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.

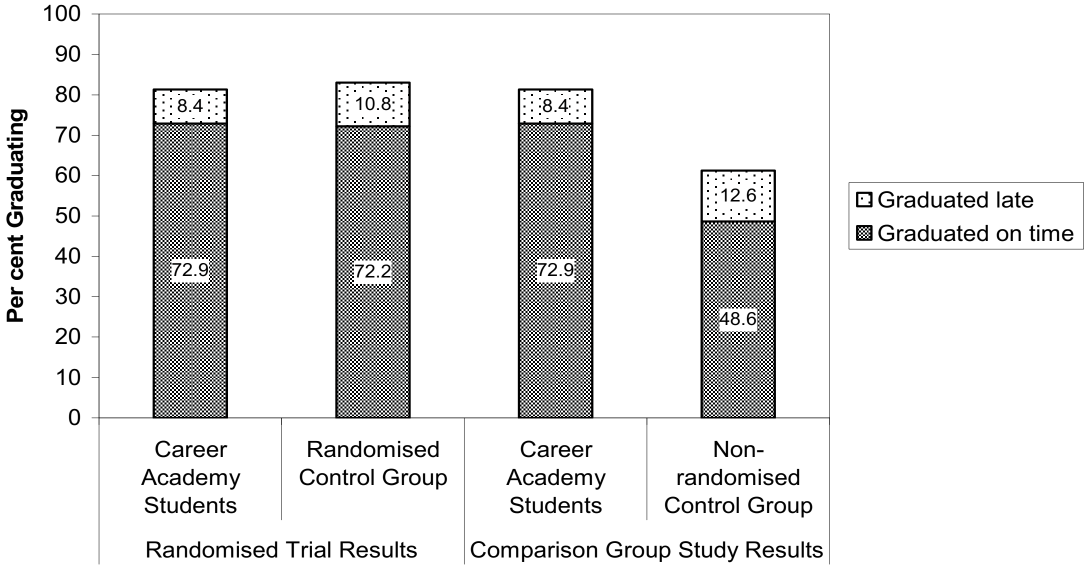
Box 2.7 **Randomised controlled trials: A ‘perfect counterfactual’**

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).

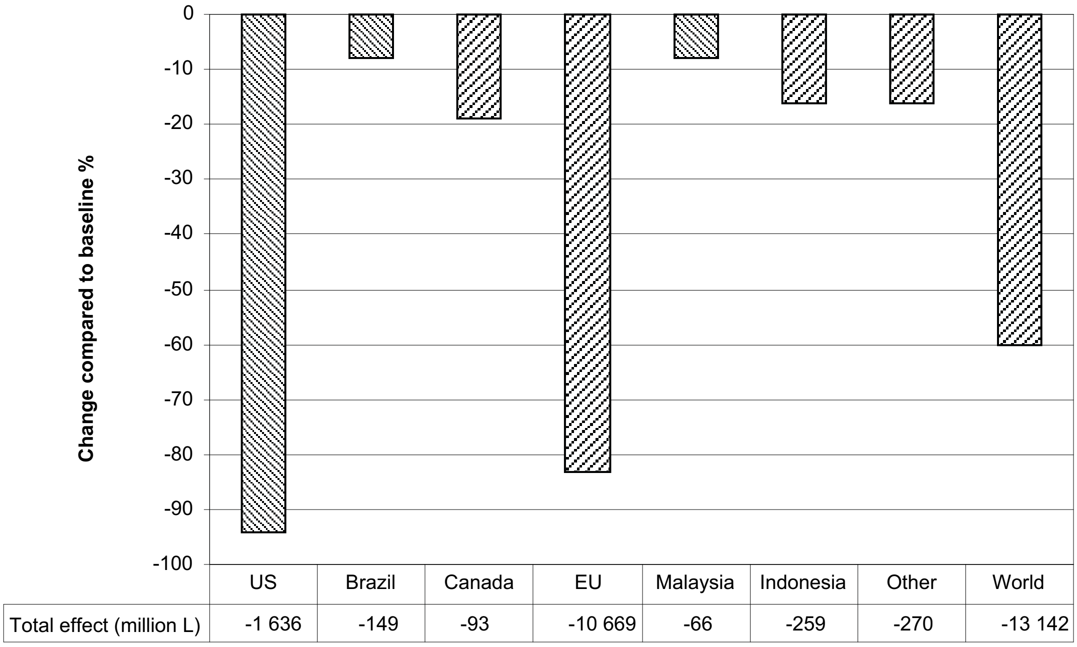
The key rationales for public support have been the ability of renewable biofuels to reduce greenhouse gas emissions and fossil fuel use. Secondary objectives include creating new market outlets for agricultural products, stimulating regional development and securing energy supply.

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).