said in an address to senior public servants in April 2008:

> Policy design and policy evaluation should be driven by analysis of all the available options, and not by ideology. …We're interested in facts, not fads. (Rudd 2008)

In the United States, President Obama (2009) echoed some language of evidence-based policy in his inauguration address:

> The question we ask today is not whether our government is too big or too small, but whether it works … Where the answer is yes, we intend to move forward. Where the answer is no, programs will end.

Subsequently, the White House's Office of Management and Budget has issued guidance on what constitutes strong evidence for policy evaluation, and its new director has outlined a program to build rigorous evidence to drive policy. In a separate development, the non profit and non partison US Coalition for Evidence-Based Policy works to encourage appropriate financing for evaluation, the use of state variation to improve knowledge of what works and the sharing of evaluation results.

Similar ideas have filtered through aid organisations and many developing countries, with impact evaluation now a routine requirement under aid donors' systems. International agencies, such as the OECD's Development Assistance Committee, the World Bank and the International Monetary Fund (with extensive evaluation programs of their own), have joined evaluation 'clubs' to pool knowledge, fund better evaluation and disseminate results and lessons learned.

Nevertheless, in essence, evidence-based policy is simply the latest name to describe what economists, social scientists and policy analysts have attempted to practise for decades. It is a collection of concepts and methods to reinvigorate existing efforts and encourage better approaches to the analysis of public policy, with the primary aim of improving the rigour of policy development and enhancing accountability of decision makers.

Both the demand for, and supply of, evidence have grown

On the demand side, the number of complex policy challenges on the horizon puts a premium on ensuring rigorous assessment of policy choices and evaluation of existing programs. For instance, in the human capital sphere there is often little settled evidence about what policy measures work best and there can be long lead times before results are fully evident. Single-cause explanations or answers to policy questions are likely to be rare. And, as the Council of Australian Governments (COAG) moves towards nation-wide solutions, policy experiments and opportunities to learn may decline. Policy mistakes could be national and come at a significant cost.

On the supply side, the opportunity for better analysis has grown with the emergence of large microeconomic and social data sets, cheap computing power and more sophisticated modelling, experimentation and econometric examination of micro and social policy impacts. These have combined to yield improvements in the ability to identify and quantify causal relationships (Donohue 2001, p. 2). Analysts can now examine what works to a degree previously unimaginable, and taxpayers expect government policies to work without wasting resources.

All this provides a challenge to ensure that the analytical tools and evidentiary practices appropriately serve contemporary policymaking and that the institutional framework is sufficiently robust to ensure transparent and independent evaluation of the evidence.

What is evidence-based policy?

At first glance, the term evidence-based policy appears self-explanatory — public policy based on rigorous evidence:

> Like all of the best ideas, the big idea here is a simple one – that research should attempt to pass on collective wisdom about the successes and failure of previous initiatives in particular policy domains. The prize is also a big one in that such an endeavour could provide an antidote to policy making's frequent lapses into crowd pleasing, political pandering, window dressing and god-acting. (Pawson 2002)

Nevertheless, there is considerable variation in what is understood by the term evidence-based policy. Some researchers narrowly associate it with a particular type of evidence or evaluation methodology, most notably quantitative evaluation, meta-analyses of existing research or randomised policy trials (Nutley et al. 2009, p. 5).

The broader definition offered here views evidence-based policymaking as a process that transparently uses rigorous and tested evidence in the design, implementation and refinement of policy to meet designated policy objectives. This definition stresses three characteristics that seem emblematic of the present interest:

- evidence should be broad, tested, rigorous, and ideally capable of replication;
- evidence should be robust and avoid common methodological pitfalls; and
- the entire process should be transparent and contestable.

This broader view recognises that evidence and evaluation are relevant at every stage of the policy cycle, from identifying the policy problem, through assessing policy options, to ex-post evaluation (figure 1.1). Moreover, it can encompass an array of evidence and research methods (figure 1.2).

Figure 1.1 **The role of evidence in the policy cycle**

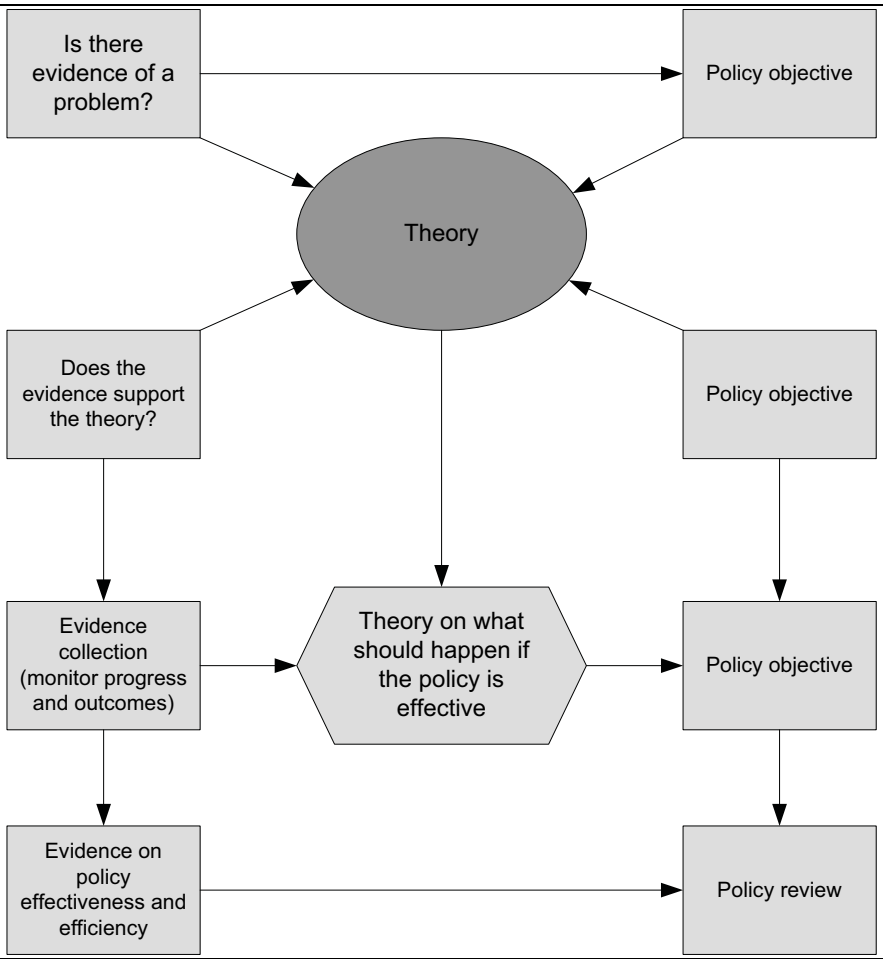
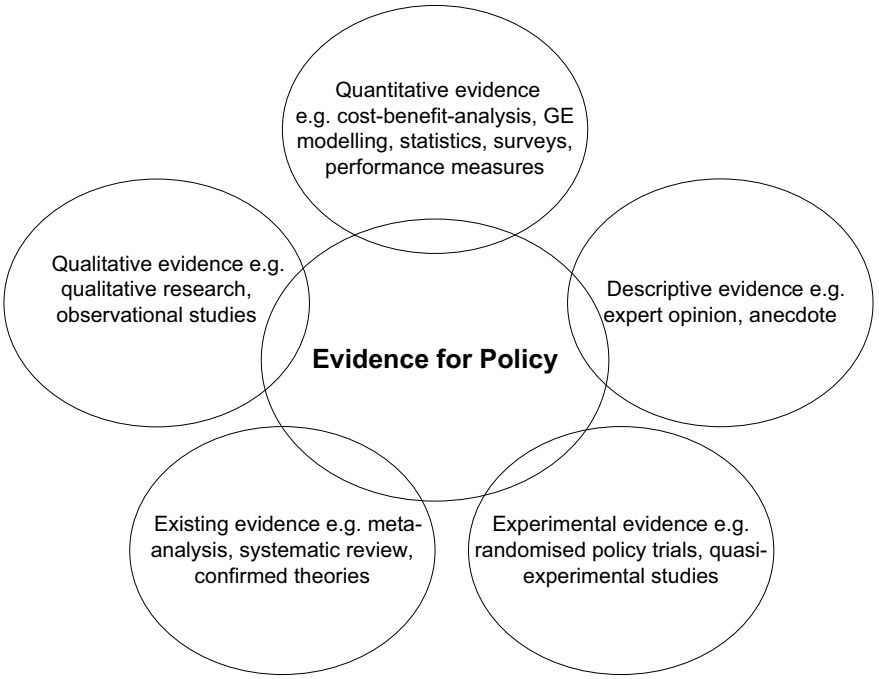


Figure 1.2 **Types of evidence**



Under this broad view, what matters for public policy development is not only what works, but also how does it work?, what are the broader ramifications?, at what cost? and who benefits or loses? For instance, increasing the school leaving age could certainly be shown to 'work' by increasing the number of children with a year 12 qualification, but it may not actually improve student outcomes. Similarly, tariffs and industry subsidies 'work' to improve the position of the target domestic industry, yet this may come at the (greater) expense of the wider economy and community. Fundamentally, evidence-based policy is about assessing whether a policy improves community wellbeing.

Why is it important?

The purpose of government is to improve the wellbeing of the community in ways that may not be possible by individuals acting alone. However, policymakers can get it wrong, be ineffective or fail to foresee unintended consequences (box 1.1). There is often considerable debate about whether government action has actually led to an improvement and, if so, the extent of the gains. An evidence-based approach to policymaking is one way to improve policy development. It is built around the belief that better quality decisions will be made if the process is informed by robust evidence.

Box 1.1 Why is evidence-based policy important? Some examples

Conventional wisdom is often wrong. Policies can be ineffective or have unintended and perverse outcomes.

Government spending on education has increased by around 260 per cent in the last four decades, largely to finance reductions in class size. But overall the evidence suggests that class size reductions do not yield significantly better student outcomes (Hanushek 2002, Leigh and Ryan 2006).

The 'United States Scared Straight' program, which aim to deter 'at risk' juveniles from future offending through first-hand observation of prison life, actually have the reverse effect, by providing 'positive' criminal role-models and attracting juveniles towards crime. A meta-analysis of randomised controlled trials demonstrated that programs like Scared Straight are likely to have a harmful effect and increase delinquency relative to doing nothing at all (Petrosino et al 2002).

Conservation regulations, introduced to protect and conserve, can have the reverse effect. For example, there was considerable pre-emptive clearing of native vegetation in Queensland in anticipation of the imposition of regulatory restrictions. Total clearing rose from around 330 000 hectares a year over the 1991-1999 period to 758 000 hectares in 1999-2000, when forthcoming restrictions were signalled (PC 2004a).

Evidence can improve clarity when dealing with complex interventions which act on complex social systems.

Childcare is often viewed in a binary fashion — it is either 'good' or 'bad' for young children. The Commission's report on Paid Parental Leave found the reality to be more complex. There is evidence of problems where non-parental care is initiated early in a child's life (three to six months), where childcare hours are extensive and care is of low quality. But, the evidence is inconclusive for babies aged six to twelve months and actually suggests positive effects from quality care between 12 and 18 months (PC 2008a).

Evidence is persuasive. It can shape the policy debate and overcome sectoral and special interest arguments vying to influence policy.

The decision to reduce tariffs in the 1970s had a strong theoretical basis. Nevertheless, it took quantitative evidence of the extent of the implicit taxes and costs on the losers from the tariff regime to galvanise support for reform. Similar evidentiary process under the National Competition Policy were pivotal in delivering the reform over the last decade (Banks 2009).

Evidence-based approaches cannot, of course, guarantee perfect policy. Evidence is sometimes difficult and time-consuming to obtain, and can be incomplete or indecisive. Even when a single study is robust, it may not be sufficient to be conclusive. And the realm of public policy and the social sciences are further complicated by the interdependencies in our society and economy. Governments are sometimes under pressure to act quickly, and politicians rightly bring their own reading of community values and objectives to bear in formulating policy. In this context, an evidence-based policy model will not provide all of the answers; nor may it be entirely welcomed. But it can improve the basis for decisions, help avoid costly mistakes and make it transparent when political trade-offs are made.

What does robust evidence-based policy look like?

There is a vast number of 'how to' handbooks and methodological guidelines for generating evidence and evaluating public policy. This paper is not intended to replicate such an exercise or provide a handbook for applying specific evaluation methodologies. Different policies and variations in data availability require analysts to assess the most appropriate evaluation technique and navigate the technical pitfalls and advantages of each.

Rather, this paper sets out some suggested principles for sound evidence-based policymaking, starting with methodological challenges, before moving to institutional and process issues. Most of the principles will be familiar to policymakers — they have been highlighted in policy manuals, Commission reports in one form or another, and reflected elsewhere such as through the Regulatory Impact Statement (RIS) process. This exercise aims to consolidate them, and to move beyond a statement of principles by illustrating their application through specific policy examples. Wider knowledge and a better understanding of such a 'checklist' of principles might serve as a practical guide to governments, officials, analysts, journalists and the general public in thinking through policy proposals.

Naturally, identifying such a set of principles necessarily involves an element of specifying best-practice, or setting a benchmark, which can appear removed from the reality facing many policymakers. For the most part, policy assessment must be done in real time, often with limited data and evidence. Even so, the aim of an evidence-based approach should be to improve the analysis for decisions, even if the evidence is not 'perfect' every time.

The challenge of gathering this evidence will vary depending on the policy area under investigation. Mulgan (2003) identifies three fields that policy challenges will typically fit. In relatively stable policy fields, where the knowledge is reasonably settled, the theoretical foundations are strong and there is a significant evidence

base, most research involves filling in the gaps and refining insights. In the second category are policy fields where the knowledge base is contested, there is disagreement over the theoretical approach, and the evidence base is uneven. And then there is a third category, where the level of uncertainty is such that there is virtually no evidence base at all (2003 pp. 6-7).

In all three fields evidence has a critical role to play. Nevertheless, there will be cases in the latter two fields where the challenge is to make the best use of evidence, and in its absence, to recognise uncertainty and take a sensible approach to policy development. Inherently new and unique policy problems, such as global warming and biotechnology, present the greatest challenge here.

In these cases, the existing evidence base will often be insufficient for decision making purposes, and can even be misleading. The latter case arises with 'black swan' problems, so named because prior to the discovery of Australia it had been concluded (with no available evidence to the contrary) that all swans were white. A modern day equivalent would be the UK Government's declaration in 1990 that there was no evidence of a threat to human health from BSE (mad-cow disease) and that eating British beef was completely safe. Tragically in this case, the initial absence of evidence of transmission did not preclude the possibility.

This does not support the view that action should be taken in advance of evidence in areas of uncertainty and risk. Such an approach reflects a rather binary view that policymakers either use evidence or not, rather than the more variable role that evidence actually plays (see box 1.2 on the precautionary principle). In cases of policymaking under uncertainty, such as global warming and public health risks, the challenge is to generate the best evidence possible, while implementing any necessary early policy action in a way that recognises uncertainties. As the OECD (1999) has argued, the objective of evaluation is not necessarily to provide an absolute truth but to provide insight and well-justified views on policy programs.

Box 1.2 **The precautionary principle and evidence-based policy**

The most widely adopted definition of the precautionary principle in Australia, based on the Rio declaration, seeks to ensure that uncertainty about potentially serious hazards does not justify ignoring them.

> Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation (United Nations Environment Program, 1992, Principle 15).

Unlike more prescriptive versions of the principle, which mandate action in response to uncertainty, regardless of the magnitude of the potential threat or the costs and benefits of action, there must be plausible, albeit uncertain evidence, relating to both the likelihood of occurrence and severity of the consequences. Scientific uncertainty alone or the possibility of minor damage will not satisfy the test for triggering application of the principle. Any precautionary action should be cost-effective.

Under such a definition, evidence has a significant role to play in:

- marshalling the evidence that is available on the size and scope of the problem
- modelling a selection of potential outcomes or a worst case scenario
- developing an adaptive approach that can respond to new information.

Source: Weier & Loke (2007).

2 Methodological principles

At the broadest level, sound policy evaluation and review generally require a number of steps: from specifying the rationale for government intervention; through examination of the nature and causes of a policy problem; setting out costs and benefits of each policy option (where possible, in both qualitative and quantitative terms); and identifying a preferred policy measure. An evidence-based approach emphasises the informational or evidentiary frameworks that may be usefully introduced at all stages of an analysis to strengthen policy outcomes.

Commission experience suggests a handful of generic methodological challenges quite frequently limit sound policy assessment. Drawing on Banks (2009), table 2.1 sets out suggested principles to address these challenges. The remainder of this section illustrates the importance of each principle, using examples that either illustrate the hazards of ignoring the principles, or cases where surprising policy insights and improvements have arisen from applying the principles.[1]

The aim of raising the principles is not to suggest that every methodological problem can be solved perfectly. In some cases there will be no practical, timely way to completely address the principle (perhaps because of data limitations). And the nature of the individual case will partly determine the degree to which evidentiary frameworks should be developed and applied. But having the principles in mind, and having thought through their application, may caution against needlessly risky policy changes where an alternative approach might achieve the ultimate objective just as quickly, and with lower risk of error and wasted resources.

[1] The policy examples in this paper have been chosen for illustrative purposes. Their use does not imply that in the chosen studies the methodological approach or conclusions drawn are without fault.

Table 2.1 **A matrix of suggested principles for evidence-based policy**

Methodological issues

Suggested principle	*Why?*
1. Define the problem carefully - understand the nature and size of the problem and the objective of policy intervention.	The size and scope of a problem will help define an appropriate policy response. In practice, policy objectives are often vague, conflicting or confuse the desired outcome with the means of obtaining it.
2. Consider all potential options for addressing the problem.	Focus on only one intervention may preclude better options and risks disproportionate or clumsy responses.
3. Rigorously assess the quality of existing evidence.	It is difficult to categorise *types* of evidence as 'good' or 'bad'. The available evidence needs to be rigorously assessed for quality and robustness. If evidence is seriously deficient, a cautious, progressive, trial-driven policy response might be warranted.
4. Consider the 'counterfactual'.	Many social and economic trends continue to some extent in the absence of policy intervention, for example because of rising incomes or education. Judging the impact of the policy requires a realistic benchmark of what would have happened without the policy.
5. Consider 'attribution' issues and design ways of handling possible multiple causation.	Most policy acts through complex economic and social systems. It is often difficult to estimate the impact of the policy compared to that of other influences that are simultaneously in play.
6. Consider possible selection bias, optimism bias, model misspecification and other sources of bias in evaluation.	Peoples' behaviour can change merely because they are being studied, and people who choose to be studied may be different than those who do not. Policy advisers want to believe their preferred policies will work, and can overestimate policy benefits and underestimate costs.
7. Account for all the effects across the community and the economy.	Policy analysis is often limited to considering only the benefits or costs of the policy, its immediate effects or the impact on a single group. There are often important indirect effects through financing the costs of the policy, scaling it up from initial trials, and other 'general equilibrium' linkages.
8. Use a cost benefit framework, even when incomplete.	Even imprecise measures of the impact of a policy can be valuable, because the process of trying to measure costs and benefits can identify those proposals that are worth proceeding with and those that are not.

Principle 1: Define the problem carefully

Understanding the nature, size and scope of the problem is a prerequisite to effective policy — failure to do that properly is a common cause of policy failure and poor regulation. Yet, this stage of policy development is often overlooked in practice. The philosopher Karl Popper criticised the natural tendency for what he called 'solutioneering': the jumping to solutions and planning without defining the problem or determining whether there is one (Maynard 2002).

For instance, the Regulation Taskforce (2006) found that in many cases the rationale for regulatory interventions had not been clearly established and that pressure on governments and their regulators to 'do something' in response to the crisis of the moment had created a 'regulate first, ask questions later' culture. A recent survey of risk-related Regulatory Impact Statements (RISs) supports this view (Austin et al. 2008). This investigation found that more than half of the RISs provided no evidence, or at best, anecdotal evidence, on the severity of the problem.

From the United Kingdom, the *Dangerous Dogs Act 1991* provides some useful lessons. A series of dog attacks, some on children, prompted widespread media coverage and calls for the government to respond. Emergency legislation, passed in record time, led to an Act that made it illegal to have dogs of various descriptions such as a Japanese Tosa (although no one had ever seen one in the United Kingdom), or vague categories such as a 'pit bull type'. Such dogs could be destroyed on an order of a court. This produced extended arguments and court cases about what a dog of 'the type' meant. Ironically, the same media that emotionally urged harsh and immediate action against vicious dogs, then published photos of doomed family pets, so the immediate regulatory action provided only temporary relief from tabloid attack (Lattimore 2009).

The Act demonstrates the risk of regulating without appreciating the nature of the problem:

- it lacked proportionality (dogs that had no history of aggression were put down)
- it badly targeted the problem (the evidence did not suggest that the barred dog types were particularly dangerous relative to many omitted breeds)
- it led to enforceability problems since the specified dog types were ill-defined.

An assessment of any problem is not limited to estimating its size and scope, but extends to identifying the potential *cause* of the problem. It could be argued, for instance, that housing affordability is a problem, but what has caused the problem? The wide range of factors influencing housing supply and demand throw up a range of possible causes: the cost and availability of finance, demographic changes, restrictive planning laws and government taxes (PC 2004b). Determining the relative contribution of these factors will have a large bearing on choosing the most effective policy solutions.

Box 2.1 Identifying the cause of the problem: The Great Chicago fire

When presented with a policy problem, the initial response is often to assume it points to a specific cause. But evidence of a problem does not equate with evidence of the cause of the problem. Even some of the most robust and intuitive 'cause and effect' chains have been found to be fallible. For example, it is generally accepted that sugar causes hyperactivity in children, but at least 12 double-blind randomised controlled trials could not detect any differences in behaviour between children who had sugar and those who did not (Vreeman and Carroll 2008, p. 1442).

DiNardo (2005, p. 12) examines this issue using the famous example of the Great Chicago fire of 1871. He asked: what does it mean to say that Mrs. O'Leary's cow *caused* the Great Chicago Fire of 1871? Even if we were to agree with this version of events:

> One dark night, when people were in bed,
>
> Mrs. O' Leary lit a lantern in her shed,
>
> The cow kicked it over, winked its eye, and said,
>
> There'll be a hot time in the old town tonight.

As to the 'ultimate' cause of the fire, you could attribute the cause of the fire to Mrs. O'Leary's cow. You could also argue that Mrs. O'Leary, and not her cow, was the cause of the fire since placing the lantern in the barn had the predictable consequence of igniting a blaze. More policy relevant perhaps, you could cite lax fire regulations as the cause: perhaps Mrs. O'Leary would have been more cautious had the placing of a lantern in a barn been illegal. More fancifully, you might even trace the cause back to US agriculture subsidies. Without the government subsidies, maybe Mr. and Mrs. O'Leary would have decided not to take up dairy farming at all.

Similarly, the challenge of deciphering the causes of childhood obesity, and in particular, whether 'junk food' advertising during children programs has a significant effect has proved a complex exercise (see box 2.2). Although there is no conclusive evidence, the available evidence suggests that junk food advertising is unlikely to be a major cause of childhood obesity, so a ban is unlikely to result in a substantial reduction in overweight children.

Box 2.2 Defining the problem carefully: Advertising and childhood obesity

The Australian Communications and Media Authority's (ACMA) recent review of children's television standards elicited considerable community concern over the contribution of junk food advertising to rising rates of childhood obesity. Around 90 per cent of submissions raised issues about food and beverage advertising. ACMA also received 20 521 postcards calling to ban junk food advertising to children as part of the Cancer Council's 'Pull the Plug' campaign (ACMA 2008, p. 10).

Australian children are reportedly exposed to more television food advertising than in the United States, United Kingdom, New Zealand or 11 other western European countries. The average estimate suggests 10 food advertisements appear per hour on children's television, with 80 per cent advertising energy-dense foods like fast food, soft drink and chocolate. This equates to the average Australian child viewing 6 074 advertisements for energy-dense food per year, or 17 per day (Carter 2006 p. 8). The impact of childhood obesity is equally alarming. Obese children are at greater risk of developing cardiovascular disease, high blood pressure and type II diabetes. Obese children also have poorer gross motor development than their peers. A large proportion of obese children (50 - 80 per cent) become obese adults (Carter 2006, p. 5).

But does junk food advertising contribute significantly to childhood obesity?

A systematic review of studies on childhood obesity found that factors affecting obesity are complex, involving the interplay of hereditary, social, cultural and environmental factors. It found that there is a correlation between advertising and children's knowledge about the nutritional value of foods, their food preferences and their requests for certain types of food (commonly labelled 'pester power' or 'kidfluence'). There is also a correlation between television viewing (as a passive activity, distinct from the advertising that it carries) and obesity in children. However, the research does not demonstrate that any of these relationships were causal, or isolate the contribution of advertising to childhood obesity (ACMA 2008, p. 11). The limited evidence that is available suggests that the strength of the association is modest, with television advertising/viewing accounting for about two percent of the variation in food choice/obesity (Ofcom 2006; Marshall et al. 2004; Wake et al. 2003).

In the absence of conclusive evidence, an alternative would be to consider a 'least cost' approach and place a restriction only on advertising energy-dense foods. However, the experience in the United Kingdom suggests that such an approach can also have unintended consequences. Its regulation bans advertising food and beverages which exceed a certain threshold for fat, sugar and salt (based on a 100 gram serve). Contrary to the intent of the regulation, the ban applies to cheese, yogurt, dried fruit and nuts, but not supermarket fish fingers, frozen chips and nuggets.

What are the objectives or goals of government action?

Once a policy problem is well understood, it is important to clearly specify the objectives of government intervention. Objectives that are clear, specific and measurable can guide policymakers in choosing from the range of policy instruments to address the problem. It also establishes the criteria on which the performance of the policy can be judged.

Some policy and legislation have no explicit objectives. In the 2001 review of the National Access Regime (Part IIIA of the *Trade Practices Act 1975*), the Commission found that the Regime lacked clarity and guidance for infrastructure owners, access seekers and those implementing and administering the legislation (PC 2001, p. 125). Without a specific 'objects clause' regulators had to infer objectives from associated regulations, ancillary material and legislature discussion — thereby widening the ambit of regulatory discretion.

Even where objectives are specified, a common shortcoming is to confuse the desired final outcome with the means of obtaining it (contrary to the caution by the Australian Government's Best Practice Regulation Handbook 2007, p. 63). For example, a broad objective of environmental regulation may be to reduce carbon emissions. This objective differs from narrower proposals such as 'increasing renewable energy production' or 'introducing an emissions trading scheme', which are two of the many means of attaining the broader objective.

In addition, policy objectives are often unclear, evolving and even conflicting. Sometimes attempts to state policy objectives simply describe what the policy will do (in an administrative sense) rather than what the policy seeks to achieve. For example the stated objectives of family assistance policies are to:

- assist families with the cost of raising children (Family Tax Benefit A)
- provide additional assistance to families with one main earner (Family Tax Benefit B), and
- recognise the legal relationship between mother and child, the role of the mother in the birth of the child and the extra costs associated with the birth or adoption of a child (Baby Bonus) (FaHCSIA 2008).

Evaluating the programs against these objectives, it would be difficult to conclude that the policies are anything but fully effective since payments will increase the resources of families with children. But more targeted measures might be superior if the underlying objectives were to improve maternal and child health and welfare, or increase the number of young children in full-time parental care.

Determining the most appropriate policy response will depend on the objectives the government is seeking to achieve. The Commission's inquiry into Paid Parental Leave found that although a paid scheme is commonly promoted as a means of achieving a wide range of objectives, only some are best targeted using such a scheme. Objectives that could be pursued through paid parental leave include enhancing maternal and child health and development, facilitating workforce participation and promoting gender equity and work/family balance. But objectives that have relatively weak rationales for paid parental leave, include financial assistance (there are more targeted ways to provide financial assistance to needy parents) and increasing population fertility (the capacity to make a significant difference to fertility levels in a cost-effective manner is small). Designing the key features of a paid scheme will depend on the government's objectives and the trade-offs that need to be made among them (PC 2008a, p. 1.1).

Clear objectives also establish a basis to assess the success or failure of a policy. For example, the clear statement of objectives for drought policy easily demonstrated that the observed pattern of expenditures was not achieving those objectives (box 2.3).

Box 2.3 National drought policy: clearly stated objectives facilitate evaluation

The objectives of the National Drought Policy (NDP) are to:

- encourage primary producers and other sections of rural Australia to adopt self-reliant approaches to managing for climatic variability
- maintain and protect Australia's agricultural and environmental resource base during periods of extreme climate stress
- ensure early recovery of agricultural and rural industries, consistent with long term sustainable levels.

The Commission found that a striking feature of the NDP was the mismatch between its policy objectives and its programs. From its inception, the policy centred on helping farmers build self-reliance to manage climate variability and preparedness to cope with droughts. Program expenditures, on the other hand, have not been directed to this end but have predominantly flowed as a series of emergency payments to a minority of farmers in perceived hardship and to farm businesses meeting eligibility criteria.

Source: PC (2008b).

Some lessons

- Understanding the policy problem is half the battle. Even where there is little definitive evidence on the size, scope and causes of the problem, rigorous investigation can clarify what, if any, government action is appropriate, or where the government can best target its efforts.
- Evidence of a problem does not equate with evidentiary support for a cause of a problem or any particular solution.
- Objectives can help policymakers design the most appropriate policy solutions. Importantly, clear objectives provide a standard to evaluate the effectiveness of a policy.

Principle 2: Consider all potential options for addressing the problem.

The existence of a policy problem, does not, of itself, justify government involvement. The case for government intervention must be based on a rigorous assessment of the relevant costs and benefits of the policy options. The first step in this process is to identify and test a range of alternative instruments such as budget measures, regulation, self-regulation, market-based instruments, providing information (e.g. educational campaigns), and taking no action.

Frequently, however, only one proposed solution is considered:

> In situations where government action seems warranted, a single option, no matter how carefully analysed, rarely provides sufficient evidence for a well-informed policy decision. The reality, however, is that much public policy and regulation are made in just that way, with evidence confined to supporting one, already preferred way forward. (Banks 2009, p. 8)

Failure to think through options and test the alternatives can produce poor outcomes for the community (see two historical examples in box 2.4).

A more contemporary example is the move to ban plastic shopping bags from supermarkets. In 2005, the Commonwealth, State and Territory governments agreed to phase out plastic bags because of the alleged problems that plastic bags pose for the litter stream and marine wildlife. But the Commission's inquiry into waste management found a wholesale ban on plastic bags was unlikely to address the problems attributed to plastic bags, or solve the litter problem more generally. This is because a ban would penalise most uses of plastic bags, whereas the potential environmental benefit would only come from the less than one per cent of bags that are littered (box 2.5).

Box 2.4 **Assessing the options for government intervention**

When the 'cure' is worse than the 'disease'…

In 1974, the then Australian Government introduced a reserve price scheme for wool to protect wool growers from market fluctuations. Under the scheme, which was largely driven by tumbling wool prices in the late 1960s and 1970s, the Australian Wool Corporation (AWC) set minimum prices for different categories of wool and then used grower funds to buy wool that did not reach the prescribed price, aiming to hold it until the market improved.

Initially, the scheme appeared to 'work' and prices stabilised from 1974 to 1987. But by the late 1980s market conditions had changed. The floor price had been set too high, and as a consequence, the AWC had amassed a stockpile of 4.75 million bales of wool, with an associated debt of $2.6 billion. The scheme joined the extended ranks of failed attempts to stabilise prices, as its key requirement — knowledge of how the long-run, market-clearing price related to observed prices — was unavailable to the scheme's administrators, who also faced systematic incentives to overestimate the price.

In 1991, the reserve price scheme was scrapped. For a short time, wool growers were paid a government subsidy to kill their sheep. It took over ten years to sell the last bale from the wool stockpile.

Or when an intervention fails to account for all the unintended consequences….

In 1967, the Australian Conciliation and Arbitration Commission issued a decision granting Indigenous pastoral workers equal access to statutory minimum wages. The reasoning was straightforward, motivated by principles of equity and directed at desirable ends, but as some Indigenous leaders such as Noel Pearson have noted, it had some perverse, even disastrous, consequences.

Immediately following the decision there was a dramatic decline in the number of Indigenous pastoral workers – estimated to be around 35 per cent, but in some areas, closer to 50 per cent (Henderson 1985 p. 109). Indigenous employees were replaced with white employees, and capital was substituted for capital, including accelerating the introduction of better fencing and helicopter mustering.

As a result, many unemployed Indigenous workers moved into government settlements. Some of those that were still employed resigned to join the settlements and maintain family ties. Without the necessary skills and few opportunities to gain alternative employment, Indigenous people had little option but seek unemployment benefits, with devastating social consequences for both the former employees and their families.

Box 2.5 **A ban on all plastic bags: the best option to tackle litter?**

In 2005, the Commonwealth, State and Territory Governments jointly announced a goal to phase out plastic bags by the end of 2008. The underlying rationale was that plastic bag litter is a particularly undesirable source of litter because it:

- can be highly visible and long lasting, since plastic bags easily become airborne, are moisture resistant and take many years to decompose; and
- has the potential to injure or kill wildlife, particularly in the marine environment through ingestion or entanglement.

However, closer investigation found that:

- Only a small proportion (0.8 per cent) of plastic bags become litter (although available estimates suggest that in absolute terms, plastic bag litter is significant).
- The commonly cited statistic that plastic bags are responsible for the deaths of 100 000 marine animals per year is not supported by the evidence. The figure was mistakenly taken from a Canadian study that found between 1981 and 1984 more than 100 000 marine animals and birds died in discarded fishing nets, not from plastic bags.
- Although the overall impact of plastic bag litter on marine life is uncertain, the available evidence suggests fishing related debris, rather than land-based sources, is the principal source of litter hazardous to marine wildlife.
- Despite successful initiatives to reduce the use of plastic bags, particularly from supermarkets (the number of retail carry bags fell by 34% from 2002 to 2005), the decline does not appear to have translated into a fall in plastic bag litter. This is because the likelihood of a supermarket bag being littered is low as people use them to carry goods to their homes. Bags supplied for away-from-home uses – such as to carry takeaway food – are much more likely to be littered.
- A cost-benefit analysis, commissioned by the Environment Protection and Heritage Council (EPHC), considered eleven regulatory options (ranging from a ban or levy through to a new code of practice for retailers) and found that all of them would impose a net cost on the community.

Source: PC (2006).

The Commission recommended an assessment of alternative policy approaches that specifically targeted litter, rather than all plastic bag uses and to focus on away-from-home sources of litter entering marine environments. The subsequent Regulatory Impact Statement (RIS), prepared by the EPHC secretariat, acknowledged that targeting litter may be a more cost-effective solution to the plastic bag problem, but rejected the option because it was inconsistent with EPHC's commitment to phase out plastic bags (EPHC 2008, p. 49).

In April 2008, the EPHC decided not to endorse uniform regulatory action to ban or place a charge on plastic bags. However, several states and territories are trialling plastic bag levies. South Australia has implemented its own ban.

Principle 3: Rigorously assess the quality of existing evidence

There are few policy problems where the evidence will be 'black or white'. And there are many cases where the challenge facing policy analysts is not one of lack of evidence, but of too much information or evidence that is contradictory or ambiguous. Sifting out the rigorous evidence from the poor quality is a challenging task. As Leigh (2009, p. 2) highlights:

> Reading solidly for 40 hours a week, 52 weeks a year, it would take a policymaker 18 months to get through the 6000 articles on 'early childhood intervention', 4 years to get through the 16,000 articles on 'teacher quality', or 5 years to get through the 20,000 articles on 'social housing'.

When it comes to assessing the quality of evidence in practice, the debate over what should count has been fuelled by a widespread interest in an evidence hierarchy, which ranks various research methods. The medical field has a long-established hierarchy of evidence, which places randomised controlled trials, or even better, meta-analyses[2] of randomised trials, at the apex (considered the 'gold standard') while qualitative evidence is assigned much lower credibility.

Some social scientists (e.g. Leigh 2009) have proposed a similar hierarchy for policy evidence, starting with randomised policy trials at the peak and descending through quantitative evidence (e.g. natural experiments or quasi-experiments) and qualitative evidence, to anecdote and expert opinion (box 2.6).

[2] Meta-analysis (or systematic review) combines the results of several studies that address similar policy questions. Meta-analysis can be performed on both quantitative and qualitative research, for example, meta-regression analysis uses statistical techniques to pool the results of all studies investigating a particular effect to provide an overall estimate of the impact of a policy. The meta-analysis result is therefore a more powerful estimate of the size of the true effect than that derived in a single study.

Box 2.6 **Leigh's evidence hierarchy for Australian policy makers**

1. Systematic reviews (meta-analyses) of multiple randomised trials

2. High quality randomised trials

3. Systematic reviews (meta-analyses) of natural experiments and before-after studies

4. Natural experiments (quasi-experiments) using techniques such as differences-in-differences, regression discontinuity, matching, or multiple regression

5. Before-after (pre-post) studies

6. Expert opinion and theoretical conjecture

All else equal, studies should also be preferred if they are published in high-quality journals, if they use Australian data, if they are published more recently, and if the issue they examine is more similar to the policy under consideration.

Source: Leigh (2009, p. 35).

The principal value of an evidence hierarchy is as a shortcut to filter potentially vast amounts of empirical evidence. It is intended to reflect the methodological strength of each research design, and its ability to generate robust evidence. For instance, a randomised policy trial can produce an unbiased estimate of the average impact of a policy in the population from which the sample has been drawn (box 2.7). Other methods of evaluation are viewed less favourably because of the potential for bias and/or their failure to prove causation rather than just correlation.

Even so, a hierarchy has some limitations. Many argue that a hierarchy is inherently more suited to medical research, which generally involves a test for efficacy measured by defined outcomes, often through universal biological processes (e.g. reductions in mortality and morbidity), but it is inadequate for assessing policy research, where policies are context-specific, may have multiple objectives, and manifest second-round, scaling, or general equilibrium effects that are not captured in small-scale tests (Nutley 2003). For instance, randomised policy trials are only useful for answering certain policy questions (box 2.8). It would not be feasible, for example, to conduct a randomised trial in relation to public goods like clean air or defence, where random assignment is not possible.

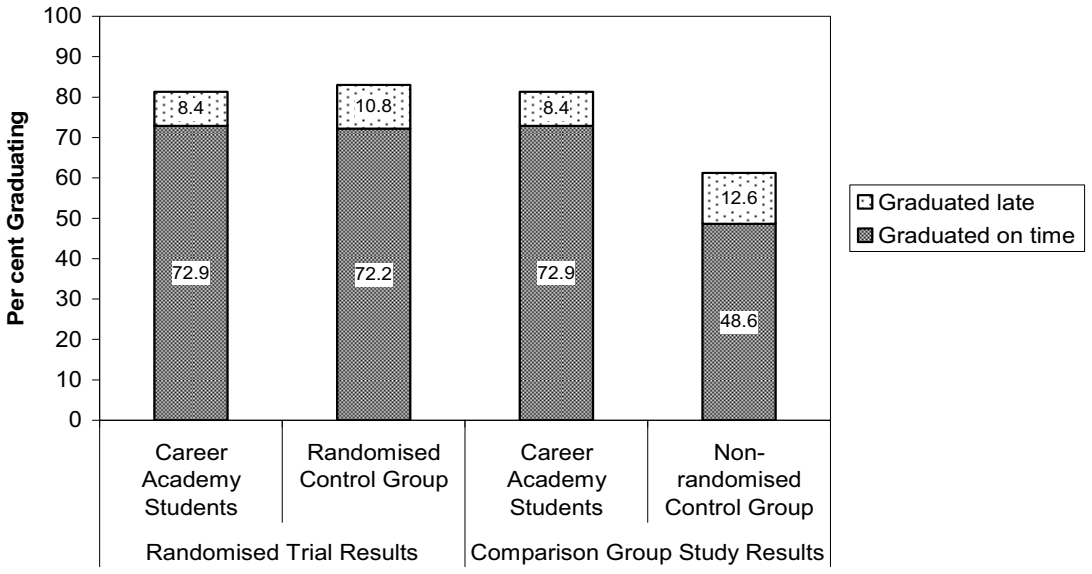
Box 2.7 **Randomised controlled trials: A 'perfect counterfactual'**

Randomised controlled trials expose an experimental group of people, and a non-experimental (control) group of people to exactly the same factors *except* the policy or program under investigation. The allocation of people to the policy intervention, or to the control group, is done purely on the basis of chance. Since random assignment should generate groups with the same average characteristics, the comparison between the two groups can be thought of as a comparison between two individuals who have the same characteristics, except for whether they are exposed to the policy. Comparing the average outcomes for the two groups will consequently estimate the causal impact of the policy on the population from which the two groups are drawn.

The unique advantage of random assignment is that it can determine whether the policy itself, as opposed to other factors, caused the observed outcomes. In effect, a randomised trial can provide the perfect counterfactual, and is, therefore considered more accurate than any other study design in measuring intended the effect of a policy.

For example, Career Academies is an educational program in the US that enrols students in academic and technical courses with a career theme in partnership with local employers. Participants' high school graduation rates are one of the outcomes of interest. A randomised trial of over 1700 students, which assigned student applicants into an academy or into a non-academy control group who continued regular schooling, found that the intervention did not result in increased graduation rates at the eight year follow-up (figure 2.1, left half). By contrast, if the evaluation had used a comparison group design comprised of like students from similar schools, the evaluation would have concluded erroneously that Career Academies increased the graduation rate by a large and statistically significant 33 percent (figure 2.1, right half).

Figure 2.1 **Impact of Career Academies on high school graduation rates**



Source: Borland et al (2005), OMB (2008), Leigh (2003).

Box 2.8 **Randomised controlled trials: The gold standard?**

Randomised trials are frequently referred to in medical and pharmacological testing as the 'gold standard'. But, despite their methodological strength, they are not always similarly powerful in social and economic policy applications.

Trials do not answer some questions of interest to policymakers. They:

- can show that a policy works but not *why* or *how* it works.
- cannot predict scaling or general equilibrium effects of a policy. (For example, a spending policy that, at full-scale, changes behaviours or requires higher taxation to finance it, may show strong benefits at small experimental scale, but fewer or no net benefits once the effects of behaviour change and taxation are taken into account.)
- cannot necessarily be extrapolated to other populations. The results of a social policy can be context specific – an education program might be effective in one country but not another.
- cannot be used to test some types of policy e.g. a government may not experimentally apply a regulation to one business and not another, or carry out a randomised trial to evaluate whether reducing carbon emissions will reverse global warming (although it may be possible to conduct a randomised trial to evaluate the cost and effectiveness of various options to reduce emissions).

Although randomised trials should avoid selection bias, particularly compared to other evaluation methods, they cannot eliminate all bias. Attrition bias, substitution bias and randomisation bias can compromise results e.g. those in the treatment group may fail to accept assignment, the control group may find a way to access the policy, and either can drop out of the experiment. The classic remedy of 'double blinding', so that neither subject nor the experimenter know which subject is in which group, is often not feasible in social experiments:

- In Project STAR, a large education experiment designed to test the effects of class size, about 10 percent of the students were moved to classes of different sizes than the ones to which they were randomly assigned at first, in part because of parental complaints and organised lobbying.
- In a subsidised meal program in Kenya, parents in over half of the control schools organised to raise funds for student meals to match what was being received in the treatment group.

Overall, randomised trials are a powerful tool and there are many circumstances in which Australian governments could use more of them, but they are not necessarily the 'gold standard' for all economic and social policy.

Source: Heckman and Smith (1995); Deaton (2009); King et al. (2007).

Another concern over evidence hierarchies is that some rank theory as the lowest form of evidence. This is a fundamental misconception. Theory is not a type of evidence. Theory establishes a testable, falsifiable framework in which to test

evidence. Theories validated by evidence provide a valuable tool to understand why and how a policy works (see box 2.9).

Second, theories that have been widely tested and validated can provide powerful policy guidance in new applications even where relevant data are scarce or non-existent. For example, it has been very widely demonstrated that an increase in the price of a good, usually leads (other factors remaining constant) to a decrease in the quantity of it demanded: demand curves slope downwards to the right. Apparent exceptions to the theory have also been widely examined and understood (for example, in terms of peculiarities in demand elasticities, cross price elasticities or income elasticities). Such a theory, validated by substantial evidence, can provide a solid basis for analysis of new cases, including through signalling whether policy formulation needs to proceed cautiously in a case where unusual circumstances might invalidate time-tested generalisations.[3]

To take another example, a theoretical benefit from reducing an industry's assistance may be that the price of the formerly protected good can fall, competitive market conditions will lead to price reductions to consumers, and consumer welfare will rise, through some combination of greater purchasing power more generally, and increased consumption of the cheaper good. If market data do not permit strong quantitative estimates of these benefits in the particular case, theory can still support robust analysis if the analyst asks questions such as: how well tested is this theory in other industry applications? Have there been other cases where the theory has not held? Is there evidence on the likely intensity of competition in the industry, and the probability of price reductions being passed through? Can we learn anything from a similar change affecting this industry in another country?

[3] The classic example in demand theory is the so-called 'Giffen good', for which quantity demanded is hypothesised to rise as the price of the good rises, because of the dominance of the income effect of the price rise over the substitution effect. Giffen thought this was observed for potatoes at the time of the Irish potato famine in 1845. Others dispute whether this effect was observed, but it remains theoretically possible (Mankiw and Taylor, 2006 p. 434).

Box 2.9 **Instrumental variables and the role of theory**

Quasi-experiments have became a prominent 'atheoretical' approach in cases where researchers have access to data from uncontrolled events, but no means to conduct an experiment. For example, the Instrumental Variable (IV) methodology estimates an 'effect' on the variable of interest (say Y, test scores) from another variable (say X, class size), using an 'instrumental variable' (say Z, thresholds of maximum class size set by schools), which is selected to be correlated with the second variable but uncorrelated with the error term in the estimating equation. Proponents argue that IV identification of 'effects' is computationally simply and robust.

However, Deaton (2009) has argued that IV is sometimes carelessly used in a way that illustrates an 'effect' of one variable on another, but gives no information on why that effect might arise. Any insight from the IV approach is difficult to generalise to other circumstances.

Recent analysis by Heckman and Urzua (2009) has shown that while well-defined IV processes can yield useful information about the average outcome from a policy, structural methods based on an economic model can identify a wider range of policy-relevant estimates, such as the distribution of outcomes as well as the average outcome.

> For under-identified structural models, it is possible to conduct sensitivity analysis guided by economic theory to explore the consequences of ignorance about features of the model. With IV, unaided by structural analysis, this type of exercise is not possible. Problems of identification and interpretation are swept under the rug and replaced by 'an effect' identified by IV that is often very difficult to interpret as an answer to an interesting economic question. (pp 20-21)

Bazzi and Clemens (2009) have recently shown that large parts of the econometric literature using IV methodology to study causes of economic growth have used invalid or weak instruments in their estimation, and lacked strong identification of causal effects. The studies cannot, therefore, support the policy interpretations that their authors draw from them. Part of Bazzi's and Clemens's response to this problem is to recommend a greater role for theory that accounts for the pattern of main findings in the literature, and to ensure that the choice of instrumental variable in the study at hand is not invalidated by its use in other studies.

Source: Deaton (2009); Heckman and Urzua (2009); Bazzi and Clemens (2009).

The reality facing policymakers in many areas, for example regulatory policy, is one of limited or patchy evidence. Perhaps the analyst has to work with theory, some information on overseas experience, the occasional policy pilot and general equilibrium or other modelling (often commissioned by a stakeholder), while also drawing on anecdotal evidence and expert and community opinion.

In these cases, an evidence hierarchy is less useful and can potentially be problematic, if researchers uncritically give more weight to flawed quantitative evidence than to the balance of all forms of evidence taken together:

> The technological and organizational factors … that have made high powered computing and large data sets widely available, coupled with the institutional efforts of the unscrupulous or untutored to offer empirical support for various policy measures, has enormously expanded the number of empirical studies of dubious quality. (Donohue 2001, p. 2)

The traditional remedies of peer review, transparency and replication can provide a good antidote to these problems, but are nevertheless incomplete (see appendix A).

In summary, the existing evidence base needs to be assessed not only based on the type of methodology used, but also its 'fitness for purpose' and robustness in application — some empirical evidence can be methodologically flawed, just as some qualitative evidence can be merely interest group opinion, and some international experience may not necessarily transfer to the Australian context. The integrity of a finding cannot be assured on the basis that certain methods have been used, but needs to be assessed in every case.

Additional options?

How might a policy adviser assess the quality of evidence? There are many formal standards and checklists for assessing research quality. Most tend to be variations on the established criteria of:

- Replicability and reliability – are the results reproducible and repeatable? Could another researcher generate the same results? Can the results be generalised to other settings and to other populations?
- Validity – does it measure what it says it does? e.g. does the IQ test really measure variation in intelligence? Do the results demonstrate a causal relationship between the policy and the outcomes?

Although these criteria were developed primarily for assessing quantitative research, several authors have extended these basic principles to develop a framework for assessing a range of evidence (box 2.10):

> As in quantitative research, the basic strategy to ensure rigour in qualitative research is systematic and … conscious research design, data collection, interpretation and communication. Beyond this, there are two goals that qualitative researchers should seek to achieve: to create an account of method and data which can stand independently so that another … researcher could analyse the same data in the same way and come to essentially the same conclusions; and to produce a plausible and coherent explanation of the phenomenon under scrutiny. (Mays and Pope 1995, p. 110)

Box 2.10 **A checklist for assessing research evidence**

Is the policy question clear?

Is the context clearly described and a literature review undertaken?

Is there an explicit account of the theoretical framework and methods used at every stage of the research?

Does the evaluation address its original aims and purposes?

How defensible is the research design?

Is there a clear description of methodology, including any data collection methods? Are the research methods appropriate to the question being asked?

Does the research make use of quantitative evidence to test qualitative conclusions where appropriate?

Is there scope for drawing wider inference – how well is this explained?

How clear are the links between evidence, interpretation and conclusions?

How reliable are the results? Does the research seek out and explain observations that may have contradicted or modified the analysis? Or test forecast performance (i.e. out-of-sample performance)?

Does the research consider the loss function (the costs of being wrong)?

Does the research address policy questions in a way that is both useful and useable?

Is the research peer reviewed? Can the evidence (including data, transcripts, recordings, submissions and analysis) be independently inspected and appraised by others?

Sources: Adapted from Mays and Pope (1995); Spencer et al. (2003).

Assessing the quality of individual pieces of evidence will not necessarily draw out definitive policy conclusions. A complementary step is to take all of the assessed evidence and information and carry out some type of process of weighing up (or triangulating) the evidence to formulate policy conclusions (see Hall 2009 for an overview of some of these methods). Public inquiries undertaken by the Productivity Commission provide one such process, by collecting and evaluating evidence from a variety of sources: from stakeholders (through visits, written submissions on an initial issues paper, public hearings and a second round of written submissions on a draft report), academic research, overseas experience and internal analysis and modelling (Banks 2007) (box 2.11).

Box 2.11 **Triangulating the evidence**

Triangulation is a method used by researchers to draw conclusions from a range of evidence. Synthesising multiple sources of data, theories and methods, assists in reconciling conflicting evidence, overcoming the weaknesses of single studies and deriving more accurate conclusions.

Triangulation can be performed over several dimensions:

- Data (the combination of different types or sources of data)
- Investigator (using several different evaluators)
- Theory (comparing different theoretical perspectives)
- Methodology (using multiple methods to study the policy)
- Mixed methods (combining quantitative and qualitative evidence, expert opinion, and anecdote).

The Inquiry into the Market for Retail Tenancy Leases in Australia (PC 2008c) stemmed from concerns among small retail tenants about leases over which they felt they had little control. A key question in the inquiry was: is there evidence of significant failings in the retail tenancy market? Or more generally, are the outcomes in the market broadly consistent with what would be expected in a competitive market?

To answer this question, the Commission weighed analysis of the market (type and scope of leases, competition for leases, incidence of business failure, formal disputes) with information collected from stakeholders through public hearings, written submissions and visits. Of the 211 written submissions that were received, 75 were received from tenants, 57 were from organisations representing tenants, 20 were from owners/operators of large shopping centres and their representatives, and 3 were received from small landlords, with the remainder contributed by government agencies, real estate agencies and other interested parties.

The Commission concluded that overall the market was working reasonably well:

- there was no convincing evidence that a systemic imbalance of bargaining position exists outside of shopping centres
- in larger shopping centres, there was stiff competition by tenants for high quality retail space and competition by landlords for the best tenants, reflected by relatively low vacancy rates and high rates of lease renewals
- the more desirable tenants and shopping locations were able to negotiate more favourable lease terms and conditions
- the incidence of business failure in the retail sector was not exceptional compared to other service activities
- formal disputes were relatively few and widely dispersed both geographically and according to shopping formats.

Source: PC (2008c).

Principle 4: Consider the 'counterfactual'

A major difficulty with assessing the impact of a policy is that there is typically a multitude of factors contributing to historical outcomes. As the Secretary to the Treasury highlighted during the Senate hearings on the $42 billion stimulus package:

> **Senator Abetz** — At what time in the future can we come back into this room and discern whether or not this package has worked? Or is that time never going to arrive?
>
> **Dr Ken Henry** — Through the course of this year and next year, as we get the figures we will do our best to make an assessment. But it will always be difficult because, in making that assessment, we will necessarily have to make a judgement about where the economy would have been without these measures, and that is even more difficult than estimating where the economy really is at. (Senate Committee Hansard 2009, p. 57)[4]

In order to decide whether the policy change of interest had any role to play, it is necessary to control for the effects of all other influences.

The counterfactual is an estimate (either quantitative or qualitative) of the circumstances that would have prevailed had a policy or program not been introduced. The counterfactual is particularly important in policy evaluation, since in most cases policymakers are trying to measure the impact of the policy in terms of change from what would otherwise have been, or the 'additionality' of a policy e.g. the increase in employment, the reduction in homelessness or the additional productivity, over and above what would have occurred without the policy.

While the outcomes in the presence of a policy should be observable (although not always easily measured), the outcomes in the absence of the policy are not obviously observable. As a result, many evaluations tend to rely on 'before and after' studies where the counterfactual is assumed to be a continuation of what was observed before the intervention. So it is reported that policy X has lead to a Y per cent decrease in electricity prices, policy Y has seen Z thousand jobs created since its introduction and so on. This assumption will only be plausible if other factors have very little influence, and is easily violated if there are other factors affecting outcomes over time (for example, economic growth or technological change).

There are numerous methods to estimate the impact of public policy that use different ways of constructing the counterfactual. A brief summary of some of these techniques is provided in table 2.2:

- The first is the pure experiment or randomised controlled trial.

[4] Forecasts based on economic modelling were published in the 2009-10 Budget Paper No 1, Statement No 4. The level of real GDP is forecast to be higher than it would otherwise have been by 2¾ per cent in 2009-10 and 1½ per cent in 2010-11.

- The second group involves natural or quasi experiments which attempt to artificially construct a counterfactual using various econometric techniques.
- The third major group are the more traditional model-based econometric methods including regression models, and their various extensions and corrections, such as the Heckman two step selection model.

There are many other ways of answering counterfactual policy questions, including structural simulation models and partial or general equilibrium analysis. However, their credibility rests critically on the quality of the empirical input and model behaviour used to generate answers.

The Commission has, over the course of many different projects, used a broad suite of qualitative and quantitative methods to answer counterfactual policy questions. Traditionally, the Commission has used structural modelling (partial and general equilibrium) to construct both historical and *future* counterfactuals [5] for trade and assistance policy questions. Some sectoral inquiries, such as the *Evaluation of the Pharmaceutical Industry Investment Program (PIIP)* (2003), have included a range of methods — simple comparisons of pre and post PIIP performance of participants and non-participants, a difference-in-differences estimate, a comparison of the levels of activity forecast by applicants versus actual levels achieved (since if the program is effective unsuccessful applicants would not achieve their targets), and case studies.

In other projects, however, the construction of any robust counterfactual has proved elusive. In quantifying the economy-wide impacts of National Competition Policy (NCP), it was not possible to disentangle the impact of NCP from other factors, such as the widespread adoption of information and communications technology. Specifying a counterfactual would have been largely based on judgements, which would in turn largely determine the results of the analysis (PC 2005, p. 2). The Commission took the approach of assessing the economy-wide effect of all productivity and service price changes, highlighting that the results obviously captured more than the impact of NCP, but that the bias was somewhat offset by the benefits of NCP that could not be captured in the analysis (PC 2005, p. 3).

[5] A future counterfactual is how different the world would look at some future point if a policy not yet in place were implemented, compared to how it would look at that same future point under a business as usual scenario (Dee 2005, p. 14).

Table 2.2 **Some methods for estimating policy impact relative to the counterfactual**

What	*How*	*Weaknesses*
1. Randomised controlled trials	A randomised trial measures the impact of a policy by randomly assigning individuals (or other units such as schools) to a treatment group, which receives the intervention, and a control group, which does not. Because the control group experiences what would have happened if there were no intervention, the difference in outcomes between the groups is the impact of the policy.	The results of randomised control trials can be compromised through attrition, substitution or randomisation bias. Trials cannot answer some important policy questions, and can be costly. Policymakers have ethical objections to trials in some cases (see box 2.8).
2. Natural or quasi experiments	Natural or quasi-experiments use various econometric techniques to artificially construct a counterfactual. These techniques include difference-in-differences, regression discontinuity and matching approaches.	Estimates are likely to be less precise compared to a randomised trial.
Differences-in-differences	The differences-in-differences technique identifies a similar population that is not affected by the policy, and tracks the outcomes of the policy and control group over time e.g. comparing the impact of a policy change in NSW with no change in Victoria.	The differences-in-differences technique assumes that differences between the control and treatment group would have remained constant in the absence of the treatment.
Regression discontinuity	Regression discontinuity compares individuals who are very close to some arbitrary cut-off such as an entry score or eligibility threshold. Individuals who fail to meet the cut-off are considered a good control group for those who narrowly exceed it.	Regression discontinuity relies on two assumptions of common time effects across groups and no compositional changes within each group. These assumptions make choosing a control group difficult. Large sample sizes are required to generate sufficient statistical power to detect a policy effect.
Matching approaches	Matching approaches attempt to control for observable differences between the treatment and control groups e.g. comparing outcomes for those children who were enrolled in pre-school, with those who were not, but who had similar observable characteristics.	Matching approaches can account for differences in observable characteristics, such as education or age, but not unobservable characteristics, such as intelligence and motivation.
3. Model-based econometric techniques	Regression models (and their various extensions) estimate how much of the variation in some outcome Y can be explained by variation in factors X and Z. If all potential influences are included in the analysis, then the estimated coefficients on each variable represents the effect of that variable on the outcome, holding the effects of all other variables constant.	Dependent on specifying the right model and assumptions. May need to account for bias e.g. self-selection, omitted variables.

Source: Leigh (2009); Borland et al. (2005); Dee (2005).

Ultimately, the appropriate method to tackle the counterfactual problem will depend on a number of factors: the type of policy under investigation; the policy outcomes of interest (e.g. the overall impact, the effect of the policy on intended recipients or extrapolation to a new policy proposal); the nature of the data available; whether it is for *ex ante* or *ex post* evaluation and the cost and importance of the policy under study. Importantly, treatment of the counterfactual needs to be explicitly addressed, and where it is not possible to credibly estimate what would have happened in the absence of a policy, appropriate qualification should be placed on any results or recommendations.

Principle 5: Consider 'attribution' issues and design ways of handling possible multiple causation

The challenges identified in the previous section on determining a counterfactual demonstrate the difficulty in establishing cause-and-effect. Seldom can a policymaker straightforwardly conclude that Policy A caused B.

A common mistake is to assume that correlation provides proof of causation (box 2.12). That is, when two events occur together they are claimed to have a cause-and-effect relationship. This can lead to some absurd conclusions: people who eat diet foods are more likely to be obese, therefore diet foods cause obesity; incarceration rates increase when crime rises, therefore incarceration causes crime.

Distinguishing causation when multi-faceted polices are acting on complex economic and social systems can be difficult.

Box 2.12 **Attribution challenges: do friends *cause* happiness?**

The British Medical Journal (BMJ) recently published a study which concluded that happiness is contagious within social networks. That is, your happiness depends on the happiness of your friends, and their friends and their friends. According to Fowler and Christakis (2008) "if your friend's friend's friend becomes happy, that has a bigger impact on you being happy than putting an extra $5000 in your pocket". This groundbreaking finding was reported in hundreds of newspapers around the globe.

Unfortunately, this happy proposition may be a statistical illusion. The study shows that your happiness is positively related to the happiness of your friends, and that this holds even after accounting for a number of other variables, including how happy you and your friends were a few years ago. This demonstrates correlation, but not causation.

Wolfers (2008) argues that there are at least three reasons why happiness is correlated within social networks. It may be that — as the study posits — happiness is contagious. Or perhaps people with similar dispositions are more likely to be friends (i.e. selection effects). The authors account for this by adding statistical controls for the past happiness of both you and your friends.

The third reason is perhaps the most likely: if you and I are friends, we are often subject to similar influences. If a friend of ours dies, we'll both be less happy. Or, less dramatically, if our football team wins, we'll both be happier (Wolfers 2008).

In the same issue of the BMJ, an article by Cohen-Cole and Fletcher (2008) demonstrates this point. They argue that caution is needed in attributing causality in studies of social networks, because current empirical methods are subject to potentially large biases that increase the likelihood of detecting social network effects where none exists.

They use Fowler and Christakis's approach on another dataset, and show that it leads to the unlikely conclusion that height, headaches, and acne are also contagious. The more likely explanation is that friends are subject to similar environmental influences.

Sources: Fowler and Christakis (2008); Cohen-Cole and Fletcher (2008); Wolfers (2008).

Case studies

Housing Affordability

In August 2003, the Government asked the Productivity Commission to conduct an inquiry into the affordability of housing for first home buyers. The driving factor was a perceived 'affordability crisis' — since 1996, house prices had more than doubled in nominal terms and had increased 80 per cent in real terms.

But there were opposing views on the causes. Were rising prices the result of a demand-induced bubble *or* a consequence of government supply-side policies:

restricted supply of land, excessive taxes/charges, and burdensome regulatory requirements?

Housing markets are large and interactive. There are many players on both sides of the market, and pervasive government influence at all levels. And there are strong cyclical as well as structural influences on market outcomes. Prices periodically rise and fall, and movements can vary across market segments (PC 2004b).

The Commission found that both demand and supply side factors had played a role. But the *dominant* cause of the price growth observed from the mid-1990s was something that could, in the first instance, have actually helped affordability: falling interest rates and rising incomes. But these factors had also driven a surge in demand, which supply could not quickly respond to (Banks 2006). Nevertheless, the Commission also found that there was scope for governments to increase the efficiency of housing markets and thereby improve price and affordability outcomes *over time*, by addressing regulatory and tax measures that unduly inflated demand or constrained the responsiveness of supply.

United States Crime Rates

When crime rates in New York city fell markedly in the 1990s, one early and popular conclusion was that Mayor Giuliani's 'zero tolerance' policing policy was having significant effect (Whaley, 1999). If policing effort was the main cause, this finding would have had obvious policy relevance all over the world — police and courts should focus more on deterring even minor crime.

But it was soon noticed that crime rates fell strongly all over the United States at much the same time. The magnitude of the fall was remarkable: experts had predicted an increase, for demographic reasons. In an 'echo' of the baby boom, the children of baby boomers were expected to enter the high-crime age brackets in large numbers, raising crime rates further, even under optimistic assumptions.[6] Instead, even cities such as Los Angeles that had not improved their policing, also experienced falling crime rates.

So the search for possible contributing causes widened.

In 2001, Donahue and Levitt published a striking hypothesis: increased abortions had significantly contributed to lowering the crime rate. They offered econometric

6 Prominent criminologist Professor James Allen Fox was commissioned in 1995 by the US Attorney General to report on crime trends, and predicted "the next crime wave will get so bad that it will make 1995 look like the good old days." Instead, juvenile homicide rates fell by more than 50 percent in the ensuing six years (Levitt 2004, p.169).

evidence using data for all US states, suggesting that after controlling for other possible influences such as differences across states and policing, incarceration policies, handgun laws and economic conditions, the *Roe v Wade* supreme court ruling in 1973, which had suddenly liberalised access to abortion, was initially estimated to have caused up to 50 per cent of the decline in criminal activity across the US twenty years later.[7] (Using later data and estimating procedures, Levitt (2004) subsequently reduced his estimate of the crime reduction attributable to abortion to about 20 per cent.) In addition, increased incarceration was initially estimated to have accounted for perhaps another 20 per cent of the fall.

The ensuing debate illustrated many of the measurement issues and attribution complexities that afflict most analysis of complex social phenomena with multiple influences. For example:

- The results suggested not only that abortion liberalisation was associated with a reduction in crime, but that most other expected influences — rates of incarceration, measures of policing intensity, and economic conditions were also statistically significant. Many influences were at work at once.
- Even though the study 'controlled for' many possible influences, perhaps other, unobserved variables were in play? Many researchers have since examined whether other factors could explain the reduction in crime. Perhaps, for example, the crack cocaine-driven crime peak was another unobserved variable distorting estimates?
- To add to the debate, other academics attempting to replicate Donahue's and Levitt's results (with the authors' encouragement, and with free access to the data), discovered a coding error that had led to some of the originally reported results not having been subject to some of the statistical tests that the authors' text claimed they had (Ananat et al. 2006; Foote and Goetz 2008). However, correction of this error did not make a significant difference to the results.
- Even after almost a decade of testing, some researchers remain unconvinced that the statistical sources and Donohue's and Levitt's methodology are sufficiently robust to support the estimates of the impact on 1990s crime of abortion liberalisation.

Appendix A provides a more detailed overview of the challenges of establishing causality in relation to US crime, and some broader analytical lessons from the US crime story more generally.

[7] For a period after Roe v Wade, abortion in the US rose to the rate of one for every two live births.

Microfinance

A final example that illustrates the difficulties of attribution is the debate on the effectiveness of microcredit in developing countries (box 2.13).

Box 2.13 **Attribution challenges: Does microfinance work?**

Microfinance has recently enjoyed great enthusiasm in development economics. One of its leading developers, Mohammed Yunus, won the 2006 Nobel Peace Prize for his work establishing the Grameen Bank in Bangladesh. The idea has spread beyond its origins in developing countries, and a microfinance bank has now opened in Queens, New York.

There are reasons why microfinance ought to work: instead of relying on 'real' or physical collateral (naturally scarce to the poor) to underpin judgements of creditworthiness and provide security for loans, microfinance relies in effect on the borrower's 'social capital'. The borrower's personal connections in a community vouch for his or her reliability. Microfinance typically also often involves close monitoring by the lender, so in effect business management services are bundled with the loan. Both factors should reduce default rates while making loans to poor people who would not normally be eligible for finance.

But hard evidence of the impact of microfinance has been surprisingly scarce. The first two decades' evidence was mainly non-experimental, and carried little weight because of problems of omitted variables, non-random access to programs, and self-selection and attrition among borrowers. Perhaps the beneficiaries of microfinance would have benefited equally or more from any access to finance, so the impact of the microfinance process itself was not rigorously tested. (To pose the question in this way is not to imply other avenues of finance would have been automatically forthcoming, but rather to focus on the relative merits of different forms of finance.)

A handful of studies appearing since the late 1990s appeared to offer quasi-experimental support for the beneficial impact of microfinance by using the instrumental variable methodology. However, a recent review of these studies, and replication and re-examination of their data by different means, has thrown doubt on their positive results. While there is no evidence to suggest microfinance does not work, the early claims that it does work are weak. The results may instead reflect reverse causality: richer, more creditworthy, better socially-conected or more entrepreneurial borrowers were receiving microfinance, and succeeding because of personal characteristics not displayed equally by non-borrowers, not because of microfinance itself.

A raft of randomised controlled trials is now underway, though the early published results show no significant beneficial impact from microfinance.

Source: Roodman and Morduch (2009).

Some lessons

- In complex issues of social policy, expert predictions and common sense explanations can be wrong. But so too can apparently scientific and quantitative approaches.
- It pays to be cautious and not over-interpret new results from single studies. Even peer review in prestigious journals and transparency are no guarantee of instant accuracy or immediate success in evaluating complex phenomena. Untangling the causes will often require on-going, sophisticated evaluation.

Principle 6: Consider possible sources of bias in evaluation.

Bias in policy evaluation can occur in a myriad of ways (see box 2.14). Some relate to methodology and have already been touched on, such as self-selection bias where individuals who choose to be studied are not representative of a random population. For example, phone-in or online polls frequently report that 90 per cent of respondents support capital punishment. But those who call in to give their opinion are self-selected rather than randomly selected. That is, people who are motivated to respond (because they have a strongly held opinion) are unlikely to be representative of the general population (box 2.15). Other biases such as optimism bias are cognitive in nature. That is, researchers and policy analysts are affected by their own value framework, past experiences and beliefs.

Box 2.14 **Some potential sources of bias**

- *Selection bias* arises from the way that data are collected. For example:
  - there may be self selection by the individuals or data units being investigated, making the participants a non-representative sample e.g. people who enrol in smoking cessation programs are likely to be more committed to quitting, and therefore more likely to succeed, than the general smoking population.
  - similarly, sample selection decisions by analysts can also result in a non-representative sample e.g. in selecting the end points of a data series a researcher could start the series at an unusually low year and end on a high one to maximise a positive trend.
- *Misspecification bias* occurs when a model is incorrectly specified e.g. the functional form is incorrect or the model omits important explanatory variables.
- *Publication bias* reflects a reluctance to publish or report results which go against a researcher's beliefs, a sponsors' interest or community expectations e.g. a 2000 survey of complementary therapy journals estimated that only 5 per cent of published articles reported a negative outcome (Schmidt et al. 2001).
- *Confirmation bias* occurs when researchers design an evaluation to only seek confirmatory evidence e.g. only including positive studies in systematic review or performing repeated experiments and reporting only favourable results.
- *Optimism bias* refers to the tendency to be over-optimistic in estimating policy or project outcomes e.g. a UK Government review of 20 years of major public procurement projects estimated average optimism biases of 17 per cent for the project's duration, 47 per cent for capital expenditure, 41 per cent for operating expenses and 2 per cent for benefits shortfall (Mott MacDonald 2002).

How an analyst will go about avoiding or accounting for bias, particularly in relation to methodology, will depend on the method and policy in question. Some of the most prominent approaches — randomised controlled trials and natural experiments — have already been highlighted. A 2006 study on the effect of migration provides a unique illustration of the relative effectiveness of these techniques in accounting for selection bias (box 2.16).

Box 2.15 **Response bias: the difficulty of estimating the number of problem gamblers**

In *Australia's Gambling Industries*, the Commission's original 1999 Inquiry into gambling, research identified the challenges of estimating the number of people with extreme gambling problems. The Inquiry commissioned an extensive telephone survey to help estimate the scale of the problem. Several overseas studies cautioned that population surveys were likely to yield underestimates, for reasons including:

- problem gamblers are less likely to be contactable at home
- financially affected gamblers are more likely to have had their phone service cut off
- those with severe gambling problems are more likely, if contacted, to refuse to participate in any survey
- of those who do participate, many do not honestly disclose their problem (a common feature where respondents feel they are engaging in any form of stigmatised behaviour).

The Commission established some dimensions to these distortions: around a quarter of problem gamblers receiving help from specialist agencies said they would not have participated in a survey prior to seeking help; and of those who would participate in a survey prior to seeking help, only 38 per cent believed they would honestly report the extent of their problem.

The Inquiry estimated that if a survey revealed a prevalence of gambling problems of 0.3 per cent, the true prevalence, correcting for the response biases above, could be about 0.7 per cent.

Source: PC (1999, pp, 6.34-6.36).

Of course, there are cases where there is no 'method' to check and correct for bias and other tools must be used. For instance, when the source of potential bias comes from the parameter values adopted in an economic model. In this case, the simplest way to test the validity of the results is to conduct a sensitivity analysis (box 2.17). More broadly, applying some kind of sensitivity analysis, scenario modelling or simulation allows researchers to test the limits of the evidence. What is the range for error? Is it possible that the policy will produce small benefits? Or are the costs large if the evidence is misleading? (Wilkie and Grant 2009).

Box 2.16 Experimental versus non-experimental measures of the income gains from migration: an illustration of selection bias

When an emigrant from a developing country works in a developed country, the potential gains are large for the sending country, the receiving country, the individual worker, and his or her dependents in both countries. On the other hand, scarce skills and initiative might be drawn from the developing country, and the receiving country incurs the cost of additional services and transfer payments.

How might one assess the personal income dimension of these effects? Migrants generally self-select in a way that makes it difficult to identify a similar 'control group' who did not emigrate, and general living standards tend to rise over time in both developing and industrial economies, thereby complicating the counterfactual.

Simple comparisons of the experiences of emigrants and 'stayers' are unconvincing, as differences in outcomes may reflect unobserved differences between emigrants and the control group (e.g. in ability, attitudes to risk and motivation).

A 2006 study (McKenzie et al.) addressed these problems with a unique experiment using the annual migration of 250 people from Tonga to New Zealand in the latter's Pacific Access Category (additional to other migration categories for skilled or family applicants). In this scheme, a lottery allocated scarce rights to emigrate. The chance of success in the lottery was about 10 per cent.

The study compared the income outcomes for the successful and unsuccessful applicants in the lottery to emigrate: differences were presumably limited to the random event of success or failure in the lottery.

In addition to this randomised experiment, the authors also sampled 'stayers' who did not apply for the lottery, and compared their experiences with the lottery winners' experiences, by five alternative, non-experimental statistical techniques: a single difference estimator; ordinary least square regression estimates; difference-in-difference regression estimates; propensity score matching; and an instrumental variables approach (see table 2.2 for an explanation of some of these methods).

The estimated gains to emigrants, using the preferred random experiment were very large: migration increased work income by about 260 per cent.

But selection bias was also shown to be a challenge that non-experimental methodologies could not wholly overcome: the other five methodologies all overestimated the gains to migration by between about 10 to 80 per cent. That is, the non-experimental methodologies falsely attributed to the fact of migration, income gains that were partly due to the earnings characteristics of the migrants themselves.

Sources: McKenzie et al. (2006).

Box 2.17 **The Stern Review and sensitivity analysis**

The Stern Review: The Economics of Climate Change concludes that climate change is a serious threat that demands urgent action. The Review contends that:

- the costs of climate change will be equivalent to losing between 5 and 20 per cent of global GDP each year, now and forever; and
- the costs of reducing greenhouse gas emissions to avoid the worst climate change impacts could be limited to 1 per cent of global GDP each year.

But, undertaking a cost-benefit analysis of dimensions so vast, long-term and uncertain is extremely difficult. As might be expected in this type of long-term analysis, small changes in critical parameters can have large impacts on final results.

Climate change is an area where damage costs are expected to initially remain small but gradually increase over time in a business-as-usual scenario. The costs of mitigation, however, would occur primarily in the near term. Discount rates are used to bring these potential long-term benefits and short-term costs together in a common time frame.

To make a unit of future consumption equivalent to a unit of current consumption a discount rate must be applied. One formula for enumerating a discount rate is:

Rate of discount = $\delta + \eta g$

Where δ is the rate of pure time preference (also called the utility discount rate); η is the elasticity of the marginal utility of consumption; and g is the growth rate of per capita consumption.

The main discount rate used in the Stern Review appears to be around 1.4 per cent per annum. The Review set: $\delta = 0.1$, which implies that the welfare of future generations should be treated roughly on par with current generations; $\eta = 1$, which assumes that people derive the same utility from an additional one per cent of consumption, irrespective of their pre-existing level of consumption; and $g = 1.3$ per cent per annum, based on historical average returns to very safe assets such as government bonds.

This low discount rate is the main reason the Review's headline estimates of damage costs are so much higher than most other studies — many times higher than the estimates by other prominent economists. Adding 1 percentage point to the discount rate reduces the damage cost estimates by more than half.

Determining the most appropriate discount rate is still a matter of debate, and may ultimately involve some degree of judgement. However, the review failed in not presenting a range of results for different discount rates. Stern did provide a limited sensitivity analysis in a postscript to the review published later, although the highest parameter values used generate discount rates that are still relatively low. When small variations in critical parameters can have large impacts on final results, a sensitivity analysis should be performed and the results reported to decision makers.

Source: Baker et al. (2008).

Principle 7: Account for all the effects across the community and the economy

Policy evaluation is often limited to considering only the benefits or the costs of a policy, the immediate effects or the impact on a single group. Taking a community-wide approach involves gauging the effects of various policy options on all parts of society – including firms and workers, consumers and taxpayers, the community sector and the environment.

This is particularly important when assessing policy proposals directed at specific industries or sectors, because what is good for part of the economy or community, need not be good for other parts (Banks 2008, p. 9). The classic example is industry protection (e.g. tariffs) and restrictions on competition (e.g. broadcasting and pharmacy regulation), but it is just as prevalent in other areas of social, environmental and regulatory policy. For instance, tighter credit regulation may help protect vulnerable and disadvantaged consumers from undue financial stress, but it may come at a cost overall, if it results in higher transaction costs, reduced availability or higher priced credit for everyone in the community.

Case Study: Biofuel Industry Support

In 2008, the OECD conducted an assessment of the economy wide impact of biofuel support policies. In most countries, biofuel production remains highly dependent on public support through budgetary measures (such as tax concessions and subsidies), blending or use mandates (which require biofuels to represent a minimum share of the fuel market) and trade restrictions (mainly in the form of import tariffs).

Australia is no exception. From 2007-08 to 2011-12, the Australian Government has committed more than $500 million to support the biofuel industry, including a production subsidy for ethanol of 38.143 cents per litre, excise-free status until 2011 (and then a 50 per cent concession thereafter) and various grant programs (PC 2009, p. 201). State and Territory governments have also introduced a number of programs, such as the escalating ethanol mandate in NSW (due to increase to 10 per cent in 2011).

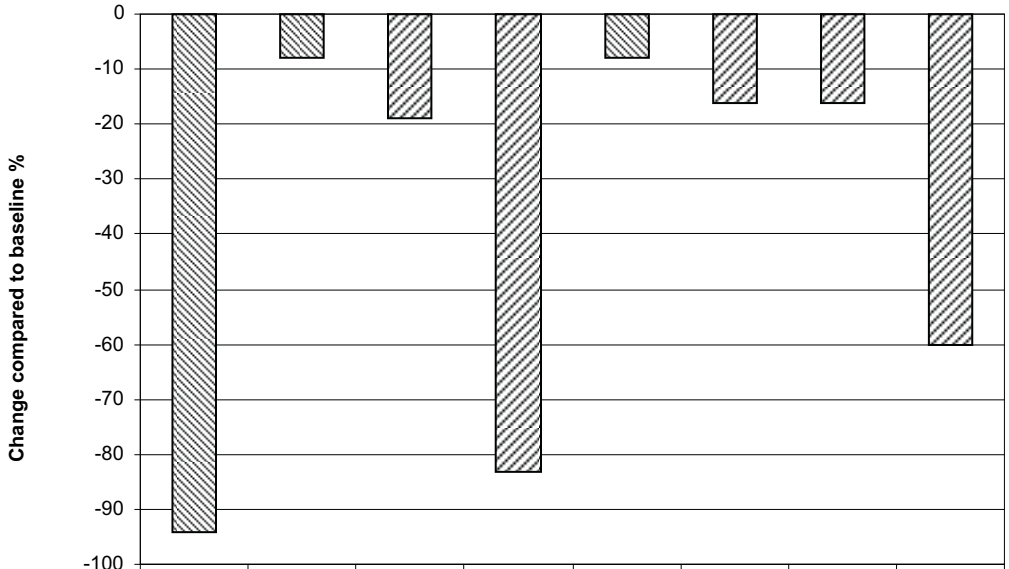
The key rationales for public support have been the ability of renewable biofuels to reduce greenhouse gas emissions and fossil fuel use. Secondary objectives include creating new market outlets for agricultural products, stimulating regional development and securing energy supply.

The OECD used a partial equilibrium model to examine the global impact of biofuel policies on production, use and trade, as well as agricultural markets. A stylised economic and natural science model was used to analyse the linkages between support policies and environmental outcomes.

The OECD found that the direct effects of biofuel support policies are indeed positive:

- Biofuels can reduce greenhouse gas emissions, although this varies significantly depending on the feedstock. Ethanol based on sugar cane generally reduces emissions by 80 per cent over the whole production and use cycle, relative to emissions from fossil fuels. However, biofuels produced from wheat, sugar beet or vegetable oils rarely provide emission savings of more than 30 to 60 per cent, while corn based ethanol generally saves less than 30 per cent.
- Support polices increase biofuel production and deliver a direct benefit to producers. The removal of these policies would lead to a substantial reduction in biofuel production and the (private) profitability of producers (see figure 2.2)

Figure 2.2 **Impact of biofuel support removal on biodiesel production**
2013-2017 average



	US	Brazil	Canada	EU	Malaysia	Indonesia	Other	World
Total effect (million L)	-1 636	-149	-93	-10 669	-66	-259	-270	-13 142

Data source: OECD (2008, p. 66). (For the impact on ethanol production also see OECD (2008, p. 66).

But, taken overall, biofuel support policies will not significantly reduce greenhouse gas emissions. An elimination of current biofuel support policies would increase net emissions from 2013-2017 by between 15 and 27 Mt of carbon dioxide — equivalent to no more than 0.5-0.8 per cent of the emissions from transport in 2015.

These relatively modest effects come at considerable costs in terms of transfers from taxpayers and consumers of US $25 billion on average for the 2013-2017 period, equivalent to between US $960 to US $1700 per tonne of carbon dioxide saved. By way of comparison, the Australian Mandatory Renewable Energy Target is estimated to cost around $30 per tonne of carbon dioxide (PC 2008d, p. 72).

In addition, biofuel support measures are estimated to increase average wheat, maize and vegetable oil price by around 5, 7 and 19 per cent respectively in the medium term. Prices for sugar and oilseed meals are actually reduced by these policies (a result of slightly lower production of sugar cane based ethanol in Brazil and significantly higher biodiesel related oilseed crush). New initiatives in the United States and European Union could further increase commodity prices by a similar magnitude.

Some lessons

Simply measuring the direct impact of a policy can give a vastly different answer than estimating the community wide effects.

Of course, the latter approach is far more complex and often requires large scale modelling. Every model is a simplification, and results will depend on the quality of the data, the credibility of the assumptions and scenarios modelled. A modelling exercise that attempts to model national or global markets will invariably be subject to data limitations and a degree of uncertainty will surround the results.

Although estimating the community-wide impacts of a policy is unlikely to give a precisely correct answer, modelling is (when conducted in a transparent way) a self-correcting exercise, in that debate and refinement over time can produce more accurate results. And it does assist decision makers to weigh up competing claims. As Alan Blinder put it:

> I often put the choice this way: You can get your information from admittedly fallible (models), or you can ask your uncle. I, for one, have never hesitated over this choice. But I fear there may be altogether too much uncle-asking in government circles. (1999)

In the biofuel case, the cost estimates would have to be out by at least a factor of twenty for support measures to be a potential cost effective policy for reducing greenhouse gas emissions.

Principle 8: Use a cost-benefit framework, even if incomplete

How to measure the impact of different policies will depend on the topic and the task — and whether it is an *ex ante* or *ex post* assessment. However, many evidence-based methodologies fit broadly within a cost-benefit framework, in that they are designed to determine whether there is an estimated (net) payoff to society.

A cost-benefit approach is useful because it can provide decision makers with quantitative information about the likely effects of a policy and does so in a broadly standardised and transparent manner, which can assist comparability and encourage consistent decision making (Australian Government 2007, p. 115). However, a cost-benefit approach should not be about quantification for quantification's sake — some policies will not be amenable to quantitative evaluation, in which case, it may be better to have no quantification rather than dubious figures. And a cost-benefit approach is more than just quantification: a rigorous qualitative identification of the costs and benefits can be useful.

Cost-benefit analysis present a significant challenge in practice, not least because it is often inherently difficult to accurately measure benefits and costs of government intervention. Consider for instance, the challenge posed by measuring the benefits to the community of clean air regulations or the intangible cultural benefits of saving a historic heritage building. But even when it is difficult to estimate costs and benefits with any precision, applying the framework is important and useful in itself. Cost-benefit analysis makes clear and transparent the assumptions and judgements made. Even imprecise measures can be valuable, because they may identify those proposals that are obviously worth proceeding with and those that are not (Australian Government 2007, p. 115).[8]

For example, box 2.18 reviews the cost-benefit analysis undertaken for the Commission's 1999 inquiry into Australia's Gambling Industries. Although this analysis could only provide 'ballpark' estimates within plausible ranges, it had some clear policy implications.

[8] Although cost-benefit analysis has a number of limitations, the alternatives such as multi-criteria analysis or triple bottom line reporting, suffer significant flaws. See, for example, Bennett and Dobes (2009) or Ergas (2008), commenting on the criteria analysis for national infrastructure project selection: 'The criteria are merely a list of questions. These criteria might be expected to invite applications for taxpayers to fund a shared tourism-wheat export inland railway, running on ethanol, with a jazz band playing on the last carriage. More seriously, they completely miss the point: it is not whether a project affects cities or regions, greenhouse gas emissions or quality of life … but whether it yields benefits that credibly outweigh the costs.'

Box 2.18 Quantifying the costs and benefits of gambling liberalisation

The Productivity Commission's 1999 inquiry into Australia's Gambling Industries attempted to quantify both the costs and benefits of gambling.

The Commission estimated that:

- The benefits to consumers are substantial – the extra value consumers derive from gambling above what it costs (i.e. consumer surplus) amounts to $4.4 billion to $6.1 billion per annum (1997-98).
- Claimed benefits of 'production-side' gains (jobs and income) from gambling liberalisation are limited. That is, gambling does not create many new jobs; what it does do is enable people to spend more on gambling and less on other things (jobs and income created in the gambling industry typically have a counterpart in jobs and income destroyed in other parts of the economy).
- As with the benefits, the costs from problem gambling are also substantial. Quantifying some of the social costs, such as family break-up and depression, is difficult, and it was necessary to use proxy measures and generate high and low estimates. But even based on conservative estimates (for example, not attempting to value the social costs of the 35–60 suicides attributed annually to problem gambling) still generated costs of $1.8 billion to $5.6 billion per annum.
- The net impact of the liberalisation of gambling could be anywhere from a net loss of $1.2 billion to a net benefit of $4.3 billion. However, there were significant differences across gambling modes, with lotteries showing a clear net benefit, whereas gaming machines and wagering include the possibility of a net loss.

Clearly, this quantification exercise could not reduce the impact of gambling liberalisation to a single number, or provide conclusive support for a particular policy. What the exercise did make evident, however, was that the social costs as well as the benefits of gambling were likely to be substantial, and that the risks of net costs were higher for some forms of gambling than others. This affirmed the need for considerable care in regulating the conditions of access to gambling. It also supported the Commission's general principle that regulation should be directed at effectively limiting the costs of problem gambling, without unduly impacting on the benefits for recreational gamblers.

Sources: Banks (2002); PC (1999).

Similarly, national security regulation creates unique challenges in measuring the costs and benefits of potential government intervention, with the primary challenge relating to estimating the probability of an attack (costs) and the change in that probability given the intervention (benefits). There are also intangible costs and benefits such as the loss of freedom to citizens.

The United States Department of Homeland Security has adopted an alternative technique to analyse security regulations. Break even analysis, sometimes called inverse cost-benefit analysis, estimates the reduction in the probability of a terrorist attack that would be required for the costs of a regulation to break-even with the benefits. Thus, if a break-even analysis concludes that a 50% reduction in the likelihood of a terrorist attack is necessary for the policy to have greater benefits than costs, then the policy seems unlikely to be a good idea. On the other hand, if such an analysis shows that only a 0.01% reduction is needed, then the policy is likely to have benefits that exceed its costs (Shapiro 2008).

3 Institutional and process issues

For evidence and evaluation to contribute materially to the selection of policies, it must be supported by institutional frameworks that incorporate evidence into the decision making process and encourage, disseminate and defend good evaluation (Banks 2009). The institutional framework should also ensure that the resources allocated to the evaluation are commensurate with the benefits it produces.

The experience of the United Kingdom is informative. Its evidence-based policy movement, which began in 1997, has seen greater investment in research, in the evaluation capabilities of the public service, and in the use of pilots and randomised policy trials (Mulgan 2003). Yet, over twelve years later, a House of Commons report on health initiatives suggests there is still some way to go to embed evidence in policy development:

> The most damning criticisms of Government policies we have heard in this inquiry have not been of the policies themselves, but rather of the Government's approach to designing and introducing new policies which make meaningful evaluation impossible. Even where evaluation is carried out, it is usually "soft", amounting to little more than examining processes and asking those involved what they thought about them. All too often Governments rush in with insufficient thought, do not collect adequate data at the beginning about the health of the population which will be affected by the policies, do not have clear objectives, make numerous changes to the policies and its objectives and do not maintain the policy long enough to know whether it has worked. As a result, in the words of one witness, 'we have wasted huge opportunities to learn'. (2009, p. 5)

Table 3.1 sets out some of the features of institutions and processes that may assist in incorporating evidence into policymaking.

Table 3.1 **A matrix of suggested principles for evidence-based policy**

Institutional and process issues

Suggested principle		*Why?*
1.	Design the most appropriate evaluation arrangements.	There are trade-offs between independent, ad hoc review, 'in-house' review, or a standing external review process.
2.	Maximise transparency — make data and evaluation public and provide for peer review and public consultation.	Provides quality assurance, improves credibility, aids government accountability, and facilitates improved evaluation over time.
3.	Establish a monitoring & evaluation program, including resourcing, at policy commencement.	Ensures data and evidence are available for evaluation. Will address the paucity of data in some policy areas.
4.	Consider sequential policy roll-out, pilot studies, or randomised trials as appropriate.	Useful for policymaking under uncertainty, where there is little settled evidence or where costs of failure are high.
5.	Disseminate evaluation and pool results across jurisdictions.	Improves evaluation practices and increases links between researchers and government. Assists in translating vast amounts of 'research' into policy 'evidence'.
6.	Ensure evidence is linked to the decision making process.	Provides an incentive for, and discipline on, government agencies to provide rigorous evidence to support policy proposals.

Principle 1: Design the most appropriate evaluation arrangements

Evaluation is likely to be more effective if there is some level of detachment, or independence, between the evaluator and the subject being evaluated. Options for providing evaluation are wide ranging and can vary from departmental (internal) evaluation units, evaluation institutions (such as the Auditor General), parliamentary committees, independent review panels, private consultancies and academics. The choice of evaluator will depend on the purposes of the evaluation task (for example, internal organisational learning, or broader policy review) and what might be at stake.

Internal evaluators have the advantage of intimate knowledge of the policy and easy access to information (both formal and informal knowledge). Working with those who designed and implemented programs, internal review can be assimilated into the organisation's operations and can disseminate improved approaches to staff in a non-threatening way. However, it can be difficult for in-house evaluation to achieve the desirable independence of perspective, where broader issues of policy are involved or if the issue cuts across agencies. External evaluators can overcome this problem, having the advantage of independence, but they may have less understanding of operational detail of the policy framework and may need long lead

times to master a new topic and marshal evidence through consultation and research.

The international Evaluation Cooperation Group[1] has drawn on the good practice standards of official audit and evaluation agencies that span across government and the corporate sector, to identify four dimensions of evaluation independence:

- *Organisational independence* ensures that the evaluation unit and its staff are not under the control or influence of decision-makers who have responsibility for the activities being evaluated and that they have full access to the information they need to fulfil their mandate.
- *Behavioural independence* enables the evaluation unit to set its work program, produce high quality and uncompromising reports and to disclose its findings without management-imposed restrictions.
- *Protection from outside interference* keeps the evaluation function free to set its priorities, design its processes and products, reach its judgments and administer its human and budget resources without intrusion by management.
- *Conflict of interest safeguards* guarantee that current, immediate future or prior professional and personal relationships and considerations are not allowed to influence evaluators' judgments or create the appearance of a lack of objectivity (Picciotto 2008, p 6).

An example of how institutions can design evaluation units to combine the advantages of internal insight to benefit organisational learning for the institution, with independence of analysis comes from the World Bank Group. Its Independent Evaluation Group is headed by a Director-General of vice-presidential rank, but instead of reporting like other Vice-Presidents to the President of the Bank, the Director-General reports directly to the Board. The position has only administrative links with the President. The Director-General is appointed by recommendation of the Board for a fixed five year term. While the occupant may come from the ranks of Bank management, the contractual appointment is not renewable and the occupant may not return into Bank employment — provisions which, together with the other arrangements noted, reduce the risk of conflicts of interest and bolsters behavioural and organisational independence (World Bank 2003).

[1] The Evaluation Cooperation Group is composed of the heads of evaluation of the multilateral development banks and observers such as the OECD's Development Assistance Committee's Evaluation Network head and the Director of Independent Evaluation at the International Monetary Fund.

Some suggest that for evaluation to be successful, it has to be performed in several places, which serves to cross-check outcomes. There are a number of examples in Australia of such arrangements, with one of the more prominent operating under national competition policy (box 3.1).

Overall, evidence-based policy cannot rest on any one evaluator or institution. Evaluation should to be decentralised and embedded in the way every policy agency works, but with the evaluator chosen depending on the policy review task at hand. The OECD (1999) recommends:

- internal evaluation when the main objectives are organisational learning and improved process;
- external evaluation by researchers and consultants when seeking new perspectives on policy or where there is a need for specialist evaluation skills;
- independent evaluation when assessing the effectiveness and efficiency of a policy.

Box 3.1 **National Competition Policy: A hybrid model**

Under the national competition policy (NCP), State and Territory governments were responsible for implementing NCP reforms, with the National Competition Council (NCC) acting as an independent monitor and assessor of performance. For example, State and Territory governments were responsible for reviewing legislation to assess whether regulatory restrictions on competition were in the public interest and the NCC would assess the rigour of the review and report to the Australian Government on compliance. State and Territory performance was linked to competition payments.

This accountability framework established pressures, via inter-jurisdictional demonstration effects, to maintain a commitment to reform over time and to adhere to agreed review processes.

As a by-product of the assessment/monitoring process, the collection and dissemination of information on reform processes and outcomes promoted 'learning by doing' experiences across jurisdictions. It has also aided the process of fine-tuning implementation processes — such as the development of better guidelines for the legislation review process — as well as identifying problem areas in need of follow-up work to progress reform satisfactorily, such as for road and water reform.

Source: PC (2005, p. 144).

Principle 2: Maximise transparency

Transparency can help raise public awareness of policy problems, lead to better-informed analysis of policy options and build support for reform. In a sense, transparency is a 'safety net' for evidence-based policy — a form of quality control that provides opportunities for correction or refinement when the evidence is not complete. It can elicit new sources of information and alternative analysis, expose weaknesses in prevailing analysis, and shed light on how the positions of sectoral interests relate to overall community impact, thereby helping achieve better policies and outcomes.

Transparency is often thought of in terms of public access to relevant information, such as freedom of information laws. But ideally, transparency goes further and provides for contestable policy review with:

- the opportunity for stakeholders to comment;
- access to data, assumptions and methods; and
- peer review and accountability provisions that, for example, require policymakers to publish their decisions and the evidence which supports them.

Transparency in this broader sense creates an important opportunity to replicate results and test alternative assumptions. This is particularly important in areas of complex economic modelling.

For decision makers, however, transparency can be a double-edged sword. Even effective policies and programs can have flaws that present communication challenges for government. And unfettered transparency is simply not possible in certain areas like defence or when commercial-in-confidence information is supplied to government. It is not surprising, therefore, that many jurisdictions struggle to balance the need for transparency with the time and risk associated with transparent policy development and evaluation. Some international initiatives aimed at increasing transparency are outlined in box 3.2. An historical Australian example is provided in box 3.3.

Box 3.2 **Transparency: Some international examples**

United Kingdom

The UK Government has established a central website where all impact assessments (similar to RISs) are publicly available. This enables full access to the evidence base used to justify the need to regulate, including details of the options that were considered and discarded as well as assessment of the impacts on equity and the environment.

New Zealand

New Zealand allows access to the full text of Cabinet documents and Cabinet decisions, unless the benefits of doing so are outweighed by a public interest in keeping the information confidential. (In Australia, Cabinet information is automatically confidential and only released under the '30 year rule').

United States

The Office of Management and Budget, together with federal agencies, are required to assess every budget program and publicly report the results, demonstrating if programs are successful or if they fall short. To date, 1015 programs have been assessed, finding 193 effective, 326 moderately effective, 297 adequate, 26 ineffective and 173 programs did not demonstrate results.

International

In 2005, the World Bank commissioned an independent, public evaluation of its research activities between 1998 and 2005. Chaired by Princeton economist Angus Deaton, the evaluators were asked to examine Bank research paying particular attention to its reliability, rigour, completeness and relevance and, in addition, to give their overall assessment of the strengths and weaknesses of Bank research.

Box 3.3 **Transparency: an Australian example**

The Wallis Inquiry into financial sector trends and regulatory reform: 1996 - 1997

The Wallis Inquiry was set up on 30 May 1996 to examine the consequences and implications of the deregulation of the Australian financial system flowing from the Campbell Report of 1981. It still stands as a creditable example of how special-purpose, one-off inquiries can be set up according to the best principles of transparency, even when dealing rapidly with issues of great sensitivity (the prudential supervision of the financial system) and important conflicting commercial interests (the respective roles of banks, insurance companies and the like). The Inquiry was tasked to recommend future regulation in the light of likely financial market developments, and with a view to best serving the interests of consumers, while preserving system stability and fairness.

The Inquiry was comprised of a panel of three senior businesspeople with experience in both the financial sector and the broader economy, and two highly-regarded professors of finance. It was served by a Secretariat led by a senior Commonwealth Treasury official.

The Inquiry published an issues paper to elicit discussion and submissions, and received 268 submissions from financial institutions, academics, State governments, the Reserve Bank of Australia, the Australian Competition and Consumer Commission and private citizens. (All major institutions' submissions were public.)

It also undertook a program of public consultations for which transcripts were made available, further enhancing the transparency with which evidence was shared with the public and interested analysts. Inquiry members tested ideas through further public engagement via 10 public addresses. The Inquiry was one of the first in Australia to use the Internet to disseminate submissions and its work.

In April 1997, the Government received the final Wallis report, and took immediate decisions to somewhat liberalise the restriction that prevented mergers among the six major banks and insurance companies, but affirmed that it would not allow mergers among the four major banks. By September 1997, the Government had completed its consideration of the Inquiry's recommendations, and set the regulatory architecture for the Australian financial system that prevails today: the Reserve Bank of Australia with macro-prudential responsibility for overall financial and payment system stability; the Australian Prudential Regulation Authority to supervise financial institutions regardless of their business specialisations; and the Australian Securities and Investment Commission to supervise market integrity, disclosure and consumer protections issues.

A dozen pieces of legislation were introduced or amended in 1998 to give effect to the new regulatory architecture. The Wallis Inquiry foresaw and successfully addressed the problems of regulatory fragmentation, gaps and overlaps that have been an element of the financial sector crisis in the United States.

Source: http://fsi.treasury.gov.au/content/default.asp

Genuine transparency provides opportunities for adequate input and review. But experience suggests that consultation sometimes occurs only after a policy decision has been made. Edwards (2001) recount of the development of the Child Support Scheme in the mid-1980s is a characteristic example:

> Having made in-principle decisions by the middle of 1986, ministers did not want these decisions undone. They therefore endorsed a discussion paper which assumed the major planks of the scheme as effectively 'non-negotiable' and sought views on 'second-order, but complex' issues (2001, p. 77).

Greater upfront consultation through the use of traditional Westminster approaches to broad new areas of policy or major policy review, such as 'Green Papers' to scope approaches, and 'White Papers' to propose intended Government policy can significantly improve policy development. Recent Australian examples include Green Papers on national aviation policy, financial services and credit reform and the Carbon Pollution Reduction Scheme.

The Productivity Commission and its predecessors have evolved a range of tested working processes that enhance transparency and improve the evidence base in policy consideration. Some of these processes are enshrined in the *Productivity Commission Act 1998*:

- Elicit evidence on key points for deliberation by way of a public issues paper.
- Proceed by seeking submissions and conducting public hearings or roundtables to gather evidence.
- Make all evidence and submissions publicly available and contestable, unless there are powerful reasons to the contrary.
- When using quantitative modelling, use several models if they are in reputable use, or use modelling roundtables to ensure key model characteristics and parameters are chosen in the light of professional best practice.
- Engage external experts as consultants where necessary, or as referees for research.
- Test tentative conclusions and recommendations by publishing a draft report for comment and review.
- Publish the final report for debate and government deliberation.

Principle 3: Establish a monitoring & evaluation program, including resourcing, at policy commencement.

The options for evaluation are restricted, and sometimes severely limited, if there has been no prior attention to monitoring, evaluating and reporting on a policy over

time. Establishing a systematic monitoring and evaluation program for significant policies at commencement, including committing resources, can play a major part in tackling the evaluation problem when the time comes for review (see box 3.4). This is because the ability to measure whether there has been progress (or not) depends on taking an upfront decision on what to measure, gathering baseline data (the measurement of initial conditions) and setting up a system to monitor results:

> Government should seek to implement policies in a way that facilitates policy evaluation; it should invest in data collection for program evaluation; it should be willing to release that data externally; and it should support research by funding external researchers and sponsoring in-house research for public release. (Borland et al. 2005, p. 114)

Box 3.4 **Evaluating the family law reforms**

Evaluating family law has in the past been patchy, piecemeal and complicated by some very difficult counterfactual questions — such as whether the family law system helps maintain relationships with children and whether the services are effective in reducing conflict and helping parents to agree on child-focused, post-separation parenting arrangements.

In July 2006, the Australian Government introduced wide-ranging reforms to the family law system, including amendments to the Family Law Act recognising shared parental responsibility and providing alternative dispute resolution mechanisms, and integrated and new services, such as Family Relationship Centres. The reform package cost almost $400 million over four years.

Prior to the reforms, the Attorney-General's Department and the Department of Families, Housing, Community Services and Indigenous Affairs commissioned the Australian Institute of Family Studies (a statutory authority) to undertake an evaluation of the family law reforms over the next ten years.

The evaluation comprises three studies:

- the families project which will assess how families are faring under the reforms
- the service provision project will assess the extent to which the new system of service delivery is meeting the objectives of reform
- the legislation and courts project will assess the implementation of legislative changes that govern post-separation parenting arrangements and processes in the family law courts.

The evaluation will build on baseline research gathered pre-reform, including a general population of parents survey (5000 families), a caring for children after parental separation survey (500 families) and a survey of family law service providers (400 participants). Other sources of data being collected are program data (throughput, service use etc), service provider feedback and a longitudinal survey of separated parents to improve the understanding of the long-term effects of family law policy.

Source: AIFS (2007).

Almost every existing policy evaluation report stresses the need to obtain better data to improve the quality of evaluation. Insufficient data is particularly a problem for regulatory, social and environmental policy, and where it is collected, administering departments can fail to collect the data most useful for evaluation, or be reluctant to release it for external analysis. For instance, the lack of baseline data for the COAG Indigenous trials meant it was very difficult for subsequent evaluations to assess the impact or outcomes of the trials. And in the field of education policy, much research on student outcomes in Australia has had to rely on the OECD's PISA, a test administered to an sample of 15 year-old students in OECD and non-OECD countries. The use of PISA reflected, at least until recently, a lack of publicly reported, timely data on student outcomes and characteristics in Australia, rather than any qualitative advantages that PISA offers (Leigh and Thompson 2008, p. 65).

One approach to encourage better monitoring and data collection is to place a requirement on agencies proposing any kind of significant policy to identify how it will be monitored and assessed, such as in the RIS process. Built-in requirements for review, like sunset clauses, have similar objectives. However, their effectiveness is sometimes questionable since there is no sanction if agencies fail to monitor the performance of the policy, or monitor it only perfunctorily.

There are also international examples of tying a certain proportion of program funds to monitoring and evaluation, with public reporting obligations to government. For example, the United States *Second Chance Act 2008*, contains a provision to set aside two per cent of program funds for evaluation of programs that facilitate prisoner re-entry into the community. Evaluations must 'include, to the maximum extent feasible, random assignment ... and generate evidence on which re-entry approaches and strategies are most effective'. This type of hypothecation can encourage rigorous evaluation, but does not guarantee it.

The United States Office of Management and Budget has recently outlined a program to build rigorous evidence to drive policy evolution (box 3.5). The program links funding under certain programs to building rigorous data and evaluation processes into the programs from the outset. In other experiments, it allocates funding in some policy areas to programs with proven results, while giving initial funding to other programs with weaker evidence of success, but on condition of generating more robust evidence which will determine their funding in future budgets.

Box 3.5 The Office of Management and Budget's encouragement of evidence-based policy

President Obama's director of the OMB, Peter Orszag, has outlined a program to place more emphasis on evidence-based policy decisions. The program seeks to:

- Design new initiatives to build rigorous data about what works, and then act on evidence that emerges — expanding the approaches that work best, fine-tuning the ones that get mixed results, and shutting down those that are failing.
- Invest more in program evaluation, with early attention to education, labour, community service and youth programs.
- Design 'new initiatives with evaluation built into their DNA', such as the Department of Education's 'Grow what works and innovation fund', and programs of home visitation (to provide support to vulnerable parents) and the prevention of teen pregnancy.

These last initiatives use a two-tier approach, with one tier providing more funds to programs with strong evidence that they deliver good results, and the second, more experimental tier conditionally funding promising approaches with some evidence of success, but with future funding conditional on rigorous evaluation and strong results:

> Organisations will know that to be considered for funding, they must provide credible evaluation results that show promise, and be ready to subject their models to analysis. As more models move into the top tier, it will create pressure on all the top-tier models to improve their effectiveness, so they continue to receive support.

Source: Orszag (2009).

An incentives-based approach is developing in Australia under the COAG Reform Agenda, whereby Australian Government funding to the States and Territories is tied to monitoring, evaluation and reporting obligations.

These policy reforms are to be implemented in the context of National Agreements/National Partnerships, containing the objectives, outcomes, outputs and performance indicators for policy reform, and clarifying the roles and responsibilities that will guide governments in the delivery of services across the relevant sectors. The performance of all governments in achieving mutually-agreed outcomes and benchmarks is to be monitored and assessed by the independent COAG Reform Council and reported publicly on an annual basis.

Principle 4: Consider sequential roll-out, pilot studies, or randomised trials as appropriate.

In Australia, most policies and programs are implemented on at least a state-wide, if not national scale. Very few policies are subject to sequential roll-out, pilot studies

or randomised trials to test what works, although there have been some exceptions (see box 3.6). Time and resource constraints, the legislative process and political pressures have been seen as obstacles to genuine policy trials, because once a policy proposal has been announced for universal application (for example, a computer for every school child) it is politically difficult to change policy content or offer a program to some but not others. Some policymakers also raise ethical and moral objections to some trials, particularly randomised controlled trials, because they withhold or delay a potentially beneficial policy from some individuals.

Box 3.6 **Two Australian randomised policy trials**

Between 1999 and 2001, the then Department of Family and Community Services conducted two randomised trials on the Job Network, examining the effect of interviews and follow-up contact from professional staff on workforce participation by the long-term unemployed. They found that the intervention led to a reduction in the number of hours worked but an increase in the number of hours spent in studying or training (Farrelly 2008)

In 1998, the NSW Government established a policy trial to test the concept of a Drug Court. During the previous decade, several countries had experimented with drug courts, but evidence on the effectiveness of these courts was limited. The NSW Drug Court trial studied 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. The evaluation found that individuals processed through the Drug Court had 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 (Leigh 2003).

Leigh (2003) responds to some of the common objections to policy trials. He notes that governments never provide assistance to all those who would benefit from it. The objection that those in need will miss out is ameliorated by the fact that researchers, while they may believe the intervention will be positive, genuinely do not know whether it is preferable to be assigned to the treatment or control group — otherwise they would not conduct the trial. He cites a review of medical trials showing that the treatment outperformed the control only about half the time and quotes 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' (Leigh 2003, p. 343).

Trials will not always be suitable (see section 2, box 2.8), particularly if an immediate response is important or where the costs are low, where it may be better to rely on post-implementation review. But, Australia is in little danger of 'overusing' trials, only having undertaken a handful of significant policy trials to date.

Greater use of policy trials would enable policy effectiveness to be tested prior to full scale implementation. Trials will be most valuable when the evidence base is thin, but even when the policy prescription is generally well-accepted, trials can be useful to explore the practicality of implementing policy in a particular ways, and discovering what delivery methods work best or are most cost-effective.

Trials can also act as insurance policies, helping to protect governments and taxpayers from expensive failures, and preventing good policies from being discarded because of implementation problems. They can also help build support for unproven policies.

For example, giving universal high school student access to laptop computers gives rise to complex issues of IT support structures and curriculum evolution to gain a benefit from enhanced access.[2] Resolving these issues takes time. If the policy were conceived as a series of trials that used natural differences in state support and curriculum approaches, it might be possible to initiate the program at least as promptly, learn from continuous feedback from well-designed evaluation, and achieve total coverage at much the same time, with better educational outcomes.

It is commonly thought that well designed roll-out and evaluation of complex programs is a daunting task, with large costs and complex institutional designs. Such views mean trials are often reserved for small, incremental projects rather than fundamental policy redesign. But these views are not supported by the experience of Mexico in undertaking sweeping, progressively evaluated welfare reform in the late 1990s (box 3.7). Since the initial pilot program, there has been a systemic and sustained effort to monitor and evaluate the Progresa-Oportunidades program using sequential rollout. Moreover, the program's continuous, publicly demonstrated success though evaluation gave the program widespread popular support, and helped insure it against partisan political attack in subsequent election campaigns and a change of government.

2 Clotfelter et al. (2008) provide some analysis as to why the impact of computer use on educational outcomes is not straightforward.

Box 3.7 **Mexico's Progresa-Oportunidades program**

In March 1995, Finance Ministry officials proposed to then President Ernesto Zedillo and his Cabinet that the government replace in-kind distributions of milk and tortillas, and price subsidies for bread and tortillas with targeted cash transfers for mothers, contingent on household members' regular attendance at health clinics. (A proposal to expand transfers and condition them on school attendance was added in 1996.) The objective was to have a single comprehensive program to transfer income to all poor households, whether urban or rural. At the time, that was a radical proposition in developing countries' welfare design, and it was met with scepticism by some Cabinet members. Three questions arose: Would the substitution of cash for in-kind transfers lead to less spending on food and more on goods like cigarettes and alcohol? Would giving cash to mothers lead to family disruption and potentially to family violence? Was making cash transfers contingent on compliance with requirements operationally feasible in a relatively poor country with weaker administrative capacities than many other OECD economies?

To answer those questions, the Finance Ministry implemented a pilot project in the state of Campeche involving 31,000 households, and commissioned an external evaluation. On one hand, the evaluation allowed the officials to reject the idea that use of cash transfers would lead to inappropriate use of the funds and to family disruption. A large majority of households preferred cash to in-kind subsidies and actually valued the link to health services; furthermore, cash transfers did not diminish households' food consumption. On the other hand, the evaluation pointed out that before program scale-up could be contemplated, targeting and selection methods had to be substantially revised and more and better data collected. In addition, the pilot project operated on an *ad hoc* basis, hiring health personnel under temporary contracts and working in premises that were not entirely suitable. A different administrative set-up would be needed for a large-scale effort.

The pilot project was essential for three reasons. First, it provided assurance to the government that the change being completed was not unduly risky, given its potential benefits. Second, it made clear that none of the existing government agencies had any incentive to coordinate a much larger-scale operation, forcing program designers to deal explicitly with that issue. And third, it established a decision making mechanism whereby the results of external evaluation, although perhaps not decisive, would nevertheless be seriously weighed in conjunction with other considerations. Learning from the pilot project and having time to adapt the program accordingly were instrumental in solving many unexpected operational issues. These factors contributed to the elaboration of a more solid proposal prior to program scale-up.

Source: Levy (2006).

Scope for greater use of trials could come from the core social policy fields of education, welfare, health and justice, where it is possible to construct control and treatment groups. In education, where the educational career of children has been referred to as '15 000 hours of compulsory treatment' (Davies et al 2000, p. 7), trials could help inform policymakers on how teacher quality and merit-based pay

affect student performance. In health, different funding models could be usefully tested. And in justice, trials could help inform the debate over the impact of different forms of incarceration and release programs on subsequent recidivism and employment patterns (Leigh 2003, p. 351).

Principle 5: Disseminate evaluation and pool results across jurisdictions

There is a significant amount of policy evaluation undertaken in Australia. Yet, a considerable amount never goes beyond the sponsoring government agency or academic institution. Other government agencies, governments and researchers are unaware or unable to access the evaluation (box 3.8). And, in cases where the evidence base in thin, lack of technical skills and resources impede the generation of robust evidence.

Box 3.8 Why isn't existing evidence used more by policymakers?

Some of the most commonly cited reasons why policymakers do not use existing research evidence include:

- research is not made available
- research is not accessible i.e. it is difficult to access ('hidden' in a morass of research databases) or not presented in an understandable way
- research does not ask policy relevant questions – academic researchers are rewarded for publishing in highly ranked (mainly theoretical) journals as opposed to engaging in policy relevant research
- research does not provide policy relevant answers i.e. did the program work? what are the costs and benefits or a policy? what level of uncertainty is associated with the estimates?
- research is inconsistent with current or proposed policy
- lack of communication (interaction) between researchers and policymakers.

Sources: Lavis et al. (2005); Innvaer et al. (2002); Edwards (2004).

Various 'club-like' arrangements have emerged to disseminate evaluation and learn from other jurisdictions. Such clubs can be organised around principles that lead members to raise the quantity and quality of data and evaluation, improve independence and transparency, and improve the dissemination of results. They also capitalise on scarce evaluation resources and exploit economies of scale.

Case studies

The OECD's Evaluation Network of the Development Assistance Committee seeks to increase the effectiveness of international development programs by supporting robust, informed and independent evaluation. The Network brings together evaluation managers and specialists from OECD members' development cooperation agencies and multilateral development institutions. It runs an online evaluation resource centre, disseminates influential evaluations, develops guidelines for evaluation good practice, and conducts peer review of evaluation functions in its multilateral member organisations.[3]

The Evaluation Cooperation Group is a similar club, composed of the heads of evaluation of the multilateral development banks, the Director of Independent Evaluation at the International Monetary Fund, and such observers as the head of the OECD's Development Assistance Committee's Evaluation Network. As well as developing evaluation practices among its own members, it aims to develop evaluation capacity in the borrowing country members of the multilateral development banks.[4] Another recent development in the field of aid evaluation is the International Initiative for Impact Evaluation (box 3.9).

Moving to the realm of developed countries, the United States Coalition for Evidence-Based Policy represents a useful example in a federal setting of the propagation of funding and the principles of robust evaluation to state and central government programs. The Coalition was founded in 2001 and has a small staff sponsored by the Council for Excellence in Government. (The Council, in turn, is made up of former senior government officials, private sector companies and philanthropic foundations.)

The Coalition seeks to:

- Fund rigorous studies, particularly randomised controlled trials, to build the number of social interventions proven to produce sizeable, sustained benefits to participants and/or society.
- Provide strong incentives and assistance for federal funding recipients to adopt such proven interventions, and put them into widespread use.

3 For evaluation guidelines, see http://www.oecd.org/dataoecd/12/56/41612905.pdf . For examples of recent peer reviews of evaluation in various UN agencies, see http://www.oecd.org /document/57/0,3343,en_21571361_34047972_36056761_1_1_1_1,00 .html.

4 https://wpqp1.adb.org/QuickPlace/ecg/Main.nsf/h_Toc/73ffb29010478ff348257290000f 43a6/?OpenDocumnt

Box 3.9 An evaluation club: the International Initiative for Impact Evaluation

Aid evaluation has traditionally evaluated project outputs – whether a school was built – but not aid impact (education outcomes and how this affected economic and other outcomes). The aid sphere is also frequently bedevilled by complex attribution problems. An African country may have over 20 significant bilateral and multilateral donors, so the question is not just whether aid worked, but whose aid worked, and how.

The International Initiative for Impact Evaluation, '3ie', recognises that evaluating development assistance involves the policies (and policy evaluations) of both aid donors and the aid recipient. (For example, whether one donor's aid to a recipient's education sector (say to build a school) has a successful impact on education outcomes, depends in part on other donors' education aid to the recipient (say to train teachers), and importantly, the recipient's own education policies (say teachers pay and curricula)).

3ie was founded in 2007 by the Centre for Global Development to channel funds to independent impact evaluations around an agenda of key questions that confront policymakers. The members are government officials from developing countries with a strong interest in issues of effectiveness and accountability (Mexico and Uganda are founding members), as well as representatives of bilateral donor agencies, multilateral agencies, non-governmental organisations, and foundations or corporations such as the Bill and Melinda Gates Foundation and Google. 3ie is also partnered with the Campbell Collaboration, the collaborative devoted to conducting and disseminating systematic reviews in the realm of social policy.

Membership entails a preparedness to commit to collaborating and dedicating significant funds to impact evaluations. 3ie funds proposals for impact evaluations that meet standards specified in the Principles for Impact Evaluation. (See: http://www.3ieimpact.org/doc/principlesforimpactevaluation.pdf)

The Coalition manages a website on evidence-based programs, *Social Programs that Work* (www.evidencebasedprograms.org), which provide policymakers and practitioners with concise, actionable information on 'what works' in social policy, based on evaluations. It also operates a 'help desk' for federal agencies providing practical resources to advance rigorous evaluation.

Through its advocacy, the Coalition has helped implement measures including new guidance by the Office of Management and Budget to federal agencies on *What Constitutes Strong Evidence of Program Effectiveness* (OMB, 2008) and various evaluation commitments in new legislative programs.

The Coalition has been externally evaluated as very effective, albeit lightly staffed, and as having dealt so far mainly with the OMB and the Departments of Education

and Justice, all of whom were already well-disposed to randomised controlled trials and other rigorous evaluation. The evaluation reported the Coalition's real test will be in extending and deepening its influence into less supportive policy areas. [5]

In the United Kingdom, much of the activity surrounding the evidence-based policy movement has focused on establishing a number of organisations to improve evaluation practice and disseminate results. Some focus on particular policy fields, such as the NHS Centre for Reviews and Dissemination (Health), but there are several large initiatives that cover a broad suite of policy areas (see box 3.10).

Opportunities in Australia?

Could models such as those above work in Australia? One promising opportunity to build a better machinery for evidence-based policy is the CoAG reform agenda, where complex policy reform issues affecting Commonwealth, State and Territory governments abound: education and health; early childhood education and care; mental health; education and training measures to meet the challenge of rising unemployment; and binge drinking, to name a few.

This framework of policy and financing provides the opportunity to build a collaborative process to finance and undertake evaluation, learn from any differences in state and territory approaches, and build evaluative skills. The COAG Reform Council, which will evaluate the progress of each jurisdiction on an annual basis, may fill part of this brief. The COAG Working Groups and Ministerial Councils could facilitate data collection and further evaluation, but have not done so to date.

An alternative vehicle that could build on existing arrangements is a COAG 'evaluation club', made up of senior Commonwealth, State and Territory policy officials, the Australian Statistician, and key official and academic evaluators from each jurisdiction. Through such a club, the greatest needs for evaluation could be identified, evaluation plans could be developed, and funding and evaluation feedback could be arranged. The process could be developed to integrate policy development, data collection, evaluation and feedback of lessons learned into a continuous process. The club members could share and develop technical expertise, resolve data deficiencies, ensure appropriate funding for evaluation, and provide peer review and peer support for the quality evaluation.

[5] The external evaluation is available at http://prod.ceg.rd.net/admin/FormManager/filesuploading/indep_evaln_for_WT_Grant.pdf .

Such a club could go beyond present support groups (such as the Australasian Evaluation Society) and inject a higher level policy and resource commitment to strengthening evidence-based policy.

Box 3.10 **Evaluation bodies in the United Kingdom**

The *National School of Government* (a government department) promotes practical strategies for evidence based policy making through:

- the development of 'knowledge pools' to promote effective sharing of information
- training officials in how to interpret, use and apply evidence
- maintaining a 'policy hub' website providing access to a wide range of current research and guidance on the use of research and evidence in the evaluation of policy
- seconding academic researchers into government agencies to carry out research projects.

The *UK Centre for Evidence Based Policy and Practice* is an initiative funded by the Economic and Social Research Council (an independent statutory authority). The Centre, together with an associated network of university centres of excellence, aims to foster the exchange of social science research between policy, researchers and practitioners by:

- improving the accessibility, quality and usefulness of research
- developing methods of appraising and summarising research relevant to policy and practice
- informing and advising those in policy making roles, through its dissemination function.

The *Evidence for Policy and Practice Information and Coordinating Centre* conducts systematic reviews of research across a range of different topic areas and provides support for others who are undertaking systematic reviews or using research evidence. Currently it has a large number of systematic reviews in the fields of education, health promotion and public health.

Source: Nutley (2003).

Principle 6: Ensure links with the decision making process.

Evaluation can only successfully influence policy where there are incentives for government agencies to provide rigorous analysis and opportunities for decision makers to consider it. That is, evidence needs to be linked to the decision making process, rather than viewed as an adjunct to it.

Most countries have adopted a mixture of 'carrots and sticks' to try and integrate evidence into the policy making process, with many introducing formal requirements to evaluate budget spending programs or prepare impact statements for proposed regulation. Ultimately, however, a predetermined policy agenda can have a far greater influence than any form of requirement, and as such, evaluation requirements can only be fully effective when they are directly tied to decision making.

For instance, the Regulation Taskforce (2006, p. 156) found that although the regulatory impact assessment process should in principle provide a strong evidence base to support policymaking, the requirements were often been circumvented or treated as an afterthought in practice. In its view, the single most important way of strengthening compliance was to link evaluation to the decision making process. It recommended that regulatory proposals that fail to met RIS requirements should not be permitted to proceed to Cabinet (unless granted an exemption by the Prime Minister).

Linking evidence with decision making will require a range of factors: the right institutional structure, adequate resourcing, consultation and transparency and, most importantly, a genuine commitment to evidence-based policy from governments and their agencies (Banks 2009). Nevertheless, there are some specific accountability provisions that may help (see box 3.11), such as the changes to the regulatory impact process mentioned above or the ministerial declaration requirement in the United Kingdom, whereby Ministers must declare that they are satisfied that a) a policy impact assessment represents a fair and reasonable view of the expected costs, benefits and impact of a policy and b) that the benefits justify the costs. Australia is also in a position to take advantage of the federal system, by using COAG arrangements to tie funding to evaluation of initiatives under the COAG reform agenda.

Box 3.11 Linking evidence to decision-making: OECD suggestions for regulatory impact analysis

The OECD identifies a range of institutional and procedural factors that can help maximise the influence of regulatory impact analysis (RIA) on decision making to improve regulatory quality outcomes:

Governments should require impact analysis to be:

- integrated into policymaking, beginning as early as possible in the process
- released for public consultation
- peer reviewed or undergo independent assessment

Governments should provide high level political support, including formal authority for RIA requirements and quality assurance mechanisms, such as assessment and approval from a central regulatory reform authority (an independent 'gatekeeper')

- In the United Kingdom, regulatory proposals that are likely to impose a significant new burden on business, must first be assessed by the Cabinet Office Better Regulation Executive, and then cleared by the Panel on Regulatory Accountability, which is chaired by the Prime Minister, before proceeding to Cabinet.

Governments should require ex-post review of impact analysis, both in terms of assessing whether the RIA met procedural requirements AND the regulatory impacts (i.e. a net benefit test). For example

- In the United Kingdom, the National Audit Office reviews a sample of impact assessments published in the previous year
- Germany requires an ex-post RIA to be completed once experience is available on the new regulation

Source: OECD (2009).

A Why did the US crime rate fall in the 1990s? Evaluation lessons from a cause celebre

When crime rates in New York city fell markedly in the 1990s, one early and popular conclusion was that Mayor Giuliani's 'zero tolerance' policing policy was having significant effect (Waley, 1999). If policing effort were the main cause, that could have had obvious policy relevance all over the world: police and courts should focus more on deterring even minor crime.

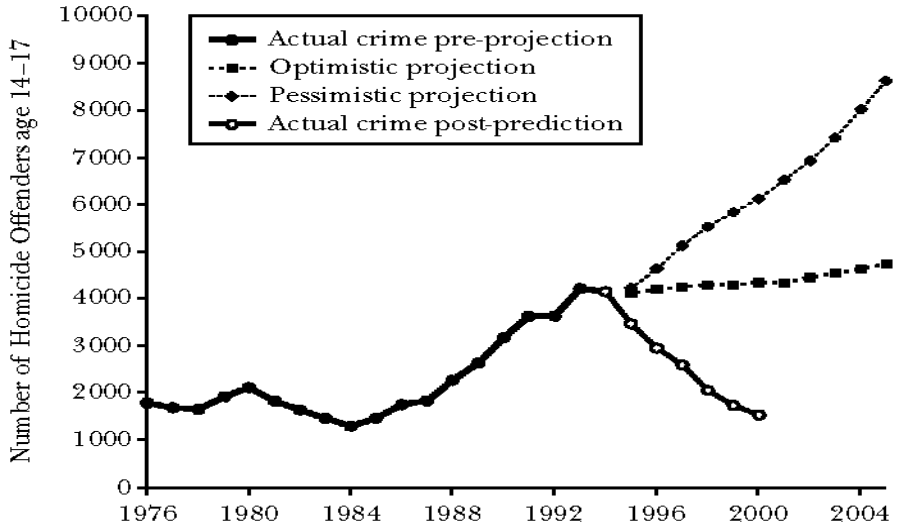
But it was soon noticed that crime rates fell remarkably all over the United States at much the same time: homicides fell by more than 40 per cent, and other violent crime and property crime by more than 30 per cent. Even the direction of change was unexpected — experts had been predicting a crime increase (figure A.1).[1] Instead, even cities such as Los Angeles that had not improved their policing, also experienced falling crime rates.

A.1 Exploring a conundrum

So the search for possible contributing causes widened: perhaps more criminals were being jailed for longer, or perhaps the peaking of the crack cocaine 'epidemic' reduced drug-related homicides and property crime. Table A.1 shows the most popular explanations in the major US newspapers over the period 1991-2001.

[1] Prominent criminologist Professor James Allen Fox was commissioned in 1995 by the US Attorney General to report on crime trends, and predicted 'the next crime wave will get so bad that it will make 1995 look like the good old days'. Instead, juvenile homicide rates fell by more than 50 per cent in the ensuing six years (cited in Levitt 2004, p. 169).

Figure A.1 **1995 Teen Homicide Forecasts Compared to Actual Teen Homicides, 1995-2000**



a Report on crime trends commissioned by US Attorney General from Professor James Alan Fox.

Source: reproduced from Levitt (2004, p. 169)

Table A.1 **Common Media Explanations for the Decline in Crime in the 1990s, Ranked by Frequency of Mention**

Explanation	*Number of mentions*
Innovative policing strategies	52
Increased reliance on prisons	47
Changes in crack/other drug markets	33
Aging of the population	32
Tougher gun control laws	32
Strong economy	28
Increased number of police	26
All other explanations	34

Notes: Based on a Lexis-Nexis search of articles written about the national decline in crime in leading newspapers over the period 1991-2001.

Source: Levitt (2004 p. 164)

In 2001, following wide circulation of a draft paper among the US economics and legal professions, Professors John Donohue and Steven Levitt published a striking hypothesis in a peer-reviewed journal. They suggested increased abortions in the early 1970s had largely contributed to lowering the crime rate a quarter-century later. They concluded the Roe v Wade supreme court ruling in 1973 had so suddenly and greatly liberalised legal access to abortion that the overall abortion rate greatly increased. (For a period after Roe v Wade, abortion in the US rose to the rate of one for every two live births.) Through channels explained below, they initially estimated liberalisation to have caused up to 50 per cent of the decline in criminal activity across the US twenty years later. In addition to this factor, increased incarceration was initially estimated to have accounted for perhaps another 20 per cent of the fall (Donahue and Levitt 2001).

They offered a mixture of natural experimental, quasi-experimental and other econometric evidence using data for all the US states, that also controlled for some other possible influences on crime such as differences across states and over time in policing, incarceration policies, handgun laws and economic conditions. (They could not econometrically control for crack cocaine trends, which had not at that time been quantified at the necessary level of state detail and over time.) [2]

A later work by Levitt, using new estimates of the likely effect of crack cocaine trends, lowered the estimated contribution of abortion law liberalisation, but suggested that just four factors could explain virtually all of the fall in violent crime (table A.2). In order of declining significance, the factors were: increased incarceration, legalised abortion, the decline of crack cocaine and more police (Levitt 2004, pp. 176-184).[3]

2 The natural experiment aspect of the study examined the differential changes in abortion and crime in the five states that had individually liberalised their abortion laws in 1970, compared to those whose abortion rates only increased after Roe v Wade in 1973. Econometric analysis also benefitted from the remarkable diversity of abortion rates across the US States, both before and after Roe v Wade. For example, in the years after Roe v Wade, the average US abortion rate was about 300 per 1000 live births, but ranged from 10 per thousand in West Virginia to 1793 per thousand in Washington DC.

3 The estimated relative significance of police numbers and the decline of crack varies by type of crime, with the crack decline having no impact on property crime, and most impact on homicide.

Table A.2 **Estimated Contributions to the US Decline in Crime in the 1990s**

Factor	*Percentage change in crime that this factor accounts for over the period 1991-2001:* Homicide	Violent Crime	Property crime	*Certainty level of estimated impact*
Strong economy	0	0	-2	High
Changing demographics	0	-2	-5	High
Better policing strategies	-1	-1	-1	Low
Gun control laws	0	0	0	Medium
Concealed weapons laws	0	0	0	High
Increased usage of capital punishment	-1.5	0	0	Medium
Increases in the number of police	-5.5	-5.5	-5.5	Medium
Increases in the prison population	-12	-12	-8	High
The decline of crack	-6	-3	0	Low
Legalized abortion	-10	-10	-10	Medium
Total of all factors considered	-36	-33.5	-31.5	
Actual change in FBI crime reports	-43	-34	-29	
Actual change in crime according to survey of victims	-	-50	-53	

Notes: The estimated impacts in the table are based on the discussion presented throughout the text of Levitt 2004. The last column of the table is Levitt's appraisal of how speculative the estimates are for each of the factors considered

Source: Levitt (2004, p. 184)

Abortion was argued to contribute to the crime decrease through two channels:

- First, the liberalisation of abortion policy caused a drop in unwanted births, so that some 20-25 years later, there was dip in the proportion of 18-24 year old males in the population. This age bracket is the period of highest crime by males. (Donahue and Levitt (2001) called this 'Shrinking the size of the cohort', and it can be considered a 'pure' demographic factor.)
- Second, as a result of women gaining another legal tool in 1973 to manage their fertility, some probably chose to bear children later in life at times of their preference, perhaps when children could be raised with more support (for example, because the mother was older, in a permanent relationship, or had a better job or higher education). So even when mothers who had had abortions subsequently chose to have children (and even if their lifetime fertility remained constant), their sons might grow to be young men with a lower statistical propensity to commit crime. (Donahue and Levitt (2001) called this 'Lowering the average crime propensity of those who are born through positive selection.') They initially estimated this selection effect contributed to about half the overall

impact on crime of the abortion changes, that is, about one-quarter of the total reduction in crime.

– Subsequent research suggested that in the US (but perhaps not in Canada - see Sen (2007)), the demographic effect dominated the selection effect. Most US women who had an abortion in the years after Roe v Wade had lower lifetime fertility, rather than just postponing the timing of their children (Ananat et al., 2006). In support of the view that selection effects may not be large, the only direct attempt to study whether unwanted pregnancies in the US, before Roe v Wade, predicted greater criminality of the offspring, produced only qualified support for the idea of 'positive selection'.[4]

A useful hypothesis explaining the 1990s' crime trends should offer or imply tests that could verify the hypothesis' application in other periods of time. Donahue and Levitt predicted the effects of Roe v Wade on crime would continue to subtract about 1 percent a year from the US crime rate for this decade and the next (other influences remaining the same) before reaching a stationary state (2001, p. 415). Essentially, criminals born pre-Roe v Wade would by 2020 have 'retired', the 'cohort effect' would have largely passed through the demographic structure, and the entire economically active population (including criminals) would have been born since the 'positive selection' effect of Roe v Wade.

This estimated contribution of liberalised abortion to reducing crime was popularised as a chapter of the best-selling *Freakonomics* (Levitt and Dubner 2005).

Donohue and Levitt stressed they were trying to establish the facts about the relative importance of many contributing factors to the crime rate reduction, not to offer any normative comment on abortion. But nevertheless their ideas seemed explosive in the US context. One magazine branded the work 'racist, genocidal stupidity' (cited in Abramsky 2001, p. 25). They were suspected 'from the right' of implied support

4 A few Scandinavian studies suggest the children of unwanted pregnancies have a higher crime rate in later life. But a study found a sample group of US children born between 1964 and 1969 (i.e. before Roe v Wade) from pregnancies which their mothers recalled to be unwanted or mistimed, exhibited greater delinquency over ages 11 to 17 (especially if male), but no significantly greater disposition to more serious crime over ages 17 to 23 (Hay and Evans, 2006, pp 57-61). The study could not test Donohue's and Levitt's exact hypotheses about positive selection. This was because data limitations forced it to examine the self-reported misbehaviour of all children from 'unwanted' or 'mistimed' pregnancies in its sample, not just the subset whose mothers might have chosen legal abortions if they had then been available. (Such information could not have been reliably gathered in sampling of voluntary respondents, even if the interest in abortion intentions had been foreseen at the time the data was gathered.)

for easier abortion to reduce crime (the demographic effect), and 'from the left', of tacitly supporting eugenics (the selection effect).[5]

A.2 Contested explanations: improving the evidence base and removing errors

The ensuing lively debate over Donahue's and Levitt's hypothesis illustrated many of the measurement issues, evaluation complexities and policy ambiguities that afflict most analysis of complex social phenomena with multiple influences. For example:

- Even though the study statistically examined and econometrically 'controlled for' many possible influences, perhaps other, unobserved variables were in play? Perhaps, for example, politically liberal and more urbanised states had higher pre-abortion crime rates and then had more abortions, both caused by other factors?[6]
- There was, of course, no direct measure of illegal abortions, and perhaps legal abortions mostly took the place of what would otherwise have been illegal abortions (Joyce 2003)? However the evidence confirms a large increase in the overall rate of abortions, and a significant fall in their price (Donahue and Levitt 2001, pp. 383-385; Donahue and Levitt 2003, pp. 32-33).
- Perhaps liberalising abortion also has the indirect effect of increasing out-of-wedlock births and creating more single-parent families, thus indirectly inducing more births from 'unwanted pregnancies' than abortion prevents? Some initially estimated that the net effect might actually have slightly increased crime, rather than reduced it. But subsequent study confirmed the net effect had been to reduce the size of the cohort born immediately after Roe v Wade (Ananat, Gruber, Levine and Staiger, 2006; Ananat, Gruber and Levine, 2007).

---

5 The reason for racial sensitivity in this debate is the estimate by Donahue and Levitt that fertility decline for black women following abortion liberalisation were about 3 times greater than for whites (2001 p. 390). A Minneapolis official who had suggested independently and earlier (in 1990) that abortion had reduced crime was heavily criticised, and subsequently defeated at the polls (Levitt and Dubner 2005, p. 142).

6 An example of an unobserved variable in the Donohue and Levitt analysis was the progressive impact on fertility of the widening availability of the birth control pill, first released in 1957 but still restricted in some states until the early 1970s. Initial marketing and distribution had been hindered in 45 states by the effect of the 1873 'Comstock Act', which prohibited 'obscenities' and anything used 'for the prevention of conception'. Paradoxically, California and Washington state, which both liberalised abortion before Roe v Wade, still restricted the availability of the pill. (Untangling the impact of these restrictions on the use of the pill is itself a fascinating example of evaluation using the variability of practices across the US states.)

- Perhaps inadequate treatment of how much 1980s crime was driven by the crack cocaine-driven 'epidemic' distorted estimates of the effects of other factors, including abortion (Joyce 2003)?[7] A recent study (Fryer, Heaton, Levitt and Murphy, 2005) improved knowledge of the extent of the crack epidemic and its costs, and has provided evidence for the view that the likely effect of crack was to contribute to underestimation of the effect of abortion in reducing crime in the late 1980s, and to only modestly overestimate its effect in latter years. Estimates derived over the period as a whole should not be biased (Donahue and Levitt 2003, pp. 39-40).
- To add confusion to the debate, other academics attempting to replicate Donahue's and Levitt's results (with the authors' encouragement, and with free access to the authors' database), discovered a coding error that meant some of the originally reported estimates had not been subjected to some of the statistical tests that the authors' text claimed they had (Ananat et al. 2006 p 34; Foote and Goetz 2008). However, correction of this error made little difference to the original results. Data improvements to address cross-state mobility better[8], and some econometric improvements, showed the original estimates still stood (Donahue and Levitt, 2008).
  - The identification and correction of error in this complex body of work carries an important lesson, drawn by Donohue himself in an earlier, different context:

    > increased complexity means that it will be harder for researchers and readers to have a feel for the data, and it will be more difficult to critique an ostensibly well-crafted empirical study that isn't marred by one of the obvious pitfalls. As econometrics becomes increasingly technical, the rewards from the academy will more often go to those who have mastered complex mathematical techniques, rather than those who are alert to the relatively low-grade (but often pivotal) issues of data quality. Studies now pour out with conclusions based on increasingly fancy statistics, which their authors accept uncritically and which reviewers are at a loss to

7 Crack cocaine effects were potentially large – crack was a technological innovation virtually unknown prior to the mid 1980s, and its distinctive characteristics changed the nature of the drug trade. Its distribution through street gangs, and ensuing gang battles over 'turf', caused US youth homicide rates to briefly more than double in the late 1980s, overwhelmingly because of a surge in homicides among young blacks.

8 Donahue's and Levitt's methodology maps data recording the changing rates of abortion by state with changing recorded crime by people in the high-crime age bracket decades later. But women residing in one state sometimes have abortions in another, and children born in one state may migrate to another state in which they may commit crimes as young adults. The original 2001 study dealt with this issue only roughly, whereas a different data source allowed better treatment in the 2008 study.

appraise effectively absent an enormous amount of work through attempted replication. (Donohue 2001, p. 4)

– Even after almost a decade of testing, some researchers remain unconvinced that the statistical sources and Donohue's and Levitt's methodology are sufficiently robust to support the current (reduced) estimates of the impact on 1990s crime of abortion liberalisation. They observe that the original Donohue and Levitt estimates of the impact of abortion on crime were so large that its effect should also be apparent in the pattern of age-specific crime rates, but it is not. Donohue and Levitt respond that the reason such pair-wise correlations are not apparent is that other factors such as the crack cocaine epidemic are simultaneously at work. The statistical and econometric arguments on these issues are beyond the scope of the present appendix, but interested readers are referred to the work of Joyce (2006, 2009 (a) and 2009 (b)).

A.3 Similar international experience?

Once the Donohue and Levitt hypothesis gained attention in the US, others wondered whether it had application in other countries. Looking to other countries might help clarify causalities, because for example crack cocaine and handgun availability were not such important disturbances to crime rates elsewhere as in the US.

Both Canada and the UK enjoyed falling crime rates in the mid 1990s, following liberalisation to their abortion laws that took effect in the late 1960s.

Canadian research suggests that abortion liberalisation that took effect in 1969 did contribute to a reduction in violent crime some 20 years later, but no impact on property crime. The Canadian data and methodology also suggested the effect was through selection effects (in particular, the decline in children born to teenage mothers) more than cohort effects (Sen, 2007).

In contrast, UK research found no evidence that liberalised abortion in England and Wales led to decreased crime with a plausible lag: property crime and overall crime did fall in the 1990s, but about 23 years after abortion liberalisation, not the 18-19 years observed in the US (and corresponding there with the peak years of male criminality). Moreover, violent crime rose steadily in England and Wales, and total crime rates did not fall in England and Wales relative to Northern Ireland and the Republic of Ireland, which did not liberalise abortion laws (Kahane, Paton and Simmons, 2008).

Another potential international test of the hypotheses arose in Romania, where abortion had been freely allowed until a pro-natalist policy in 1966 criminalized abortion and taxed childless couples. In the context of restricted access to birth control measures, the birth rate rose sharply, and the subsequent demographic bulge of young people did indeed corresponded to higher crime rates in the late 1980s and 1990s. However, the case is judged to shed little light on the effect of abortion law changes alone, because of interactions with the significant economic disruptions that attended the last years of Ceausescu and the fall of communism (Kahane, Paton and Simmons, 2008, p. 1).

A brief Australian study using more limited data than available in the US pointed to some of the tell-tale signs that abortion law liberalisation in the most populous Australian states over the period 1969 to 1974 might have caused some of the declines in homicide rates 20 years later (Leigh and Wolfers 2000). For example, homicides fell most in states that had liberalised abortion, and fell at times that lagged the individual state's liberalisation with the expected delay. Moreover the Australian homicide declines preceded the US reductions by about the times that Australian abortion liberalisations preceded Roe v Wade. Nevertheless, it is impossible to be confident of the linkages, because of Australian data limitations (such as the lack of extended data recording the age at which criminals offended, and of comprehensive abortion statistics), and perhaps also because the liberalisations of Australian states' laws were not as dramatic or compressed in time as the impact of Roe v Wade in the US. One recent overview concludes the linkage in the Australian case is unproven (Zimring 2007, pp. 218-221).

An interpretation of these diverse studies is that while there is some suggestive evidence of possible international equivalents to the US experience with abortion and crime, there has been no robust confirmation of similar links.

It seems now reasonably settled that the US experience of sudden, major liberalisation of legal abortion in 1973 had a significant contributing impact on the fall in the crime rate a quarter century later. Corresponding effects have not been robustly demonstrated in other countries that liberalised abortion laws, perhaps because the legal restrictions on abortions (and other forms of birth control) in other countries studied were never as strict as in the US, and the liberalisations of those restrictions were not so sweeping.

A.4 Some policy evaluation lessons illustrated in the US crime debate

Attribution of changes to underlying causes

1. Untangling causes of a complex social problem will usually require formal and sophisticated evaluation. Both expert predictions (e.g. of a crime explosion) and common sense explanations (e.g. increased policing was the dominant cause of falling crime rates) can be wrong.
   - As one of the prominent contributors to the abortion-crime debate has reflected in a broader context:

     > The single most important advance in the social sciences in the last 25 years has been the enormous improvements in the ability to analyze microdata in order to identify and quantify causal relationships. Some of these gains have come from technological progress, as the vast increases in computing power have enabled the analysis of larger data sets using more sophisticated statistical techniques. Some of these gains have been organizational, as government and private entities have funded the collection of an extensive array of data sets that provide fertile grounds for academic researchers. But the most intellectually intriguing developments have been scientific, as the tools and methods of statistical analysis have been developed and sharpened. (Donohue 2001, p. 2)

2. Distrust mono-causal explanations and simple correlations; seek evaluative processes that explicitly estimate the separate contributions of all major influences.
   - Work before 2001 that sought to explain the fall in crime without including the impact of abortion law changes were subject to 'omitted variable bias', and the estimates of the impact of other factors were overstated.
3. Beware of overemphasis on factors that can be measured, at the cost of study of factors that may be important but are not quantified.
   - It was expected from the early 1990s that the crack cocaine epidemic must have had something to do with the surge and then decline in crime rates, but because there was no good quantitative estimate of those effects until the work of Freyer, Heaton, Levitt and Murphy in 2005, initial quantitative work by Donohue and Levitt in 2001 and 2003 overestimated the impact of other factors (such as imprisonment, policing and abortion policies) – another example of omitted variable bias.
4. Look for unintended linkages and indirect effects.
   - It took a leap of imagination to hypothesise the radical change in US abortion policy in 1973 had an unintended consequence for crime rates in the 1990s.

Data issues

5. Data quality matters: most social phenomena are imperfectly measured, and small differences in measurement quality may prove material to estimating the relative contribution of different causes.
   - The attention attracted by the 2001 Donahue and Levitt article led to more relevant data being estimated (e.g. abortion by state of woman's residence, rather than by state in which the abortion was performed). This data strengthened the original results (Donahue and Levitt 2003).[9]
   - The absence of good proxies for crack cocaine usage and its impacts made it impossible to initially factor that cause into an explanation of declining crime rates. The creation in 2005 of an index of the usage of crack cocaine by state and large city, and a study of its impacts (not only on crime but on low birth weights, foetal death rates, etc) allowed a significant advance in analysing the fall of crime.[10] This index and associated data has been made freely available to all researchers (Freyer et al., 2005, p. 6).

Transparency

6. Publish policy evaluation and underlying data freely, to facilitate checking and replication of results, and trigger improvements and corrections.
   - Notwithstanding peer review for publication in a prestigious journal, a potentially significant but arcane coding mistake affected part of Donahue's and Levitt's original 2001 results. It is unlikely that mistake would have been discovered and corrected without the transparency of the research and publication process, and the freedom with which Donahue and Levitt made their data available and engaged in dialogue with their critics.[11]
7. Be cautious not to over-interpret new results. Even peer review and transparency are no guarantee of instant accuracy or immediate success in evaluating complex phenomena.
   - Notwithstanding intense interest in this subject, it has taken the best part of a decade to resolve disputes about the estimation of the abortion impact.

---

9 Measurement error usually leads to 'attenuation bias' and understatement of the estimated causal relationships under study (Donahue and Levitt 2003, pp. 36-37).

10 The study of why crack usage exploded so quickly and was so devastating for a period to young blacks is itself a fascinating analysis of the economics of illicit drugs, though beyond the scope of this paper. Freyer et al (2005) provides an excellent account.

11 Foote and Goetz explicitly acknowledged the openness with which data and analysis was shared by Donahue and Levitt (2008, p 407).

8. Vibrant academic engagement greatly strengthens evaluation.
   - With the exception of one important contribution by two economists employed in the Federal Reserve system (Foote and Goetz 2008), all the improvements in understanding cited above arose from academic contributions.

Policy

9. Good evaluation requires independence and sometimes, courage.
   - In the ethically and politically fraught environment of debate over abortion policy in the US, it is perhaps surprising that the abortion/crime hypothesis emerged at all, and that it steadily progressed through careful, high quality and freely contested evaluation of the evidence.

10. National policy analysis requires national evaluations because of country-specific factors, but evidence-based policy can learn also from international comparisons and contrasts.
    - Widening study beyond the US to other countries that had experienced liberalisation of abortion laws allowed further testing of the hypothesis in circumstances where national data strengths and weaknesses were different, and other potential influences were of lesser importance.

11. Ethically charged policy topics will inevitably generate passion, but transparent evaluation can help clarify policy improvements that all would support.
    - The US and Canadian abortion experiences can be interpreted as showing that any effective method of improving a woman's control of her fertility, or any effective means of improving assistance to the carers of children born into difficult circumstances, can pay surprising dividends not only in better realisation of those children's potential, but also in less crime.[12]
    - As Leigh and Wolfers note:

      > When the political dust settles, we might – surprisingly - learn something far more interesting about child rearing than about abortion. (2000, p. 28)

12 Voluminous evidence that 'being unwanted' leads, on average, to much poorer life outcomes for children is beyond the scope of this paper, but see Gruber, Levine and Staiger (1999) for a summary.

References

Abramsky, S. 2001, 'Did *Roe v. Wade* abort crime? And why hardly anybody wants to talk about it', *American Prospect,* 1 January.

ACMA (Australian Communications and Media Authority) 2008, *Review of Children's Television Standards 2005*, Canberra, August.

AIFS (Australian Institute of Family Studies) 2007, *A framework for the evaluation of the family law reform package*, Canberra, March.

Ananat, E.O., Gruber, J., Levine, P. and Staiger, D. 2006, *Abortion and Selection*, National Bureau of Economic Research Working Paper 12150, March.

Ananat, E.O., Gruber, J. and Levine, P. 2007, 'Abortion Legalisation and Life-Cycle Fertility', *Journal of Human Resources*, spring, vol XLII no. 2, pp. 375–397.

Austin, S., Hartigan, J., Palmer, A., Ritman, F. and Sturges A. 2008, 'Risk in regulation: Achieving better outcomes', Report to the Australian and New Zealand School of Government (ANZSOG), unpublished.

Australian Government 2007, *Best Practice Regulation Handbook*, Canberra.

Baker, R., Barker, A., Johnston, A. and Kohlhaas, M. 2008, *The Stern Review: an assessment of its methodology*, Productivity Commission Staff Working Paper, Melbourne, January.

Banks, G. 2002, *The Productivity Commission's gambling inquiry: 3 years on*, Presentation to the 12th Annual Conference of the National Association for Gambling Studies, Melbourne, 21 November, Productivity Commission, Canberra.

—— 2006, 'Explaining the housing market puzzle' Presentation to the Centre for Independent Studies *Consilium*, Coolum, Queensland, 12 August.

—— 2007, 'Public inquiries in policy formulation: Australia's Productivity Commission', Address to an International Workshop, China-Australia Governance Program, Beijing, 3 September.

—— 2008, *Industry Policy for a Productive Australia*, Colin Clark Memorial Lecture, Brisbane, 6 August. Productivity Commission, Melbourne.

—— 2009, 'Evidence-based policymaking: What is it? How do we get it?', Speech to the Australian and New Zealand School of Government, Canberra, 4 February.

Bazzi, S. and Clemens, M.A. 2009, 'Blunt Instruments: On establishing the causes of economic growth', Center for Global Development Working Paper No 171, May.

Blair, T. and Cunningham, J. 1999, *Modernising Government*, Prime Minister and Minister for the Cabinet Office, London, UK.

Blinder, A. 1999, *Central Banking in Theory and Practice*, MIT Press.

Borland, J., Tseng, Y. and Wilkens, R 2005 'Experimental and quasi-experimental methods of microeconomic program and policy evaluation' in *Quantitative Tools for Microeconomic Policy Analysis*, Productivity Commission Conference Proceedings, 17–18 November 2004, Canberra.

Carter, O.B.J., 2006, 'The weighty issue of Australian television food advertising and childhood obesity', *Health Promotion Journal of Australia*, 2006: 17 (1) pp. 5–11.

Clotfelter, C.T., Ladd, H.F. and Vigdor, J.L. 2008, 'Scaling the digital divide: Home computer technology and student achievement', Duke University, 29 July, http://econrsss.anu.edu.au/pdf/seminars/2009juldec/vigdor09.pdf (accessed 28 July 2009).

Cohen-Cole, E. and Fletcher, J. 2008, 'Detecting implausible social network effects in acne, height and headaches: Longitudinal analysis', *British Medical Journal* 337:a2533, December.

Davies, H., Nutley, S. and Smith, P. 2000, 'Introducing evidence-based policy and practice in the public services' in Davies, H., Nutley, S. and Smith, P. (eds) *What Works? Evidence-Based Policy and Practice in Public Services*, Policy Press, Bristol.

Deaton, A.S. 2009, 'Instruments of development: Randomisation in the tropics and the search for the elusive keys to economic development', National Bureau of Economic Research Working Paper 14690, January.

Dee, P. 2005, *Quantitative Modelling at the Productivity Commission*, Productivity Commission, Melbourne.

Dinardo, J. 2005, *A Review of Freakonomics*, University of Michigan, December.

Dobes, L. and Bennett, J. 2009, 'Multi-criteria analysis: "Good enough" for government work?', *Agenda*, forthcoming.

Donahue III, J.J. 2001, *The Search for Truth: In Appreciation of James J. Heckman*, Stanford Law School, John M Olin Program in Law and Economics, Working Paper No 220, July.

—— 2003, 'Further Evidence that Legalised Abortion Lowered Crime: A reply to Joyce', *Journal of Human Resources*, vol. 39 no.1, pp. 29–49.

—— 2008, 'Measurement Error, Legalised Abortion, and the Decline in Crime: A Response to Foote and Goetze', *Quarterly Journal of Economics*, vol. 123, no. 1, February, pp. 425–440.

—— and Levitt, S.D. 2001, 'The Impact of Legalised Abortion on Crime', *Quarterly Journal of Economics*, vol. 116, no. 2, May, pp 379–420

Edwards, M. 2001, *Social Policy, Public Policy: From Problems to Practice*, Allen and Unwin, Sydney.

Edwards, M. 2004, 'Social Science research and public policy: Narrowing the divide', Occasional Paper 2/2004, Academy of the Social Sciences in Australia, Canberra.

EPHC (Environment Protection and Heritage Council) 2008, 'Decision Regulatory Impact Statement: Investigation of options to reduce the impact of plastic bags', April.

Ergas, H. 2008, 'Miracle cure that wastes tax dollars', *The Australian*, 8 October.

FaHCSIA (Department of Families, Housing, Community Services and Indigenous Affairs) 2008, Family Assistance Guide – 1.2.1. Family Tax Benefit (FTB) — Description, http://www.facsia.gov.au/Guides_Acts/fag/faguide-1/faguide-1.2.html (accessed 8 November 2008).

Farrelly, R. 2008, 'Policy on Trial', *Policy,* vol. 24, no. 3 pp. 7–12, Centre for Independent Studies.

Foote, C.L. and Goetz, C.F. 2008, 'The Impact of Legalised Abortion on Crime: Comment', *Quarterly Journal of Economics*, vol. 123, no. 1, February, pp. 407–423.

Fowler, J. and Christakis, N. 2008, 'Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study', *British Medical Journal*, 337:a2338, December.

Freyer, R.G., Heaton P.S., Levitt, S.D. and Murphy, K.M. 2005, 'Measuring the Impact of Crack Cocaine', National Bureau of Economic Research Working Paper 11318, May.

Gruber, J., Levine, P., and Staiger, D. 1999, 'Abortion legalisation and child living circumstances: Who is the 'marginal child'?' *Quarterly Journal of Economics*, Vol 114, pp. 263–291.

Hall, R. 2009, 'Qualitative research methods' in Argyrous, G. (ed), *Evidence for Policy and Decision-Making: A Practical Guide*, UNSW Press, Sydney, pp. 218–239.

Hanushek, E.A. 2002, 'Evidence, politics and the class size debate' in Mishel, L. and Rothstein, R. (eds), *The Class Size Debate*, Economic Policy Institute, Washington.

Heckman, J.J. and Smith, J.A. 1995, 'Assessing the case for social experiments' *Journal of Economic Perspectives*, vol. 9, no. 2, pp. 85–110.

—— and Urzua, S, 2009, 'Comparing IV with structural models: What simple IV can and cannot identify, National Bureau of Economic Research Working Paper 14706, February.

Henderson, G. 1985, 'How to create unemployment: the Arbitration Commission and the Aborigines' in Hyde, J. and Nurick, J. (ed), *Wages Wasteland: A Radical Examination of the Australian Wage Fixing System*, Hale and Iremonger, Sydney.

House of Commons Health Committee (UK), 2009, *Health Inequalities: Third Report of Session 2008-09*, vol 1, Stationery Office, London.

Imbens, G. 2009 'Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)', Working Paper, Harvard, 10 April.

Innvaer, S., Vist, G., Trommald, M. and Oxman, A. 2002, 'Health policymakers' perceptions of their use of evidence: a systematic review', *Journal of Health Services Research and Policy*, vol 7, no. 4, pp. 239–244.

Joyce, T. 2003, 'Did legalised abortion lower crime?', *Journal of Human Resources*, vol. 38, no. 1, pp. 1–37.

—— 2006, 'Further tests of abortion and crime: A response to Donohue and Levitt (2001, 2004, 2006)', National Bureau of Economic Research Working Paper 12607, October.

—— 2009a, 'A simple test of abortion and crime', *Review of Economics and Statistics*, vol. 91, no.1, February, pp. 112–123

—— 2009b, 'Abortion and Crime: A Review', National Bureau of Economic Research Working Paper 15098, June.

Kahane, L.H., Paton D. and Simmons, R. 2008, 'The abortion-crime link: Evidence from England and Wales', *Economica*, vol. 75, February, pp. 1–21.

King, G., Gakidou, E., Ravishankar, N., Moore, T., Lankin, J., Vargas, M., Tellez-Rojo, M., Avila, J., Avila, M. and Llamas, H. 2007, 'A politically robust experimental design for public policy evaluation, with application to the

Mexican Universal Health Insurance Program', *Journal of Policy Analysis and Management*, vol. 26, no. 3, pp. 479–506.

Lattimore, R. 2009, 'Economic perspectives on regulation', Presentation to the Australian and New Zealand School of Government (ANZSOG), May, unpublished.

Lavis, J., Davies, H., Oxman, A., Denis, J.L., Golden-Biddle, K. and Ferlie, E. 2005, 'Towards systematic reviews that inform health care management and policymaking', *Journal of Health Services Research and Policy*, vol 10, suppl. 1, pp. 35–48.

Leigh, A. 2003 'Randomised policy trials' *Agenda*, vol 10(4), pp. 341–354

—— 2009, 'What evidence should social policymakers use?', *Economic Roundup*, issue 1, pp. 27–43.

—— and Ryan, 2006 'How and why has teacher quality changed in Australia?' ANU Discussion Paper, no. 534, September.

—— and Thompson, H. 2008, 'How Much of the Variation in Literacy and Numeracy can be Explained by School Performance'*?*, *Economic Roundup*, issue 3, pp. 63–78.

—— and Wolfers, J. 2000, 'Abortion and Crime', *AQ* August-September, pp. 28–31.

Levitt, S.D. 2004, 'Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not', *Journal of Economic Perspectives*, vol. 18, no. 1 Winter, pp. 163–190.

—— and Dubner, S.J. 2005, *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, Harper Collins, New York.

Levy, S. 2006, *Progress Against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*, Brookings Institution Press, Washington.

Mankiw, G.N. and Taylor, M.P. 2006, *Microeconomics*, Cengage Learning College, United States.

Marshall, S., Biddle, S.J.H., Gorely, T., Cameron, N. and Murdey, I. 2004, 'Relationships between media use, body fatness and physical activity in children and youth: a metaanalysis' *International Journal of Obesity*, 28(10), pp. 1238–46.

Maynard, A. 2002, 'Barriers to evidence-based policymaking in healthcare' Productivity Commission and Melbourne Institute of Applied Economic Research, *Health Policy Roundtable*, Conference Proceedings, Canberra.

Mays, N and Pope, C. 1995, 'Qualitative research: rigour and qualitative research', *British Medical Journal*, vol. 311, pp. 109–112.

McKenzie, D., Gibson, J. and Stillman, S. 2006, 'How important is selection? Experimental versus non-experimental measures of the income gains from migration', World Bank Policy Research Working Paper 3906, May.

Mott McDonald 2002, *Review of Large Public Procurement in the UK*, Report prepared for HM Treasury, July.

Mulgan, G. 2003, 'Government, knowledge and the business of policy making', *Canberra Bulletin of Public Administration*, vol. 108, pp. 1–5.

Nutley, S. 2003, 'Bridging the policy/research divide: reflections and lessons from the UK', *Canberra Bulletin of Public Administration*, vol. 108, pp. 19–28.

—— Walter, I. and Davies, H. 2009, 'Past, present, and possible futures of evidence-based policy', in Argyrous, G. (ed), *Evidence for Policy and Decision-Making: A Practical Guide*, UNSW Press, Sydney, pp. 1–44.

Obama, B. 2009, *Inaugural Address*, transcript, 20 January 2009.

OECD (Organisation for economic Cooperation and Development) 1999, *Improving Evaluation Practices: Best Practice Guidelines for Evaluation*, Background Paper, OECD, Paris.

—— 2008, *Biofuel Support Policies: An Economic Assessment*, OECD, Paris.

—— 2009, *Regulatory Impact Analysis: A Tool for Policy Coherence*, OECD, Paris.

Ofcom (Office of Communications UK) 2006, *Television Advertising of Food and Drink Products to Children: Options for New Restrictions*, London.

OMB (Office of Management and Budget) 2008, *What Constitutes Strong Evidence of a Program's Effectiveness* http://www.evidencebasedpolicy.org/docs/What_Constitutes_Strong_Evidence.pdf (accessed 2 February 2009).

Orszag, P. 2009 'Building rigorous evidence to drive policy' http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy/ (accessed 21 July 2009).

Pawson, R., 2002, 'Evidence-based policy: In search of a method', *Evaluation*, April 8(2), pp. 157–181.

Picciotto, R. 2008, *Evaluation Independence at DFID*, http://iacdi.independent.gov.uk/wp-content/uploads/iadci-evaluation-independence-at-dfid-final.pdf (accessed 1 May 2009).

Productivity Commission 1999, *Australia's Gambling Industries*, Report No. 10, Canberra.

—— 2001, *Review of the National Access Regime*, Report no. 17, AusInfo, Canberra.

—— 2003, *Evaluation of the Pharmaceutical Industry Investment Program*, Research Report, Canberra.

—— 2004a, *Impacts of Native Vegetation and Biodiversity Regulations*, Report no. 29, Melbourne.

—— 2004b, *First Home Ownership*, Report no. 28, Melbourne.

—— 2005, *Review of National Competition Policy Reforms*, Report no. 33, Canberra.

—— 2006, *Waste Management,* Report no. 38, Canberra.

—— 2008a, *Paid Parental Leave: Support for Parents with Newborn Children*, Draft Inquiry Report, Canberra.

—— 2008b, *Government Drought Support*, Draft Inquiry Report, Melbourne.

—— 2008c, *The Market of Retail Tenancy Leases in Australia*, Inquiry report no. 43, Canberra.

—— 2008d, *What Role for Policies to Supplement an Emissions Trading Scheme?*, Productivity Commission Submission to the Garnaut Climate Change Review, May.

—— 2009, *Trade & Assistance Review 2007-08*, Annual Report Series, Productivity Commission, Canberra, May.

Petrosino, A., Turpin-Petrosino, C., and Buehler, J. 2002, 'Scared Straight and other juvenile awareness programs for preventing juvenile delinquency', *Cochrane Database of Systematic Reviews*, Issue 4.

Regulation Taskforce 2006, *Rethinking Regulation: Report of the Taskforce on Reducing Regulatory Burdens on Business*, Report to the Prime Minister and Treasurer, Canberra, January.

Roodman, D. and Morduch, J. 2009, 'The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence', Center for Global Development Working Paper, no 174, June.

Rudd, K. (Prime Minister) 2008, Address to Heads of Agencies and Members of Senior Executive Service, Great Hall, Parliament House, Canberra, April (http://www.pm.gov.au/media/Speech/2008/speech_0226.cfm).

Sen, A 2007, 'Does Increased Abortion Lead to Lower Crime? Evaluating the Relationship between Crime, Abortion and Fertility', *The Berkeley Electronic Journal of Economic Analysis and Policy*, vol 7, issue 1, article 48, pp 1–36.

Schmidt, K., Pittler, M.H. and Ernst, E. 2001 'Bias in alternative medicine is still rife but is diminishing', *British Medical Journal*, 323:1071, November.

Shapiro, S. 2008, 'Analysis of Homeland Security Regulations, Small Steps Forward, Giant Leaps to Go', *Regulatory Analysis*, 08-03, April.

Spencer, L., Ritchie, J., Lewis, J. and Dillon, L. 2003, 'Quality in Qualitative Evaluation: A framework for assessing research evidence', Government Chief Social Researchers Office (UK), August.

Standing Committee on Finance and Administration 2009, *Hansard Transcript*, Monday 9 February, p. 57.

Stern, N. 2007, *The Economics of Climate Change: The Stern Review*, Cabinet Office – HM Treasury, Cambridge University Press, UK.

Vreeman, R. and Carroll, A. 2008, 'Seasonal medical myths that lack convincing evidence', *British Medical Journal*, 337:a2769.

Wake, M., Hesketh, K. and Waters, E. 2003, 'Television, computer use and body mass index in Australian primary school children', *Journal of Paediatric Child Health,* 39(2), pp. 130–4.

Waley, J. 1999, 'Zero Tolerance New York Style' Sunday, Channel 9, 7 March http://sunday.ninemsn.com.au/sunday/cover_stories/transcript_309.asp

Weier, A. and Loke, P. 2007, *Precaution and the Precautionary Principle: Two Australian Case Studies*, Productivity Commission Staff Working Paper, Melbourne, September.

Wilkie, J and Grant A. 2009, 'Using evidence well', *Economic Roundup*, issue 1 2009, pp. 17–25.

Winston, C. 2006, *Government Failure versus Market Failure: Microeconomic Policy Research and Government Performance*, Brookings Institution Press, Washington.

Wolfers, J. 2008, 'Is Happiness Contagious?', *New York Times*, 9 December.

World Bank 2003, *Operations Evaluation Department: The First 30 Years*, Washington DC.

Zimring, F.E. 2007, *The Great American Crime Decline*, Oxford University Press, New York.