Australian Government
Productivity Commission

Strengthening Evidence-based Policy in the Australian Federation

Roundtable Proceedings

Canberra, 17-18 August 2009
Volume 1: Proceedings

ISBN 978-1-74037-311-1

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

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

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

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

An appropriate citation for this paper is:

Productivity Commission 2010, *Strengthening Evidence Based Policy in the Australian Federation*, *Volume 1: Proceedings*, Roundtable Proceedings, Productivity Commission, Canberra.

JEL code: Q28

The Productivity Commission

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

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

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

Foreword

The Productivity Commission's 2009 roundtable was organised around the topic *Strengthening Evidence-Based Policy in the Australian Federation*, and held at Old Parliament House in Canberra on 17-18 August. Participants included government officials, academics, consultants and representatives of non-government organisations. Keynote addresses were presented by Dr Ron Haskins, Senior Fellow of the Brookings Institution, and Professor Jeffrey Smith of the University of Michigan.

The Commission has long grappled with how best to bring available evidence to bear in informing policy. Some months after coming to power, Prime Minister Rudd said: 'evidence-based policy making is at the heart of being a reformist government', foreshadowing an intensified interest in making the best use of evidence, and implying questions about how to deepen the pool of experienced evaluators and build institutions to facilitate good use of evidence.

The roundtable commenced by discussing the principles of the evidence-based policy movement and reviewed how well Australian use of evidence conformed to best practice. It then considered how to improve the availability of quality evidence, and reviewed possible institutional developments to embed good use of evidence more firmly into policy-making.

The roundtable proceedings are being published to enable a wider audience access to the information and insights that emerged. This volume includes papers by the speakers and a summary of the key points covered in the discussion sessions. A second volume is a background paper prepared by Commission staff and provided to roundtable participants.

The Commission is grateful to the speakers and other participants whose contributions made the roundtable such a valuable exercise.

Gary Banks AO

Chairman

March 2010

Contents

Foreword **III**

1 Introduction 1
Gary Banks

Session 1 Evidence-based policy: Its principles and development

2 Evidence-based policy: principles and requirements 13
Brian Head

3 With a scope so wide: using evidence to innovate, improve, manage, budget 27
Ron Haskins

4 Putting the evidence in evidence-based policy 59
Jeffrey Smith and Arthur Sweetman

General discussion and dinner address 103

Session 2 How robust is our evidence-based policy making?

5 Reflections on four Australian case studies of evidence-based policy 109
Bruce Chapman

6 Evaluating major infrastructure projects: how robust are our processes? 127
Henry Ergas and Alex Robson

7 Evidence-based policy: reflections from New Zealand 169
Grant Scobie

General discussion 185

Session 3 From rhetoric to practice — how do we improve the availability and quality of evidence?

8 Facilitating better linkages between evidence and health policy: the role of the Cochrane Collaboration 189
Sally Green and Miranda Cumpston

9 Learning from the evidence about evidence-based policy 195
Patricia Rogers

10 Evidence-based policy: summon the randomistas? 215
Andrew Leigh

General discussion 227

Session 4 Institutionalising an evidence-based approach — how can an evaluation culture be embedded into policy-making

11 Institutionalising an evidence-based approach to policy making: the case of the human capital reform agenda 231
Peter Dawkins

12 Drawing on Powerful Practitioner-Based Knowledge to Drive Policy Development, Implementation and Evaluation 249
Robert Griew

13 Intelligent federalism: accountability arrangements under COAG's reform of federal financial relations 259
Mary Ann O'Loughlin

General discussion 275

Session 5 What have we learned and where to from here?

14 Rapporteur's comments 281
Jonathan Pincus

General discussion 291

Appendices

A Roundtable program 295

B Roundtable participants 297

Box

10.1 US federal legislation that specifically funds randomised evaluation 224

10.2 A possible evidence hierarchy for Australian policy makers 224

Figures

2.1 The policy process 20

2.2 Wicked as combination of complexity, uncertainty and divergence 22

3.1 Factors that influence policy choice 29

5.1 Earned income by age and education (persons) 114

5.2 Projections of long-term unemployment 116

5.3 The unemployment–vacancies curve 117

5.4 Relative annual earnings with Year 12 certificates or TAFE diplomas 120

5.5 Comparison of the household incomes of the Youth Allowance and non-Youth Allowance groups 123

6.1 Time path of speeds: scenario A 140

6.2 Time path of speeds: scenarios B and C 140

6.3 Time path of willingness to pay curves: scenarios A and B 141

6.4 Time path of willingness to pay: scenario C 141

6.5 Path of willingness to pay under the baseline and the National Broadband Network: scenario A 142

6.6 Path of willingness to pay under the baseline and the National Broadband Network: scenario B 142

6.7 Path of willingness to pay under the baseline and the National Broadband Network: scenario C 143

6.8 Time path of speeds under scenario D 146

6.9 Time path of annual willingness to pay curves: scenario D 147

6.10 Path of annual individual annual WTPs under the baseline and National Broadband Network: scenario D 148

6.11 Take-up paths: scenario D 148

8.1 A healthcare decision-making model 192

9.1 Evidence uptake as a 'leaky pipeline' 197
9.2 Evidence uptake as an ongoing, knowledge building process 200
10.1 Costs and benefits of more randomised policy trials 222
11.1 Progress towards Year 12 completion rate (Victoria) 238
11.2 Outcomes and evaluation framework structure 240
13.1 Structure of the National Education Agreement 266
13.2 Proportion of Year 5 students achieving at or above the national minimum standard for reading 268

Tables

2.1 Types of knowledge relevant to evidence-based policy 19
3.1 Congressional agencies that provide non-partisan analyses 33
3.2 Temporary Aid for Needy Families program requirements 39
3.3 Changes in US population and employment, 1960 to 1990 42
3.4 Is low-wage work a good deal? 44
6.1 Incremental benefits under various scenarios 144
6.2 Incremental benefits under various scenarios: enhanced WTP 145
6.3 Present value of the net incremental benefits of the National Broadband Network: scenario D 149
6.4 Present value of the cumulative 12-year change in GDP due to construction of the National Broadband Network 153
9.1 Approaches to evidence-based policy 201
13.1 Commonwealth payments to the States and Territories 262
13.2 National Partnerships supporting the National Education Agreement 271

1 Introduction[1]

Gary Banks
Chairman, Productivity Commission

This annual Roundtable has a special place among the Commission's many activities.

- it provides an opportunity for us to step outside our project work and focus collectively on what we see as a key policy issue or theme;
- in a setting which allows for frank discussion among a cross-section of influential people;
- affording 'time out' that will hopefully benefit all of us back on the job and, ultimately, promote the cause of good public policy — our principal objective.

As with past Roundtables, we have been very fortunate in our final list of attendees — including our keynote speakers from overseas, Ron Haskins from the Brookings Institution and Jeff Smith from the University of Michigan. The senior ranks of public service, both at the Commonwealth and State level, are well represented here, as are academia, private research and consultancy organisations.

The topic for this year's Roundtable seemed almost to choose itself. In a well-publicised address early last year to senior public servants, some months after coming to power, Prime Minister Kevin Rudd said: 'evidence-based policy making is at the heart of being a reformist government'. Other Ministers in the new government echoed similar sentiments, foreshadowing a change in policy-making that was widely welcomed, particularly in Canberra.

A number of initiatives bore early testimony to the government's convictions, including radical changes to the framework for national policy-making under COAG, the refocussing of the national reform agenda (NRA) and a suite of evaluations and public reviews in key policy areas. A succession of policy decisions since then have excited considerable controversy, however, including on the very

1 Opening remarks to the Productivity Commission Roundtable, *Strengthening Evidence-Based Policy In the Australian Federation*, Old Parliament House, Canberra, 17 August 2009.

question of whether they could be justified by analysis and evidence. These include the 'alcopops' tax; the linkage of Indigenous welfare payments to school attendance; fuel watch; grocery watch; the Green Car Innovation Fund and, more recently, the National Broadband Scheme.

There was, of course, similar public questioning of a number of policy initiatives by the previous government, such as the Alice-to-Darwin rail link; the Australia-US Free Trade Agreement; the Baby Bonus; the banning of filament light bulbs; Work Choices and the National Water Initiative, and others. And, under both governments, the evidence supporting policies to reduce carbon emissions has been a matter of great contention — both in relation to the science linking global warming to anthropogenic activity (people) and, more frequently, in relation to the instruments chosen to reduce Australia's own emissions.

Moreover, where public reviews informed such initiatives, they have themselves been subjected to considerable criticism — in relation to their makeup, their processes and the quality of their analysis.

This too is obviously not a new phenomenon. But it illustrates the challenges of properly implementing an evidence-based approach to public policy — and of being *seen* to have done so, which can be crucial to community acceptance of consequent policy decisions.

The degree of difficulty has of course been elevated considerably over the past 12 months by the global financial crisis and its attendant threats to jobs and living standards. A speedy response was called for, and some tradeoffs necessarily made with normal processes.

The signs thus far appear to be vindicating this approach to dealing with the global crisis. But the very basis for Australia's success in the short term presents some longer term policy challenges, in which an evidence-based approach will again need to come into its own. For one thing, it will be important to disengage from those 'crisis' measures that turn out to be misplaced, unnecessary or unsuitable for the longer haul. For another, the stimulus spending has left a legacy of major fiscal pressure that will call for careful prioritisation of spending programs, based on a good understanding of their relative payoffs. The macro policy induced pressures will compound the existing fiscal pressures of our inexorably ageing population. These call for micro reforms to enhance workforce participation and productivity in policy areas like education, health and welfare — areas that pose trickier challenges than many of the microeconomic reform areas of the past.

And of course the 'green elephant' in the room — the global warming issue — must be addressed in a way that can meet multiple objectives in a state of great

complexity and uncertainty. This demands an adaptive policy approach in which monitoring and evaluation of novel regulatory frameworks and institutions must play a central role.

So how well prepared are we to deliver the evidence and analysis that can indeed be at the heart of our governments' forward agenda? The truth is, that while there has been much talk about evidence-based policy, far less attention has been paid to how we actually go about it and how we might do it better.

These questions provide the main focus for this Roundtable, with four sessions devoted to key aspects and a final one to possible ways forward.

Evidence-based policy — its principles and development

The first session will seek to develop a common understanding of what evidence-based policy is, its main principles and the role evidence should play in the policy process.

Brian Head will provide an overview of the prevailing currents in critical thinking about evidence-based policy among analysts of government and our American guests, Jeff Smith and Ron Haskins, will reflect on their experiences in the application of evidence to policy, on some recent methodological trends and institutional experiments.

While there is room for debate about various dimensions of these questions, the notion that public policy decisions can benefit from a process of deliberation informed by facts and analysis is unlikely to be contentious. Nor, I think, would we disagree that throughout history practice has often fallen short of this modest ambition.

In this respect, I have become fond of citing Florence Nightingale who, over a century ago, admonished the English Parliament thus:

> You change your laws so fast and without inquiring after results past or present, that it is all experiment, seesaw, doctrinaire; a shuttlecock between battledores.

I don't say that this depiction would be typical today. But I suspect many of us can relate to the sentiments she so colourfully expressed.

That said, it would be idle to pretend that policy decisions could ever be determined by evidence or analysis alone. As many of us in this room will also know first-hand, the realpolitic of public policy involves a much richer array of influences.

But evidence and analysis can nevertheless play a useful role in informing policy-makers' judgements. Importantly, they can also condition the political environment in which those judgements need to be made.

We can all cite instances where attention to evidence has helped achieve better policy outcomes, and where its absence has led to 'misfires' and unintended consequences. And I'm sure we can also think of policy reforms that became politically more palatable because the community had the opportunity to learn about the tradeoffs involved.

Some of the questions that we might consider in the opening session this afternoon, therefore, include the following:

- *Given the realities of political decision-making, how should we best define or characterise an 'evidence-based' approach?*
  - *What degree of influence can or should we expect it to have?*
- *More basically still, what constitutes 'evidence'?*
  - *What forms of evidence are there?*
  - *Is there a role for qualitative analysis and opinion?*
  - *Should multiple sources of information be used?*
- *How does evidence relate to 'theory'?*
  - *Does the Blairite mantra, 'what counts is what works', mean that pragmatism should take precedence over principle?*
- *Are there differences that need to be recognised when contemplating a new policy initiative, as opposed to assessing an existing one?*
- *Who is best placed to provide the evidence needed for public policy?*
  - *In particular, what are the relative merits of evidence generated within government and that generated or commissioned externally?*
- *And how important is transparency and the ability of third parties to scrutinise the analysis or replicate the results?*

How robust is our evidence-based policy making?

The second session will examine where we have been successful and where we are falling short, and why. Bruce Chapman and Henry Ergas will reflect on lessons from recent policy reforms, and Grant Scobie will offer some observations on New Zealand practice.

Australian Governments have directed significant resources and effort towards data collection and analysis, and policy evaluation.

But, overall, the 'evidence' on the extent to which evidence-based policy is actually applied is mixed.

- Often policy-related research and evaluation efforts have focused on areas where there is good evidence and tended to avoid those that are more challenging.
- Similarly, some parts of government are open, provide access to data holdings and actively invest in the evidence base, whereas others hold data more tightly and resist efforts to build an information base that could be used to evaluate their programs.
- And, we know from experience that, despite best intentions, policies can ultimately be shaped more by guesswork and political drivers than rigorous analysis.

Some of the questions and issues that are relevant to this session are:

- *To what extent do we observe evidence and analysis being effectively used to inform political decision–making?*
  - *Are some ingredients more often lacking than others?*
  - *Is evidence used to assess the relative merits of different feasible policy options?*
- *The Commission's experience with the Office of Regulation Reivew, and for a time, the Office of Best Practice Regulation, as watchdogs on good regulatory practice, revealed many instances of regulation impact statements being concocted to support a regulatory decision that had already been made.*
  - *Is it common for evidence to be marshalled to support predetermined policies (policy-based evidence) in other areas?*
- *When evidence is used, how good is it, and where does it typically come from?*
  - *Is it often 'tested' publicly or subjected to peer review?*
  - *Do we see much quantification and empirical investigation in policy development? Where do we see it most and least?*
- *Even when evidence has been properly marshalled, how influential has it been on policy decisions and outcomes?*
  - *Have governments revealed a preference for using evidence when it really counts?*

- *Do we see 'proportionality' in the evidence that is brought to bear on decisions with varying potential impacts?*
- *Has Australia capitalised sufficiently on its federal system of government, to learn from the policy experiments and experiences of different jurisdictions?*
  - *How have Ministerial Councils performed in this respect?*
  - *What about the more recent experience with COAG working groups?*

From rhetoric to practice: how do we improve the availability and quality of evidence?

The third session will examine ways to improve the evidence for policymaking.

Sally Green will recount the role of the Australasian Cochrane Centre in marshalling and disseminating high quality evidence, including beyond traditional clinical and pharmacological work. Andrew Leigh will be considering the case for randomised policy trials, and their varying suitability to different classes of social and economic policy problems. And Patricia Rogers will reflect on her wide analytical and practical experience of evaluation challenges.

There are two main dimensions to discuss: data and methodologies.

On the first, Australia has been well served by the ABS and the integrity and breadth of the data bases that it has generated. Data is particularly good (comparative, consistent over time) in the economic and demographic domains.

But we lack good data in many social and environmental areas, including some that are at the centre of the COAG Reform Agenda, like education and Indigenous policy. We have suffered from a lack of longitudinal data in particular (though the HILDA project has helped remedy this since 2002).

Where official data is lacking, there are a number of choices available to policy makers, including special purpose surveys, focus groups and overseas studies. All are fraught with difficulty and can pose risks for policy makers, some of which have been satirised to devastating effect in *The Hollowmen* TV series.

So some questions here include:

- *What constitutes 'good' data for the purposes of building evidence to inform policy?*
- *Can data be developed in the (truncated) 'real time' of a policy development process?*

- *Where should we be collecting more data?*
- *Where data to assess new programs cannot be generated automatically, should we design programs to fill the gap? In particular, do we need to collect more baseline data to enable 'before and after' comparisons?*

Towards better methodologies?

The data we need or use are often related to the methodologies available and there is considerable debate about their relative merits.

That said, all good methodologies have a number of features in common.

- Most fundamentally, they test a theory or proposition as to why policy action is needed and will be effective.
- They have a considered treatment of the 'counterfactual': what would happen in the absence of any action?
- They involve quantification of impacts, both direct and indirect (often it's the *indirect* effects that can be most important).
- They set out the uncertainties and control for other influences that may impact on observed outcomes.
- They have the ability to be tested and, ideally, replicated by third parties.

However, best practice approaches will not always be practicable. Policy advice often has to be provided within tight timeframes, and can be subject to significant constraints. We need to develop practical ways to provide the most robust evidence in these cases.

- *What can be achieved through econometrics, modelling, trials and other evaluation methods in the context of these constraints? How can these be designed to manage validity, cost, ethical and other considerations?*
- *Where can we invest in rigorous evaluation and when should we rely on post-implementation monitoring and review?*
- *How can we learn from overseas experiences, and when should we be wary about the applicability of overseas evaluation/information?*

Institutionalising an evidence-based approach

For evidence and evaluation to contribute materially to the selection of policies, it must be supported by institutional frameworks that embed the use of evidence and encourage, disseminate and defend good evaluation.

The institutional framework should also ensure that the resources allocated to evaluation are commensurate with the potential benefits.

The fourth session will explore what institutional frameworks and government processes might best support evidence-based policy.

Our presenters in this penultimate session are Peter Dawkins, who will reflect on experiences in the NRA including through his leadership of Victoria's Department of Education and Early Childhood Learning, Robert Griew, who will draw on both his state and federal experience, and Mary Ann O'Loughlin from the COAG Reform Council who will consider the national architecture for implementing reforms.

Even the best evidence is of little value if it's ignored or not available when it is needed. An evidence-based approach requires a policy-making process that is *receptive* to evidence; a process that begins with a question rather than an answer, and that has institutions to support such inquiry.

Ideally, we need systems that are open to evidence at each stage of the policy development 'cycle': from the outset when an issue or problem is identified for policy attention; to the development of the most appropriate response, and subsequent evaluation of its effectiveness.

The ongoing struggle to achieve effective use of regulation assessment processes within governments, to which I alluded to before, tells us how challenging that can be in practice.

Admittedly, an evidence-based approach can make life harder for policy makers and politicians. Lord Keynes, whose ideas appear to have made a bit of a come-back recently, said in the 1930s:

> There is nothing a Government hates more than to be well-informed; for it makes the process of arriving at decisions much more complicated and difficult.

I think we can see what he meant. But, against this, are the undoubted political *benefits* that come from avoiding policy failures or unintended 'collateral damage' that can rebound on a Government, and from enhancing the credibility of reformist initiatives.

- *How can the real politic of public policy be made more compatible with evidence-based approaches?*
- *Is there scope to strengthen existing institutions within each government and across our federation?*

- *How can we ensure that we get the best out of the resources already being devoted to research and policy advice?*
- *Do we need to (re-)build research capacity and capability within government or should we continue to rely more on external sources of research, analysis and advice?*
  - *If we need both, how do we get the balance right?*
- *How do we create incentives for quality analysis, whether within government or through contractors/consultants?*
- *How can we limit an inherent tendency for second-guessing and 'policy-based evidence', that can mean superior policy options being ignored?*
- *Should the Australian Government play a stronger role in promoting necessary data collections and evaluations within and across jurisdictions?*
  - *Could it make more use of its funding leverage with the States and Territories?*
- *Is there scope for COAG to establish an 'evaluation club' to help propagate and disseminate evidence in key policy areas? (An encouraging development of this kind has emerged in the Indigenous policy area; namely the 'national clearing house' on best practice and success factors).*
- *In the TV series 'Yes Minister', Sir Arnold confides to Sir Humphrey "If people don't know what you're doing, they don't know what you're doing wrong". Can data and analysis that are not able to be scrutinised by third parties really be called 'evidence'?*
  - *How do we achieve greater transparency in the data that is collected and in the evaluations that are conducted?*
- *Given the time lag between data collection and analysis and policy development, how good are we at anticipating the policy questions of the future?*
  - *While many policy questions have been around for a while, the impetus and timing for policy reform in particular areas is often hard to predict. How do we ensure that the currency/availability of necessary evidence matches the contemporary policy issues being addressed by government?*

Where to from here?

The fifth session will conclude proceedings by drawing out some of the more important implications for public policy in Australia that emerge from previous sessions and the address by Terry Moran.

Jonathan Pincus, a former Principal Advisor Research at the Commission and now Visiting Professor at the University of Adelaide, will introduce this final session with a summary of the main issues and insights from the Roundtable over the two days. A panel comprised of David Tune from Prime Minister and Cabinet (now Secretary of the Department of Finance and Deregulation), as well as Ron Haskins, Jeff Smith and Mary Ann O'Loughlin, will then reflect on the discussions and draw their own conclusions.

SESSION 1

EVIDENCE-BASED POLICY: ITS PRINCIPLES AND DEVELOPMENT

2 Evidence-based policy: principles and requirements

Brian Head
University of Queensland

Abstract

Evidence-based policy (EBP) is an aspiration rather than an accomplished outcome. The advocates of EBP urge the incorporation of rigorous research evidence into public policy debates and internal public sector processes for policy evaluation and program improvement. The primary goal is to improve the reliability of advice concerning the efficiency and effectiveness of policy settings and possible alternatives. This is attractive to pragmatic decision makers, who want to know what works under what conditions, and also to those professionals concerned with improving information bases and improving the techniques for analysis and evaluation. Some concerns are raised by professionals whose knowledge-discipline or whose policy focus is not well served by quantitative analytical techniques, and who worry that important qualitative evidence may be overlooked. Scientific experts may reasonably disagree about methods, instruments and impacts. Whatever methodologies are employed, EBP requires good data, analytical skills and political support. Hence there are inherent limitations, even where government officials are able to draw on the results of reliable information and sound analytical skills. The politics of decision making inherently involves a mixing of science, value preferences, and practical judgments about feasibility and legitimacy. Outside the scientific community, the realm of knowledge and evidence is even more diverse and contested. Competing sets of evidence and testimony inform and influence policy. The professional crafts of policy and program development require 'weaving' these strands of information and values. The cutting-edge issues in modern EBP debates focus on problem-framing, methods for gathering and assessing reliable evidence, communicating and transferring knowledge into decision making, and evaluating the effectiveness of implementation and program delivery in complex policy areas.

2.1 Introduction

There are three crucial enabling factors that underpin modern conceptions of evidence-based policy (EBP): high-quality information bases on relevant topic areas, cohorts of professionals with skills in data analysis and policy evaluation, and political incentives for utilising evidence-based analysis and advice in governmental decision-making processes. The precursors of modern EBP thus have a long, if patchy, history over many decades, inspired by a desire to improve social, economic and environmental outcomes through the application of reliable knowledge. The story of EBP is as much about institutional development as about data and skills.

Australia, along with other prosperous Western countries, has developed a strong institutional foundation for nurturing EBP capacities. However, those enabling factors have developed unevenly, being more prominent in some periods than in others. For example, postwar reconstruction — a major theme in the federal government's policy and planning concerns from about 1943 — arguably galvanised those factors. However, there was a political retreat from comprehensive 'planning' discourses under the Menzies government in the 1950s, although the infrastructure for data collection and skills development continued in various ways. Policy development and innovation, especially in social and urban policy, again became a strong theme in the early 1970s under a reformist federal government, and there was another boost from the mid-1980s with a stronger policy emphasis on economic productivity, regulatory liberalisation, and new approaches to social equity and environmental protection. As higher education expanded, the general culture of business and government became more favourable to the creation of 'policy intellectuals' in various institutional niches (Head 1988; Withers 1981).

Over time, specialised governmental organisations devoted to systematic data collection, the analysis of information and the evaluation of policy options, including the Productivity Commission, have grown in size and capability. In particular, long-term public investment in economic and social statistics, along with the development of more specialised units for policy and regulatory analysis, have provided a solid foundation for contemporary EBP capacities (Banks 2009). This investment has been driven by broad and diverse policy needs — for example, the need to understand and influence population trends, to assess and improve environmental sustainability, to plan and fund effective human services and social security, to provide better communications and transport infrastructure, to assess tax revenue capacities, and to meet the economic productivity challenges of international competition. (I omit here the highly distinctive research/policy needs of foreign policy, defence and intelligence organisations.) The Australian Government's commitment to good data and sound analysis has also been reinforced by its growing involvement in international organisations (such as the

Organisation for Economic Co-operation and Development) and its endorsement of international agreements that require sophisticated reporting on comparative performance trends.

Beyond the sphere of the federal government, there have also been some important investments in EBP capacities within State governments, but on a much smaller scale and with generally lower levels of political support. Much of the impetus for States to invest in EBP capacity has been linked to their involvement with the powerful intergovernmental policy reform processes driven through the Council of Australian Governments and to a lesser extent other ministerial councils. Other new inputs to policy development have been associated with the recent proliferation of policy-oriented consultancy firms, and the emergence of independent think tanks (Marsh and Stone 2004) outside the public sector, sharpening the ongoing debates about the quality and timeliness of advice.

The United States has long been the major global location for policy analysis and evaluation professionals, both within government and in other policy-relevant sectors (see, for example, Lerner and Lasswell 1951; Nathan 1988; Wilson 1981). The refinement of methodologies for evidence-based assessment of program implementation, and for analysing alternative policy options, has been driven substantially by large cohorts of US scholars and policy managers. The mandating of particular forms of program appraisal as a condition of program funding has also proceeded further in the United States than in most other nations (Boruch and Rui 2008, Haskins 2006), although the practical experiences of evaluation remain diverse and somewhat fragmented across agencies and levels of government.

The British Government under Prime Minister Blair attempted to develop a coherent approach to policy development, championing EBP as a major aspect of the increased policy capability and the fresh thinking required by a reformist government (UK Cabinet Office 1999a, 1999b). This increased respect for research and evaluation was generally welcomed by policy researchers (for example, Davies et al. 2000), although some were troubled by the implicit preference for quantitative precision and for technical expertise over other forms of knowledge (for example, Parsons 2002, 2004). One of the positive outcomes has been a more comprehensive investment in policy-relevant research and a stronger commitment to evaluation.

In Australia, Prime Minister Rudd announced on 30 April 2008 that his government, not unlike the Blair government a decade earlier, saw a strong link between EBP and good governance:

> A third element of the Government's agenda for the public service is to ensure a robust, evidence-based policy making process. Policy design and policy evaluation should be driven by analysis of all the available options, and not by ideology. When preparing

> policy advice for the Government, I expect departments to review relevant developments among State and Territory governments and comparable nations overseas. The Government will not adopt overseas models uncritically. We're interested in facts, not fads. But whether it's aged care, vocational education or disability services, Australian policy development should be informed by the best of overseas experience and analysis. In fostering a culture of policy innovation, we should trial new approaches and policy options through small-scale pilot studies.
>
> Policy innovation and evidence-based policy making is at the heart of being a reformist government. (Rudd 2008)

These sentiments have been well received. However, the practical implications remain open to interpretation and debate, especially as to whether EBP in Canberra will entail an incremental approach building on best practice in professional analysis and advice, or will require greatly enhanced professional practices and wider adoption of specific skills (for example, cost–benefit analysis: Argyrous 2009). The initial lack of explicit guidance concerning preferred methodologies may have been a matter of either serious concern or great relief for different sections among the policy professionals. The overall level of commitment to investments in policy-relevant research, program evaluation and policy skills training in Australia has been disappointing, especially at State level. It remains to be seen whether the reinvigorated commitment to EBP will lead to measurably greater investment in policy research and evaluation over the coming years. In the following sections, I raise some basic issues about the political and institutional context in which EBP is pursued, as well as the internal debates about methods and reliable evidence.

2.2 Knowledge and rigour

The advocates of EBP urge the incorporation of rigorous research evidence into public policy debates and internal public sector processes for policy evaluation and program improvement. The primary goal is to improve the reliability of advice concerning the efficiency and effectiveness of policy settings and possible alternatives. The quest for rigorous and reliable knowledge, and the desire to increase the utilisation of rigorous knowledge within the policy process, are core features of the EBP approach.

Two main kinds of critical commentary have been expressed by observers and participants. The first can be termed 'internal' critical commentary, focusing on the suitability of various preferred *methodologies* for collecting, interpreting and applying evidence as a basis for understanding — and perhaps improving — particular programs or policies. The second can be termed an 'external' or *contextual* commentary, focusing on how and where the EBP contributions (based on rigorous evidence) can be most influential, and how they fit into the wider

picture of policy debate and evaluation — a public canvas largely painted by partisan viewpoints.

The quest for rigour is vital. There are many sophisticated sources of guidance for methodological questions of data validity, reliability and objectivity. Those who focus on these fundamental issues are usually specialists in the design of information capture and analysis (such as the Australian Bureau of Statistics and the Australian Institute of Health and Welfare) and/or specialists in the design of applied research on specific problems or programs. Two of the most widely debated matters in recent years are the significance of 'qualitative' evidence and a possible 'hierarchy' of reliability in different models of applied research. These are matters discussed by other papers and are noted here only briefly.

Concerns about the value of qualitative evidence stem from different research traditions in the social sciences. Some disciplines (for example, social anthropology and history) have usually tended to be concerned with accounting for the 'experience' of participants — meanings, motives, contexts — rather than seeking behavioural generalisations (which are more typical of quantitative approaches relying on economic and social statistics). Bridges are being built between the advocates on both sides. Program evaluation professionals tend to use mixed methods as appropriate. Large-*N* qualitative studies are increasingly seen as open to the techniques of quantitative analysis. Longitudinal panel studies are an increasingly rich source of several types of evidence. In the United Kingdom, the early champions of very rigorous EBP quickly found it necessary (for example, Davies 2004) to soften the apparent bias towards randomised controlled trials (RCTs); and the central agencies have recognised that qualitative studies are important, provided they are conducted with appropriate methodological rigour (UK Cabinet Office 2003, 2008; UK Treasury 2007). Mixed methods are increasingly championed by analysts who are attempting to explain complex problems and assess complex interventions (for example, Woolcock 2009).

Turning to the related debate on 'evidence hierarchy', the underlying question is trust in the reliability of research findings. Some argue that there is a research-quality hierarchy, based on the types of methodological rigour used to design and interpret field studies. In particular, it is claimed that the RCT approach pioneered in medical research can and should be applied in the social sciences (Leigh 2009). A variant of this is the argument that single-study findings are misleading, and that a better understanding of causes and consequences emerges from 'systematic reviews' of all available research (Petticrew 2007; Petticrew and Roberts 2005), taking into account the rigour of the methods followed. The counterarguments turn on the difficulty of implementing RCTs in sensitive areas of social policy; the difficulty of transplanting quasi-experimental results to large-scale programs

(Deaton 2009); and the tendency to downplay the knowledge and experience of professionals with field experience (Pawson 2006; Schorr 2003). It is also highly likely that politicians, policy managers, scientists and service users may have very different perspectives on what kinds of evidence are most trustworthy (see, for example, Glasby and Beresford 2006).

2.3 Knowledge for policy: how many lenses?

The movement to improve the evidence base (or bases) available for policy analysis and for program improvement is of crucial importance and is widely supported in many quarters. Encouraging organisational cultures that support more systematic evaluation of initiatives and interventions is also crucial. But building capability can be expensive. Moreover, providing transparent evaluations of program initiatives can be politically risky. Governments do not relish being exposed to strong public criticism for poor program outcomes or for pilot schemes that produce weak results. The culture of evaluation is best understood as a culture of learning, and therefore needs to be embedded as bipartisan good practice.

The knowledge base for EBP is diverse. Systematic research (*scientific* knowledge) provides an important contribution to policy making, and is undertaken in external institutions as well as in the public service. But science is only one of the inputs for EBP. The larger world of policy and program debate comprises several other types of knowledge and expertise that have legitimate voices in a democratic society. It has been argued elsewhere (Head 2008b; Shonkoff 2000) that these other 'lenses' or knowledge bases may include the following:

- The *political* strategies, tactics and agenda setting of political leaders and their organisations set the 'big picture' of priorities and approaches. The logic of political debate is often seen as inimical to the objective use of policy-relevant evidence: ideas and values are instead mobilised to support political objectives and to build coalitions of support; spin may become more important than accountability.
- The *professional* knowledge of service delivery practitioners and program coordinators is vital for advising on feasibility. They have crucial experience in service delivery roles and field experience in implementing and monitoring client services across social care, education, health care, etc. They wrestle with everyday problems of effectiveness and implementation, and develop practical understandings of what works (and under what conditions), and sometimes improvise to meet local challenges.

- In addition to the above *institutional* sources of expertise, the experiential knowledge of *service users* and stakeholders is vital for 'client-focused' service delivery. Ordinary citizens may have different perspectives from those of service providers and program designers; their views are increasingly seen as important in program evaluation for ensuring that services are appropriately responsive to clients' needs and choices.

As illustrated in table 2.1, rigorous and systematic science seeks a voice in a competitive struggle for clarity and attention, jostled by many players in the wider context of public opinion and media commentary. To the extent that rigour is valued, it therefore needs to be protected by strong institutions and robust professional practices.

Table 2.1 **Types of knowledge relevant to evidence-based policy**

Political knowledge	Scientific rigorous knowledge	Professional–managerial knowledge	Client and stakeholder knowledge

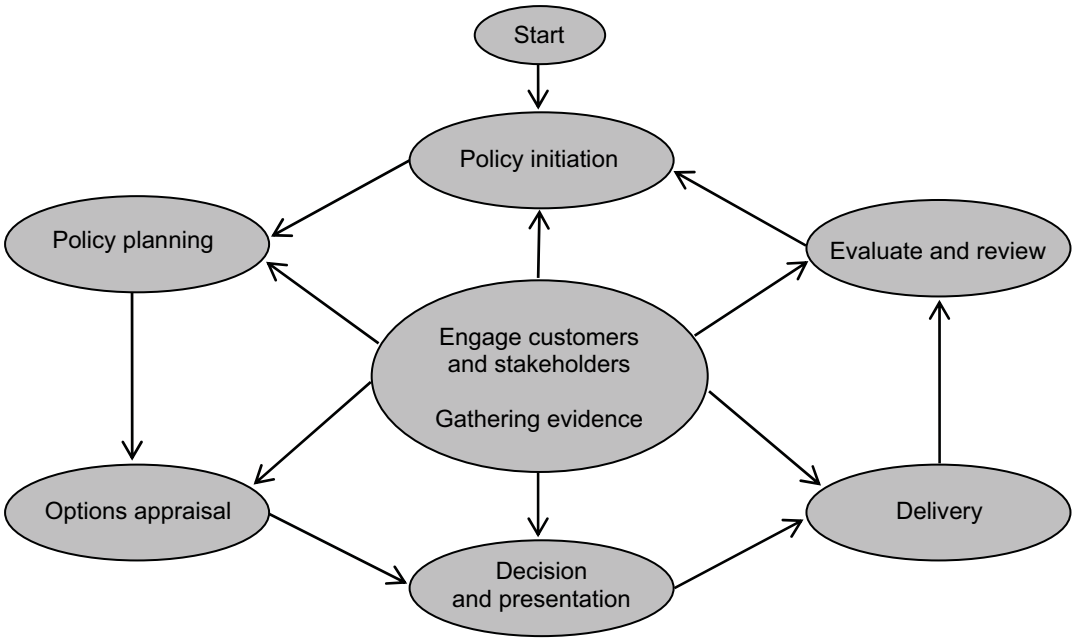
The mass media and political culture

2.4 The policy process

The value proposition for EBP is that policy settings can be improved on the basis of high-quality evidence. How does reliable knowledge actually flow between producers and users? How strong are the channels and relationships that improve those flows? Unfortunately, the channels through which rigorous evidence may influence policy making are readily disrupted by external pressures, and therefore need specific care and attention (Landry et al. 2001; Lavis et al. 2003; Nutley et al. 2007). Supply-side provision of good research about 'what works' is not enough. Potential users will pay closer attention only when they are better aware of these inputs, understand the advantages and limits of the information, and are in a position to make use of the findings either directly or indirectly (Edwards 2004; Nutley et al. 2007). How can the social and institutional foundations for EBP be improved? What capabilities need to be built within and across organisations? Considerable research on such matters is already being undertaken across a range of social policy areas but cannot be summarised here in detail (for example, see Boaz et al. 2008, Dopson and Fitzgerald 2005; France and Homel 2007; Jones and Seelig 2005; Mosteller and Boruch 2002; Lin and Gibson 2003; and Saunders and Walter 2005).

A related matter is to determine at which points in the policy development and policy review 'cycle' EBP contributions (based on rigorous evidence) can be most influential. Based on the few available studies of policy development in Australia, it would appear that there are no general answers (Edwards 2001). The formal expectation might be that policy-relevant research about the effectiveness of various options (what works under what conditions) might be most closely linked into the evaluation phase of the policy cycle (for example, Roberts 2005). However, the notion of a rational and cyclical process of policy development, implementation and review does not correspond closely with political realities (Colebatch 2006). A more realistic and complex model is conveyed in a diagram published by the Scottish Executive (see figure 2.1), which allows for reiteration of process steps and continual processes of further consultation and gathering of evidence.

Figure 2.1 **The policy process**



Source: Scottish Executive (2006).

Rigorous evidence can therefore be relevant at several points in the development and review processes. But not all matters are genuinely open to rethinking. Some areas of policy are tightly constrained by government priorities, electoral promises and ideological preferences. There is perhaps less scope for changing these as a result of evidence about 'what works'. It is also useful to identify program areas that appear to be more settled than others over a period of time (Mulgan 2005). It is possible that evidence-based arguments about 'fine-tuning', based on careful research about effectiveness, might be more likely to gain traction in those areas if

they are away from the political heat. On matters of deep controversy, however, research findings are more likely to be mobilised as arrows in the battle of ideas, and sometimes in ways that the original authors may find distasteful. In this sense, the policy process is a patchwork quilt of arguments and persuasion (Majone 1989). However, policy adjustments, and opportunities for new thinking, can emerge in unexpected ways in response to incidents, crises and conflicts.

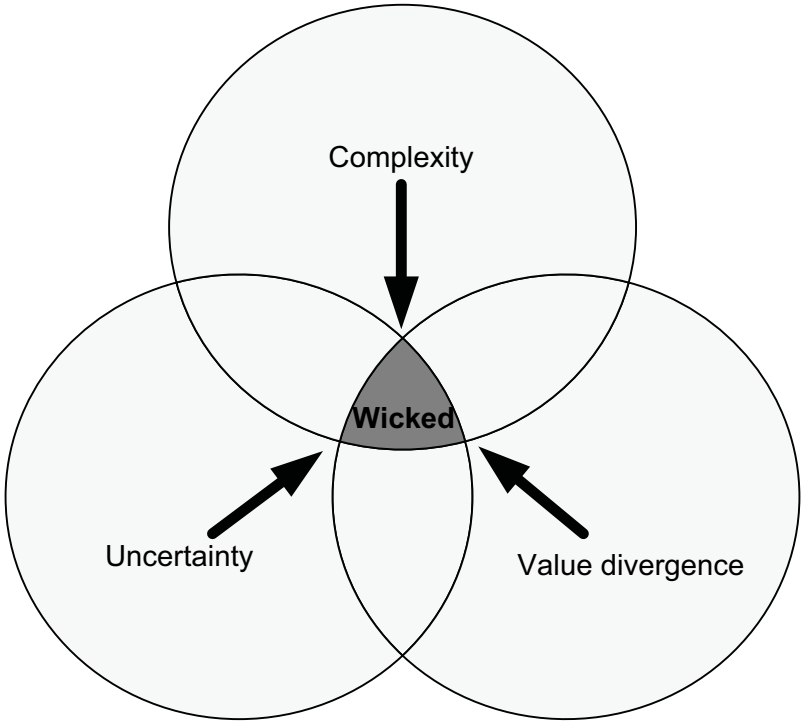
2.5 The tangles of complexity

Problems can be conceptualised at a variety of scales and with varying degrees of complexity. This is inherent in the nature of politics and policy debate. The scale or unit of analysis (for example, micro or macro level) and complexity (one issue or a nest of related issues) makes a big difference to how policy problems are framed, debated and researched. One feature of modern policy making is that a lot of 'big' problems are being addressed at the same time. Examples in Australia include specific programs to tackle Indigenous disadvantage, navigate the global financial crisis, respond effectively to the challenges of climate change, reform the health and education systems, and promote frameworks for social inclusion and nurturing early years development. Such large focus areas of policy attention pose challenges for evidence, analysis and recommendation. The political requirement for solutions will sometimes encourage broad-brush responses, rather than detailed bottom-up research as a basis for program trials and for careful evaluation prior to larger-scale implementation.

Some of these large problems have been termed 'wicked', owing to their resistance to clear and agreed solutions. These systemic and complex problems are marked by value divergence, knowledge gaps and uncertainties, and complex relationships to other problems (see figure 2.2, and discussion in Head 2008c and APSC 2007). It is not clear that traditional bureaucratic structures, or even the sophisticated managerial approaches of modern outcome-focused governments, are able to tackle these intractable problems successfully through standardised approaches.

One of the features of complex social problems is that there are underlying clashes of values, which are sometimes not adequately recognised and addressed (Schon and Rein 1994). Policy analysts have therefore tended to drift instinctively into two camps — one group tends to look for simple technical solutions (for example, inserting specific conditions into the funding for individual recipients of a program), whereas others seek to focus on identifying the underlying value conflicts as a basis for dialogue, mediation and conflict reduction prior to discussion of next steps toward solutions (Lewicki et al. 2003).

Figure 2.2 **Wicked as combination of complexity, uncertainty and divergence**



The attractiveness of recent behavioural approaches based on incentive theory (for example, Thaler and Sunstein 2008) is that judicious 'nudging' of citizens through incentives and penalties can potentially produce positive outcomes, with less need for intensive long-term case management or other expensive oversight and compliance mechanisms. In effect, the citizens are 'nudged' towards voluntary behavioural change arising from the 'choice architecture' embedded in the program design. Similarly, the attraction of quasi-market mechanisms for allocating scarce resources (such as irrigation water) and the perceived advantage of voluntary codes of conduct for industry is that detailed prescriptive regulation can be minimised (APSC 2009b).

The alternative widely favoured approach for addressing complex social problems is participatory collaboration, partnering and devolution (see APSC 2007, 2009a). The difficulties of such approaches are well known in terms of time, energy and ambiguity. Multiple stakeholders certainly complicate the challenges both of designing clear programs with defined roles and responsibilities, and of assessing the effectiveness of outcomes to be achieved through collaboration (Head 2008a). The social science challenges of evaluating complex programs are significant. Nevertheless, there is a very strong case for persisting in the face of complexity, since the underlying problems are of enormous importance for governments and citizens alike.

Some elements of these policy puzzles may be amenable to close scrutiny via rigorous appraisal and even through commissioning more RCTs. This is desirable. But the place of high-quality case studies in the broader context of complex policy challenges will need to be carefully contextualised. The professional crafts of policy and program development will continue to require 'weaving' together the implications of case studies with the big picture, and to reconcile the strands of scientific information with the underlying value-driven approaches of the political system.

References

APSC (Australian Public Service Commission) 2007, *Tackling Wicked Problems: A Public Policy Perspective,* APSC, Canberra.

—— 2009a, *Policy Development through Devolved Government*, APSC, Canberra.

—— 2009b, *Smarter Policy: Choosing Policy Instruments and Working with others to Influence Behaviour*, APSC, Canberra.

Argyrous, G. (ed.) 2009, *Evidence for Policy and Decision-making*, UNSW Press, Sydney.

Banks, G. 2009, 'Evidence-based policy-making: What is it? How do we get it?', ANZSOG Public Lecture, 4 February, http://www.pc.gov.au/speeches/cs20090204. Also reprinted as 'Challenges of Evidence-based Policy', Australian Public Service Commission.

Boaz, A., Grayson, L., Levitt, R. and Solesbury, W. 2008, 'Does Evidence-based Policy Work? Learning from the UK experience', *Evidence & Policy*, vol. 4, no. 2, pp. 233–53.

Boruch, R. and Rui, N. 2008, 'From randomized controlled trials to evidence grading schemes: current state of evidence-based practice in social sciences'. *Journal of Evidence-Based Medicine*, vol. 1, no. 1, pp. 41–9.

Campbell Collaboration, http://www.campbellcollaboration.org/

Colebatch, H.K. (ed.) 2006, *Beyond the Policy Cycle*, Allen & Unwin, Sydney.

Davies, P. 2004, 'Is evidence-based policy possible?', The Jerry Lee Lecture, Campbell Collaboration Colloquium, Washington, 18–20 February.

Davies, H.T, Nutley, S.M. and Smith, P.C. (eds) 2000, *What Works? Evidence-based Policy and Practice in Public Services*, Policy Press, Bristol.

Deaton, A.S. 2009, 'Instruments of development: randomization in the tropics and the search for the elusive keys to economic development', Working Paper

14690, National Bureau of Economic Research, Cambridge, Massachusetts, http://www.nber.org/papers/w14690 (accessed 5 January 2009).

Dopson, S. and Fitzgerald, L. (eds) 2005, *Knowledge to Action? Evidence-based Health Care in Context*, Oxford University Press, Oxford.

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

Edwards, M. 2004, *Social Science Research and Public Policy: Narrowing the Divide*, Policy Paper 2, Academy of Social Sciences in Australia, Canberra.

France, A. and Homel, R. (eds) 2007, *Pathways and Crime Prevention: Theory, Policy and Practice*, Willan Publishing, Cullompton, Devon.

Glasby, J. and Beresford, P. 2006, 'Who knows best? Evidence-based practice and the service user contribution', *Critical Social Policy*, vol. 26, no. 1, pp. 268–84.

Haskins, R. 2006, Testimony on the Welfare Reform Law, 19 July 2006, Committee on Ways and Means, US House of Representatives, Washington.

Head, B.W. 1988, 'Intellectuals in Australian society', in Head, B.W. and Walter, J. (eds), *Intellectual Movements and Australian Society*, Oxford University Press, Melbourne.

—— 2008a, 'Assessing network-based collaborations: effectiveness for whom?' *Public Management Review*, vol. 10, no. 6, pp. 733–49.

—— 2008b, 'Three lenses of evidence-based policy', *Australian Journal of Public Administration*, vol. 67, no. 1, pp. 1–11.

—— 2008c, 'Wicked problems in public policy', *Public Policy*, vol. 3, no. 2, pp. 101–18.

Jones, A. and Seelig, T 2005, *Enhancing Research–Policy Linkages in Australian Housing,* final report 79, Australian Housing & Urban Research Institute, http://www.ahuri.edu.au/publications/projects/p20216/

Landry, R., Amara, N. and Lamari, M. 2001, 'Utilization of social science research knowledge in Canada', *Research Policy*, vol. 30, no. 2, pp. 333–49.

Lavis, J.N., Robertson, D., Woodside, J.M., McLeod, C.B. and Abeldon, J. 2003, 'How can research organisations more effectively transfer research knowledge to decision makers?', *Milbank Quarterly*, vol. 81, no. 2, pp. 221–48.

Leigh, A. 2009, 'What evidence should social policymakers use?', *Economic Roundup*, no. 1, pp. 27–43.

Lerner, D. and Lasswell, H.D. (eds) 1951, *The Policy Sciences*, Stanford University Press, Stanford.

Lewicki, R.J., Gray, B. and Elliott, M. (eds) 2003, *Making Sense of Intractable Environmental Conflicts*, Island Press, Washington.

Lin, V. and Gibson, B. (eds) 2003, *Evidence-based Health Policy: Problems and Possibilities,* Oxford University Press, Oxford.

Majone, G. 1989, *Evidence, Argument and Persuasion in the Policy Process*, Yale University Press, New Haven.

Marsh, I. and Stone, D. 2004, 'Australian think tanks', in Stone, D. and Denham, A. (eds), *Think Tank Traditions*, Manchester University Press, Manchester.

Mosteller, F. and Boruch, R. (eds) 2002, *Evidence Matters: Randomized Trials in Education Research*, Brookings Institution, Washington, DC.

Mulgan, G. 2005, 'The academic and the policy-maker', presentation to Public Policy Unit, Oxford University, 18 November.

Nathan, R.P. 1988, *Social Science in Government*, Basic Books, New York.

Nutley, S., Walter, I. and Davies, H.T. 2007, *Using Evidence: How Research Can Inform Public Services*, Policy Press, Bristol.

Parsons, W. 2002, 'From muddling through to muddling up — evidence based policy making and the modernisation of British government', *Public Policy and Administration*, vol. 17, no. 3, pp. 43–60.

—— 2004, 'Not just steering but weaving: relevant knowledge and the craft of building policy capacity and coherence', *Australian Journal of Public Administration*, vol. 63, no. 1, pp. 43–57.

Pawson, R. 2006, *Evidence-based Policy: A Realist Perspective*, Sage, London.

Petticrew, M. and Roberts, H. 2005, *Systematic Reviews in the Social Sciences*, Blackwell, Oxford.

Petticrew, M. 2007, 'Making high quality research relevant and accessible to policy makers and social care practitioners', presentation to Campbell Collaboration Colloquium, 16 May.

Roberts, H. 2005, 'What works?', *Social Policy Journal of New Zealand*, no. 24, pp. 34–54.

Rudd, K. (2008) *Prime Minister: Address to Heads of Agencies and Members of Senior Executive Service*, 30 April, http://www.pm.gov.au/node/5817 (accessed 5 January 2009).

Saunders, P. and Walter, J. (eds) 2005, *Ideas and Influence: Social Science and Public Policy in Australia*, UNSW Press, Sydney.

Schon, D.A. and Rein, M. 1994, *Frame Reflection: Toward the Resolution of Intractable Policy Controversies*, Basic Books, New York.

Schorr, L.B. 2003, 'Determining "what works" in social programs and social policies: towards a more inclusive knowledge base', Harvard University.

Shonkoff, J.P. 2000, 'Science, policy and practice: three cultures in search of a shared mission', *Child Development*, vol. 71, no. 1, pp. 181–7.

Thaler, R. and Sunstein, C. 2008, *Nudge: Improving Decisions about Health, Wealth and Happiness*, Yale University Press, New Haven.

UK Cabinet Office 1999a, *Modernising Government*, Cabinet Office, London.

—— 1999b, *Professional Policy Making for the Twenty First Century*, Cabinet Office, London.

—— 2003, *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*, Cabinet Office, London.

—— 2008, *Think Research: Using Research Evidence to Inform Service Development for Vulnerable Groups*, Social Exclusion Taskforce, Cabinet Office, London.

UK Treasury 2007, *Analysis for Policy: Evidence-based Policy in Practice*, Government Social Research Unit, Treasury, London.

Wilson, J.Q. 1981, 'Policy intellectuals and public policy', *The Public Interest*, no. 64, pp. 31–46.

Withers, G. 1981, 'University centres for policy research', *Vestes*, vol. 24, no. 2, pp. 3–8.

Woolcock, M. 2009, 'Toward a plurality of methods in project evaluation', *Journal of Development Effectiveness*, vol. 1, no. 1, pp. 1–14.

3 With a scope so wide: using evidence to innovate, improve, manage, budget

Ron Haskins
Brookings Institution

Abstract

Evidence from social science research and evaluation is used for at least two broad purposes in improving program effectiveness. One purpose is to influence the decisions of policy makers; a second purpose is to contribute to continuous program improvement by influencing program management and implementation. This paper examines the role of evidence in three recent episodes of policy making in the United States: the welfare reform legislation of 1988 and 1996; funding for after-school programs in 2003; and the ongoing debate over establishing a new federal home-visiting program. The examples demonstrate that, although evidence is often not a primary factor in policy debates, in some cases its role can be important if not decisive. In reviewing these examples, characteristics of evidence that make it most likely to be useful to policy makers are examined. The paper concludes with a brief review of several guidelines for using evidence to improve program management and implementation.

3.1 Two uses of evidence in policy formulation

This paper has a split personality. The dominant personality wants to examine the influence of high-quality evidence from program evaluations on the formation and enactment of social policy. Three interesting examples from policy making in the United States are examined, although the paper focuses on welfare reform legislation enacted in 1996 after a long and bitter debate in Congress that ended the entitlement to cash welfare and required work by all welfare recipients. The examples serve two purposes. First, they illustrate how evidence from social science research and evaluation can have an impact on policy — or not. Second, the examples suggest several generalisations about how good evidence can be developed, communicated, and used to influence policy.

But another, less confident, personality wants to examine whether evidence can be built into a broad management system that a government can use to bring continuing attention to what it is trying to do, how it is trying to do it, how to improve it, and how to pay for it. Confining attention to the United States, I am well aware of the many thoughtful schemes that have been previously implemented to improve government effectiveness and efficiency. These include 'program planning and budgeting', 'zero-based budgeting', the Government Performance and Results Act, and the Program Assessment and Rating Tool. In every case, there seems to be general agreement that results have fallen short of expectations (Radin 2000).

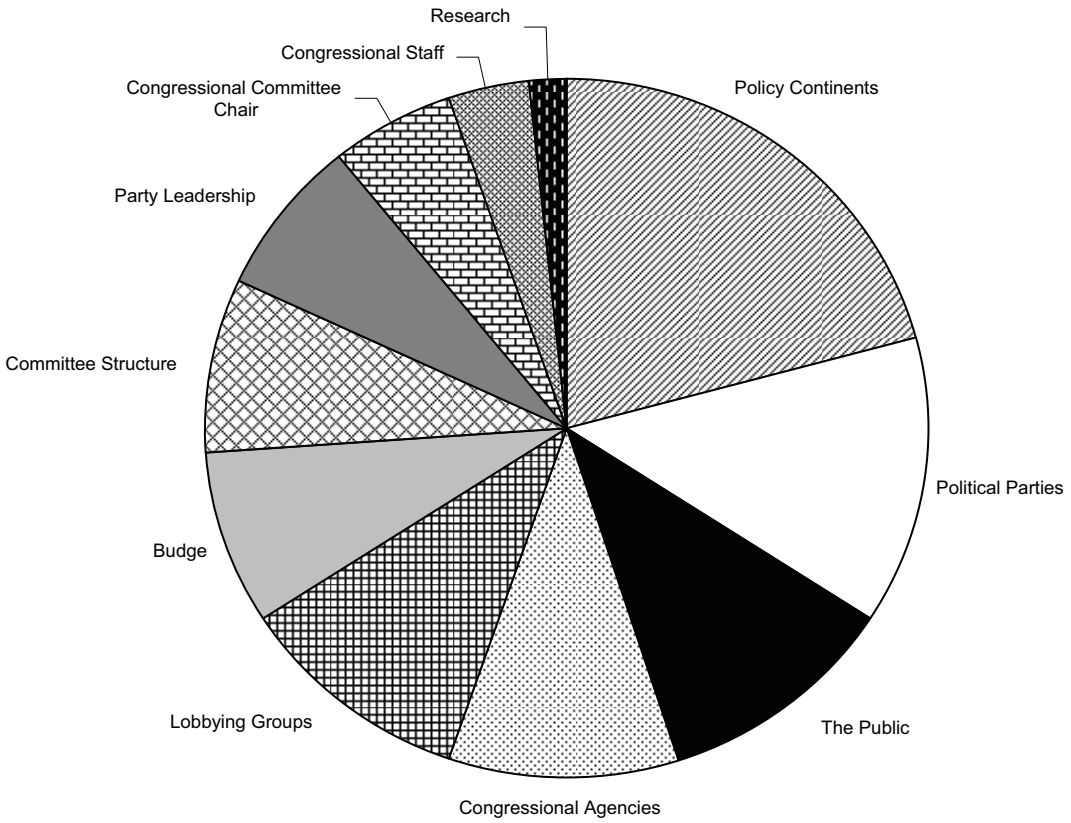
Throughout, I pursue a strategic question that reformers must face: Is it better to try to launch a government-wide initiative that attempts to bring order across a huge range of agencies and programs using a top-down initiative or to focus most attention on individual programs and how to improve them in an initiative that starts from the bottom and leaves open the question of when and whether the initiative should try to build towards something that involves many agencies and programs?

In the end, I argue that we should devote most of our attention to program evaluation and using evaluation results to enact, implement or improve individual programs. Although broader schemes seem reasonable, experience shows that putting them into practice has certain costs and uncertain benefits. Even so, experience and research also suggest that improvements may be possible and that evidence could become an important part of a broad movement to manage and budget wisely.

Applying evidence to policy choice

Consider the somewhat whimsical portrayal of the factors that influence policy formulation portrayed in figure 3.1. Although this paper focuses on research, the policy process often does not. Having participated in policy formulation and enactment directly from inside the US House of Representatives and the White House for a decade and a half, and having observed — and occasionally tried to influence — from outside for more decades than I would care to discuss, experience leads me to observe that research rarely drives policy. Of the factors that compete with research as prime influences on policy debates, at least three are typically of much greater importance than research: policy continents and inertia; the philosophy of political parties; and powerful politicians.

Figure 3.1 **Factors that influence policy choice**



Data source: Illustrative proportions only; Ron Haskins, The Brookings Institution.

Policy continents demand a word of explanation. All the industrial democracies have highly developed social policies for the elderly, children, the disabled, and the poor and destitute. Those policies are typically embodied in thousands of pages of legislation and regulation. In the United States, many of our most important social programs — including cash welfare, adoption and foster care, medical care, child support enforcement, welfare-to-work programs and programs for the disabled — are located in the Social Security Act (CWM 2005). Each line on every page of this wonderful statute is watched by hawk-like individuals and interest groups. Many of them are looking for an opportunity to change the statute; many are guarding the statute so no-one else can change it. Thus, the statutes, usually developed over many decades, own a kind of energy to ensure that they are not changed without a fight. If someone wants to change them, they had better be armed. Like continents, important statutes move slowly.

Perhaps an even more powerful influence on policy formulation is political philosophy. I am not well versed in the politics of other industrial democracies, but

in the United States the two political parties offer a great contrast in philosophies, which translates to a striking difference in agendas — and a continuing feast of arguments in the federal capital and state capitals all over the nation. There are exceptions, but in general Democrats are the party of big government and high taxes, especially on the rich — the protestation of President Bill Clinton notwithstanding (CNN 1996). At this moment, Democrats are trying to reform health care to achieve universal coverage and to feature a government insurance plan that would be available to all. Government already pays for half of the medical care in the United States, but if reform passes the share of health care paid for by government will expand even faster. Conservatives — and not just conservatives — are fearful that, if the reforms include a government insurance program open to all, Democrats will subsidise the plan with public funds and drive private insurance companies out of the market. Government would then be bigger, taxes would have to be raised, and government could have a health care monopoly after a decade or so. In effect, the long hoped-for introduction of competition in the American health market would be dead. All this is taking place in an atmosphere in which the federal government has already taken control of two of the nation's three largest car companies and major parts of the financial system, and is attempting to exert much greater control over the nation's energy infrastructure.

Democrats are now putting another one of their philosophical tendencies on display. Despite the fact that the upper 5 per cent of earners already pay 60 per cent of federal income taxes and 75 per cent of corporate taxes (CBO 2007), Democrats plan nonetheless to increase tax rates on people earning over $250 000. In addition, despite the fact that the bottom 40 per cent of households already pay negative federal income taxes (they get a cheque from government in the mail each year), Democrats hope to lower the taxes of those households further (or send them a bigger cheque) by creating a new tax credit for workers (Tax Policy Centre 2009).

In contrast to Democrats, Republicans want smaller government and lower taxes for everyone (some would add 'especially the rich'). The reputation of Republicans for smaller government took a major hit under President Bush when they enacted a huge and unfunded expansion of the Medicare program by providing the elderly with a drug benefit, but their reputation for tax cuts was dramatically reinforced. Bush cut taxes more than any other president, even when it became obvious that the federal government would once again run very large deficits. Republicans even went so far as to produce one-sided budget rules that would apply strict limits to spending but not to tax cuts. One could conclude that Republicans were willing to sacrifice their reputation for fiscal responsibility in order to cut taxes.

The point is that, in the face of these fundamental goals of the two major parties, evidence from social science research has little to no chance of shifting decisions.

Budget projections, for example, are a kind of social science, although the artistic side of projections must be admitted — especially, as Mark Twain observed, when projections involve the future. Yet the long-term projections of fiscal doom for the federal budget have been largely ignored by elected officials from both parties (Antos et al. 2008). Even the Obama administration, headed by a president who has pledged to cut the deficit in half and a budget director who has been a fiscal hawk his entire career, has increased the federal deficit to the previously unheard of level of over $1.17 trillion for 2010 (OMB 2009, table S-2, p. 114) (partly justified by Keynesian spending to rescue the nation from the worst recession since the Great Depression) and at least $1 trillion per year thereafter (with no justification known to man or God) (Auerbach and Gale 2009). Evidence that these deficits are out of all proportion to any reasonable definition of what government should be doing have no impact. Evidence that past deficits have been associated with inflation and high interest rates is ignored. Democrats want to expand programs and add new ones; Republicans want to cut taxes. Who cares that our children and grandchildren must pay the bill? Nothing could provide stronger evidence of the extent to which our political parties are driven by their philosophies than their joint willingness to bankrupt the nation to achieve their philosophical ends.

In addition to the power of the statutory continents and the political philosophies and historical goals of our major parties, another force that is more powerful than evidence is the personal views and political strength of individual politicians. Most people familiar with the American system of government have an appreciation for the power of the president, but congressional party leaders and committee chairs, many of whom are barely known to the public, are also immensely powerful — sometimes so powerful they can modify or even defy the agenda of the president. Committee chairs have two powers that are especially useful for exercising their legislative muscle. First, they can call a public hearing of the committee and then invite witnesses that will present the views and policies favoured by the chair. In this way, a savvy chair can shape the political debate to tilt toward her favoured outcome. Second, when chairs decide to write a bill, they can author the first draft of the bill themselves. If they craft their bill properly so that it reflects (or at least does not flagrantly violate) the public will and the views of members of their own party and if they are clever in building support for their legislative goals, the bill can often survive its long legislative journey with many or even most of its major provisions more or less intact. Thus, if the chair wants a particular provision because his constituents favour it or because his financial backers want it or because his political philosophy demands it, the state of evidence for or against the provision is of modest, if any, concern.

I note in passing that a president, chairman, or other powerful political figure who thinks evidence should be a major consideration in program enactment and funding

could greatly increase the role of evidence in policy formulation. Indeed, as we will see, President Obama and his budget director, Peter Orszag, appear to illustrate this point in spectacular fashion.

My goal in this opening section is to lower expectations about the potential role of evidence in the policy process because I think it should be clearly understood that evidence will only rarely be a dominant force in a debate over policy formulation. As suggested by figure 3.1, there are simply too many powerful forces that operate above and beyond evidence. When I left the scholarly world to work in the US Congress, I thought it was important for the social science community to try to convince elected officials that their decisions should be shaped by evidence whenever possible. But it did not take long to realise that even the members of Congress most disposed to pay attention to social science evidence simply regarded it as one small room in the mansion that is political debate and decision making. When occasions arose on which I had the opportunity to advise members about votes on social programs, my approach was to make them aware of whether there was evidence for or against a particular program and some idea of the quality of the evidence. Over the years, I dropped the idea that, if only those rascal politicians had their heads screwed on straight, they would listen to evidence and even seek it out. I regarded my role as being one of slightly expanding the purview of members to include at least some attention to evidence.

Arguably, the best way to bring evidence to the attention of members of legislative bodies is to have experts in organisations that elected officials trust report to them on a regular basis (table 3.1). In the US capital and in nearly all the state capitals, there is at least one organisation that helps legislative bodies keep track of their budgets and the cost and budget impacts of specific pieces of legislation. Either that budget organisation or other agencies also provide advice to members and committees about whether programs are working and whether new legislation is consistent with evidence about previous programs.[1] The US Congress has three agencies of this type that provide reliable and nonpartisan advice: the Congressional Budget Office, the Congressional Research Service and the Government Accountability Office (formerly the General Accounting Office; see table 3.1). These organisations specialise in bringing evidence to bear on the policy process. Many of their senior staffers have advanced degrees and know a great deal about the programs in their purview, as well as about legislative procedure. All three

[1] A good example of a state organisation of this type is the Washington State Institute for Public Policy. Created by the Washington state legislature in 1983, the institute 'carries out practical, non-partisan research — at legislative direction — on issues of importance to Washington State'. Issues studied by the institute include education, criminal justice, welfare, children and adult services, health, and general government. See www.wsipp.wa.gove/default.asp.

organisations have established a tradition of being nonpartisan and of presenting neutral analyses of budget and program issues. Importantly, most members of Congress trust them. These organisations, in short, are in position to bring neutral and evidence-based positions to the attention of Congress. But, of course, it is well beyond their power to make members pay much attention to the evidence. Even so, democracies need these neutral and respected agencies in order to know the financial implications of pending legislation and to maximise the chance that evidence will find an important role in policy debates and decisions.

Table 3.1 **Congressional agencies that provide non-partisan analyses**

Agency	*Year founded*	*Mission*
Congressional Research Service (CRS)	1914	CRS supports an informed national legislature by developing creative approaches to policy analysis, anticipating legislative needs, and responding to specific requests from legislators in a timely manner. CRS provides analysis that is authoritative, confidential, objective and nonpartisan.
Government Accountability Office (GAO)	1921	GAO helps Congress improve the performance and ensure the accountability of the federal government by providing Congress with timely information that is objective, fact-based, nonpartisan, nonideological, fair and balanced.
Congressional Budget Office (CBO)	1974	CBO provides Congress with objective, nonpartisan and timely analyses to aid in economic and budgetary decisions on the wide array of programs covered by the federal budget and with any information and estimates required for the Congressional budget process.

Sources: See http://www.loc.gov/crsinfo/; http://www.gao.gov/about/index.html; http://www.cbo.gov/aboutcbo/factsheet.shtml.

As it happens, the United States is now in the midst of an episode that nicely illustrates the vital role played by one of these nonpartisan analysis agencies. President Obama and the Democrats are in the midst of a serious attempt to bring universal health care to the nation. Because of the nation's yawning deficit, the President and his budget director, Peter Orszag, have promised to pay for any health care expansions that are enacted. However, as almost always happens, when the bills began to work their way through Congress in June and July 2009, it became increasingly obvious that the Democratic majority — precisely like the Republican majority when it enacted the elderly drug benefit in 2002 — found it too difficult to include the tough measures that would provide the financing for their new health benefits. When the first bill emerged from committee, amidst great claims of success by Democrats and the President, Douglas Elmendorf, the Director of the

Congressional Budget Office (and himself a Democrat who had served in the Clinton administration), testified that the bill was not fully financed and would increase the nation's already spectacular deficit by nearly $240 billion over 10 years. For this piece of honest analysis, he was berated by leading Democrats. Harry Reid, the Democrats' Majority Leader in the Senate, said that Elmendorf should run for office if he wanted to play such a decisive role in legislative battles — never mind that making cost estimates is the single most important job of the CBO director. Reid was probably upset because he knew that Elmendorf and the CBO have much more credibility on budget issues than he and other elected officials do. Not surprisingly, the media accepted Elmendorf's estimate and told the nation that the Democrats were trying to pull a fast one (Montgomery and Murray 2009; Pear 2009). It is precisely for situations like this that 'Anonymous' invented the old aphorism about the king's new clothing.

Having sufficiently lowered expectations about the political power of evidence, I now argue that evidence can nonetheless play an important role in policy choice under some circumstances. More specifically, in what follows I describe three legislative struggles in which evidence from social science research played an important role in policy debate. The first example is from the seminal welfare reform legislation of 1996, in which several ways that evidence influences policy were on display. The second is from a much more concise episode, in which evidence formed the basis for a policy choice by the Bush administration and in which the evidence was quickly swept aside and cast into a deep and ignominious grave. The third is still ongoing and, like welfare reform, shows the power of evidence to play an important role in a major congressional debate.

The Welfare Reform Law of 1996

Much of the discussion in this section is based on *Work over Welfare* (Haskins 2006). I had the good fortune of being a Republican staffer on the Ways and Means Committee[2] in the US House of Representatives when Republicans won control of both houses of Congress in the elections of 1994. Thus, I had a ringside seat for the festivities that followed. A small group of Ways and Means Republicans had been working on welfare reform for three or four years before 1994, had formed a coalition that involved Republican leaders in the House and on other committees, and had introduced several bills. Of course, House Republicans had been in the minority in the House for four decades and, as usual, their bills were ignored. But

2 When referring to the Ways and Means Committee, scholarly and press reports often use the phrase 'the powerful Ways and Means Committee'. I was on the committee staff for four months before I figured out that the committee's official title was not 'Powerful Ways and Means Committee'.

once Republicans took over control of both houses of Congress in 1994, the importance of their bills and the ideas they represented took a quantum leap forward.

The opportunity for Republicans to fundamentally reform welfare was created by none other than the great moderate Democrat, Bill Clinton. In the presidential election of 1992, Clinton had led the nation to believe that he was going to 'end welfare as we know it' (Weaver 2000), but then failed to follow up on his promise during his first two years in office. Instead, he squandered much of his prestige and power on a health care reform bill that failed completely. He did send a thoughtful and sweeping welfare reform bill to Congress late in 1994, but the senior leadership of the Ways and Means Committee, joined by the House Democratic leadership, did not like the bill and simply ignored it — a clear example of the power of senior congressional leaders to shape the fate of legislation, even when it is sponsored by the President.

Unlike Clinton, after the 1994 congressional elections that led to such a surprising Republican victory, Ways and Means Republicans had a welfare reform bill ready to introduce. Even more to the point, Republicans in the House were united on almost all the major welfare reform issues, leaving no doubt that they would be able to pass a bill in the House. This they did in short order, although the bill received hardly any support from Democrats (194 House Democrats voted against the bill; 9 voted in favour). In the record time of less than three months from the opening of the congressional session, House Republicans sent their bill to the Senate. After an exciting and drawn-out drama typical of the Senate, a bill somewhat more moderate than the House bill was passed on a surprisingly strong bipartisan vote in the autumn of 1995 (only 11 Democrats voted against the bill). A House–Senate conference committee then worked to resolve the differences between the two bills, and a version of the bill was sent to President Clinton in December as part of a huge budget bill designed to reduce the deficit. This compromise welfare reform bill, closer to the House than the Senate bill, lost much of its bipartisan support in the Senate. In addition, the budget bill of which the welfare bill was only one part was strongly opposed by Democrats. Thus, the lack of Democratic support for the huge bill in Congress gave cover to President Clinton to veto it. After the President's veto in December 1995, Republicans extracted the welfare reform bill from the budget bill and sent welfare reform separately to the president. Clinton vetoed the second bill as well. After the president's second veto, congressional Republicans worked with governors on a bipartisan basis, made some important changes in their bill, and then passed the revised bill on strong bipartisan votes in both the House and the Senate (about half the Democrats in both the House and the Senate supported the bill). President Clinton then signed the bill into law on 22 August 1996 (Haskins 2006). The third time was the charm.

I know of no example in which evidence played such a decisive role in a major legislative battle. By far the most important role played by evidence was already obvious when Democrats, joined by a large majority of Republicans and President Reagan, enacted a welfare reform bill called the Family Support Act in 1988. The 1988 Act, like the more radical welfare reform legislation passed eight years later, was designed primarily to boost work by welfare applicants and participants. Even in the highly contentious debate of 1996, as well as in the calmer debate of 1988, almost everyone believed that programs designed to increase work would actually lead more welfare mothers into the labour force and save money. Why did members of Congress believe work programs could be so effective?

The answer is that a body of high-quality research had accumulated, both before and after the 1988 reforms, showing that programs that emphasise helping mothers on welfare find a job and communicate the clear expectation that mothers should work would increase employment and reduce welfare costs. These studies are nicely reviewed by Judy Gueron (2003), the former president of MDRC, a widely-respected research firm that conducted many of the pre-1988 studies.[3] I draw attention to four characteristics of this line of research because the studies are vital to understanding how research could have such influence and perhaps suggest ideas for creating future occasions on which research is as pivotal as it was during the welfare reform era in the United States.

The first characteristic, and perhaps the most remarkable, is that the research on welfare reform was of such high quality. The remarkable thing about the quality is that the field of random-assignment evaluations under real-world conditions was in its infancy at the time the studies began in the late 1970s. Before that time, there were no major random-assignment studies on welfare demonstrations to test whether work programs changed the behaviour of recipients.[4] But, starting with large-scale demonstrations to test various approaches to increasing work levels, a style of research that had immense influence on subsequent welfare evaluations began to take shape (Gueron and Pauly 1991). Over the next decade and a half or so, numerous studies were initiated that were large-scale; featured random-assignment; were conducted in cooperation with states and with welfare offices within states; were conducted by professional companies that developed great skill in both the design of research of this magnitude and in the human-relations and organisational capacity needed to conduct complex studies of this type; were often

3 For a thorough review of the welfare reform studies, see Besharov (2009b, in press). In the spirit of full disclosure, I am now a member of the MDRC board of directors.

4 The income maintenance experiments were an exception, but they did not test job search and similar welfare-to-work programs; see Munnell (1986).

funded by a coalition of government and foundations; and featured tests of reforms that could be generalised to other settings.

These studies often involved changes in programs that were not specifically allowed by federal welfare statues. What made such studies possible without violating the law was that in 1962 Congress had enacted an obscure statutory provision[5] that allowed the Secretary of the Department of Health and Human Services to waive provisions of welfare law in order to test innovative programs. Eventually, about 40 states conducted demonstration programs under this waiver provision, and most of those demonstrations experimented with ways to increase work by parents on welfare. The partnership between researchers and state officials who applied for the waivers and who often contributed to financing the demonstrations offered a durable model for additional studies of research on this scale. One especially desirable outcome of this approach is that the research results have a ready-made audience of senior government officials who not only support the research but have a great interest in learning from the demonstrations because they think the results can have wider application and help them achieve the political goals of elected officials.

Another important characteristic of this research is that the demonstrations gave states — and in some cases local government — the opportunity to develop the practical experience and skills needed to implement these big and unwieldy reforms. Imagine two scenarios. In the first, states are required by the federal government to do something they had not done before and with which many people disagree. The states often use the term 'unfunded mandate' to refer to federal legislation that requires them to do something that imposes costs on their budgets. Given the level of invective states direct at such federal mandates, one way to ruin a policy that states must implement is to impose a federal mandate on them. In the second scenario, states not only agree with the approach but actually have some experience with developing similar approaches and implementing them by their own bureaucracies. Welfare reform clearly fell into the second scenario. The demonstrations produced a bottom-up movement of people who believed in helping mothers work, had the skills necessary to communicate the work requirement to mothers, and were learning how to build programs that would successfully help mothers find work. Thus, not only did almost every state support the radical reforms under discussion in Washington in 1995 and 1996, but they were already in the process of developing their own programs that, perhaps with some adjustments, would meet the new federal requirements.

5 *Social Security Act 1962*, section 1115.

A third characteristic of the demonstration movement was that governors themselves began to take notice of their own state programs. Here is some shocking news: governors don't know everything that is going on in their states, even when the government they head is sponsoring the activity. Most of the demonstrations produced positive impacts on work and saved money — outcomes that governors were bound to like. That got their attention. Many governors, both Republicans and Democrats, began to laud their own programs and to think about and discuss how much more they could do if they had more flexibility from federal requirements. Thus, the governors themselves became an important constituency for welfare reform — and for implementing aggressive programs once the federal legislation passed in 1996.

Finally, it should not go unremarked that MDRC and the other research companies conducting the welfare-to-work experiments developed great skill in bringing attention to their findings. Not only did MDRC give snappy testimony and briefings to important congressional committees, members, and staffers in Washington, DC, and state capitals, but they also developed expertise at working with the media to bring attention to their results. MDRC and a few of the other research companies thereby fulfilled two important goals that are difficult for individual researchers to achieve: they kept up a steady stream of information about successful demonstrations to key players in Washington and state capitals, and they stimulated media stories about the growing successes of welfare reform.

By the time President Reagan and the federal Congress decided to reform welfare in 1988, a background assumption of both Democrats and Republicans was that work programs could help mothers get jobs and leave welfare. There was even a widespread belief that these programs could save money (Long 1988). So popular were these findings that both Democrats and Republicans framed their welfare reform arguments around the claim that their fundamental goal was to advance the work agenda. Never mind that the 1988 bill actually did little to require anyone to work (Haskins 1991). This same understanding of the effectiveness of job search programs also served as a background condition for the debate leading to the 1996 reforms that really did require mothers on welfare to work or suffer rather serious consequences. The 1996 debate produced lots of arguing, some of it bitter, but not about whether programs could successfully help mothers get jobs and thereby fulfil the major goal taxpayers had for welfare reform — requiring work and not permitting long stays on welfare that reinforce idleness. The single greatest achievement of research in the 1988 and 1996 welfare reform debates was to create this background belief, backed by solid data everyone accepted and trusted, that it was possible to increase the employment rates of poor mothers and save public dollars in the process.

I think that almost any reasonable person studying the events I have summarised or who participated in them would agree that research and demonstrations on welfare reform created an important predicate for both the 1988 and 1996 reforms. In a real sense, research was one of the most important factors that created the opportunity for the Aid to Families with Dependent Children (AFDC) program, the nation's major cash assistance program for destitute families since the New Deal of the mid-1930s, to be repealed and replaced by the Temporary Aid for Needy Families (TANF) program. Every elected president from Nixon to Clinton, except the first Bush, had proposed major welfare reform initiatives, but all the initiatives had foundered because they cost too much or failed to find a compromise between encouraging work and allowing continued welfare dependency (Haveman 1995). But the experimental evidence that welfare mothers could and did find work when helped or cajoled into doing so allowed Republicans to argue that they and their children would be better off leaving welfare and, eventually, allowed Democrats to believe that Republicans might be right.

Moreover, the work requirements in the Republican bill were based on the assumption, well supported by research, that if ways could be found to get more mothers to find work, the goal of helping mothers become self-sufficient would be advanced. As shown in table 3.2, the major characteristics of the TANF program designed by Republicans in 1996 were intended to encourage or, when necessary, force mothers into the labour force. Ending the entitlement (the legal right to a cash benefit) sent a strong signal that mothers had to work and also made it easier for states to establish tough work requirements without violating the mothers' legal rights.

Table 3.2 **Temporary Aid for Needy Families program requirements**

1	End Cash Entitlement
2	Block Grant Funding
3	Work Requirements
4	Sanctions
5	5–Year Time Limit

Source: Haskins (2006).

To make sure that states mounted programs that would require work and that mothers would conform to the work expectation, the law included numerical work standards that states had to meet. To ensure that states would meet the work standards and that individual mothers would comply with the work requirements, the law included financial sanctions on states that did not meet the percentage requirement (50 per cent of their caseload had to be in work programs when the

requirement was fully phased in) and mandatory financial sanctions on mothers who did not cooperate with the state work requirements. None of these specific provisions by which the work requirements were to be implemented had been carefully tested by research, but all were justified indirectly by research because they were designed to make sure states established programs that would implement the work requirement and that welfare recipients would have a legal duty to participate in the work programs.

While research established a common base of knowledge that programs could help mothers on welfare find work, another goal of research is to influence political debate so that data and research become a vital part of the discussion on important public issues. Most people who have reflected on this goal of research would not expect research to be dispositive in settling any particular dispute, but it should insert itself into the debate in such a way that all sides must contend with facts about the policy problem and potential solutions. Let us now turn to a consideration of whether research fulfilled this goal in the case of the 1996 reforms.

A useful approach is to begin by identifying the major arguments Democrats made against the Republican bill and then examine those claims in the light of evidence that was available in 1995 and 1996, when the debate was in full bloom. Democrats used four major arguments against the approach to mandatory work taken in the Republican bill. Perhaps the Democratic argument with the most potency was that welfare reform requiring work would harm children. The original purpose of the AFDC program was precisely to protect children from the ravages of destitution. If Democrats could convince the public that requiring single mothers to work rather than entitling them to benefits, as the AFDC program had done, would harm children, they would strike a major blow against the bill.

During the entire two years of congressional debate, however, I never heard a Democrat cite any study or other kind of evidence that welfare reform would in fact hurt children. There was a rather substantial child development literature on whether mothers' work harmed children, but there were no random-assignment studies and the correlational evidence was not consistent in showing harm to children, even preschool children (Ainslie 1984; Gruber et al. 1994). Further, little of this research had been done on poor mothers. The most potent 'evidence' among developmentalists was essentially a theoretical argument about disrupting the attachment between mother and child if they were separated too often when children were young (Bowlby 1969). But even developmental scientists were not in agreement that mothers' work actually disrupted the attachment bond, and the empirical evidence was modest to weak. In any case, it was rare to see this body of evidence marshalled against the Republican bill on editorial pages, and I do not think any member of Congress cited it during committee meetings or during floor

debates. Democrats simply asserted that children would be hurt by the Republican bill.[6]

This issue shows yet again that evidence is not necessarily a trump card in debates about important social issues. Under any reasonable set of rules for rational debate, assertions such as harm to children should be backed up by evidence. There was some evidence (albeit equivocal) available, but it was not used very often, if at all. Without evidence, Democrats had a hard time shaming Republicans into changing their strong work requirement — especially because millions of American mothers, as well as mothers in virtually all the industrial democracies, had been voting with their feet on this issue for three decades. If millions of mothers, including college-educated mothers and single mothers, were flocking to the workforce, how could Democrats successfully argue that welfare mothers were hurting their children by working when they would also be arguing by implication that millions of other mothers were hurting their children (and when many of these middle class working women were themselves members of Congress or the spouses of members of Congress)? The mass movement by women into the labour force — and even into elected office — had been underway for decades, making it all but impossible to use the harm to children claim as a reason to kill welfare reform.

A second major argument Democrats made against the bill was that there were not enough jobs available for all the mothers Republicans expected to leave welfare for work. Prominent labour economists, such as Rebecca Blank (1995) and Gary Burtless (1995), both Democrats, had written about this issue in some depth. Blank and Burtless agreed that the availability of jobs was not likely to be a problem, but that mothers would be forced by their low level of education to take low-wage jobs, and that their advancement to higher wages would be painfully slow. Blank analysed the growth in jobs generated by the American economy and compared the growth in numbers of people who had jobs during the 1960s, 1970s, and 1980s with the percentage increase in the population (table 3.3). In each decade, the growth in jobs had exceeded population growth. In other words, the fraction of adult Americans with jobs had been increasing for at least three decades. Thus, the best evidence indicated that the Democrats' claim that jobs would not be available was exaggerated. But again, evidence on job availability was not often cited and the evidence did not prevent many Democrats from simply asserting that there were not

6 A host of good studies, many based on random assignment, published after the 1996 welfare reform law had passed showed that preschool children were probably helped when their mothers went to work (because they went to decent child care centres where they learned more than they would have learned at home); children of elementary school age were neither helped nor hurt when their mothers went to work; and adolescents may have been modestly harmed, as indicated by slight increases in trouble at school when their mothers went to work. See Morris et al. (2005).

enough jobs, while perhaps adding an anecdote about how hard it was to find jobs in their state or district.

Table 3.3 **Changes in US population and employment, 1960 to 1990**

Year	*Percentage population increase*	*Percentage employment increase*	*Increase in employed workers (millions)*
1960–70	16.9	19.6	12.9
1970–80	22.4	26.2	20.6
1980–90	12.1	18.7	18.6

Source: *Economic Report of the President* (1994, table B-33).

The third argument Democrats made against the bill was that mothers on welfare should get more education before being thrown into the job market where they would earn only low wages. It was certainly obvious, as Blank and Burtless showed with persuasive evidence, that the overwhelming majority of welfare mothers would receive low wages. Even so, despite low wages, most welfare mothers who worked even half-time would be financially better off than they had been on welfare because of cash benefits from the Earned Income Tax Credit, health coverage, food stamps and child care subsidies from the nation's growing work support system (see below), which provided cash and in-kind subsidies to low-income working families, especially those with children (Coe et al. 1998; Haskins and Sawhill 2009, ch. 9).

But there was an even more fundamental problem with the argument that mothers on welfare should get education and training before being required to work. Despite around $240 billion spent on employment and training programs between 1968 and 1995 (Burke 2003), reviews of this literature showed that most of the programs produced modest impacts, if any. As Burtless (1995, p. 100) concluded, based on a careful review of these studies, 'Even though training can improve the earnings prospects of women who are dependent on AFDC, it will not cause enough of an improvement to remove many low wage single mothers from poverty.' Other reviewers were even less generous in their conclusion about the impacts of employment and training programs than Burtless (LaLonde 1995).

There were other problems with the plea by Democrats for more education and training. Demanding that welfare mothers be trained before they could be required to work overlooked the fact that the public had already paid many thousands of dollars providing a public education to welfare recipients. True, many mothers had dropped out of high school, but dropping out was a personal decision. Moreover, there were substantial public subsidies available for both education and training that mothers on welfare could get on their own. If mothers on welfare wanted education and training, there were many federal and state programs that could help them

(Crawford 1994). But most welfare mothers did not want training. Perhaps they would select education or training if the alternative was work at $7 or $8 an hour, but mothers on welfare were not beating down a path to get training on their own. Thus, the evidence that training resulted in better jobs was weak, lots of other Americans were working for low wages, and mothers on welfare were not clamouring for more education and training. What taxpayers and many policy makers wanted was for them to fulfil their duty to avoid welfare by finding work.

Even if the federal Congress had decided to spend additional billions on employment and training, it is doubtful that most welfare mothers would have gotten jobs that paid much higher wages. Nonetheless, even if they knew the research literature on the tepid results from employment and training programs, most Republicans were reluctant to cite this literature because it seemed to place them in the position of arguing against education. The typical Republican response when Democrats urged more education and training was simply to assert that welfare reform was about employment. Mothers should be required to work and gain some work experience.

The debate over work vs. education was fascinating because Republicans could accurately have made a research-based argument that Democrats were simply wrong to think that employment and training programs would provide much of a boost to the economic value of welfare mothers in the labour market. Although a few Republicans did make this argument, mostly Republicans preferred to simply assert that the goal of welfare reform was to increase work, not education. Indeed, Republicans often argued, education got in the way of work.[7]

A fourth argument made by Democrats was that, even if mothers found work, their wages would be so low that they could not survive. The problem with this argument was that it ignored the mathematics of the value of benefits available to low-income working mothers outside the welfare system compared with the very modest benefits available to mothers while they were on welfare. Consider the figures in table 3.4. The typical mother on welfare who did not work in 1995 on the eve of the welfare reform debate would receive $7774 in cash and food stamps and she and her children would be covered by Medicaid. By contrast, if the mother took a job for $7.25 an hour, she would have earnings of $15 000, a cash payment of $2842 from the Earned Income Tax Credit, and $1538 in food stamps, for a total income of $19,280 ($17 812 after deducting taxes). A little known fact before the welfare debate, often discussed publicly by Republicans during debate on the legislation,

[7] Even so, the final bill did allow some education to count as work. Ignoring lots of complications, states were allowed to fill roughly 30 per cent of their work requirement by placing mothers in education and training programs. Despite this provision of the bill, few states chose to put very many mothers in education and training programs.

was that since the mid-1980s or so the federal government had created or modified a series of programs that provided support to low-wage working parents. The Earned Income Tax Credit had been dramatically expanded, Medicaid had been modified to cover children who were not on welfare, food stamps were available to low-wage workers, and the 1996 welfare reform legislation approximately doubled the amount of money available for child care. Thus, mothers could greatly increase their income by taking low-wage jobs, even if they had $3000 or $4000 in work expenses.

Table 3.4 **Is low-wage work a good deal?**
Work vs. welfare, 1996

Income category	*Mother on welfare*	*Mother working ($7.25 an hour)*
	$	$
Earnings	0	15 000
Earned Income Tax Credit	0	2 842
Social Security Tax	0	1 148
State Income Tax		420
Federal Income Tax		0
Aid to Families with Dependent Children	5 052	0
Food Stamps	2 722	1 538
Medicaid	Yes	Partial[a]
Total (net of taxes)	7 774	17 812

[a] The mother had one year of Medicaid coverage, and her children were covered as long as she had low income.

Note: Even if the mother has $4000 or $5000 of work expenses, she would still be better off working, especially because states had extensive funds available to pay for child care.

Source: Committee on Ways and Means, (*1996 Green Book*, p. 399).

Research played a major role in setting the stage for welfare reform by convincing nearly everyone that good welfare-to-work programs could help mothers find employment and save federal and state welfare dollars. But the impact of evidence on the specific arguments used by those who opposed or supported the federal reform legislation in 1996 was modest at best. Most members did not use evidence, even when it was available, to buttress their case, let alone carefully consider research evidence in arriving at their position on welfare reform and its various features. Plato's republic of philosophers who make wise policy decisions for society has not yet arrived.

Using muscle to overcome a negative evaluation: the case of after-school programs

If welfare reform provides a mostly positive example of evidence playing an important role in creating agreement on the efficacy of welfare-to-work programs, the 2003 Mathematica evaluation of the 21st Century Community Learning Centers program provides a far less positive example of evidence influencing policy. Initiated by the federal government in 1998 and funded at $1 billion per year, the 21st Century program provides after-school care that includes the opportunity for homework and recreational activities for elementary school and middle school students in the afternoon after regular school hours. The typical 21st Century program is open five days a week for around three hours a day. Eighty or so children are enrolled in most of the programs, which are operated by certified teachers and teachers' aides. A typical schedule is that students arrive after school, receive a snack, participate in supervised homework, and then participate in organised activities. Some centres offer activities in martial arts, fitness, dance, art and music. The goals most frequently cited for the program are to provide a safe place for children in the afternoon and to help students improve their academic skills. Several reports based on a large-scale random-assignment evaluation by Mathematica showed that students in the program, as compared with controls, were more likely to be under adult supervision in the afternoon but did not improve their academic performance in school. Worse, students in the after-school group were more likely to get into trouble in school than control children, and teachers were less likely to report that program group students got along well with peers (Dynarski et al. 2002; James-Burdumy et al. 2005). Thus, the study found at least as many negative as positive impacts.

When the initial Mathematica study was released in 2003, the Bush administration's Office of Management and Budget (OMB) was preparing the President's budget for 2004. Under strong pressure to cut spending in order to reduce the projected 2004 deficit, OMB decided to use the study as justification to reduce spending on the 21st Century program by about 40 per cent, from $1 billion to $600 million. The President approved this proposal, and OMB put the cut in the President's budget. Here was a clear case in which a random-assignment evaluation found that a program produced poor outcomes and was therefore used as a justification to reduce funding for the program.

Under the budget process followed by the federal government, the President proposes a budget in February of each year, Congress passes its own budget resolution around April, and then committees write legislation to meet the numbers in the congressional budget resolution. If the President and Congress are from the same political party, which they were in this case, the President's budget and the

congressional budget are often close cousins. The 21st Century program, however, was popular among many members of Congress and enjoyed an effective group of child advocates arguing against the cut proposed by President Bush. In addition, as improbable as it might sound, the famous body builder and actor, Arnold Schwarzenegger, entered the action just as the congressional committees were making decisions about spending programs. Schwarzenegger was widely known for his advocacy of after-school programs and had led the fight on a state-wide referendum, enacted by California voters in 2002, which provided funding for a state after-school program (Rivera 2002, p. 6). Adding to the fascinating political situation, Schwarzenegger was in the final stages of deciding to run for governor of California as a Republican and was receiving encouragement from the Bush White House to run. Thus, the White House may have been somewhat conflicted about its recommendation to cut funding for the 21st Century program. In May 2003, the Senate Appropriations Committee held a hearing on its 2004 budget and, knowing full well what he would say, invited Schwarzenegger to testify on the administration's proposed cut in after-school programs. When an official representing the administration testified in favour of the cut by citing the findings from the Mathematica study, Schwarzenegger told the committee:

> It would be a mistake, let me repeat, a big mistake, to use that study as justification to reduce current funding levels for after-school programs. Instead of cutting back the funding for after-school programs, we should begin to work together to focus on finding ways to improve them. (O'Keefe 2003)

Schwarzenegger did not feel it necessary to refute the study with any data or cogent arguments. Rather, he simply said it would be a mistake to use the study to cut the program. This is analysis by assertion, but it killed the Bush administration's proposal to cut the program. The Republican Congress ignored both the Mathematica study and the Bush administration and went ahead with full funding for the program. The lesson here, yet again, is that good evidence does not speak for itself in the policy process and is only one — sometimes a rather puny — element in a policy debate. Perhaps a few more studies and a continuing string of claims that after-school programs are not working would eventually cause Congress to take notice. But there was no chance in 2003 that a single study, even a high-quality study like Mathematica's, would overcome the political forces supporting after-school programs. In this case, star power and political exigency trumped evidence — and this particular data-based justification for a cut in the 21st Century program has not been heard since.

Home visiting programs and child abuse

The United States has many programs that aim to reduce child maltreatment and improve parenting by having trained staff visit pregnant mothers or mothers of newborns in their homes and offer them instruction about healthy living, financial management, child rearing and similar issues. [8] There are several prominent home-visiting models, many with written curriculums, trained staff and elaborate financing arrangements. Individual programs vary with respect to children's and mothers' age, the risk status of families served, the range of services offered, and the intensity of the intervention as measured by the frequency and duration of the home visits. A recent review of these studies for the journal *The Future of Children* found that seven home-visiting programs have evidence from random-assignment studies showing that they produced at least one significant effect on parenting behaviour or child outcomes (Howard and Brooks-Gunn 2009).

Arguably, the most notable of these programs is the Nurse–Family Partnership program developed by David Olds of the University of Denver. First tested by a random-assignment design in rural New York on a sample of poor white teen mothers beginning in 1977, the program was later evaluated by random assignment in Memphis and Denver. In both replications, some characteristics of the original program, as well as the types of participating families, were varied. The original and both replications produced significant impacts on several maternal and child outcomes and have been reported in refereed journals. In 1996, Olds began expanding the program by working with state officials and others while trying to ensure fidelity to his program model. By 2008, Nurse–Family Partnership programs were being conducted in 25 states. Seldom has an intervention program been so carefully tested and expanded with such serious attention to getting new sites to maintain fidelity to the program model (Olds et al. 1998).

The success of the Olds program did not go unnoticed by senior officials in the Obama campaign and subsequently the Obama administration. President Obama's 2010 budget Blue Print, released in February 2009, included funds for 'Nurse Home Visitation' modelled on the Olds program (OMB 2009). With this proposal, the administration served notice that it intended to fund only programs based on the Olds model.

The president's intention to fund only the Olds program startled the worlds of early childhood education in general and home visiting in particular, because it meant that other nationally prominent programs, such as Parents as Teachers and Healthy Families America, would be left out. The concerns of these groups were not without

8 The summary used here follows Haskins et al. (2009, especially pp. 4–5).

merit. Some of them had, as we have seen, been subject to random-assignment evaluations. Furthermore, within the scholarly world, some believed that the Olds program required further evaluation: there were inconsistencies in the results from the three evaluations; the programs had not been subject to evaluation by researchers outside of Olds's team; and the program focused on a narrow group of mothers — notably, low-income first-time mothers who agreed, while pregnant, to participate in a two-year program.

With the emphasis on 'nurse home visiting' in Obama's budget Blue Print, the debate left the pristine confines of academic journals and conferences and leaped into the rough and tumble forum of federal policy making. In this venue, the home-visiting programs that felt slighted by the President's budget Blue Print initiated a lobbying campaign to broaden the President's proposal to include additional home-visiting programs. Many of the programs not singled out by the President were part of a long-established coalition of influential and effective Washington child advocacy groups that included the Center for Law and Social Policy, the Children's Defense Fund, the Child Welfare League of America and others. The general line taken by these programs and their advocates was that Obama's emphasis on home visiting was an important advance for children and families, but that his proposal to single out one program for support was ill-advised. All high-quality, evidence-based programs, they argued, should be eligible for funding. Not surprisingly, groups favouring the Olds program started lobbying, too. All this is standard fare for federal policy making.

Two entries in the debate are especially worthy of note. The Coalition for Evidence-Based Policy, an influential Washington lobby for high-quality program evaluation, declared its support for the President's decision to fund research-proven home-visitation programs such as the Nurse–Family Partnership. Run by Washington veteran Jon Baron, the coalition has assembled an advisory board that includes several noted scholars and others with an interest in applying high-quality evidence to policy choice, including a Nobel laureate. In April, the coalition issued a well-reasoned brief that emphasised its nonpartisan nature as an organisation focused on promoting the development of rigorous evidence. Indeed, Baron and his coalition have almost single-handedly succeeded in getting many pieces of federal legislation to designate funds for program evaluation. Citing an authoritative evidence review from *The Lancet*, a respected medical journal, that found the Olds program to have the 'best evidence for preventing child abuse and neglect', the Coalition for Evidence-Based Policy expressed unqualified support for funding of programs, such as the Nurse–Family Partnership, that meet the highest standards of evidence. A six-page attachment to the brief reviewed evidence from the three randomised controlled trials (RCTs) by which the Nurse–Family Partnership had shown its

strong impacts, while pointing to deficiencies in the RCTs by which five other home-visiting programs had been evaluated.

Perhaps spurred by the coalition's brickbat against the non-Olds programs, four highly respected scholars, including Deborah Daro of the University of Chicago, Kenneth Dodge of Duke, Heather Weiss of Harvard and Edward Zigler of Yale, sent a public letter to President Obama. Their soundly argued letter praised his proposal for investing in home-visitation programs, but criticised the focus on one program model. The impressive quartet argued that a single program model would leave out too many at-risk parents. They also cautioned against a sole reliance on evidence generated from RCTs, which do not provide guidance on how to scale up a model program to serve national needs. Finally, they expressed the view that although at-risk families merit the most intensive services, all families should have access to early child development programs. The world of social science, it appears, does not speak with one voice, and even the best evidence can lead to multiple—and sometimes directly opposing—conclusions.

By the time Congress approved its budget resolution in late April, the forces supporting the broader language appeared to be making headway, because the budget supported home-visiting programs that 'will produce sizable, sustained improvements in the health, well-being, or school readiness of children or their parents' and contained no mention of nurse visiting. Similarly, the Obama language on nurses was gone from the final administration budget released in early May.

The next and critical step was for congressional committees to begin writing the new program into law. The chairman of the Human Resources Subcommittee of the House Ways and Means Committee, Jim McDermott (a Democrat), was the first out of the box. He circulated draft legislation in early June 2009, and then held a hearing on his bill on 9 June. Like the budget resolution, the McDermott draft bill represents a compromise between the contending forces. Specifically, it would give priority funding to programs that 'adhere to clear evidence-based models of home visitation that have demonstrated significant positive effects on important program-determined child and parenting outcomes, such as reducing abuse and neglect and improving child health and development'. Preferred programs must also have 'well-trained and competent staff' and include training, technical assistance and evaluation.

Perhaps the most important sign of the central role being played by evidence in this debate is the 8 June blog posting by Peter Orszag, the director of the federal Office of Management and Budget and President Obama's closest adviser on budget policy. Orszag asserts that he and the President are placing evidence of program success from 'rigorous' evaluations (by which he appears to mean RCTs) at the

centre of decision making. He states emphatically that the Obama administration will evaluate as many programs as possible, cut off funding for those that are not working, and expand those that are. In the case of home-visiting programs, he endorses the two-tier approach of giving more money to the programs with the strongest evidence of success and some but less money to programs that have 'some supportive evidence but not as much'. Orszag also cites several examples of how the Administration is expanding funds for conducting rigorous program evaluations and then using the evidence to make funding decisions.

As this episode unfolds, there is a lot to like for those who want to see quality evidence play a more prominent role in policy choice. It must be counted as a victory for the forces that favour evidence-based policy that the federal policy process on home visiting hinges importantly on evidence, a clear sign that both the Administration and Congress are giving a prominent place to high-quality evidence on successful programs. It also augurs well for evidence that the McDermott bill requires continuing evaluation of programs that receive the bill's funding. Indeed, the bill sets aside $10 million in guaranteed funding, mostly for program evaluation.

Regardless of the outcome, social scientists have taken an important step towards the goal of getting policy makers to consider high-quality evidence when making program funding decisions. That is a signal achievement for the research community — and, in the long run, for the improvement of public programs for children and families. As we have seen, the policy process does not often show any special deference to arguments based on evidence, but the US Congress and executive branch are showing new appreciation for how information can be used to fund the best programs that are likely to produce the best benefits for taxpayer investments.

Even so, it cannot be concluded that in recent years social science evidence has suddenly become a dominant force in legislative debates, but its status does seem to be improving. If the Obama administration actually delivers on the promise by the President and his budget director to fund programs that have strong evidence of success and to end programs that fail to produce impacts, the importance of evidence in political decisions in Washington will take a major leap forward. But before fans of evidence-based decisions get too excited, I suggest that they follow one round of the annual congressional appropriations process and see how many decisions are based on any appeal to evidence.

Applying evidence to program management

I now arrive at the less assertive part of my paper's split personality. I am by no means an expert on management issues. Rather, I have dipped into the management literature from time to time over the years, usually when I was curious about how data could inform program management. It is also impossible to be in Washington, DC, and not notice the many efforts to improve government efficiency, especially the Government Performance and Results Act passed in 1993 and the Bush Administration's Program Assessment Rating Tool implemented in 2002. As befits a man with only a modest amount to say, I have made this section much shorter than the first. My intent is to discuss five straightforward ideas for creating and applying program performance data to management decisions in order to build a system of continuous program monitoring and improvement.

Methods other than random assignment

In all three examples discussed above, RCT evaluations conducted under field conditions played a prominent role. There is impressive — albeit not unanimous — agreement in the social science world that RCTs are the most reliable method of program evaluation. The Coalition for Evidence-Based Policy, referred to above, was established specifically to promote the use of rigorous evidence in assessing program effectiveness, especially through the use of RCTs.[9] Recently, the prestigious National Academies concluded that 'the highest level of confidence' in evidence of program effectiveness 'is provided by multiple, well-conducted randomized experimental trials'. Indeed, the National Academies went so far as to assert that when evidence from randomised trials is not available, 'evidence for efficacy or effectiveness cannot be considered definitive, even if based on the next strongest designs' (O'Connell et al. 2009, p. 371).

It is difficult not to agree that RCTs provide the strongest evidence of program effectiveness. However, it would be naive not to recognise that there are several problems with RCTs. The first is cost. It is not at all unusual for a multi-year, multi-site RCT evaluation on a single intervention to cost $3 million or more.[10] The US Government, sometimes in cooperation with the states, operates more than 1000 programs, including around 70 that are major social intervention programs (Brodsky

9 See the Coalition for Evidence-Based Policy's website (http://coalition4evidence.org/wordpress/?page_id=6/).

10 The cost of an RCT depends in large part on the type of data used to measure outcomes. If participants must be interviewed or tested, the costs increase sharply. If, on the other hand, administrative outcome data such as data from unemployment records, school records or government program records can be used, costs decline.

2008). A single evaluation of all 1000 programs could cost as much as $3 billion. Even if someone were to present a strong argument that the $3 billion would be well spent on evaluations, the current fiscal difficulties of the US federal government, which are sure to last for a decade or more, all but preclude finding $3 billion to conduct RCT evaluations of all 1000 federal programs. Even if only the 70 major social programs identified by the Congressional Research Service were evaluated, the price tag could be $210 million (Burke 2003). It follows that there should be some set of rules for picking the programs that promise to produce the highest return from RCT evaluations.

Another reason that an exclusive focus on RCTs would be unwise is that there are many other designs that have produced results that many social scientists think are worthy of attention (Besharov 2009a). Even if RCTs are the most reliable design, other designs and methods might be used more cheaply to identify promising initiatives that could then be subject to RCTs. Given the difficulty and expense of conducting RCTs, it might be argued that they should be saved for only the most promising intervention programs. How do we know the most promising programs? By using second-best designs and methods. At a minimum, these include regression discontinuity, propensity matching, difference-in-difference, fixed effects, instrumental variables and interrupted time series, all of which have been productively used in evaluating the effects of social programs in the United States and abroad (Besharov 2008; Smith 2009).

Although I remain a huge fan of RCTs, cost considerations plus the availability of worthy alternatives leads me to conclude that non-optimum methods will find a role in any broad system of evaluating large sets of social programs.

Test more than one program

Agencies can almost always identify several potential programs that could address a problem under their purview. Almost all preschool and K-12 programs, for example, have more than one curriculum designed to help children reach major goals such as proficient reading and math or cooperative social behaviour. Both RCTs and most of the other methods listed above can handle the simultaneous testing of more than one program. The great advantage of simultaneous testing is that more can be learned in a fixed period of time, and probably at a lower average price. Another advantage is that testing more than one intervention raises the odds of finding one that works. Neither politicians nor the public are happy to sit around waiting a decade or so while evaluators determine whether a particular program can produce reliable impacts. Concurrent testing of differing program models with the same goals should be pursued whenever possible.

Test current programs against alternative programs

One of the great problems with RCT designs is that they can leave program operators and managers without obvious next steps. In high-stakes evaluations, the most reliable information is simply whether the programs produce better results than are produced by no program. If there are no significant program-control differences, we are left knowing that the particular approach used by one intervention program does not work, but with little else. By contrast, if we compare a current program with several improved versions of the program or with an entirely new approach, the odds of learning something constructive are improved. In programs that are already operating at multiple sites, continuous monitoring of data on outcomes of interest will often reveal programs or program features seen to be producing superior results. Careful study of these outliers that produce impressive results can lead to the generation of hypotheses about effective program variations that can be tested against less successful programs. In the final stages of testing superior programs, after experience with second best but less expensive and time-consuming methods shows good outcomes, RCTs can be used to provide definitive evidence.

Know the program; study its implementation

Managers and their staff are often located in government buildings far away from the field in which their programs are implemented. In talking with bureaucrats about their programs, I often have the feeling that they know very little about the specific circumstances under which the programs are implemented, the people carrying out the implementation, or the children or adults the program is intended to help. If managers are to play a role in identifying ways to improve programs, they must do more than read evaluation reports. They must get out into the countryside and get up close and personal with their programs and the people implementing and participating in the programs. It is the combination of direct observation and evaluation reports that will help managers make good use of evaluation results and participate in searching for ways to alter the intervention and then test the new innovations to determine whether they produce better results.

Evaluation expert Richard Nathan has recently argued that program evaluations should devote more attention to program implementation and the institutions responsible for program implementation (Nathan 2009). In the case of national reforms, Nathan thinks researchers should visit and study multiple sites using whatever methods seem appropriate to find out exactly how field offices are implementing reforms. He cites as an example a remarkable study of welfare reform implementation following the 1996 legislation by his colleague Irene Lurie (2006).

Lurie and her team observed over 1000 interactions between caseworkers and clients in 12 local sites to determine how caseworkers conveyed the work requirement to families applying for welfare. Summarising a complex set of results, a central finding was that, unlike previous iterations of welfare reform that typically failed to have impacts on local welfare institutions, caseworkers observed by Lurie strongly enforced the work requirement of the 1996 law by telling applicants that they had to look for work as a condition of applying for welfare. Nathan's point is that trying to understand policy means trying to understand how the policy is actually implemented — if at all — at the local level. He argues further that knowledge provided by implementation studies of this type can be a useful complement to evaluation studies in allowing program managers to determine the effectiveness of programs under their purview and to continuously monitor and improve their programs.

3.2 Summary

The United States and other Western democracies are developing a worthy tradition of subjecting many social intervention programs to RCTs to determine their effectiveness. The prestigious National Academies in the United States has determined that only RCTs provide 'definitive' evidence of program effects. Not all researchers agree that RCTs are definitive, but there does appear to be all but universal agreement that social science has now developed effective methods of determining whether social intervention programs are having their intended impacts.

It is useful to think of using this evidence about program impacts in at least two ways. The first is to inform legislative decisions about whether to fund programs. This paper shows that there are interesting examples from the United States of the application of evidence from program evaluations to legislative decisions. All the examples show that many factors besides evidence inevitably play an important role in legislative debates, but in some (albeit not all) cases the role of evidence in the political debate can be of great importance in determining the outcome.

The second productive use of evaluation evidence is to provide information to program managers about whether their programs are being implemented effectively. In this case, since democratic governments tend to be organised by large administrative agencies with responsibility for many programs, the question arises of whether evaluation information can be built into a centralised system of continuous evaluation for the large number of social intervention programs under the control of specific agencies. While remaining somewhat sceptical of building large, bureaucratic mechanisms that encompass many programs, this paper offers

several generalisations about how managers can use evaluation information, both from RCTs and from other methods, to improve their programs.

Program evaluation, especially the RCT, is a reliable and valuable weapon in the ongoing effort by democratic governments to provide their citizens with effective social programs. There is no question that evaluations of single programs provide valid information about program impacts that can be effectively used to continuously improve programs. Whether evaluation information can be efficiently incorporated into broader, system-wide schemes for monitoring and improving many programs remains an open question.

References

Ainslie, R. (ed.) 1984, *The Child and the Day Care Setting*, Prager, New York.

Antos, J. et al. 2008, 'Taking back our fiscal future', *Brookings–Heritage Fiscal Seminar*, Washington, DC, April 2008, Brookings Institution and Heritage Foundation, Washington, DC.

Auerbach, A.J. and Gale, W.G. 2009, *The Economic Crisis and the Fiscal Crisis: 2009 and Beyond, An Update*, Brookings, Washington, DC.

Besharov, D.J. 2008, 'From the great society to continuous improvement government', *2009 APPAM Presidential Address*, Washington, DC, 7 November, College Park, Maryland, University of Maryland, Welfare Reform Academy.

—— 2009a, 'From the great society to continuous improvement government: shifting from "Does it work?" to "What would make it work better?"', *Journal of Policy Analysis and Management*, vol. 28, no. 2, pp. 199–220.

—— (ed.) 2009b, *Poverty and Welfare: Readings from Journal of Policy Analysis and Management*, Wiley, New York, in press.

Blank, R.M. 1995, 'Outlook for the US labor market and prospects for low-wage entry', in Nightingale, D.S. and Haveman, R.H. (eds), *The Work Alternative: Welfare Reform and the Realities of the Job Market*, Urban Institute, Washington, DC, pp. 33–69.

Bowlby, J. 1969, *Attachment and Loss,* vol. 1: Attachment, Basic Books, New York.

Brodsky, R. 2008, 'Commanding performance', *National Journal*, 19 April, pp. 65-7.

Burke, V. 2003, *Cash and Noncash Benefits for Persons with Limited Income: Eligibility Rules, Recipient and Expenditure Data, FY2000-FY2002*, CRS Report

to Congress, RL32233, Congressional Research Service, 25 November, Washington, DC.

Burtless, G. 1995, 'Employment prospects of welfare recipients', in Nightingale, D.S. and Haveman, R.H. (eds), *The Work Alternative: Welfare Reform and the Realities of the Job Market*, Urban Institute, Washington, DC, pp. 71–106.

CBO (Congressional Budget Office) 2007, *Historical Effective Federal Tax Rates: 1979–2005*, http://www.cbo.gov/ftpdocs/88xx/doc8885/12-11-HistoricalTax Rates.pdf, (accessed 4 September 2009).

CNN, 1996, 'The era of big government is over', http://www.cnn.com /US/9601/budget/01-27/clinton_radio/ (accessed 3 September 2009).

Coe, N.B., Acs, G., Lerman, R.I. and Watson, K. 1998, 'Does work pay? A summary of the work incentives under TANF', Series A, No. A-28, Urban Institute, Washington, DC, http://www.urban.org/uploadedPDF/anf28.pdf, (accessed 2 March 2010).

Crawford, C.C. 1994, *Multiple Employment Training Programs: Major Overhaul Is Needed*, T-HEHS-94-109, Government Accountability Office, Washington, DC.

CWM (Committee on Ways and Means) 2005, *Compilation of the Social Security Laws Including the Social Security Act, as Amended, and Related Enactments through January 1, 2005*, vol. I, WMCP 109–2, US Government Printing Office.

Dynarski, M. et al. 2003 *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program; First Year Findings*, PR02-82, Mathematica Policy Research, Princeton.

Gruber, C.P., Clarke-Stewart, K.A. and Fitzgerald, L.M. 1994, *Children at Home and in Day Care,* Erlbaum, Hillsdale, NJ.

Gueron, J.M. 2003, 'Presidential address: fostering research excellence and impacting policy and practice: the welfare reform story', *Journal of Policy Analysis and Management*, vol. 22, no. 2, pp. 163–74.

—— and Pauly, E. 1991, *From Welfare to Work*, Russell Sage, New York.

Haskins, R. 1991, 'Congress writes a law: research and welfare reform', *Journal of Policy Analysis and Management*, vol. 10, no. 4, pp. 616–32.

—— 2006, *Work over Welfare: The Inside Story of the 1996 Welfare Reform Law*, Brookings, Washington, DC.

—— and Sawhill, I. 2009, 'Supporting and encouraging work', *Creating an Opportunity Society,* Brookings, Washington, DC.

—— Paxson, C. and Brooks-Gunn, J. 2009, 'Social science rising: a tale of evidence shaping public policy', *The Future of Children*, Policy Brief, Princeton University – Brookings Institution, Princeton, New Jersey.

Haveman, R.H. 1995, 'The Clinton alternative to "Welfare as we know it": is it feasible?', in Nightingale, D.S, and Haveman, R.H. (eds), *The Work Alternative: Welfare Reform and the Realities of the Job Market*, Urban Institute, Washington, DC.

Howard, K.S. and Brooks-Gunn, J. 2009, 'The role of home visiting programs in preventing child abuse and neglect', *The Future of Children*, vol. 19, no. 2, Brookings – Princeton, Princeton, forthcoming.

James-Burdumy, S. et al. 2005, *When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program*, Final Report, PR05-12, Mathematica Policy Research, Princeton, NJ.

LaLonde, R.J. 1995, 'The promise of public sector-sponsored training programs, *Journal of Economic Perspectives*, vol. 9, no. 2, pp. 149–68.

Long, D.A. 1988, 'The budgetary implications of welfare reform: lessons from four state initiatives', *Journal of Policy Analysis and Management*, vol. 7, no. 2, pp. 289–99.

Lurie, I. 2006, *At the Front Lines of the Welfare System: A Perspective on the Declines in the Welfare Caseload*, Rockefeller Institute Press, Albany, New York.

Montgomery, L. and Murray, S. 2009, 'Lawmakers warned about health care costs', *The Washington Post Online*, 17 July, http://www.washingtonpost.com/wp-dyn/content/article/2009/07/16/AR2009071602242.html (accessed 29 July 2009).

Morris, P.A., Gennetian, L.A. and Duncan, G.J. 2005, 'Effects of welfare and employment policies on young children: new findings on policy experiments conducted in the early 1990s', *Social Policy Report*, vol. 21, no. 2, MDRC, New York.

Munnell, A.H. 1986, *Lessons from the Income Maintenance Experiments: An Overview,* Conference Series No. 30, Brookings, Washington, DC.

Nathan, R.P. 2009, 'Social science methods and government effectiveness', paper prepared for the National Association for Welfare Research and Statistics Annual Conference, Albany, New York, 12–15 July.

O'Connell, M.E., Bost, T. and Warner, K.E. (eds) 2009, *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities*, National Academies, Washington, DC.

O'Keefe, E. 2003, 'Terminator touts after-school programs, *ABC News Online*, 14 May, http://abcnews.go.com/US/Story?id=90638&page=1.

Olds, D. et al. 1998, 'Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial, *JAMA: The Journal of the American Medical Association,* vol. 280, no. 14, pp. 1238–44.

OMB (Office of Management and Budget) 2009, *A New Era of Responsibility: Renewing America's Promise*'.

Pear, R. 2009, 'Housing committee approves health care bill', *The New York Times Online*, 16 July, http://www.nytimes.com/2009/07/17/us/politics/17cbo.html (accessed 29 July 2009).

Radin, B.A. 2000, 'The Government Performance and Results Act and the tradition of federal management reform: square pegs in round holes?', *Journal of Public Administration Research and Theory*, vol. 10, no. 1, pp. 111–35.

Rivera, C. 2002, 'Schwarzenegger's star power drives prop. 49, *Los Angeles Times*, 22 October.

Smith, J. 2009, Putting the evidence in evidence-based policy, paper presented at Productivity Commission Roundtable on Strengthening Evidence-based Policy in the Australian Federation, Canberra, 17–18 August, 2009, Productivity Commission, Melbourne.

Tax Policy Center 2009, '"Making work pay" tax credit', Brookings – Urban Tax Policy Center, Washington, DC, http://www.taxpolicycenter.org/taxtopics/conference_makingworkpay.cfm (accessed 4 September 2009).

Weaver, R.K. 2000, *Ending Welfare as we Know it*, Brookings, Washington, DC.

4 Putting the evidence in evidence-based policy

Jeffrey Smith[1]
Department of Economics, University of Michigan

Arthur Sweetman[2]
School of Policy Studies, Queen's University

Abstract

Evidence-based policymaking presumes good evidence. This chapter considers what policymakers can do to enable and encourage the production of such evidence. The core of the chapter reviews the current state of knowledge on alternative ways of estimating the causal effects of programs and policies. While we highlight the value of social experiments, we also make clear that opportunities exist for increasing the quality of the evidence provided by non-experimental evaluations through improvements in policy design and implementation, in the collection of survey data and in administrative data systems. We lay out the role that policymakers can play in exploiting these opportunities. The chapter also considers programs that affect non-participants as well as participants, cost–benefit analysis, potential substitutes for econometric evaluation such as performance management and customer satisfaction, and institutional changes that could improve the quality of evaluation evidence.

4.1 Introduction

The success of evidence-based policymaking depends on the quality of the evidence that underlies it. In this chapter we consider how policymakers can improve the quality of the evidence they use in making policy decisions. We focus almost

1 econjeff@umich.edu.

2 sweetman@qsilver.queensu.ca.

exclusively on evidence regarding the effectiveness of programs at changing the outcomes of the individuals, firms or local governments they serve. Such evidence necessarily plays a key role in cost–benefit analyses designed to guide decisions about program initiation, expansion, contraction and termination. The different, but important, concerns addressed by audits and process evaluations lie outside our scope.

Our discussion emphasises that policymakers largely determine the quality of the evidence that they have available for making policy decisions. They exercise this control in a variety of ways, both direct and, more importantly, indirect. Good evidence depends on much more than just a demanding client, a topnotch evaluator and an adequate budget when commissioning an evaluation. It also depends on broader decisions about program design and implementation prior to evaluation, on the design and funding of general social science data sets, on the quality of administrative data systems, on peer review and on institutions that encourage the development of informed evaluation consumers within government. Policymakers must also avoid the temptation of thinking that performance management or customer satisfaction can substitute for rigorous impact evaluation.

The rapid pace of methodological development in program evaluation is the secondary theme of this chapter. Evaluations that looked good 15 years ago sometimes look mediocre now. Policymakers need to have some sense of the existence and nature of these developments. At the same time, we have endeavoured to keep the discussion accessible to relatively non-technical readers while providing numerous references to the literature for readers who want more depth.

We organise the chapter as follows:

- Section 4.2 considers parameters of interest in impact evaluation, building on the basic insight that programs and policies have effects that often differ substantially among those they serve, or over time or over space.
- Section 4.3 forms the main course of our intellectual meal. It reviews the basic experimental and non-experimental strategies for estimating the 'partial equilibrium' impacts of programs. We make the case for doing more experiments, but at the same time we highlight for each non-experimental methodology what policymakers can do to increase the quality of the evidence produced.
- Section 4.4 considers the difficult problem of accounting for spillovers and other 'general equilibrium' effects that arise when programs have effects on non-participants.

- Section 4.5 considers the notion of a 'hierarchy of evidence' that attempts to rank the different evaluation methodologies.
- Section 4.6 discusses ways to improve the practice of cost–benefit analysis while
- Section 4.7 critiques alternatives to impact evaluation that sometimes distract policymakers.
- Section 4.8 focuses on the role of data quality and what policymakers can do to improve it.
- Section 4.9 provides some suggestions for institutional improvements we think would increase evaluation quality.
- Section 4.10 concludes.

4.2 Parameters of interest

Policy discussions often casually refer to 'the effect' of a particular program as though it constitutes a universal constant. In fact, program effects often vary substantially along several important dimensions. The recent literature on evaluation has made remarkable progress both conceptually and empirically in clarifying the nature of heterogeneous program effects and tracing out the implications of heterogeneous effects for the design and execution of evaluations.

Before delving into the substantive issues raised by heterogeneous treatment effects, we need to lay down some conceptual and definitional foundations. We sometimes use 'units' as a generic term for participants to emphasise that programs may serve, say, firms or local governments, rather than individuals. We use the term 'treatment' as a generic term for programs and policies.

A treatment effect (sometimes called an 'impact' or just an 'effect') refers to the difference a treatment makes in the outcome of a unit. In this regard, we can think of each unit as having two outcomes: first, a treated outcome, realised in the (possibly counterfactual) world wherein the unit gets treated; and, second, an untreated outcome, realised in the (possibly counterfactual) world wherein the unit does not get treated. Some readers will recognise this as the so-called 'potential outcomes framework'. The treatment effect (sometimes called the 'causal effect') is the difference between these two potential outcomes. Put differently, the treatment effect consists of the value added to (or sometimes, subtracted from) the outcome as a result of treatment.

Most analyses focus on estimating averages of treatment effects across units. The average treatment effect on the treated (ATET), which provides an estimate of the

expected difference between the treated outcome and the untreated outcome *for those who receive the treatment*, receives the most attention in the literature. This parameter informs a cost–benefit analysis that addresses the question of whether to keep or scrap a program in its present form. In contrast, the average treatment effect (ATE) estimates the expected effect of a program on all eligible units, whether or not they actually participate. This parameter informs a cost-benefit analysis that considers whether or not to make a program mandatory. For a program that is already mandatory, the ATET and the ATE coincide. For voluntary programs where potential participants have some idea of their own treatment effect, we expect positive selection on the treatment effect, so that ATET > ATE; put differently, we expect that in voluntary programs, participants will have higher impacts, on average, than all eligible units, while non-participants will have lower than average impacts. For voluntary programs, we might also be interested in impacts at the margin of participation; that is, impacts for those units for which a small change in costs or benefits would change the participation decision. The mean impact for these units (and not the ATET or ATE) should guide decisions about marginal expansions or contractions in the number of participants served. These impacts at the margin relate to what Imbens and Angrist (1994) named local average treatment effects (LATEs), which we discuss in Section 4.3.

In addition to varying by participation status, treatment effects may also vary among subgroups defined by observable characteristics, such as men and women, older or younger individuals, larger or smaller firms and so on. This variation may result from different patterns of selection across groups due to, for example, differences in the cost of participation. If two groups have the same distribution of treatment effects, but different costs of participation, the group with a lower cost of participation should have a higher participation rate but a lower average treatment effect conditional on participation. Variation across groups may also result from differences in the appropriateness of the treatment, what we might call the match between the treatment and the group. For example, a textbook-based job search course may have a larger treatment effect on more educated participants due to their presumably greater facility at absorbing written material.

In some evaluation contexts, such as educational interventions and active labour market programs, presentation of subgroup impacts has become fairly standard. Differences in impacts across groups that result from differences in the quality of the match between the treatment and the group may illuminate aspects of program operation not obvious from the overall impacts and so suggest where to focus efforts at program reform. Such differences in impacts can also guide efforts to use statistical treatment rules to target program services, as described by Lechner and Smith (2007) and Blattman (2008), or as in the survey by Smith and Staghøej (2010). Such rules formalise the assignment of treatments based on characteristics

that predict larger treatment effects. Where sample sizes allow it, policymakers should want to see, and evaluators should want to provide, subgroup impacts.

Average treatment effects may also vary in other ways. For example, the impacts of active labour market programs may vary over the business cycle, as described by Lechner and Wunsch (2009). They may remain stable over time after participation as with the US Job Training Partnership Act (US General Accounting Office 1996) or they may fade out over time as with the California Greater Avenues to Independence (GAIN) program (Hotz, Imbens and Klerman 2006). Impacts may vary by local or regional office due to local social or economic conditions or due to variation in the quality of program management or program staff. These types of variation should interest policymakers as well; they imply a concern at the evaluation design stage with ensuring the availability of adequate sample sizes for precise estimation of differential impacts over time or across offices or regions. Finally, treatment effects may vary across units even within subgroups, or time periods, or local offices; Heckman, Smith and Clements (1997), Bitler, Gelbach and Hoynes (2006) and Djebbari and Smith (2008), among others, address the related conceptual and econometric issues.

To summarise, the fact that treatment effects vary across units has important implications for evaluation design, execution and interpretation. Different mean treatment effects address different policy questions and often suggest different econometric evaluation strategies. Careful evaluation design should lead to harmony among these elements. In addition, designing in sufficient sample size to capture variation in treatment effects across key dimensions often adds great value to the findings from evaluations.

4.3 Partial equilibrium evaluation methods

This section lays out the standard econometric methods used to estimate the impact of interventions in a 'partial equilibrium' context. Partial equilibrium is economist-speak for operating under the assumption that the program only affects participants, and so does not affect non-participants via spillovers such as displacement in the job market or changes in market prices. Section 4.5 considers such spillover effects.

Social experiments

Random assignment and the selection bias problem

To see the problem that random assignment solves, think about estimating the ATET for some voluntary program. The ATET consists of the difference between the observed mean outcome of program participants and the counterfactual (and thus unobserved) mean outcome that program participants would have experienced had they not participated. The first of these presents little trouble to the evaluator, as it requires only collection of outcome data on a random sample of participants. The second of these presents the evaluator with a much more difficult problem: how to estimate what would have happened to participants in the imaginary world in which they did not actually participate. We cannot simply draw a random sample of eligible non-participants and estimate their mean outcome because we worry (quite rightly and with much evidence to back up our concerns) that individuals select non-randomly into programs. As a result, a comparison of the outcomes of participants with the outcomes of eligible non-participants will conflate the effects of the program (if any) with other differences between participants and non-participants that would have shown up in outcomes even if the program did not exist and the participants experienced their non-participation outcomes. For example, the participants might have higher levels of education, ability or motivation, or be better looking, or just have fewer other things, like young children, to keep them busy and so away from the program. The literature calls bias that results from non-random program participation 'selection bias'.

Randomisation solves the selection bias problem by taking a group of would-be program participants and randomly forcing some of them to realise their untreated outcome by excluding them from the treatment. In samples of reasonable size, the units randomly assigned to the treatment group and allowed to receive treatment will have the same pre-program characteristics, both observed and unobserved, as the randomly assigned control group that gets excluded from treatment. As a result, the mean difference in outcomes between the experimental treatment and control groups provides an unbiased estimate of the ATET. Random assignment makes the two groups statistically equivalent in all aspects other than access to treatment, with the result that only the difference in treatment can cause a difference in outcomes between them.

In the United States, experiments have been applied to policy areas as diverse as health insurance, welfare-to-work programs, the handling of calls to the police reporting domestic violence, electricity pricing, the negative income tax, and abstinence-only sex education. Greenberg and Schroder (2004) document all but the

most recent US experiments. The last few years have also witnessed an explosion in experiments in developing countries (see for example, Banerjee and Duflo 2009). As a result of all these experiments, a large body of knowledge regarding design and implementation exists as well as many organisations capable of pulling off high-quality experimental evaluations. Thus, policymakers in countries with little experience with experimental evaluation have little to fear and much to gain.

Issues with random assignment

Burt Barnow likes to say that random assignment is not a substitute for thinking; on this theme see the article by Barnow (2010) and also the humorous but pointed contribution by Smith and Pell (2003). Indeed, experiments present a more difficult evaluation challenge than their basic conceptual simplicity suggests. Experiments accomplish one very important thing: they solve the selection bias problem in partial equilibrium evaluations in a simple and compelling way. As noted by Heckman and Smith (2000), experiments remain subject to all the other issues that make empirical program evaluation so much fun, such as outliers, survey non-response and attrition, error-filled and poorly documented administrative data, and Hawthorne effects. Experimental evaluations may also have issues with external validity, particularly when relying on volunteer sites. 'External validity' refers to the extent to which program impact estimates obtained at a given time and in a particular place plausibly carry over to other times and places (while 'internal validity' refers to the applicability of the estimates to the time and place of the evaluation).

As discussed in detail in section 5 of Heckman, LaLonde and Smith's chapter (1999), experimental evaluations also face some issues typically not faced in non-experimental evaluations. Non-experimental evaluations compare treated units to untreated units using the various methods discussed under 'Selection on observed variables' and 'Regression discontinuity' in this section. In contrast, experiments often randomly assign *access* to treatment, rather than treatment itself, with the result that in many contexts, not all experimental treatment group members actually receive treatment and, less often, some control group members obtain the same or similar treatments from other sources. In the presence of treatment group dropout, the difference in outcomes between the treatment and control groups now estimates the mean impact of the offer of treatment (called in the literature the 'intention to treat') rather than the mean impact of treatment itself. Things get even more complicated with control group substitution into similar services from other sources. The articles by Heckman, Smith and Taber (1998) and Heckman, Hohmann, Smith and Khoo (2000) and the 'Instrumental variables' section below discuss these issues in greater depth.

An experimental evaluation may require a program to dig deeper into its eligible population than it normally would in order to fill up the control group while still maintaining its normal scale of operations. In such cases, external validity concerns arise because the participant population during the experiment differs from the usual participant population. Also, randomised rather than deterministic access to treatment may deter complementary investments prior to treatment or may change the composition of participants by deterring the risk-averse and attracting the risk-loving, again raising issues of external validity.

Though very real and of serious concern, thoughtful experimental design can often reduce the practical importance of these concerns; only occasionally do they become severe enough to outweigh the general case for random assignment.

Variants of random assignment

Random assignment has many uses beyond the estimation of the ATET for use in cost–benefit analyses of whether to keep or drop a program. Such uses address different questions that sometimes possess equal or greater policy relevance and often avoid or reduce political, practical and ethical (see below) concerns related to a no-treatment control group. Consider two illustrative real world examples.

Black, Smith, Berger and Noel (2003) document the clever use of randomisation in the Unemployment Insurance (UI) system in Kentucky. Like all other US states, Kentucky employs a statistical model to predict the fraction of the (usually) 26 weeks of UI benefit entitlement each new claimant will consume as a function of claimant and local area characteristics. The state then converts this predicted duration into a score between one and twenty, with twenty indicating benefit eligibility exhaustion and one indicating a very short predicted duration. In each local UI office in each week, the state assigns new UI claimants to receive (or not) mandatory reemployment services based on their score. Assignment starts with the highest score in a given office and a given week and proceeds until it runs out of slots or claimants. In many cases, for the marginal score (the one where the slots run out) the number of claimants with that score exceeds the number of remaining slots; these slots are randomly assigned. This scheme passed the scrutiny of sceptical state officials who were concerned that the alternative of randomly assigning all claimants, including those with long predicted durations on UI, would break the state budget.

The ‘randomisation at the margin’ approach used in Kentucky has many positive aspects, including low cost, no direct caseworker involvement, and staff perceptions of fairness. Moreover, it provides compelling experimental evidence that addresses the question of the effects of the mandatory reemployment services requirement on

claimants just at the margin of having it imposed. As the primary policy question in this area concerns small increases or decreases in the budget rather than program termination, this evidence corresponds to the cost–benefit analysis of greatest current policy interest.

McConnell, Decker and Perez-Johnson (2006) experimentally evaluate three alternative ways of structuring the 'Individual Training Accounts' (ITAs) provided to some participants in the US Workforce Investment Act (WIA) program. The three alternatives 'vary in whether counseling is mandatory, whether the counselor is asked to direct the participant in their training choice and can veto the participant's ultimate choice, and whether the value of each ITA is preset or determined by the counselor.' Everyone receives services but important aspects of the service delivery process differ among the three treatment arms. The policy question addressed in this evaluation concerns not keeping or scrapping the WIA program, nor expanding or contracting it, but rather how to operate the ITA component of the program most effectively. Other variants of random assignment include randomised rollout of programs too big to put in place in all locations at the same time, and randomised encouragement designs, as described by Hirano, Imbens, Rubin and Zhou (2000), that randomly assign not treatment but an incentive to participate in the treatment.

In short, given the tremendous variety of possible randomised designs, we can hardly overemphasise the potential to conduct persuasive yet inexpensive (and relatively uncontroversial) experimental evaluation.

Ethics, politics and experiments

Policymakers sometimes express ethical concerns with the random service denial inherent in random assignment designs with 'no treatment' or even 'less treatment' control groups. In our view, these ethical concerns often simply provide cover for policymakers who prefer not to have clear evidence on program effectiveness, perhaps because they think the program would not pass a benefit–cost test even though it succeeds in transferring public resources to favoured groups such as providers or clients.

While noting the potential for ethical misrepresentation, advocates of experiments can also address such concerns directly. First, evaluation efforts should focus on programs whose impacts and cost–benefit performance remain uncertain. In such cases, there is no way to tell in advance whether the control group is being randomly punished through denial of valuable services or randomly saved from having its time and effort wasted on an ineffective treatment. Second, the government can always compensate experimental participants for contributing to

the public good of knowledge creation. Unlike the case of some medical treatments, only modest payments should quell any ethical concerns for most social policies. Third, an alternative and perhaps weightier ethical concern militates in favour of random assignment where possible. Is it really ethical for policymakers to spend public money (implicitly taken by force from taxpayers) on programs without a compelling evidentiary basis, when they could easily bring about the production of such evidence?

Selection on observed variables

Consider the case where non-random selection into treatment occurs but the analyst observes all the variables with important effects on both participation and on the outcome of interest in the absence of participation. Economists call this case 'selection on observed variables' while statisticians call it 'unconfoundedness'.

Selection on observed variables represents a very strong assumption indeed! In our view, most evaluations that rely on this assumption fall far short of this standard, sometimes because of data limitations and sometimes, more broadly, because we simply lack the knowledge in many policy contexts of what variables to condition on. Successful application of this strategy requires careful thought about the institutions and the economics of the situation in order to make the case that all of the variables that both theory and existing empirical knowledge suggest should appear among the conditioning variables in fact do so. Making this case requires much more than just saying that the evaluation uses 'rich' data containing a large number of variables, though many evaluations offer up only this unconvincing justification. It is not the number of conditioning variables that matters, but rather having the ones that make the 'selection on observed variables' assumption plausible.

When relying on the selection on observed variables assumption, analysts typically employ either a parametric linear regression model or else some sort of weighting or matching estimator, such as inverse probability weighting or propensity score matching. In general, weighting and matching estimators represent the first choice for various technical reasons, provided the sample size justifies their use. See, for example, the methodological discussions by Heckman, Ichimura, Smith and Todd (1998), Angrist (1998), Smith and Todd (2005), Caliendo and Kopeinig (2008) and Busso, DiNardo and McCrary (2009a, 2009b).

Policymakers and evaluators can take many steps to make the evidence provided by evaluations based on the selection on observed variables assumption more compelling. The design of the program can include explicit guidance regarding the

factors that gatekeepers should use in making access decisions, which serves to clarify important matching variables. Process evaluations can provide further information about the factors influencing participation decisions. Collecting data on factors that often go unmeasured, such as the attitudes toward work, future orientation (that is, discount rate), risk aversion, motivation, social and other non-cognitive skills, and cognitive ability of potential program participants, could also make the selection on observed variables assumption more credible.

A larger literature suggests the value in many substantive contexts of flexibly conditioning on past outcomes measured at a relatively fine level of temporal detail. In the context of active labour market programs, see, for example, the articles by Card and Sullivan (1988), Heckman, Ichimura, Smith and Todd (1998) and Dolton and Smith (2010). Collecting such data, or obtaining it from administrative records, is often a more or less necessary condition for relying on methods that assume selection on observed variables. In addition, policymakers can require formal sensitivity analyses along the lines of those in the articles by Altonji, Elder and Taber (2005) and Ichino, Mealli and Nannicini (2008) that indicate the inferential consequences of departures from the selection on observed variables assumption. Finally, and perhaps most importantly, policymakers can fund basic social science research on the determinants of participation and outcomes that provide the foundation for choices about data collection and analysis and for arguments about the plausibility of the selection on observed variables assumption for particular combinations of treatment, data and outcomes.

Instrumental variables

Instrumental variables (IV) can sometimes provide consistent estimates in contexts where selection into a program occurs on variables unobserved by the analyst, rendering the methods described above under 'Selection on observed variables' inappropriate. An 'instrument' (nothing to do with marching bands) is a variable that affects participation in the program but is not correlated with outcomes other than through its affect on participation. The classical bivariate normal selection model for which Heckman (1979) developed a famous estimator represents a close cousin to IV; all of the same comments apply.

A good instrument has two properties. First, it strongly predicts treatment receipt, where the recent technical literature precisely defines how strong is strong enough. On this point, see the oft-cited paper by Bound, Jaeger and Baker (1995) and the literature it spawned. This property has the pleasant feature that it lends itself to easy testing using the available data. The better the instrument predicts participation, the more powerful (in the statistical sense) the analysis for a given

sample size or, put the other way, the stronger the instrument the smaller the sample required to obtain a given level of statistical power.

Second, a valid instrument affects outcomes only through its effects on the treatment, conditional on the included covariates. For example, intellectual ability does not represent a good instrument for schooling in an analysis of labour market outcomes, because while intellectual ability has a strong positive relationship with schooling attainment, and so possesses the first property of a good instrument, it also affects labour market outcomes directly, conditional on years of schooling. Put differently, even within groups with the same amount of schooling, intellectual ability will still predict labour market outcomes, and so it lacks the second property of a valid instrument. In contrast, random assignment yields an ideal instrument in the form of the indicator variable for belonging to the treatment group. By construction, this variable predicts treatment but has no relationship to outcomes other than through its effect on treatment. The search for instrumental variables in policy evaluation represents a search for variables that embody similarly random variation in program participation.

In general, there is no way to test the second property of a good instrument short of running an experiment. Instead, the analyst must make the case for the instrument using the relevant theory, along with information about the institutional context and prior knowledge regarding the determinants of treatment and outcomes. This process of argumentation renders instrumental variable estimates controversial in many contexts. Chapter 4 of Angrist and Pischke's book (2009) provides a good conceptual introduction to instrumental variables. Blundell, Dearden and Sianesi (2005) explicate and apply IV methods (as well as the bivariate normal selection model) in the context of a study of the effects of schooling on labour market outcomes. Heckman, Tobias and Vytlacil (2001) provide a broad conceptual framework for thinking about instruments.

Where do good instruments come from? Sometimes nature provides instruments, as when Kochar (1999) uses weather as an instrument for agricultural income or when the sex composition of the first two children serves as an instrument for the total number of children in the Angrist and Evans study (1998) of the effect of children on women's labour supply. Sometimes social events provide an instrument as in the Evans and Lein study (2005) that uses a bus strike in Philadelphia to study the impact of prenatal care on low income mothers. In other contexts, nature and institutions combine as in the paper by Evans and Kim (2006) that uses random variations in emergency room admissions on the weekend to study the impact of nurse-to-patient ratios on patient outcomes. And sometimes government itself provides an instrument, as with the variation in funding levels between jurisdictions that cut across the same local labour market employed in the study by Frölich and

Lechner (2010). In each of these cases, the researchers can make a good case that their instrument has both of the properties of a good instrument described above.

The literature on applied econometrics has spent the last decade or so coming to grips with the fact that analyses using instrumental variables generally estimate a somewhat unusual treatment effect parameter. In particular, under some (usually innocuous) assumptions they estimate the impact of the treatment on those whose treatment choice depends on the value of the instrument. Economists call this the 'local average treatment effect' (LATE) and statisticians call it the 'complier average causal effect' (CACE) where the compliers are those who change their treatment status when the instrument changes.

It helps to consider a couple of examples. In an experiment with treatment group dropout and control group substitution, the LATE is the impact on those who would receive treatment if assigned to the treatment group but not if assigned to the control group. In the context of the bus strike paper cited above, the analysis estimates the mean impact of prenatal care on those who would obtain prenatal care when there is not a bus strike, but who do not obtain it when there is a bus strike. Or consider the literature that uses variation over time or over jurisdictions in the compulsory schooling age to estimate the labour market impact of additional schooling, such as the article by Oreopoulos (2006). Changes in the compulsory schooling age induce variation in schooling levels only for a particular subset of the population. For example, increasing the age from 16 to 17 years in the North American institutional context will affect only those individuals contemplating dropping out prior to high school completion. The resulting treatment effect of additional schooling refers only to those individuals whose schooling changes as a result of the policy change, and not to individuals who would go to college or university regardless of the value of the compulsory schooling age.

It follows quickly from the insight that each instrumental variable estimates a LATE to the insight that different instrumental variables will estimate LATEs corresponding to different complier groups. Some instruments will estimate LATEs of great relevance to policy, while others will not. In general, no instrument will estimate the ATET parameter, which means that instrumental variables estimates typically cannot directly answer the 'keep it or cut it' question that underlies most cost–benefit analyses. On the other hand, an instrument that varies, say, the costs of program participation at the margin, may provide exactly the parameter of interest if the relevant policy dimension consists of modest spending increases to reduce the costs of program access (or small cuts that would increase those costs). Recent papers by Deaton (2009), Heckman and Urzua (2009) and Imbens (2009) debate this and related issues.

The quality of the estimates obtained by applying IV methods depends on the quality of the instrument. A weak or invalid instrument may be worse than no instrument. Good instruments can be obtained in one of three ways: clever data collection, exploitation or creation of useful institutional variation, and randomisation. Obtaining good instruments is facilitated by careful planning at the program design and implementation stage (to produce that useful institutional variation) and at the time of evaluation design. Collection of high-quality data aids in instrumental variables analyses as well, whether because the data contains potential instruments or because having better conditioning variables available makes it more plausible to assume that an instrument generated outside the data (that is, from institutional variation) satisfies the second property discussed above.

Longitudinal methods

Longitudinal methods use variation over time in treatment status to estimate the impact of treatment. The simplest longitudinal method consists of a comparison of outcomes before and after treatment. This before–after estimator can be applied to individuals, as when comparing outcomes before and after participation in a training program, or to a jurisdiction, as when comparing alcohol-related fatalities at the state level before and after a change in the minimum legal drinking age. The implicit assumption underlying before–after comparisons is that in the absence of the treatment or policy change, outcomes in the 'after' period would have been the same as (at least in expected value terms) the outcomes in the 'before' period. Sometimes this assumption makes sense and other times it does not. It fails when other factors affecting outcomes also change over time. For example, in the training program case, an individual might choose to participate in training following job loss. If the individual would have found a job reasonably quickly even without training, then a before–after comparison that includes the period of unemployment prior to the start of training produces an upwardly biased estimate of the effect of training on earnings. In the case of the minimum legal drinking age, a change in the fraction of the population between the ages of 18 and 22 or changes in related policies, such as the blood alcohol level used to define drink-driving, at around the same time might confound a causal interpretation of the before–after outcome difference.

Concerns about the plausibility of simple before–after comparisons have led many researchers to prefer the 'difference-in-differences' estimator. This estimator compares the before–after change in outcomes of the treated units to the before–after change in the outcomes of a sample of untreated units. This estimator is a special case of a more general class of panel data estimators that rely on within-unit variation over time to estimate the impacts of programs or policies, using untreated

units to control for common trends in outcomes. Both difference-in-differences and more general panel data studies rely on the assumption that, in the absence of the program or policy, the beforeafter change in outcomes for the treated units would equal (at least in expectation) that for the untreated units. Put differently, any differences between the treated units and the untreated units must remain constant between the before and after periods or, in the case of more general panel models, over the period covered by the data. Some parts of the literature refer to this situation (perhaps a bit misleadingly) as a 'natural experiment'; for further discussion see for example, Meyer's article (1995).

In certain contexts, the assumption of a common change in expected outcomes between treated and untreated units in the absence of treatment will make sense when an assumption of no change in outcomes in the absence of treatment for the treated units would not. At the same time, difference-in-differences is not a panacea. In cases where the treated units select into treatment based on transitory outcome changes, the difference-in-differences assumption fails. Thus, much of the intellectual action when considering evaluating a program or policy using these methods centres on learning about how the treated units came to be treated *when* they did. The analyst must also worry about anticipatory effects in the form of changes in behaviour prior to a treatment actually starting but as a direct result of its impending arrival, as when customers rush to buy prior to a sales tax increase.

Some examples of studies from the literature that use this method will help to clarify the picture, and to illustrate the many different types of comparison groups employed within this estimation framework. Heckman and Smith (1999) apply difference-in-differences in the context of a job training program. The comparison group consists of eligible non-participants in the sample local labour markets as the participants. Using an experimental benchmark, they find that that difference-in-differences performs poorly in this context, exhibiting both bias and strong sensitivity to the choice of particular before and after periods. This poor performance results from the fact that training program participants select (in part) into training based on transitory labour market shocks — typically job loss.

The famous minimum wage paper of Card and Krueger (1994) provides an example of difference-in-differences applied at the jurisdictional level. Their paper, as well as the companion paper by Neumark and Wascher (2000) that uses (arguably) better data and obtains a somewhat different answer, compares the changes in employment in a set of fast food restaurants in a local labour market that straddles the New Jersey and Pennsylvania border before and after an increase in the minimum wage that affects only New Jersey. The focus on a single labour market plays a key role in the plausibility of the estimates, though it also raises the possibility of spillover effects. Milligan and Stabile's evaluation (2007) of changes

to Canada's National Child Benefit using both differences-in-differences across provinces provides another example using jurisdictional policy variation.

In the United States, state level policies ranging from right-to-carry (a gun) laws to minimum legal drinking ages have had their effects estimated via panel data models applied to state level data on policies and outcomes. Many of these studies fail to do much to justify the application of these methods, which is to say that they do little to convince the reader that the timing of policy changes at the state level does not depend on transitory changes in the outcomes of interest. The United States has something of an advantage in the application of panel data methods to policy evaluation compared to countries with smaller numbers of jurisdictions because it has 50 states rather than six states as in Australia or 10 provinces as in Canada. This additional variation provides useful degrees of freedom and leads directly to a recommendation to the governments of countries like Australia and Canada to break up large states and provinces into smaller ones so as to facilitate policy experimentation and evaluation (!).

Discussions of the econometrics of longitudinal evaluation methods can be found throughout the literature. Moffitt (1991) provides an accessible introduction. Wooldridge (2002), Cameron and Trivedi (2005) and Angrist and Pischke (2009) provide textbook treatments. Bertrand, Duflo and Mullainathan (2004) highlight important issues regarding calculation of the standard errors when applying longitudinal methods. Heckman and Hotz (1989) highlight the use of additional periods of data to do tests of the assumptions underlying longitudinal evaluation methods. Heckman (1996) critiques the application of difference-in-differences methods.

Policymakers have almost complete control over the ability of researchers to apply longitudinal methods to the evaluation of treatments at the jurisdictional level. Many programs roll out over time rather than all at once to avoid administrative overload and to allow later implementing jurisdictions to learn from the early movers. Randomly assigning the order in which jurisdictions implement a program, as in the rollout of the PROGRESA conditional cash transfer program in Mexico, represents a gold standard. Absent random assignment, trying to avoid rolling out programs in a way that is correlated with the outcomes it is designed to affect is a second best. Regardless of what is done, carefully documenting the decision rule used to order the rollout and the actual timing of implementation on the ground at least gives the evaluator a fighting chance.

Beyond program implementation, the ongoing collection of large social science panel datasets that include information on the geographic location of respondents, along with detailed information on program participation and outcomes at a

relatively fine level of temporal detail, facilitates the application of longitudinal methods to the evaluation of both individual level treatments and jurisdiction level treatments. Panel data sets represent a complement to, rather than a substitute for, quality administrative outcome data at the jurisdictional level, as with data on alcohol-related traffic deaths or receipt of income assistance. A key in both cases is having data on outcomes that begins prior to the treatment under study.

Regression discontinuity

Regression discontinuity (RD) designs exploit discontinuous changes in treatment receipt that result from discontinuities in program rules. The RD estimator has the great virtue of conceptual simplicity. In situations where assignment to treatment depends on a continuous variable, such as a test score or proposal rating, and where the probability of treatment changes abruptly at a particular value of the continuous variable, a comparison of mean outcomes just above and just below the cut-off value can provide a compelling source of information about treatment effects. The literature calls the continuous variable that determines treatment assignment the 'running variable' and the particular point at which the probability of treatment changes abruptly the cut-off value or discontinuity (from which comes the name regression discontinuity). The econometric literature defines a number of different estimators for the RD case, but they all just represent different ways of taking averages of outcomes on the two sides of the discontinuity.

In thinking about exactly what treatment effect gets estimated in the context of a particular discontinuity, it helps to distinguish between what the literature calls 'sharp' and 'fuzzy' RD designs. In a sharp design, the probability of treatment moves from zero to one the discontinuity point. In this case, RD identifies the average treatment effect for units whose characteristics put them *at the discontinuity*. In a fuzzy design, the probability of treatment need not equal zero or one on either side of the cut-off but it must vary discontinuously at the cut-off. For example, a senior citizen discount on publicly provided flu shots could induce a discontinuity in the probability of receiving a flu shot at age 65. In the United States, the distribution of ages at which children start the first year of primary school corresponds to a fuzzy design at the nominal age cut-off due to selective choices by parents and administrators to advance or delay particular children relative to the norm. In the fuzzy case, under certain pesky but often plausible additional assumptions, one can estimate the LATE on those units who change their treatment status at the cut-off value. For example, in the case of the flu shots, a comparison of health outcomes on either side of the cut-off at age 65 would yield the mean impact of receiving a flu shot on individuals aged exactly 65 who would not get a shot without the discount. It does not provide information about the impact

of a shot on those who would get one with or without the discount, or who would not get one with or without the discount.

In both the sharp and fuzzy cases, generalisation of the estimated impacts to units with values of the running variable other than the value at the cut-off requires additional assumptions; the plausibility of such assumptions will depend on both prior knowledge, such as how the mean outcome varied with the running variable in periods prior to the implementation of the treatment, and on the institutional context.

A few examples will help to clarify the mechanics and the usefulness of RD. Perhaps the most well-known RD evaluation in the US context is that of the Reading First program commissioned by the Institute for Education Sciences of the US Department of Education and executed by Abt Associates and MDRC; see the final report by Jackson et al. (2007) for more information. This evaluation relied on the discontinuity created by the use of an index score to assign Reading First grants and found no real effect of Reading First on the outcomes of primary school children. Of course, these results apply only to schools near the discontinuity, a point missed in much of the discussion surrounding the evaluation, including the Dillon article (2008) in the *New York Times*.

Lee and McCrary (2009) examine the effect of punishment severity on criminal behaviour using the discontinuity in the US legal system between the punishment regime for juveniles (age less than 18 years) and adults (age 18 and above). Their setup has two main virtues: lots of data around the discontinuity and a very strong treatment due to large differences in severity between the juvenile and adult punishment regimes. Indeed, much to the surprise of pretty much everyone, they find only a very small change in criminal behaviour at age 18, indicating, at least for this age group, either a very present-oriented outlook or a very small response to anticipated punishment or both.

The foundational papers in economics are by Goldberger (1972a, 1972b, 2008), and consider a compensatory education program allocated according to a test score, with students scoring below a cut-off assigned to the program and those scoring above the cut-off not. Cook (2008) gives a broad history of RD in the social sciences. For methodological details on RD see the surveys by van der Klaauw (2008), Imbens and Lemieux (2008), and Lee and Lemieux (2009) and chapter 6 of Angrist and Pischke's book (2009). McCrary (2008) provides a useful test of the assumption of no manipulation of the running variable around the cut-off.

The opportunity to estimate impacts using RD methods depends almost entirely on program design decisions made by policymakers and program managers. Many of

the existing evaluations using RD methods rely on the 'luck' of having available institutions that happen to embody useful discontinuities. Policymakers and program operators should think prospectively about how to design programs to embody discontinuities that will yield useful impact estimates.

Successful use of a discontinuity design in program evaluation demands more than just a discontinuity in program eligibility rules or in the costs of program participation. The discontinuity must build on a variable that both the program and the evaluators can measure without much error and that potential participants or program staff cannot easily manipulate in order to change their status. For example, a generous subsidy to firms with 10 or fewer employees will induce some firms to change their number of employees from 11, 12 or 13 down to 10 in order to qualify for the subsidy. Such behaviour invalidates the regression discontinuity design, as the firms on one side of the margin (with 10 employees) no longer look like the firms on the other side of the margin (with 11 employees) due to the self-selection.

When relying on an age cut-off in a discontinuity analysis, the analyst must address the potential for spillovers, as in De Giorgi's study (2008) of the British New Deal for Young People (NDYP). His analysis relies (in part) on comparing the labour market outcomes of young unemployed people just above and just below the age cut-off for NDLP eligibility. To the extent that these young people represent close substitutes in the labour market, we would expect the existence of the program to have effects on both. Using either calendar time or age to define a discontinuity also raises the potential for anticipatory behaviour that has the potential to bias the estimated treatment effects.

Program designers need to locate the discontinuity at a point with many potential participants, so that sufficient data will exist to estimate a treatment effect with reasonable power; a sometimes difficult standard to reach given that sample sizes required for discontinuity designs typically exceed those for randomised trials with comparable power, as documented by Schochet (2008). Finally, the discontinuity in the policy variable must generate a corresponding discontinuity in treatment receipt. These criteria represent a tall order for program designers, and even when satisfied the evaluation still yields (as noted earlier) an estimated treatment effect only at the discontinuity. At the same time, a well-executed evaluation using regression discontinuity methods has nearly the same credibility as a well-executed experiment.

Other partial equilibrium evaluation methods

It is worth briefly commenting on some other partial equilibrium evaluation approaches. Process evaluations, such as the fine examples by Doolittle and Traeger (1990) and Kemple, Doolittle and Wallace (1993) from the US Job Training Partnership Act (JTPA) experiment, examine the flow of money and participants within programs. They have great value, but represent a complement to, rather than a substitute for, the sort of impact evaluation considered in this chapter. We have much the same view of comparative case studies, which can add richness to our understanding of outcome differences between programs or between sites within a program but cannot substitute for large sample econometric evaluations. Lab experiments have also started to play a small role in the evaluation literature (see, for example, Eckel, Johnson and Montmarquette 2005, Eckel et al. 2007, and Falk and Fehr 2003). In our view, lab experiments have the potential to play a small but useful role in evaluation going forward, though it will take some time for the lab experimenters to learn to think like evaluators and for evaluators to learn that lab experiments present more challenges than it might appear from outside. The hierarchical linear models, sometimes called multilevel models, widely used in education research (see, for example, Raudenbusch and Bryk 2001) represent not a separate method but rather a particular framework within which to apply the methods already described. This approach has the advantages of focusing attention on correct calculation of the standard errors for group (usually classroom or school) level treatments, of encouraging careful thought about causal relationships across institutional levels and of highlighting heterogeneity in the effects of treatments. Finally, structural methods (as economists use that term) rely on economic theory and related functional form assumptions to fill in for missing data. In the right hands the structural approach can add powerfully to the approaches already described. Todd and Wolpin (2005) provide an excellent example of the partial equilibrium structural approach.

4.4 Spillovers and general equilibrium effects

We now consider the effects of programs on persons or organisations or markets that do not directly participate in them. Such spillovers may accrue directly, as when a training program improves life for the family members of the trainee or an educational intervention reduces crime, or indirectly, via the operation of labour and product markets (or even changes in norms). Economists refer to indirect spillovers as general equilibrium effects. For example, a job placement program that helps one group of people find jobs may simultaneously make job finding more difficult for another group for whom they represent substitutes in production. Ethanol subsidies in the developed world may drive up the price of food in developing countries.

These external costs and benefits have proven, in general, quite difficult to pin down, but we argue that, contrary to the belief implicit in much of the literature, 'difficult to estimate' does not imply 'equals zero'.

Evaluations can often pick up direct spillovers via thoughtful data collection. For example, an educational intervention increasing the amount of classroom time devoted to mathematics in primary school should collect outcome data not only on math achievement but on achievement in the subjects whose classroom time gets reduced. Evaluations of labour market programs should collect data on criminal behaviour, as in the US National Job Corps Study, where reductions in crime represent an important component of program impacts, and on children, as in the Morris and Michalopoulos analysis (2003) of Canada's Self-Sufficiency Project.

Evaluators can sometimes obtain estimates of indirect spillovers by assigning treatment at the group level and then measuring outcomes for both participants and non-participants. For example, Dahlberg and Forslund (2005) use variation across municipalities to estimate the displacement effects of wage subsidies (large) and training (small) in Sweden. Many educational interventions affect some but not all students in a classroom; assigning the intervention to classrooms rather than students and then measuring outcomes of all students in both treated and untreated classrooms allows estimation of any spillovers. Finally, the clever village-level random assignment in the experimental evaluation of the PROGRESA conditional cash transfer program in Mexico, combined with the collection of data on both eligible and ineligible households in both treatment and control villages allows Angelucci and De Giorgi (2009) to provide a subtle analysis of within-village spillovers from the program.

In many cases, obtaining estimates of general equilibrium effects will require writing down and either estimating or calibrating a model of the relevant market. This approach represents a major investment of evaluator time and energy and requires a different skill set, more like that of modern macroeconomics, than that possessed by many in the evaluation business. This means it does not make sense to undertake such ventures for every evaluation of every program. Instead, general equilibrium evaluation analyses of this type should address important cases in terms of program size or program design and proceed on a somewhat separate track (that is, with more attention from academic economists and more funding from research funders rather than policy funders). Most evaluations should simply draw on this broader literature when discussing the possible nature and extent of such effects in a given context and when examining the sensitivity of cost–benefit calculations to likely general equilibrium effects.

Three examples highlight the power of this sort of analysis, along with its effort costs and heavy reliance on economic theory in general and specific functional form assumptions in particular. Davidson and Woodbury (1993) looked for displacement effects in one of the US Unemployment Insurance (UI) bonus experiments. In these experiments, the treatment consisted of the offer of a cash bonus to claimants who found a job early in their UI spells. They estimate that the displacement of workers not in the experiment cancelled out about 20 per cent of the employment impact of the program estimated in the experiment. In a study of tuition subsidy programs for university students, Heckman, Lochner and Taber (1998) find much larger general equilibrium effects. In their study, the partial equilibrium estimate of the impact of treatment on the treated is ten times larger than a general equilibrium impact that accounts for the decline in the relative wage of persons with a university degree resulting from their increased supply. Finally, Lise, Seitz and Smith (2010) examine the general equilibrium effects of Canada's Self-Sufficiency Project, an earnings subsidy for single parents on income assistance for at least a year who find a job during the second year of their benefit receipt spell. They find that taking account of the program's effects on the job search behaviour of other workers (and of the single parents themselves early in their spells of income assistance receipt) in the labour market leads to a reversal of the positive cost–benefit conclusions reached in the partial equilibrium experimental evaluation.

Policymakers play a limited but important role here. Discussion of possible direct and indirect spillovers should take place when designing an evaluation's basic identification strategy and when laying out plans for data collection. Policymakers should insist on such discussions at the start of each evaluation and should make sure that spillovers play a role in the interpretation of the impact estimates and in the related cost–benefit calculations at the end of the evaluation as well.

4.5 Comparing and ranking econometric evaluation methods

Leigh (2009) draws on a literature that proposes various 'hierarchies of evidence' and proposes his own hierarchy for Australia (see his Box 3). A generic version of such a hierarchy would have random assignment studies on top, followed by regression discontinuity designs, followed by instrumental variables or difference-in-differences designs, followed by studies relying on selection on observed variables, followed by before–after comparisons, expert opinion and, at the very bottom (what would our theorist colleagues say?), 'theoretical conjecture'.

We do not dispute that if one did a serious, impartial quality ranking according to well-defined and generally agreed-upon criteria that the average quality of

published evaluation studies using each method would likely correspond to this ordering. Nor do we dispute that this information has some value. Our concern lies in two not uncommon misinterpretations of such rankings. First, this ranking focuses on the 'between' variation rather than the 'within' variation, which leads some observers to forget the 'within' variation entirely. In fact, the relative importance of differences in the average quality of evaluations using the various different methods and variation in quality conditional on method is an empirical question, one well worth investigating and one for which we know of no available systematic quantitative evidence.

Second, the differences in mean quality across methods represent an equilibrium relationship; they need not be causal in the sense that, in a given context, moving up the hierarchy may make things worse rather than better. A given study, for example, may rely on cross-sectional data and an assumption of selection on observed variables because, in its context, no good instrumental variables suggest themselves and, looking at the time series of outcomes, there appears to be important selection into treatment based on transitory shocks. In this case, moving 'up' the hierarchy will likely lower the quality of the evaluation because it will mean using an invalid instrument or applying longitudinal methods when the assumptions that underlie them fail to hold in the data.

This second concern leads us directly to the misguided literature set in motion by LaLonde (1986). This literature seeks the holy grail of non-experimental evaluation: a non-experimental method that always and everywhere solves the selection problem. Dehejia and Wahba's works (1999, 2002) represent the most famous papers in this literature, which many (not necessarily including the authors) have interpreted as showing that matching 'works' in the sense of always solving the selection problem. Their work in turned spawned a large literature addressing the question of 'does matching work?' by comparing matching estimates to experimental estimates, sometimes using laughably weak sets of conditioning variables in the matching. In fact, the question 'does matching works' is ill posed. As noted in Section 4.3 under 'Selection on observed variables', matching works in the sense of providing consistent estimates when the available variables suffice to make the conditional independence assumption hold in a given context and not otherwise. Thus, we know the answer to the generic 'does matching works' question in advance; it is 'sometimes, but only when the data support it.'

Put in the context of our discussion of hierarchies, sometimes matching will outperform methods ranked above it in the hierarchy of evidence, as in a context where the analyst observes all the relevant conditioning variables but no instruments, and sometimes not. Rather than searching for a non-existent magic bullet estimator the literature should seek to build a body of knowledge on what

methods work for particular combinations of parameter of interest, available data, and program institutions. Rather than relying on a hierarchy to choose an identification strategy, researchers should seek to use the particular strategy best suited to providing a compelling impact in a given context given the nature of the program institutions and the data at hand. In our view, the main role of evidentiary hierarchies is to give policymakers an extra nudge in favour of experiments and to encourage them to push hard on evaluators who claim that a strategy low on the hierarchy represents the best choice in a given context.

Two final points on hierarchies: First, one should, of course, use all of the available high-quality evidence rather than just relying on one study. Meta-analysis represents a very useful tool for combining evidence, but it does not create any new evidence. Thus, it seems out of place in Leigh's (2009) hierarchy of evidence for Australia. Moreover, as poorly done meta-analyses often obscure the high-quality evidence by assigning all studies with qualities above some relatively low threshold equal weight, in some contexts the evidence from a meta-analysis may actually provide less guidance than would a handful of the best studies on their own. Second, as we discussed at length above, experiments do not solve every problem or answer every question. Putting them at the top of an evidentiary hierarchy makes it easier to forget that they too have quality variation and can sometimes, as in the presence of substantively important general equilibrium effects, provide quite misleading policy guidance.

4.6 Cost–benefit analyses

Cost–benefit analysis exposes the full range of costs and benefits associated with a policy or program by requiring their itemisation, justification and valuation. For reasons of time and space, we do not attempt a full consideration of cost–benefit analysis here; for that we refer the reader to the vast array of journals, textbooks and conferences on the subject: see, for example, Gramlich's *Guide to Cost Benefit Analysis* (1997) or the *Journal of Benefit-Cost Analysis*. Instead, we highlight a small number of important issues often ignored in the cost–benefit analyses associated with evaluations of active labour market programs and educational interventions. Our discussion draws in part on section 10 of Heckman, LaLonde and Smith's chapter (1999).

First, we want to reiterate the importance of doing a full-blown cost–benefit analysis, especially for large programs, expensive programs and politically important programs. We do not have in mind here the sort of 'cost effectiveness' analysis that compares one program to another or one service strategy to another without a no program or no service option; in our experience these usually arise in

contexts where politicians or program operators fear that the no-program or no service option will dominate the competition.

Second, we highlight the importance of considering multiple possible outcomes, including other outcomes for participating units and, as noted in the preceding section, spillovers to related units. For example, employment and training programs may have impacts on outcomes other than earnings and employment, such as participation in transfer programs, health, marital and family behaviours, and crime. Lechner and Wiehler (2010) find effects of German training programs on fertility. Both the original non-experimental Mathematica evaluation of the US Job Corps program, summarised by Long, Mallar, and Thornton (1981), and the more recent experimental one, summarised by Burghardt et al. (2001), stand out on this dimension, in particular for their important findings regarding the impacts of that program on participants' criminal activities. Some outcomes present real challenges to the analyst who must convert them to dollar terms, as with the primary school test scores in the Krueger cost–benefit analysis (2003) of the experimental class-size reduction in Tennessee. But, as noted in Section 3.4 in relation to general equilibrium effects, 'hard to measure' does not imply 'equals zero', despite what one might infer from reading the existing literature.

Third, a complete cost–benefit analysis should account for what economists call the deadweight costs of taxation or the marginal cost of public funds. These costs raise the social cost of one dollar of program budget to well over one dollar. They combine the direct costs of operating the tax collection system and the indirect costs imposed on society via the effects of (distortionary) taxes on behaviour. For example, income taxes lead workers to consume more leisure than they otherwise would, and so lower their utility relative to a world without income taxes. Resources spent in tax avoidance also figure into these costs. These costs will vary across countries depending on the mix of tax types (for example, income, consumption, value-added or excise) and tax rates (and perhaps local differences in behaviour conditional on these). The exact magnitude of these costs remains controversial in the scholarly literature, which suggests the wisdom of using two or three defensible values in a given cost–benefit analysis to give a sense of the sensitivity of the results to this parameter; see Dahlby's recent monograph (2008) on this topic.

Fourth, evaluations typically have available only a few years of follow-up data. For programs expected to have impacts in the medium and long term, this implies the need to project the impact estimates to time periods outside the data. In some cases, the cost–benefit performance of a program may depend critically on the persistence of impacts observed in the period covered by the available data in future periods. As such, the results of the cost–benefit analyses can be presented conditional on multiple assumptions about the persistence of any estimated program impacts.

Heckman, LaLonde and Smith (1999) provide an example of a cost–benefit analysis that does this. The assumptions about benefit persistence should build on findings on the persistence of impacts in similar programs drawn from the literature.

Fifth, most programs incur costs in the short term but reap their benefits, if any, in both the present and the future. Taking proper account of the timing of benefits requires the discounting of future benefits (and costs, if any) back to the present. Doing this, in turn, requires a well justified social discount rate, as described by Burgess (2010).

Sixth, we both often experience a sense of wonder when we learn in response to questions about the cost of particular public programs, as we often do, that no good data exist on this score. Even some quite modestly sized businesses know their average and marginal cost structures in great detail, as they recognise the critical role these costs play in making sensible management decisions. Public managers, in contrast, often know little beyond their total agency budget and a couple of major line items, such as labour costs, and can offer only a sad face when asked about the costs associated with the marginal or average participant, let alone the costs of particular service components. Serious benefit–cost analysis requires good data on costs, data that public agencies ought, in any event, to have handy both to guide their decisionmaking and to justify their activities to the taxpaying public.

Finally, as discussed in Section 3.5, a complete cost–benefit analysis should take account of general equilibrium effects when possible. This may require a separate evaluation component or it may rely on estimates from the literature for similar programs. Once again, a sensitivity analysis including alternative estimates of the general equilibrium effects drawn from the literature may be in order.

What can policymakers do in regard to cost–benefit analysis? To start, they can demand thorough cost–benefit analyses in those cases — new or unusual programs, expensive programs, big programs, and politically popular programs — where a cost–benefit analysis likely passes its own cost–benefit test. They can also make sure that the pieces necessary for a high-quality cost–benefit analysis get incorporated into the evaluation from the beginning, particularly in regard to the collection of data on outcomes (including longer term outcomes) and on treatment costs. They can also fund, most likely via traditional research grant programs, the work required to obtain good estimates of the marginal cost of public funds and the social discount rate.

More broadly, policymakers should face reality and accept Peter Rossi's (1987) famous 'iron law' of program evaluation, which states, 'The expected value of any net impact assessment of any large scale social program is zero.' Of course, calling

it a law overstates the case to make a point, as does the quip that the US Department of Education's 'What Works Clearinghouse' should really be called the 'Nothing Works Clearinghouse.' At the same time, examples of seemingly promising treatments associated with compelling estimates of no impact litter the programmatic ground. Experimental evaluations in the United States, for example, have found no impact of the quite expensive (well over $10 000 per participant in current dollars) National Supported Work Demonstration on men (Couch 1992); no impact of youth programs under the Job Training Partnership Act (Bloom et al. 1997; Orr et al. 1996); and no impact of politically popular abstinence-only sex education curricula relative to its traditional competitors (Trenholm et al. 2007). A similar fate befell the Bush Administration's highly touted Reading First program in Abt Associates' regression-discontinuity based evaluation (Gamse et al. 2008). Policymakers should treat these results as good news, as they allow them to free up resources for promising new programs (or even to return some resources to the long-suffering taxpayer).

4.7 Alternatives to econometric program evaluation

In this section we briefly address some of the leading alternatives to serious econometric program evaluation. At the bottom we put charlatans of the sort who fill in a 'sites of oppression matrix', as described by Gregory (2000). A few steps up the ladder reside the 'guns for hire' consulting firms that cater to the crowd that knows the answer it wants in advance, as with the sorts of ex ante evaluations of professional sports facilities that rely heavily on magic multipliers; see, for example, the critique by Crompton (1995) and the papers by Noll and Zimbalist (1997).

At the top of the heap sit performance management and customer satisfaction or participant self-evaluation. The performance management 'movement' got going with Osborne and Gaebler's book *Reinventing Government* (1992). Since that time, it has exploded within the public sectors of many developed countries, including in the United States under the Clinton Administration (recall the National Performance Review) and under the Bush II administration (with its Program Assessment Rating Tool, or PART). We have no objection to many aspects of performance evaluation, such as thinking seriously about program goals, collecting good data on program inputs, outputs and outcomes or simply riding people to get them to work harder and think harder about their jobs.

The trouble comes when performance management systems confuse outcomes with impacts. For example, in many countries, something analogous to what the United States calls the 'entered employment rate' constitutes a core performance measure

for active labour market programs. This rate consists of the fraction of the program's enrollees employed at some particular point (for example, 13 weeks) after leaving the program. In the terminology of the treatment effects literature, the performance measure consists of the mean of the treated outcome for the treated units. The performance measure omits any explicit counterfactual; put differently, it says nothing about what the employment rate would have been among participants had they not participated. This omission encourages, often with the help of misguided or mendacious program managers, the idea that the counterfactual equals zero, so that the outcomes summarised by such performance measures also represent impacts. Rather obviously, this interpretation encourages an overly positive view of program effectiveness. In addition, using performance measures that capture outcome levels means that high performance reflects both value-added and selection on untreated outcome levels. The literature frames this as outcome-based performance measures providing programs with an incentive to 'cream-skim' by differentially serving individuals who would have good labour market outcomes whether or not the program helps them.

In short, commonly used performance measures do not correspond to program impacts and so cannot substitute for econometric evaluation methods that do, in fact, estimate program impacts. They also provide an incentive for strategic behaviour on the part of the organisations that face them. For more on these points, as well as broader discussions of the plusses and minuses of performance management, we recommend the works of Heckman, Heinrich and Smith (2002), Barnow and Smith (2004), Radin (2006), Heinrich (2007), and Courty et al. (2010). For policymakers, the key lies in not asking performance management to do things it cannot do, like provide impact estimates.

Finally, we have what one might call participant self-evaluation. This can range from the sort of generic customer satisfaction questions used by many firms to questions that implicitly suggest some sort of counterfactual. For example, the New Chance evaluation in the United States had this question: 'Using the 0 to 10 scale, where zero is completely dissatisfied and 10 is completely satisfied, how satisfied were you overall with the New Chance program?'. The National JTPA Study had this one: 'Do you think that the training or other assistance that you got from the program helped you get a job or perform better on the job?'.

Sometimes the responses to such questions get highlighted in impact evaluations when the econometric estimates turn out badly, as if the fact that some large percentage of the customers say they liked the program makes up for a low mean impact on earnings. Recent research by Smith, Whalley and Wilcox (2010) indicates that program participants do not do a very good job of estimating impacts relative to a counterfactual, at least not with the sorts of questions presently used in

evaluations. Like them, we favour additional research with alternative question designs. For the present, though, we suggest not relying on participant self-evaluations as substitutes for econometric impact estimates. We do think that participant reports have an important role to play in aspects of process evaluations, such as rating the courtesy and helpfulness of program staff.

4.8 Data quality

The quality of the underlying data plays a crucial role in determining the quality of both experimental and non-experimental evaluations. Policymakers can exert real influence here, in a very non-political, 'good government' sort of way, to improve the quality of evaluation research.

Administrative data

Administrative data increasingly form the basis of econometric program evaluations. They have a number of advantages relative to survey data. Generally, they cover long time periods, allowing the use of longitudinal methods or allowing conditioning on rich histories of outcomes and program participation in evaluations that assume selection on observed variables. They also reduce the cost of looking at longer-term impacts. Administrative data also typically, though not always, do a better job of measuring the extent and nature of treatment received than surveys; see the related discussion by Smith and Whalley (2010). Administrative data usually allow access to a whole population, which means no issues of survey non-response and larger numbers of data points for analysis than with surveys, whose higher marginal costs typically limit data collection.

At the same time, as described by Hotz and Scholz (2002), administrative data have some weaknesses for evaluation purposes — weaknesses that relatively inexpensive administrative changes can in part ameliorate. For one, administrative data often lack key covariates required in 'selection on observed variables' evaluation strategies. In some cases, the field exists in the file but remains empty for many observations and contains measurement error when filled in. Changes in software to encourage reliable data entry, along with data audits and links with other, more reliable data sets can solve these problems, and greatly improve the value of administrative data not just for impact evaluation but for process evaluation and everyday management and monitoring tasks. The main ingredient required is prioritisation of administrative data quality by policymakers, including making the effort to design institutions that allow linkages across data sets (and reasonably quick access to data) for research and evaluation purposes, while maintaining

reasonable privacy protection. Policymakers can also encourage the collection of valuable data not already collected, such as the caseworker evaluations of the employability of unemployed workers routinely collected in Swiss administrative data.

Survey data

Survey data remains an important (if somewhat reduced) component of evaluation research. Surveys allow the collection of data not likely to show up in administrative data sets, including conditioning variables such as pre-program attitudes and outcome variables such as self-reported health or customer satisfaction. Survey data can also make up for problems with existing administrative data on variables such as schooling.

Policymakers can help maintain and improve the quality of survey data in three main ways. First, they can insist on response rates high enough to avoid criticisms about non-random non-response (and non-random attrition from longitudinal surveys) as well as over-reliance on statistical corrections for these problems. The US Office of Management and Budget requires response rates of 80 per cent. The leaders in the US survey industry, such as the National Opinion Research Center at Chicago and the Institute for Social Research at Michigan, routinely attain this level of response in their major research data sets. They do so despite the widespread view in the survey world that obtaining such rates has become more difficult due to lifestyle changes combined with 'survey fatigue' resulting from frequent use of surveys by commercial and advocacy groups. It just requires some money and, perhaps more importantly, the expertise. Policymakers can provide the first and hire the second.

Policymakers can also support methodological research on the potential value of new types of variables in program evaluation. This includes work on new types of conditioning variables such as the measures of risk aversion, time preference, financial knowledge, and 'ability' developed for use in the US Health and Retirement Study as well as on low cost biomarkers such as the hand traces that indicate testosterone levels based on finger length. It also includes research on how best to collect information on sensitive outcomes such as crime and sexual activity targeted by some programs. It makes sense to collect these variables first as part of large general social science data sets, then do research to determine their value, and then to add them to evaluation surveys.

4.9 Institutions

In this section we offer some ideas for relatively modest institutional changes that have the potential to improve the quality of evaluation work.

Public use data

Subject to privacy concerns, a well-documented public use data set should be one of the products ('deliverables') associated with every major evaluation. Public use (or, more accurately, researcher use) data allow independent verification of the official evaluation findings. Further, these data allow additional sensitivity analyses and the application of additional econometric methods beyond those in the official evaluation. They also encourage the production of valuable additional research, much of it of direct or indirect interest to government, at little or no cost. Academics from tenured professors to lowly graduate students will jump at the chance to work with good data when it becomes available. In addition, as any researcher doing empirical work knows, the possibility of future replication, and therefore of future public embarrassment if a mistake is found, provides a powerful motivator to thought and care.

Public use data sets exist for many major US evaluations. The Upjohn Institute for Employment Research maintains a set of evaluation datasets that it checks, documents and sells at cost to researchers. Similarly, MDRC, known for its role in many of the US welfare-to-work experiments, has a formal process to provide access to some of its evaluation datasets. The amount of knowledge about low income labour markets and about how to do econometric policy evaluation generated just from re-analysis of the data from the National Supported Work Demonstration and the National JTPA Study is simply huge, particularly when compared to the small cost of preparing the data sets for research use.

Peer review

Academia relies heavily on peer review in both the publication process and the hiring process. In our view, peer review also helps to increase the quality of program evaluations undertaken by governments. Most governments already do some of this but in many cases they could do more. We have in mind four specific avenues for increased peer review. The first consists of the use of outside experts as part of 'technical working groups' to oversee the development of an evaluation as it progresses from design, to implementation, to data collection, and finally to report writing. These should include both subject area experts and methods experts. The

second consists of the presentation of evaluation results at professional meetings and conferences, ideally prior to the completion of the final report, so that comments received can affect its substance. The third consists of publication of evaluation findings in peer-reviewed academic and policy journals. Independent review by academic journals subjects technical aspects of the methods and interpretations of the official evaluation to outside scrutiny. The fourth consists of incorporating written discussant comments as part of the final evaluation report. We have in mind here what was done in Westat's evaluation of the US Employment Service and also what the *Journal of the American Statistical Association* sometimes does with important and potentially controversial articles, such as that by Heckman and Hotz (1989).

Increasing peer review improves the quality of evaluation work directly, through the comments provided by the reviewers, as well as indirectly, as the anticipation of expert scrutiny focuses and increases evaluator effort. All of these forms of additional review, particularly publication in peer-reviewed journals, have the side benefit of increasing the number of scholars, policy analysts, program managers and policymakers who learn about the methods and findings of the evaluations. This in turn should lead to increases in both the quality and quantity of related policy discussions and thereby, one imagines, to improved future policy choices. Policymakers can foster the sorts of additional review suggested here via the simple expedient of including it in the statement of work (and the evaluation budget) when commissioning evaluations from outside, or demanding it from in-house evaluators.

Encouraging interaction among academics, government and consultants

We think that more interaction among academics doing evaluation work, government evaluators and evaluation consumers, and the consultants who often produce evaluations leads to better and more useful evaluations. It also helps build a knowledgeable constituency within government for serious evaluation. This sort of interaction can take many forms, but we have personally observed the value of professional meetings such as the annual research conference operated by the Association for Public Policy and Management (APPAM, publishers of the *Journal of Policy Analysis and Management*). Consultants, as well as in-house evaluators from government agencies, get a chance to show off their work to a broader audience as well as get useful feedback. Policy oriented academics get to present their work to an audience particularly knowledgeable about the policy process and the institutions. We have also witnessed the value of having academics spend time in the government, either in roles such as the chief economist at a particular agency or in more directly research-oriented roles, wherein an economist on sabbatical

might spend a year at an agency working with their data, getting writing done and interacting with the staff.

Institute of Education Sciences

Perhaps no single organisation in the United States has had a bigger effect on the quality of evaluation work in the last decade than the Institute of Education Sciences (IES). In terms of its direct effects, it has transformed the nature of federally funded evaluation of educational programs through its emphasis on funding high-quality evaluations using random assignment or regression discontinuity designs. It has brought together experts in economics, education, and other fields along with top evaluation consulting firms to conduct these evaluations. In the process it has generated valuable evidence on the effectiveness of programs such as alternative teacher certification, teacher mentoring and computer-aided mathematics instruction.

Perhaps even more important have been the indirect effects, operating through several channels. First, the IES has funded interdisciplinary training programs for education researchers at leading universities, with all of the programs having strong components in quantitative evaluation methodology and economics. Second, the IES has changed the way it runs its research grant programs to better emphasise serious quantitative research, particularly research using random assignment designs. Third, the IES has revamped the What Works Clearinghouse, a research collection, quality rating and synthesis institution modelled on the Cochrane Collaboration in medicine and the Campbell Collaboration in the social sciences, with the goal of raising the standards of empirical work in education. More details on the theory and practice behind IES can be found in the material by Rudalevige (2008) and US Institute of Education Sciences (2008). We heartily recommend it as an example both for other countries and for other policy areas.

4.10 Summary and conclusions

Evidence-based policy, and good government more generally, rest on a foundation of serious, hard-headed program evaluation. This paper has emphasised what policymakers can do to increase the quality of such program evaluation on a variety of different dimensions. The following points summarise our views and recommendations.

1. Be clear about the policy question of interest. Be sure that the econometric evaluation methods and data collection strategies adopted provide an answer to that question, even in a world where the impacts of programs vary across

persons and where both persons and program staff may make participation choices based on their informal estimates of individual impacts.

2. Use random assignment when possible. Frequent use of random assignment signals that a government is serious about evaluation and serious about basing policy on evidence. Infrequent use of random assignment sends the opposite signal. Keep in mind that randomisation can often aid in evaluation even without a no-treatment control group.
3. The success or failure of non-experimental evaluation methods depends critically on decisions about the design and implementation of the program, and on the quality of the administrative and/or survey data used in the evaluation. Thoughtful choices about program implementation and design can create useful variation in participation across time, space or persons that allows for credible evaluation. Slick econometric methods will not, other than by chance, overcome weak data or careless program design and implementation.
4. General equilibrium effects of programs matter. Analyses of these effects require different methods, in general, than analyses of the impacts of programs on their participants. Funding such analyses makes sense for large-scale programs. When a new analysis is impossible, the literature should guide an analysis of the sensitivity of the cost–benefit performance of the program to likely levels of general equilibrium effects.
5. Cost–benefit analysis represents the final step in program evaluation. Programs cost real money that taxpayers would otherwise use for their own ends. They deserve a full and complete accounting of the success or failure of the programs operated on their behalf, one that takes account the marginal cost of public funds, the possibility of general equilibrium effects, and the possibility of effects on outcomes other than those directly targeted by the program and that makes reasonable assumptions about the persistence of program impacts beyond the data.
6. Avoid the siren call of popular alternatives (such as performance management and surveys of customer satisfaction) to serious program evaluation. Both have their uses but the literature makes clear that neither provides a reliable substitute for econometric evaluations.
7. Many relatively simple and inexpensive institutional changes can have important effects on evaluation quality. These include the creation of public use data sets, greater use of outside expertise during evaluation design and execution, and publication of evaluation findings in peer-reviewed outlets as well as the creation of institutions to encourage deeper interaction between government, academics involved in evaluation research, and evaluation consultants.

References

Altonji, J., Elder, T. and Taber, C. 2005, 'Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools', *Journal of Political Economy,* vol. 113, no. 1, pp. 151–84.

Angelucci, M. and De Giorgi, G. 2009, 'Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption?', *American Economic Review,* vol. 99, no. 1, pp. 486–508.

Angrist, J. 1998, 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica,* vol. 66, no. 2, pp. 249–88.

—— and Evans, W. 1998, 'Children and the parents' labor supply: evidence from exogenous variation in family size', *American Economic Review,* vol. 88, no. 3, pp. 450–77.

—— and Pischke, J-S. 2009, *Mostly Harmless Econometrics*, Princeton University Press, Princeton.

Banerjee, A. and Duflo, E. 2009, 'The experimental approach to development economics', *Annual Review of Economics,* vol. 1, pp. 151–78.

Barnow, B. 2010, Setting up social experiments: the good, the bad and the ugly, Manuscript, Johns Hopkins University, unpublished.

—— and Smith, J. 2004, 'Performance management of U.S. job training programs: lessons from the Job Training Partnership Act', *Public Finance and Management,* vol. 4, no. 3, pp. 247–87.

Bertrand, M., Duflo, E. and Mullainathan, S. 2004, 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics,* vol. 119, no. 1, pp. 249–75.

Bitler, M., Gelbach, J. and Hoynes, H. 2006, 'What mean impacts miss: distributional effects of welfare reform experiments', *American Economic Review,* vol. 96, no. 4, pp. 988–1012.

Black, D., Smith, J., Berger, M. and Noel, B. 2003, 'Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system', *American Economic Review,* vol. 93, no. 4, pp. 1313–27.

Blattman, C. 2008, Impact evaluation 2.0, Presentation to the Department for International Development (DFID), London, UK.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W. and Bos, J. 1997, 'The benefits and costs of JTPA Title II-A programs: key findings from the National

Job Training Partnership Act Study', *Journal of Human Resources,* vol. 32, no. 3, pp. 549–76.

Blundell, R, Dearden, L. and Sianesi, B. 2005, 'Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey', *Journal of the Royal Statistical Society: Series A,* vol. 168, no. 3, pp. 473–512.

Bound, J., Jaeger, D. and Baker, R. 1995, 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American Statistical Association,* vol. 90, no. 430, pp. 443–50.

Burgess, D. 2010, 'Toward a reconciliation of alternative views on the social discount rate', in Burgess, D. and Jenkins, G. (eds), *Discount Rates for the Evaluation of Public-Private Partnerships,* McGill-Queen's University Press, Montreal, pp. 131–56.

Burghardt, J., Schochet, P., McConnell, S., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J. and Jackson, R. 2001, *Does the Job Corps Work? Summary of the National Job Corps Study,* Mathematica Policy Research, Princeton, NJ.

Busso, M., DiNardo, J. and McCrary, J. 2009a, New evidence on the finite sample properties of propensity score matching and reweighting estimators, Manuscript, University of Michigan, unpublished.

——, ——, —— 2009b, Finite sample properties of semiparametric estimators of average treatment effects', Manuscript, University of Michigan, unpublished.

Caliendo, M., and Kopeinig, S. 2008, 'Some practical guidance for the implementation of propensity score matching', *Journal of Economic Surveys,* vol. 22, no. 1, pp. 31–72.

Cameron, C. and Trivedi, P. 2005, *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.

Card, D. and Krueger, A. 1994, 'Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania', *American Economic Review,* vol. 84, no. 4, pp. 772–93.

—— and Sullivan, D. 1988, 'Measuring the effect of subsidized training programs on movements in and out of employment', *Econometrica,* vol. 56, no. 3, pp. 497–530.

Cook, T. 2008, '"Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics', *Journal of Econometrics,* vol. 142, no. 2, pp. 636–54.

Couch, K. 1992, 'New evidence on the long-term effects of employment and training programs', *Journal of Labor Economics,* vol. 10, no. 4, pp. 380–8.

Courty, P., Heckman, J., Heinrich, C., Marschke, G. and Smith, J. 2010, *Performance Standards in a Government Bureaucracy,* W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Crompton, J. 1995, 'Analysis of sports facilities and events: eleven sources of misapplication', *Journal of Sports Management*, vol. 9, no. 1, pp. 14–35.

Dahlberg, M. and Forslund, A. 2005, 'Direct displacement effects of labour market programmes', *Scandinavian Journal of Economics,* vol. 107, no. 3, pp. 475–94.

Dahlby, B. 2008, *The Marginal Cost of Public Funds: Theory and Applications,* MIT Press, Cambridge, MA.

Davidson, C. and Woodbury, S. 1993, 'The displacement effects of reemployment bonus programs', *Journal of Labor Economics,* vol. 11, no. 4, pp. 575–605.

Deaton, A. 2009, 'Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development', NBER Working Paper no. 14690.

Dehejia, R. and Wahba, S. 1999, 'Causal effects in nonexperimental studies: reevaluating the evaluation of training programs.' *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1053–62.

——, —— 2002, 'Propensity score matching methods for non-experimental causal studies', *Review of Economics and Statistics,* vol. 84, no. 1, pp. 151–61.

De Giorgi, G. 2008, 'Long-term effects of a mandatory multi-stage program: the New Deal for Young People in the UK', Institute for Fiscal Studies Working Paper 05/08.

Dillon, S. 2008, 'An initiative on reading is rated ineffective', *New York Times*, 2 May.

Djebbari, H. and Smith, J. 2008, 'Heterogeneous impacts in PROGRESA', *Journal of Econometrics,* vol. 145, no. 1–2, pp. 64–80.

Dolton, P. and Smith, J. 2010, The econometric evaluation of the New Deal for Lone Parents, Manuscript, University of Michigan, unpublished.

Doolittle, F. and Traeger, L. 1990, *Implementing the National JTPA Study*, MDRC, New York.

Eckel, C.C., Johnson, C.A. and Montmarquette, C. 2005, 'Saving decisions of the working poor: short- and long-term horizons', in Carpenter, J., Harrison, G. and List, J. (eds), *Field Experiments in Economics: Research in Experimental Economics*, Volume 10, JAI Press, Greenwich, CT, pp. 219–60.

——. ——. —— and Rojas, C. 2007, 'Debt aversion and the demand for loans for postsecondary education', *Public Finance Review,* vol. 35, pp. 233–62.

Evans, W. and Kim, B. 2006, 'Patient outcomes when hospitals experience a surge in admissions', *Journal of Health Economics,* vol. 25, no. 2, pp. 365–88.

—— and Lein, D. 2005, 'The benefits of prenatal care: evidence from the PAT bus strike', *Journal of Econometrics,* vol. 125, no. 1–2, pp. 207–39.

Falk, A. and Fehr, E. 2003, 'Why labour market experiments?', *Labour Economics,* vol. 10, no. 4, pp. 399–406.

Frölich, M. and Lechner, M. 2010, 'Exploiting regional treatment intensity for the evaluation of active labor market policies', *Journal of the American Statistical Association*, forthcoming.

Gamse, B., Jacob, R.T., Horst, M., Unlu, F., Bozzi, L., Caswell, L., Rodger, C., Smith, W.C., Brigham, N. and Rosenblum, S. 2008, *Reading First Impact Study Final Report* (NCEE 2009-4038), National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, Washington, DC.

Goldberger, A. 1972a, Selection bias in evaluating treatment effects: some formal illustrations, Manuscript, University of Wisconsin, unpublished.

—— 1972b, Selection bias in evaluating treatment effects: the case of interaction, Manuscript, University of Wisconsin, unpublished.

—— 2008, 'Selection bias in evaluating treatment effects: some formal illustrations', in Millimet, D., Smith, J. and Vytlacil, E. (eds), *Modeling and Evaluating Treatment Effects in Economics: Advances in Econometrics,* vol. 21, pp. 1–31.

Gramlich, E. 1997, *A Guide to Benefit-Cost Analysis,* 2nd edn, Waveland Press.

Greenberg, D. and Shroder, M. 2004, *Digest of Social Experiments,* 3rd edn, Urban Institute Press, Washington, DC.

Gregory, A. 2000, 'Problematizing participation: a critical review of approaches to participation in evaluation theory', *Evaluation,* vol. 6, no. 2, 179–99.

Heckman, J. 1979, 'Sample selection bias as a specification error', *Econometrica,* vol. 47(1), pp. 153–61.

—— 1996, 'Comment', in Feldstein, M. and Poterba, J. (eds), *Empirical Foundations of Household Taxation*, University of Chicago Press, Chicago, pp. 32–8.

——, Heinrich, C. and Smith, J. 2002, 'Understanding incentives in public organizations', *Journal of Human Resources,* vol. 37, no. 4, pp. 778–811.

——, Hohmann, N., Smith, J. and Khoo, M. 2000, 'Substitution and dropout bias in social experiments: a study of an influential social experiment', *Quarterly Journal of Economics,* vol. 115, no. 2, pp. 651–94.

—— and Hotz, V.J. 1989, 'Choosing among alternative methods of evaluating the impact of social programs: the case of manpower training', *Journal of the American Statistical Association,* vol. 84, no. 408, pp. 862–74.

——, Ichimura, H., Smith, J. and Todd, P. 1998, 'Characterizing selection bias using experimental data', *Econometrica,* vol. 66, no. 5, pp. 1017–98.

—— LaLonde, R. and Smith, J. 1999, 'The economics and econometrics of active labor market programs', in Ashenfelter, O. and Card, D. (eds), *Handbook of Labor Economics,* vol. 3A, North-Holland, Amsterdam, pp. 1865–2097.

——, Lochner, L. and Taber, C. 1998, 'Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents', *Review of Economic Dynamics.* vol. 1, no. 1, pp. 1–58.

——, and Smith, J. 2000, 'The sensitivity of experimental impact estimates: evidence from the National JTPA Study', in Blanchflower, D. and Freeman, R. (eds), *Youth Employment and Joblessness in Advanced Countries*, University of Chicago Press for NBER, Chicago, pp. 331–56.

——, —— and Clements, N. 1997, 'Making the most of programme evaluations and social experiments: accounting for heterogeneity in programme impacts', *Review of Economic Studies,* vol. 64, no. 4, pp. 487–535.

——, —— and Taber, C. 1998, 'Accounting for dropouts in social experiments', *Review of Economics and Statistics,* vol. 80, no. 1, pp. 1–14.

—— Tobias, J. and Vytlacil, E. 2001, 'Four parameters of interest in the evaluation of social programs', *Southern Economic Journal,* vol. 68, no. 2, pp. 210–23.

—— Urzua, S. 2009, 'Comparing IV with structural models: what simple IV can and cannot identify', NBER Working Paper no. 14706.

Heinrich, C. 2007, 'Evidence-based policy and performance management: challenges and prospects in two parallel movements', *American Review of Public Administration,* vol. 37, no. 3, pp. 255–77.

Hirano, K., Imbens, G., Rubin, D. and Zhou, X-H. 2000, 'Assessing the effect of an influenza vaccine in an encouragement design', *Biostatistics,* vol. 1, pp. 69–88.

Hotz, V.J., Imbens, G. and Klerman, J. 2006, 'Evaluating the differential effects of alternative welfare-to-work training components: a reanalysis of the California GAIN program', *Journal of Labor Economics,* vol. 24, no. 3, pp. 521–66.

—— and Scholz, J.K. 2002, 'Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data' in *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press, pp. 275-315.

Ichino, A, Mealli, F. and Nannicini, T. 2008, 'From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?', *Journal of Applied Econometrics,* vol. 23, pp. 305–27.

Imbens, G. 2009, 'Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009)', NBER Working Paper no. 14896.

—— and Angrist, J. 1994, 'Identification and estimation of local average treatment effects', *Econometrica,* vol. 62, no. 4, pp. 467–76.

—— and Lemieux, T. 2008, 'Regression discontinuity designs: a guide to practice', *Journal of Econometrics,* vol. 142, no. 2, pp. 615–35.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., Ross, C., Schochet, P., and Swank, P. 2007, *National Evaluation of Early Reading First: Final Report*, US Government Printing Office, US Department of Education, Institute of Education Sciences, Washington, DC.

Kemple, J., Doolittle, F., and Wallace, J. 1993, *The National JTPA Study: Site Characteristics and Participation Patterns*. Manpower Demonstration Research Corporation, New York, NY.

Kochar, A. 1999, 'Smoothing consumption by smoothing income: hours-of-work responses to idiosyncratic agricultural shocks in rural India', *Review of Economics and Statistics,* vol. 81, no. 1, pp. 50–61.

Krueger, A. 2003, 'Economic considerations and class size', *Economic Journal,* vol. 113, no. 485, pp. F34–F63.

LaLonde, R. 1986, 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review,* vol. 76, no. 4, pp. 604–20.

Lechner, M. and Smith, J. 2007, 'What is the value added by case workers?', *Labour Economics,* vol. 14, no. 2, pp. 135–51.

—— and Wiehler, S. 2010, 'Kids or courses: gender differences in the effects of active labor market programs', *Journal of Population Economics*, forthcoming.

—— and Wunsch, C. 2009, 'Are training programs more effective when unemployment is high?', *Journal of Labor Economics,* vol. 27, no. 4, pp. 653-92.

Lee, D. and Lemieux, T. 2009, 'Regression discontinuity designs in economics', NBER Working Paper no. 14723.

—— and McCrary, J. 2009, 'The deterrence effect of prison: dynamic theory and evidence', Manuscript, University of California, Berkeley, unpublished.

Leigh, A. 2009, What evidence should social policymakers use?, Manuscript, Australian National University, unpublished.

Lise, J., Seitz, S. and Smith, J. 2010, Equilibrium policy experiments and the evaluation of social programs, Manuscript, University of Michigan, unpublished.

Long, D., Mallar, C. and Thornton, C. 1981, 'Evaluating the benefits and costs of the Job Corps', *Journal of Policy Analysis and Management,* vol. 1, no. 1, pp. 55–76.

McConnell, S., Decker, P. and Perez-Johnson, I. 2006, 'The role of counseling in voucher programs: findings from the individual training account experiment', Manuscript, Mathematica Policy Research, unpublished.

McCrary, J. 2008, 'Manipulation of the running variable in the regression discontinuity design: a density test', *Journal of Econometrics,* vol. 142, no. 2, pp. 698–714.

Meyer, B. 1995, 'Natural and quasi-experiments in economics', *Journal of Business and Economic Statistics,* vol. 13, no. 2, pp. 151–61.

Milligan, K. and Stabile, M. 2007, 'The integration of child tax credits and welfare: evidence from the Canadian National Child Benefit Program', *Journal of Public Economics,* vol. 91, no. 1–2, pp. 305–26.

Moffitt, R. 1991, 'Program evaluation with nonexperimental data', *Evaluation Review,* vol. 15, no. 3, pp. 291–314.

Morris, P. and Michalopoulos, C. 2003, 'Findings from the Self-Sufficiency Project: effects on children and adolescents of a program that increased employment and income', *Journal of Applied Developmental Psychology,* vol. 24, no. 2, pp. 20-39.

Neumark, D. and Wascher, W. 2000, 'Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: comment', *American Economic Review,* vol. 90, no. 5, pp. 1362–96.

Noll, R. and Zimbalist, A. 1997, *The Economic Impact of Sports Teams and Facilities*, Brookings Institution, Washington, DC.

Oreopoulos, P. 2006, 'Estimating average and local average treatment effects of education when compulsory schooling laws really matter', *American Economic Review,* vol. 96, no. 1, pp. 152–75.

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

Osborne, D. and Gaebler, T. 1992, *Reinventing Government: How The Entrepreneurial Spirit is Transforming the Public Sector,* Perseus, Boulder, CO.

Radin, B. 2006, *Challenging the Performance Movement: Accountability, Complexity and Democratic Values*. Georgetown University Press, Washington, DC.

Raudenbusch, S. and Bryk, A. 2001, *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn, Sage, New York.

Rossi, P. 1987, The Iron Law of Evaluation and Other Metallic Rules, *Research in Social Problems and Public Policy*, no. 4, pp 3-20.

Rudalevige, A. 2008, 'Structure and science in education research', in Hess, F. (ed), *When Research Matters*, Harvard Education Press, Cambridge, MA, pp. 17–40.

Schochet, P. 2008, *Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations* (NCEE 2008-4026), National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, Washington, DC.

Smith, G. and Pell, J. 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *British Medical Journal,* vol. 327, pp. 20–7.

—— and Staghøej, J. 2010, Using statistical treatment rules for assign of participants in labor market programs, Manuscript, University of Michigan, unpublished.

—— and Todd, P. 2005, 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics,* vol. 125, no. 1–2, pp. 305–53.

—— and Whalley, A. 2010, How well do we measure public job training?, Manuscript, University of Michigan, unpublished.

——, —— and Wilcox, N. 2010, Are program participants good evaluators?, Manuscript, University of Michigan, unpublished.

Todd, P. and Wolpin, K. 2005, 'Assessing the impact of a school subsidy program in Mexico using a social experiment to validate a dynamic behavioral model of child schooling and fertility', *American Economic Review,* vol. 96, no. 5, pp. 1384–1417.

Trenholm, C., Devaney, B., Fortson, K., Quay, L., Wheeler, J and Clark, M. 2007, *Impacts of Four Title V, Section 510 Abstinence Education Programs: Final Report*, Mathematica Policy Research, Princeton, NJ.

US General Accounting Office 1996, *Job Training Partnership Act: Long-Term Earnings and Employment Outcomes* (Report HEHS-96-40), US Government Printing Office, Washington, DC.

US Institute of Education Sciences 2008, *Rigor and Relevance Redux: Director's Biennial Report to Congress* (IES 2009-6010), US Department of Education, Washington, DC.

Van der Klaauw, W. 2008, 'Regression-discontinuity analysis: a survey of recent developments in economics', *Labour: Review of Labour Economics and Industrial Relations,* vol. 22, no. 2, pp. 219–45.

Wooldridge, J. 2002, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

General discussion and dinner address

The roundtable discussion following the first session centred around three themes: the role of evidence in improving public policy; differences in quality of evidence; and the important role of institutions in marshalling and making best use of evidence. The general discussion session was followed by an after-dinner address by Terry Moran, the Secretary of the Department of Prime Minster and Cabinet.

Evidence is important and can help improve public policy

Roundtable participants discussed some of the policy examples highlighted by the keynote speakers where evidence had played an important role in policy development, such as welfare reform in the United States, and raised a number of other examples.

Several participants wanted to emphasise that evidence-based policy was not confined to *ex post* evaluation, and that evidence was important at every stage of the policy development process, from identifying the policy problem, through *ex ante* assessment of policy proposal, to *ex post* evaluation. Brian Head noted 'every day spent on problem identification and discussion is worth ten days of actually doing the study'.

Another academic speaker noted that policy insights were sometimes driven by wider access to administrative data, and mentioned two examples where data availability drove improved understanding:

- A government work incentives program in the United Kingdom, which paid low income workers a welfare bonus to work additional hours, was found to be ineffective. It had a very low uptake because although the welfare payment substantially increased beneficiaries' gross income, that increase was then offset by a reduction in other welfare benefits such as rent assistance.
- Job creation figures for Indigenous programs can overstate their success. The data can mask the turnover of Indigenous employees who have simply transferred to different jobs or programs, rather than being genuine employment growth.

Not all evidence is created equal

There were differing views on the relative merits of various methods for generating evidence and their relevance to different types of policy questions. Different types of policy questions will require different forms of evidence and some evidence will be more robust than others. For instance, formulating an evidence base for responding to climate change would rely on very different techniques than those required for assessing the case for merit based pay for teachers. There is no one 'right' type of evidence. Jeffrey Smith suggested 'let a hundred flowers bloom and let the marketplace for ideas sort it out. The key is thinking hard about each of these approaches; thinking hard about what it can add and what it can't add to the discussion'.

Regardless of the particular method used to generate evidence, several speakers stressed that it should be robust and be open to scrutiny. Ron Haskins cited the American sociologist, Peter Rossi, who concluded on the basis of many *ex-post* evaluations that the expected value of any large scale social program is zero (Rossi's 'Iron Law'). He noted the frequent tension between the enthusiasm of those close to program implementation and more formal, high-quality evaluation: 'if you ask program operators, [they'll say] 'this is better than sliced bread – it changed my whole community' but then a randomised trial finds the policy has no effect, and this happens time after time'.

Speakers stressed that openness to different forms of evidence and different analytical methodologies does not mean 'lowering the bar' on the standard of evidence for policy. Some participants noted that the most rigorous evaluation sometimes seems to be reserved for smaller projects and policy questions, because they are often analytically more tractable, and politically less contentious, whereas major policy issues are sometimes subject to little rigorous analysis. Jeffery Smith noted the paradox that for some large macroeconomic questions (such as appropriate monetary rules to respond to recessions) there were relatively few data points (for example, just a handful of well-documented recessions) and no easy way of testing counterfactual propositions. Policymakers' need for guidance meant a wide variety of methodologies (for example, modelling, econometrics and case studies) were mined by analysts keen to marshal whatever evidence they could.

Institutions matter

It was broadly recognised in the roundtable discussion that factors other than evidence are often the main force driving policy development. One participant raised the question of why, for two policy areas, both with equally compelling

evidence on the most effective policy choice, the balance of evidence prevails in one case and not another. Was it possible to identify the factors that determine when evidence has an influence on policy development?

Ron Haskins' view was that it was counterproductive to try to remove other influences on the policy development process, such as lobbyists, especially since they sometimes bring important insights and evidence to the issue. Rather, one of the key factors in ensuring that objective evidence has an influence, was having the right government institutions (appropriately skilled analysts producing publicly available evidence), such as the Congressional Budget Office in the United States, and making this evidence transparent and contestable — for example, by enabling academics and other researchers to have access to data and methods.

Speakers also noted that extensive demand for high-quality evaluation built a pool of skilled analysts and institutions in the United States, pointing to the important roles of private sector organisations such as MDRC and Mathematica, with decades of specialisation in high-quality program evaluation and policy research. Australian demands for evaluation had not been large enough so far to support the growth of such expertise.

Dinner address

In his address to the roundtable dinner, Terry Moran revisited some recent history in the development of the human capital reform agenda, leading through to the reform of the structure of Commonwealth-State financial relations in November 2008. The history was replete with examples of how analysis and emergent evidence shaped policy thinking.

His talk traced the analytical stimulus provided to Victorian Government officials by the Commonwealth Treasury's initial Intergenerational Report in 2002-03 and Ken Henry's associated speeches on the contributions of the '3 Ps' — population, participation and productivity — to per capita GDP trends. Thinking about such issues as trends in health spending led officials to envisage a human capital reform agenda which could help address the participation and productivity elements of the response to the demographic challenge.

The human capital reform agenda also illustrated, in Terry Moran's assessment, the importance of institutional change to helping improve outcomes in complex government service delivery systems such as the education and health systems. The Productivity Commission's work in illustrating the 'outer envelope' of benefits from the national reform agenda had also been an important stimulant of reform thinking.

The upshot of this thinking was the transformation of Commonwealth-State financial relations in November 2008, when the Council of Australian Governments radically streamlined the system of 96 specific purpose payments down to six streams of spending, giving states the scope for policy innovation in how they delivered agreed objectives, outcomes and outputs. The agreement also proposed additional revenue and a stream of potential reward payments for State and Territory policy innovation successes, and an independent umpire of progress, in the form of the COAG Reform Council.

He observed that the new system gave the States and Territories the opportunities they had sought, and the tests of the new system would be the quality of policy reform and demonstrated improvement in outcomes over years to come.

SESSION 2

HOW ROBUST IS OUR EVIDENCE-BASED POLICY MAKING?

5 Reflections on four Australian case studies of evidence-based policy

Bruce Chapman
Crawford School of Economics and Government, College of Asia and the Pacific, Australian National University

Abstract

Policy-making should be informed by solid evidence. This paper explores four Australian examples of the way that evidence has been used to influence policy through case studies relating to higher education financing, labour market programs, TAFE funding and student income support. The paper distinguishes between instances where evidence was used as a foundation for new policy ('Ms Polyanna' evidence) and instances where evidence was used to justify a pre–existing policy agenda ('Mr Hyde' evidence). The wider role of evidence in public policy is explored through case studies, with an acknowledgement that evidence is often used to persuade or silence critics as much as it is to formulate sound policy.

5.1 Introduction

The following are reflections on what I perceive the role of evidence to have been in my experiences in Australian public policy making over the past 20 years or so. I confess that I find evidence-based policy a complicated area to think about, because of some difficulties I have with the meaning of the words. What exactly is 'evidence-based policy'?

It is inconceivable that a politician or other influential policy person would dispute the importance of evidence-based policy. To make this point absurdly clear, imagine a minister announcing a policy reform and saying at the media conference that 'An important contributing factor behind [new policy name] is that there is no research available to suggest that it will have desirable effects. Indeed, there is even some possibility that things will be made worse'.

If sanctioning evidence-based policy is as obviously trivial as being in favour of, say, efficient government or a fair go, what do its proponents really have in mind in their endorsement and advocacy of evidence-based policy? To assist in this matter I consulted speeches made by Brian Head and Gary Banks, which provide important background to the topic. Head (2009, p. 13) argues that 'The advocates of EBP urge the incorporation of rigorous research evidence into public policy debates and internal public sector processes for policy evaluation and program improvement'. A clarifying observation from Banks (2009, p. 14) is that 'If it hasn't been tested, or contested, we can't really call it evidence'. In combination, these remarks imply that a policy development is evidence based if the development process used meticulous research methods and data, and was subjected to disinterested scrutiny.

With these clarifications it is accurate to suggest that, in the examples I know about, policy has indeed been evidence based. Even so, this does not necessarily mean that the evidence usually motivated the policy, not does it imply that the available information was used only to define and calibrate the parameters for reform. To help understand the different motives for, and use of, evidence, it is useful to classify two distinct functions of rigorous and contested information in the policy process. These can be labelled as:

- Evidence Type 1 — data of which the principal benefit is to inform the policy stance. This is essentially what more innocent commentary implies by the term 'evidence-based policy' (this classification can also be called 'Ms Polyanna').
- Evidence Type 2 — data of which the principal benefit is to smooth the implementation process or silence the potential opponents of policy reform. This is what more sceptical analysts of government mean by what might be labelled, unsupportively, 'policy-based evidence' (this classification can also be called 'Mr Hyde').[1]

This chapter explains and documents important examples of both.

5.2 Background

This paper draws on my involvement as an advisor in four different areas of economic policy reform in the period from 1987 to 2009: higher education financing (1987–2009); labour market programs for the long term unemployed (1992–95); TAFE funding (2006–08); and student income support (2008–09).

1 The first time I heard this phrase, it came from Professor Richard Mulgan of the Crawford School of Economics and Government at the Australian National University.

Higher education financing

In 1987, the Minister responsible for Australian higher education, John Dawkins, invited me to prepare a paper outlining the costs and benefits of different approaches to the reintroduction of a user-pays higher education system for Australia. Critically, the Minister had already decided that tuition fees should be reintroduced; they had been abolished by the previous Labor government in 1973.

My report was delivered in December 1987. It presented analyses of several financing mechanisms, including up-front fees with scholarships, up-front fees with government subsidised bank loans, and an income contingent charge system. The paper recommended the last of these, with repayments to be made via the direct tax system. Details of how such a system might work were provided, including possible fee levels and repayment parameters.

The minister subsequently set up the Higher Education Financing Committee, chaired by Neville Wran, former Labor Premier of New South Wales. I was appointed as a consultant to the committee. Its report, delivered in May 1988, recommended the adoption of an income contingent loan to underpin higher education tuition, to be called the Higher Education Contribution Scheme (HECS).

Labour market programs

In 1991, Raja Junankar, Cezary Kapuscinski and I, as research academics, were engaged in forecasting exercises with respect to the likely future levels of Australian long-term unemployment (LTU) — that is, the number of people who are unemployed continuously for 12 months or more. Our analysis suggested that, well after recovery from the serious recession of 1990–93 had begun, LTU numbers would rise to levels that were more than double the historical peak. Using labour market theory, we explained why this situation was both very inefficient for the operation of the macroeconomy and very inequitable for the people concerned. This research gained some publicity during 1992, and was raised in parliament by the Liberal Party – National Party coalition, then in opposition, as an indictment of the Labor government's economic policy stance.

After Labor's (arguably surprising) victory in the 1993 federal election, I was asked by the Minister for Employment, Education and Training, Kim Beazley, to undertake a consultancy related to LTU (Chapman 1993a). Soon after that, the Australian Government set up a high-level committee chaired by the Secretary of the Department of the Prime Minister and Cabinet, Michael Keating. Among others, Professors Bob Gregory and Barry Hughes were appointed to the committee, and I served as a consultant at the same time as I was preparing the paper for Minister

Beazley. The committee process led to the Working Nation program, which had as its centrepiece the 'Job Compact', under which all people who had been unemployed for 18 months or longer had access to wage subsidy, training or public sector employment opportunities (Australian Government 1994).

TAFE funding

Since the late 1980s, I and many others have been involved in a series of research exercises related to the potential for radical reform of the up-front fees associated with TAFE (Technical and Further Education) courses and vocational education and training in general. Over time, academics and others have continually and publicly made the case in favour of TAFE funding reform.

This policy stance was arguably reinforced when, in 2005–06, I had the opportunity to serve as a consultant on the topic to the Victorian Government and also undertook a joint project on the topic with colleagues Mark Rodrigues and Chris Ryan.[2] In this exercise we modelled and analysed the potential for TAFE fees policy to be converted to an income-contingent loans system, based on HECS. In 2008, the Victorian Government, in partnership with the Australian Government, announced major changes to TAFE funding for associate diplomas and diplomas, using such an approach (Victorian Government 2008).

Student income support

In 2008, I and several others, including Lin Martin and David Phillips, were engaged as consultants to the Review of Australian Higher Education (the Bradley Review). Working with Professor Martin, my principal area of analysis was student income support. An important part of this role was to provide arguments for, and evidence relevant to, possible reforms to the existing system. At the completion of its process the Bradley Review recommended radical changes to Youth Allowance. The changes were adopted by the Australian Government in the 2009–10 Budget.

5.3 Case studies

In the four examples of policy reform cited above, my role was essentially that of researcher. I endeavoured to analyse the issues, using the conceptual framework of economics, and sought to bring to bear the best available statistical evidence

2 Mr Rodrigues was on secondment from the Australian Treasury to the Australian National University at that time.

relevant to the policy problem. These approaches and the particular role of evidence are now explained in the context of a case study of each policy reform.

Higher Education Contribution Scheme

Relative lifetime earnings

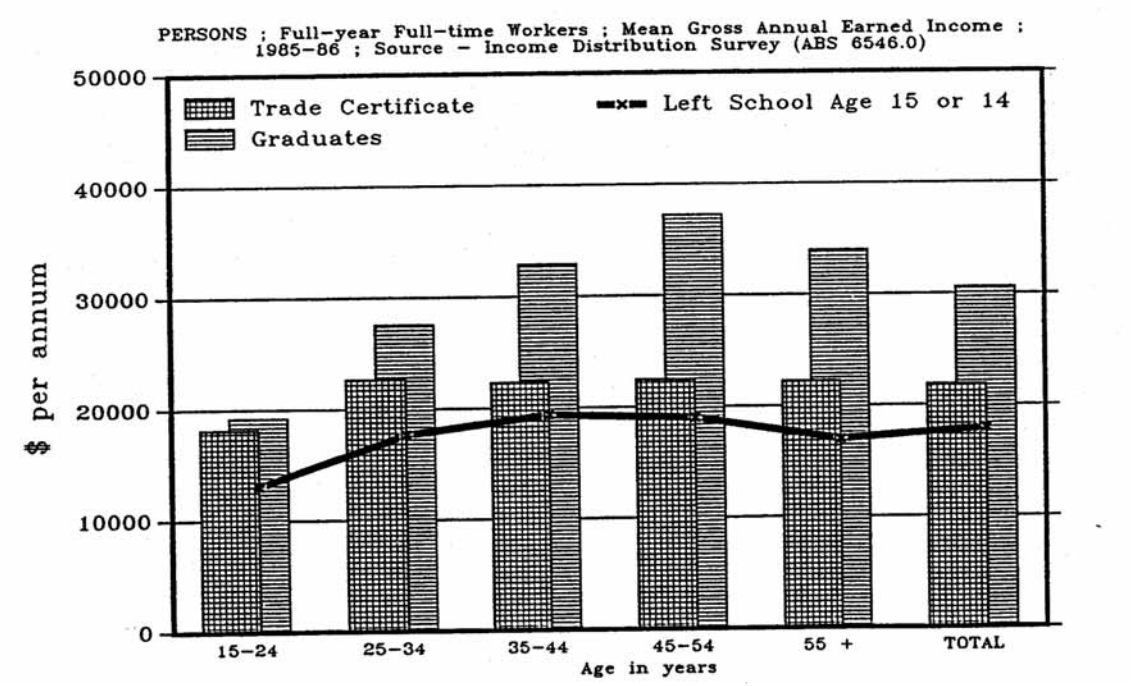
An important part of the background to HECS is the fact that in 1987 two Labor Cabinet ministers, John Dawkins and Peter Walsh, were strongly in favour of the reintroduction of student fees on the grounds of equity. Their view was that a system which did not charge tuition fees for higher education students was regressive, since at the time universities were paid for by all taxpayers, yet students on average came from relatively privileged backgrounds and as graduates received relatively high personal economic benefits.

My role in writing the options paper to set the scene for the reintroduction of tuition fees had at least two motivations. One was to consider the costs and benefits of alternative student financing policies, but another was to examine evidence relevant to the case for changing from a fully taxpayer-funded system to a system requiring financial contributions from students. To achieve this end, I employed the most commonly used evidence concerning the alleged lifetime earnings advantage of graduates. The data are shown in figure 5.1 (Chapman 1988).

There is no doubt that evidence of average lifetime income advantages of graduates, such as that presented in figure 5.1 was critical to the debate surrounding the case for the reintroduction of university tuition fees. The data showed fairly compelling support for the position already held by ministers Dawkins and Walsh, and probably also by the majority of members of Cabinet. Thus the evidence presented did not initiate the commitment to reform; instead, it was the case that the data facilitated the politically successful introduction of HECS. Therefore, in the classification system suggested in section 5.1, this aspect of the HECS exercise is Evidence Type 2, Mr Hyde.

Figure 5.1 **Earned income by age and education (persons)**

Background to the Higher Education Contribution Scheme



Data source: Chapman 1988.

Income contingent loans and government-guaranteed bank loans

Student loans systems are commonplace internationally. What was different about HECS was that, for the first time, the mechanism involved the notion that repayments would be collected through the income tax system contingent on the former student's income (an instrument known as an 'income contingent loan'). At the time of the development of HECS, most other countries with student loans schemes used banks to finance the loans, with student debts being guaranteed by government.[3] While there was no direct evidence in favour of the Australian-suggested approach (since no other country had introduced such a scheme), it was nevertheless possible to use the conceptual tools and empirical evidence from labour economics relevant to an assessment of the likely effects of this different policy stance.

The most important concern about the introduction of the policy related to the access of poor students to higher education. Many opponents of the scheme asserted that a HECS-type approach would significantly diminish the access of the poor.

3 For analysis of the economic issues relevant to this aspect of student loans policies, see Chapman (2006).

However, my view with respect to the conceptual issues and the evidence from related issues in the labour economics literature was that the new arrangement would have at most benign effects on the access of the poor to higher education. Indeed, it even seemed likely that many poor students would be advantaged by the introduction of HECS, if the promise of additional revenue also implied an expansion in the number of university places.

This suggests that research underpinning the development of income contingent loan policy in Australia was not motivated principally by a need to persuade opponents of the benefits of HECS as such, but was instead based on a view that this aspect of the policy design constituted a better economic policy approach than the alternatives.[4] This aspect of the HECS process should be classified as an example of Evidence Type 1, Ms Polyanna.

Working Nation

Helping to set the scene for the early 1990s labour market program intervention

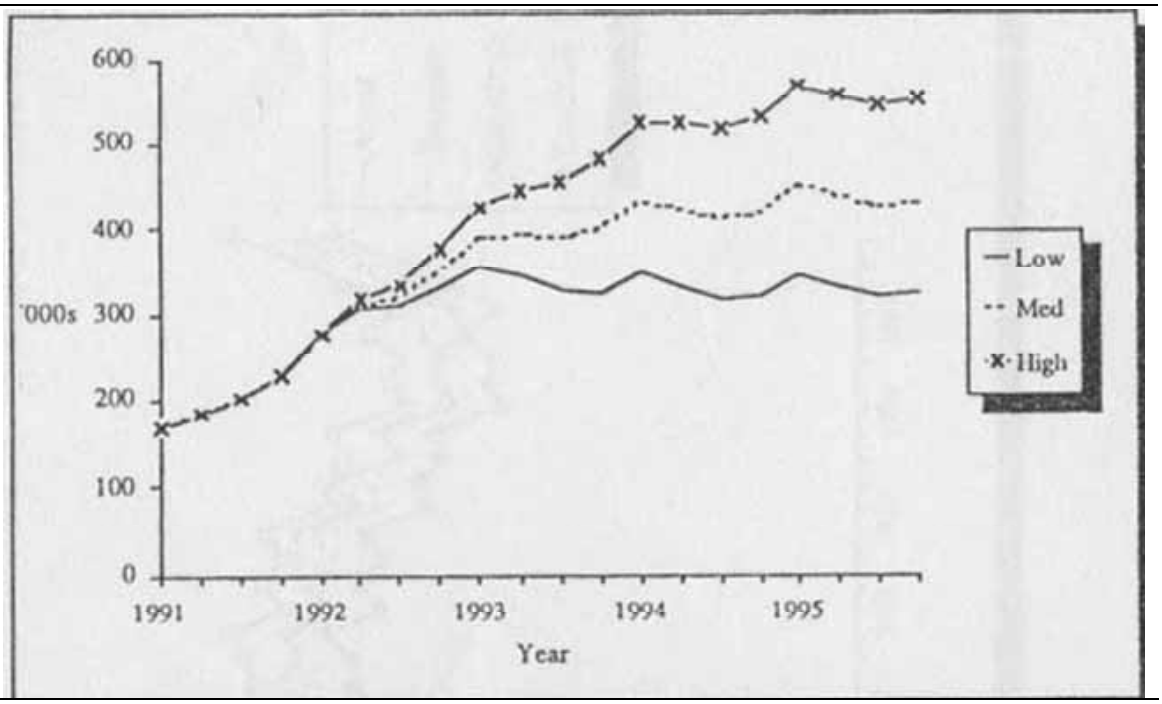
In 1992, as research academics, Raja Junankar, Cezary Kapuscinski and I developed an econometric model which related various and complicated forms of the quarterly Australian adult male and female unemployment rate to contemporaneous and future levels of the numbers experiencing LTU. We published what we considered to be realistic boundaries of the future levels of LTU; these were between about 300 000 and 500 000 people by the middle of the decade. We also pointed to the highly deleterious consequences of the LTU situation, in both equity and efficiency terms. Our projections are shown in figure 5.2. At the time the data and analysis were treated with both scepticism and alarm; however, the analysis seemed to matter in the setting-up of the Working Nation Task Force after Labor's federal election win in early 1993.

There is little doubt that the LTU projections broadly influenced the nature of the debate concerning the role of government policy with respect to the recovery from the recession. By the end of 1992, most analysts inside and outside government were expressing relief with respect to the end of the recession. The focus was very clearly on the macroeconomic aggregates — such as the unemployment rate, which had fairly clearly peaked around that time. Refocusing the debate to consider the importance of unemployment duration and not just overall levels of joblessness can

4 It is true, however, that major sections of the Australian Labor Party were against the introduction of HECS; this is best understood as coming from an aversion to there being a charge and not to the form it took.

be seen to contribute to a quite different way of understanding the effects of recession and the potential role of policy in recovery. The classification for this evidence in the policy process is a further example of Evidence Type 1, Ms Polyanna.

Figure 5.2 **Projections of long-term unemployment**
The beginnings of Working Nation



Data source: Chapman, Junakar and Kapuscinski 1992.

Providing persuasion for the implementation of the policy

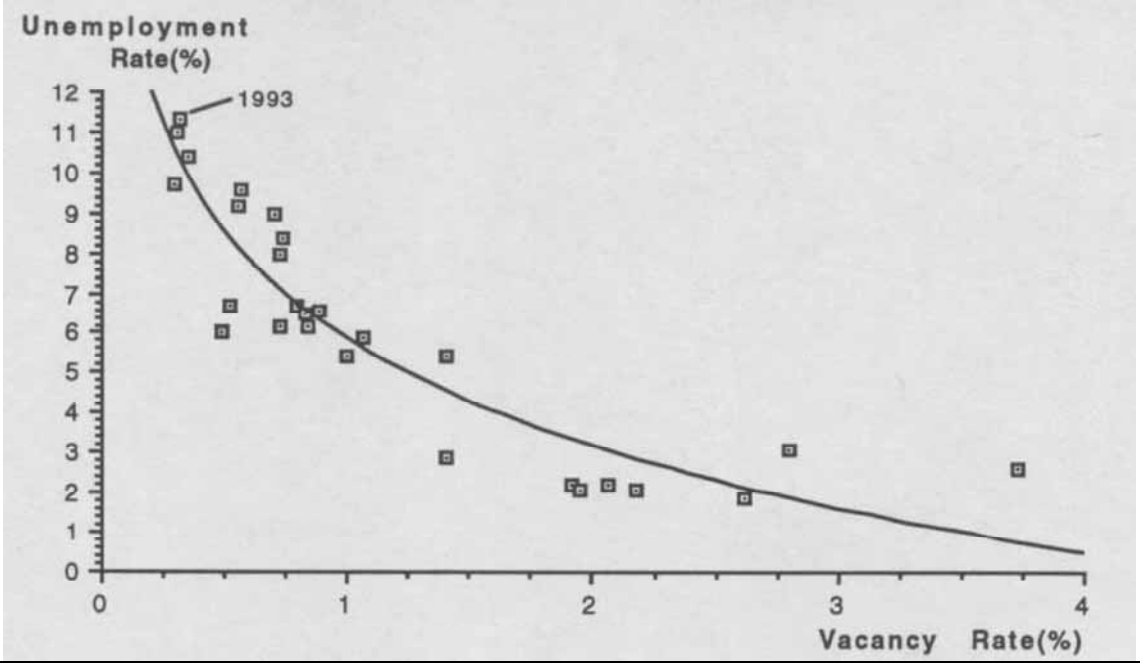
An important background aspect of the development of the Working Nation programs relates to whether or not there is an efficiency case for intervention. The basis for this perspective can be traced to both an economics conceptual framework and the availability of econometric tests of the essential propositions. The former is provided through the analysis of Chapman, Junankar and Kapuscinski (1993) and is based on the work of Lord Beveridge in the 1930s concerning the aggregate unemployment–vacancies (UV) relationship. The essential conjecture is that the UV trade-off deteriorates as LTU increases, implying that the macroeconomy operates less effectively when LTU is relatively high. There is considerable empirical evidence in support of this perspective (Budd, Levine and Smith 1988).

The nature of the apparent UV relationship in Australia became very important to the deliberations of the Working Nation Committee, for the following reason. If it is

the case that large numbers of long-term unemployed implies a mismatch between the available supply of labour and emerging job opportunities, this implies the possibility of wasted government expenditure on unemployment benefits and forgone tax revenue though lower than possible levels of employment and incomes. Depending on the evidence, there might be a case for targeted programs simply on the basis of improved budgetary costs, meaning that the 'Job Compact' being considered for policy reform could be justified without recourse to the more obvious case for intervention made on the basis of distribution and equity. The issue was clarified for the debate at the time through the use of the data shown in figure 5.3, information relevant to the issue of whether or not the Australian labour market operated less effectively in times of high LTU.[5]

Figure 5.3 **The unemployment–vacancies curve**

A critical part of the Working Nation expenditure debate



Data source: Chapman (1993b).

The relationship depicted in figure 5.3 became a critical part of the in-house debate surrounding the development of Working Nation in 1993 and early 1994. Most importantly, it became clear that officials in the Department of the Treasury were sceptical about critical empirical issues related to the effect of LTU on the efficiency of the operation of the aggregate labour market, and as part of the policy development their concern required a high-level, technical response. As a result, Barry Hughes and I spent an intense day with the Treasury officials, discussing (or,

5 The conceptual issues are considered in detail in Chapman (1993b).

more accurately, debating — perhaps even arguing about) whether or not LTU had shifted the Australian UV curve by the small amount of 0.5 per cent, or the larger amount of 1.2 per cent. The agreed size of the shift was significant because modelling had suggested that, with effective programs, expenditure on the LTU would be revenue-saving for the federal budget if, as a result, the UV curve was shifted by around 1 per cent.[6]

In the end we reached an agreement that the shift was around 0.8 per cent, and this provided a macroefficiency basis for the Job Compact of Working Nation. I believe that it would have been much more difficult to have broad government support for the size of the intervention implied by Working Nation in the absence of this agreement, and that it opened the road internally for the relatively smooth policy development and implementation process that followed. Since the data did not lead to the policy, this example should be classified as Evidence Type 2, Mr Hyde.

TAFE funding reforms

Setting the scene for TAFE funding reforms

For a very long time since the late 1980s, many education analysts were interested and disappointed in the differences between the financing approaches for undergraduate higher education and the TAFE system. TAFE had long had up-front tuition fees and, over time, these had grown to levels which were likely to act as barriers to the access of prospective students, even though there were also scholarships and concessions available to many. Yet, conceptually, there was nothing different between the capital market problems which had been recognised for university students and had led to HECS, and the sorts of financing difficulties likely to be faced by prospective students of vocational education and training.

The anomalies between having an income contingent loan for higher education undergraduates and up-front fees for other areas of tertiary education were noted and criticised by a growing number of commentators, including importantly Gavin Moodie, who wrote a regular column in *The Australian* newspaper's higher education supplement. As well, a series of academic articles, most notably Watson (2001), Wheelahan (2001), Watson, Whelahan and Chapman (2001) and Noonan, Burke and White (2004) argued the case on conceptual and equity grounds. A 2008 report from the Tertiary Education Union made oblique reference to the important need for an overhaul of TAFE funding, which was interpreted by many to be a call

6 For an application of the technical side of the modelling, see Piggott and Chapman 1995.

for the extension of HECS. This position was also taken in a high-profile presentation by David Phillips in 2002.

In combination, these arguments and the associated evidence concerning the conceptual errors inherent in maintaining the status quo eventually set the scene for radical reforms to TAFE funding, with the beginnings of change (which is still incomplete) appearing in the mid-2000s. Given the long period of inertia demonstrated by governments in this area, and the criticism of the policy stance, it is clear that this aspect of TAFE policy reform was led by the information, and the information was not instead used to reinforce a policy position that had already been taken. Therefore the classification here is Evidence Type 1, Ms Polyanna.

Defining the parameters for TAFE funding reform

In 2006, two research exercises arguably helped set the scene for policy development in this area. The first entailed the secondment to the Australian National University from the Department of the Treasury of Mark Rodrigues, to work with Chris Ryan and me on TAFE funding. The secondment was motivated by the benefits to both institutions of a shared research project.

An important part of the Chapman, Rodrigues and Ryan partnership involved analysing the Household, Income and Labour Dynamics in Australia (HILDA) dataset to determine the lifetime earnings outcomes for TAFE diploma recipients. The information mattered because it would help ensure that a HECS-type scheme could be designed in a way that might work for TAFE graduates without incurring considerable budgetary costs for the Australian Government. Figures 5.4 and 5.5 illustrate what we found based on 2005 information.

The data from the figures, which show that the lifetime incomes for those with TAFE diplomas are relatively high, imply two things:

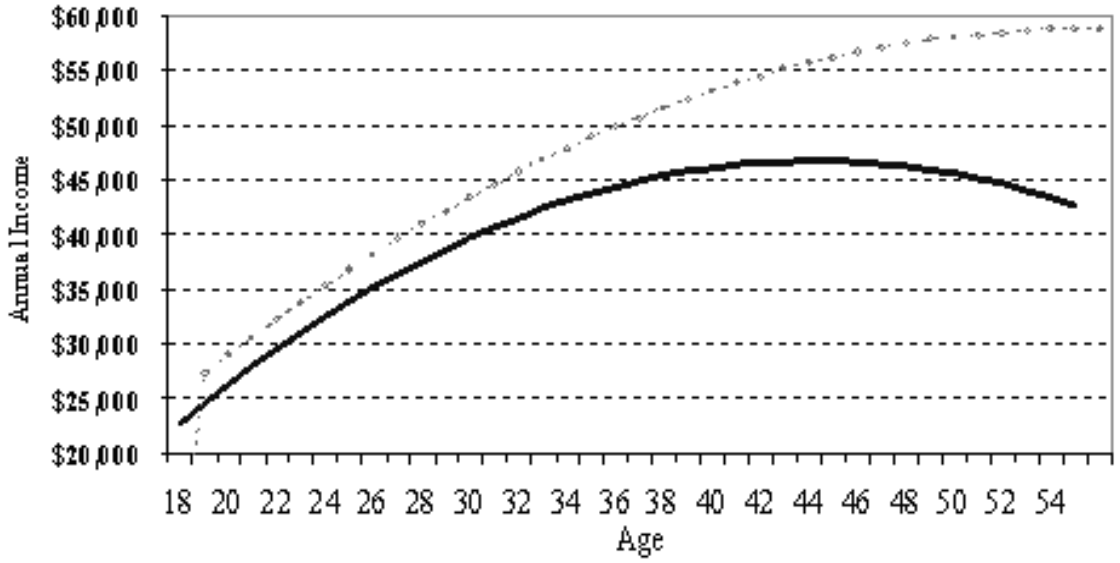
- On average, there seems to be a real private benefit from receiving a TAFE diploma, compared to only finishing Year 12.
- The application of the collection parameters of a HECS variant for loans to TAFE diploma graduates had the potential for the Australian Government to recover the debt relatively quickly.

Both possibilities were confirmed by the technical analysis we undertook reported in Chapman, Rodrigues and Ryan (2008).

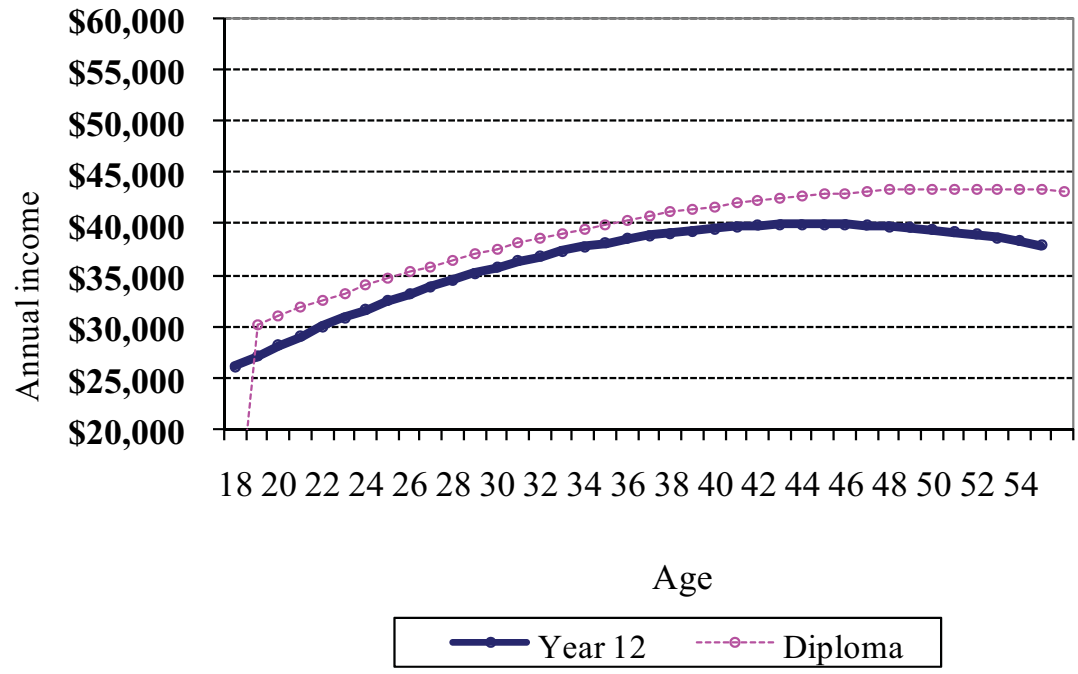
Figure 5.4 **Relative annual earnings with Year 12 certificates or TAFE diplomas**

Evidence of the feasibility of an income-contingent scheme for TAFE fees

Males



Females



Data source: Chapman, Rodrigues and Ryan (2008)

At about the same time, I was helping the Victorian Department of Employment and Training with the same issue — the introduction of an income-contingent loan for students in the TAFE system. As was the case for most State/Territory governments, TAFE up-front fees were becoming a political liability for the Victorian Government.

These research exercises were of use in helping to persuade sceptics of the feasibility of a HECS-type system for TAFE, and with respect to the design parameters of the scheme. But it is highly likely that the need for such policy reform was already a conviction in the minds of many. This certainly appeared to be the case with respect to senior members of the Victorian bureaucracy in 2005–06. Accordingly, the example should be classified as Evidence Type 2, Mr Hyde.

Youth Allowance reforms

In 1998, the Australian Government extended the basis under which full-time tertiary students under the age of 25 years would be considered to be 'independent' of the financial circumstances of their parents and thus eligible for income support grants even if living at home. The additional criteria included working a given number of hours in paid employment over a specified period of time, or earning $18 850 (in 2008 dollars) in a recent 18-month period. This opened the possibility that students could receive non–means tested income support after having a 'gap year' or after being employed at an exceptionally high wage rate for a short period by a family member or friend. It is very likely that this policy development was in part a response to the possible inequities associated at that time with the increase in the 'age of independence' to 25 years.

The number of students in receipt of the 'independent-at-home' (IAH) allowance increased very rapidly in the period from 1999 to 2003, from around 1000 to around 21 500. In absolute terms, the figure has since remained virtually unchanged; it stood at 22 689 in 2007 (Commonwealth of Australia 2008). This represented about 18 per cent of all recipients of Youth Allowance (YA) in 2007.

Possible reforms to YA, including with respect to IAH, were canvassed in detail as part of the Bradley Review. A critical issue for policy was whether or not IAH income support recipients were in fact financially disadvantaged. This is more complicated than it might seem at first blush, because the actual government assistance provided to those in the IAH category is well below the amounts delivered to those in other categories of YA, in which the recipients live away from

their parent or parents.[7] Thus the issue concerns whether or not those on IAH assistance are receiving help from YA in addition to the help that is implicitly assumed to be transferred from parents or guardians in various forms. Unfortunately for our analysis, there was no evidence available to allow confident conclusions with respect to the distribution of resources within households.

Kiatanantha Lounkaew (a PhD student at the Australian National University) and I set about the task of determining the true relative household income situation of students in receipt of IAH. This was made arduous by the fact that there is only poor information available from the government concerning all the relevant economic circumstances of people receiving YA. We were required to access five waves of the HILDA panel data set, adjust the income data for wage inflation, and make assumptions concerning who in the broad category of YA recipients was likely to be in the IAH category. The detail of our process is explained in Chapman and Lounkaew (forthcoming, 2010).

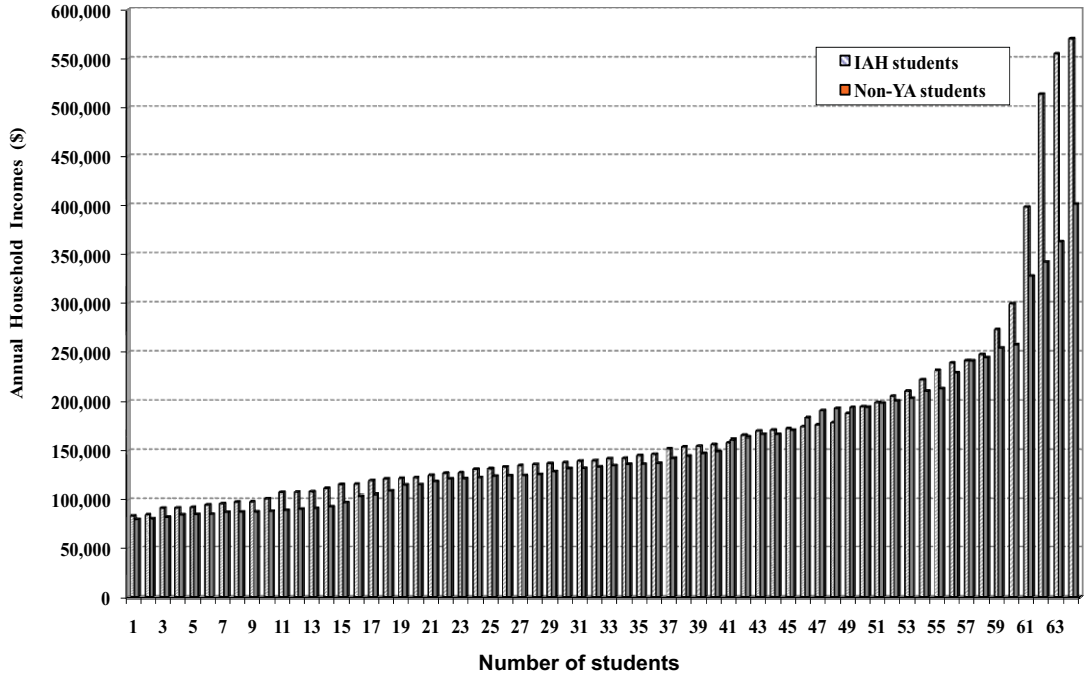
This research led to the critical finding that is illustrated in Figure 5.6. The data show that the household incomes of full-time university students aged less than 25, living with a parent or parents and in receipt of IAH, were essentially the same as the household incomes of otherwise similar young people not receiving YA. This was taken by members of the Bradley Review committee to constitute compelling evidence that this category of YA was poorly targeted and inequitable.

It is a matter of opinion as to whether or not the evidence presented above led the policy change, or instead was highly useful ammunition to form the basis of a policy decision that was inevitable. My view is that the evidence in this case was more in the former category, a perspective influenced importantly through an interpretation of the position taken on the issue in the Bradley Review process (Commonwealth of Australia, 2008). I think a fair assessment of this position is that, without the data, the recommendation for the policy reform was unlikely to be forthcoming, or at least would have meant that a much more nuanced stance was taken in the report from the committee. Accordingly, on balance, I think the right classification for this example is Evidence Type 1, Ms Polyanna.

7 For example, the amount paid to those on IAH is around $220 per fortnight, but the maximum received by those on YA living away from home is about $350 per fortnight and is supplemented by rent assistance for students residing in high-rent areas.

Figure 5.5 **Comparison of the household incomes of the Youth Allowance and non-Youth Allowance groups**

Reforms to the 'independent-at-home' Youth Allowance oxymoron



5.4 Conclusion

Important data in support of policy reform are frequently used as part of the policy-making process. Indeed, it is hard to think of significant policy changes that are not characterised by the use of what is often referred to as 'evidence'. But even if the information can rightly be referred to as evidence, in the sense of having been tested and contested, this does not necessarily mean that a policy procedure should be considered to be entirely consistent with what its proponents refer to as 'evidence-based policy'.

Policy makers — politicians in particular — use research for different reasons. In some circumstances the evidence provides a fundamental basis for policy change; in others, the data can be employed to persuade, or quieten, opponents of a policy which has already been decided for reasons not related directly to the data. This distinction has been used above in a classification of the role of evidence in case studies of four policy matters that I have been involved in.

All of the examples described here used what would generally be considered to be meticulous and scrutinised research methods and data, but it was not always, or even generally, the case that the policy was initiated in response to the evidence. This should not be seen to be criticism of the way policy reform takes place, nor should it undermine the important role played by research in the process. But it is useful for those engaged in policy-related research to be aware of the constraints and limitations inherent in what we are trying to do.

References

Australian Government 1994, *Working Nation*, Canberra.

Banks, G. 2009, 'Evidence-based policy-making: What is it? How do we get it?', ANZSOG Public Lecture, 4 February, http://www.pc.gov.au/speeches/cs20090204. Also reprinted as 'Challenges of Evidence-based Policy', Australian Public Service Commission

Budd, A., Levine, P. and Smith, P., 1988, 'Unemployment, Vacancies and the Long Term Unemployment', *The Economic Journal*, vol. 98, no. 393, pp. 1071 - 1091

Chapman, B.J. 1988, 'An Economic Analysis of the Higher Education Contribution Scheme of the Wran Report', *Economic Analysis and Policy*, vol. 18, no. 3, September, pp. 171–85.

—— 1993a, *Long Term Unemployment in Australia: Causes, Consequences and Policy Responses*, Paper prepared for the Minister for Employment, Education and Training, Australian Government, Canberra.

—— 1993b, 'Long Term Unemployment: The Case for Policy Reform', *The Economic and Labour Relations Review*, vol. 4, no. 2, December, pp. 218–40.

—— 2006, *Government Managing Risk: Income Contingent Loans for Social and Economic Progress*, Routledge, London.

—— and Junankar, P.N. and Kapuscinski, C. 1992, 'Projections of Long-term Unemployment', *Australian Bulletin of Labour*, September, pp. 176–88.

—— and Lounkaew, K. 2009, 'Reforming Youth Allowance: The "Independent-at-Home" Category', Discussion Paper no. 623, Centre for Economic Policy Research, Research School of Social Sciences, Australian National University, Canberra.

——, Rodrigues, M. and Ryan, C. 2008 'An Analysis of FEE-HELP in the Vocational Education and Training Sector', *Australian Economic Review*, vol. 41, no. 1, March, pp. 1–14

Commonwealth of Australia 2008, *Review of Australian Higher Education*, Final Report, December.

Head, B. 2009, 'Evidence-based policy: principles and requirements', Paper presented at Productivity Commission Roundtable on Strengthening Evidence-based Policy in the Australian Federation, Canberra, 17–18 August, 2009, Productivity Commission, Melbourne.

Noonan, P., Burke, G. and White, P. 2004, Policy Developments in VET: analysis for selected countries, CEET Working paper No. 54.

Phillips, D. 2002, 'Policy Reform Ideas for Australian Higher Education', paper presented to CEPR and NISS conference, *The Future of Australian Higher Education*, Australian National University, Canberra, November.

Piggott, J. and Chapman, B. 1995, 'Costing the Job Compact', *The Economic Record*, vol. 71, no. 215, December, pp. 313–28.

Victorian Government 2008, *Skills for the Future,* Melbourne.

Watson, L. 2001, *Who Pays of Lifelong Learning?* Paper presented to the fourth Annual Conference of the Australian Vocational Education and Training Research Association, 28-30 March, Adelaide.

——, Wheelahan, L., and Chapman, B. 2001, 'Gross-Sectoral Funding Issues for Australian Post-Compulsory Education, vol. 45, no. 3, pp. 249-262.

Wheelahan, L. 2001, Bridging the Divide: Developing the institutional structures that most effectively deliver cross-sectoral education and training. National Centre for Vocational Research, Adelaide.

6 Evaluating major infrastructure projects: how robust are our processes?

Henry Ergas and Alex Robson[1]
Concept Economics

Abstract

Australian Government spending on infrastructure projects has increased rapidly in recent years, especially over the course of 2009. In this paper, we examine the processes for project evaluation in the light of the Government's commitment in the 2008–09 Budget to '[infrastructure] decision making based on rigorous cost–benefit analysis to ensure the highest economic and social benefits to the nation over the long term' and to 'transparency at all stages of the decision making process'. We find that, contrary to this commitment, significant projects have been approved either with no cost–benefit analysis or with cost–benefit analysis that is clearly of poor quality. Moreover, despite the commitment to transparency, very little information has been disclosed as to how most projects were evaluated.

To better assess the quality of project evaluation, we examine the largest single project the Australian Government has committed to — the construction of the new National Broadband Network (NBN) — and find that, in present value terms, its costs exceed its benefits by somewhere between $14 billion and $20 billion, depending on the discount rate used. We also find that it is inefficient to proceed with the project if its costs exceed $17 billion, even if the alternative is a world in which the representative consumer cannot obtain service in excess of 20 Mbps

1 We are grateful to Jason Soon for research assistance. The views expressed in this paper are strictly those of the authors and should not be imputed to any of their clients. This is a shortened version of a longer paper that is available on the web at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465226.

(megabits per second) and even if demand for high-speed service is rising relatively quickly. This amount of $17 billion is well below current estimates of the costs the NBN will involve, especially if (as the Government has pledged) the NBN is to serve non-metropolitan areas.

In the longer version of this paper we also examine the cost–benefit assessment undertaken for the second largest infrastructure project the Government has committed to, which involves the construction of a rail link in Victoria. We find that lower-cost alternatives to the project were not taken into account in the evaluation, in particular the option of increasing capacity through improved efficiency and better governance of the rail network. Even taking that exclusion on board, we find that the appraisal that was approved by Infrastructure Australia (or, at least, the only version of that appraisal that has been made available) is seriously flawed, including errors of double counting and manifestly incorrect estimates of project benefits. Absent these errors, the project would generate benefits that fall well short of its costs.

We conclude by noting that high-quality project evaluations will not be made if governments do not see value in them. This appears to be the case in Australia, especially with respect to major projects. Nonetheless, we advance a number of proposals for improving the process, including transparency (which is now largely lacking), serious audits and reappraisal of projects at predetermined milestones, and steps to introduce greater rigour into key aspects of the analysis.

'The core of public finance', as Jurgen von Hagen (2006, p. 464) has succinctly put it, 'is that some people spend other people's money'. This separation of spenders and payers gives rise to a wide range of problems of accountability and control (which economists typically analyse under the rubric of 'principal–agent' problems), reflecting divergences of interest between these parties and the inability of voters and taxpayers to costlessly and perfectly discipline the behaviour of those who spend money on their behalf.

There are broadly three sets of control mechanisms that are commonly used to limit these risks: *ex ante* rules that shape taxing and spending powers; budget processes that signal the opportunity cost of public funds and manage resource allocation so as to control, if not prevent, externalities between spending agents (including those associated with common pool problems); and political competition and accountability that, however effectively or ineffectively, discipline 'poor' uses of resources and reward 'good' uses. The role of formal project appraisal within these control mechanisms, and the effectiveness with which it is implemented within the Australian federation, is the central concern of this paper.

The specific focus is on the processes used in the economic evaluation of major infrastructure decisions. Particularly since the election of the Rudd Labor Government in 2007, very significant increases have occurred in public infrastructure outlays. Many of these decisions involve individual projects whose costs exceed a billion dollars; if those projects' costs exceed their benefits, the result is to make future generations poorer. The public stake in proper project evaluation is therefore great and, indeed, has been stressed by the Government itself. Thus, in its 2008–09 Budget, the Government committed to '[infrastructure] decision making based on rigorous cost–benefit analysis to ensure the highest economic and social benefits to the nation over the long term' and to 'transparency at all stages of the decision making process'.[2] However, serious concerns have been expressed about the extent and quality of project evaluation in Australia.

So how robust are our project evaluation processes? In examining this question, we start by setting out the nature and role of cost–benefit analysis, and especially its bearing on efficient resource allocation and on the control of principal–agent problems in government. That discussion highlights just how important cost–benefit analysis is to serious project appraisal, and to helping to control the risks inherent in a situation where very large projects, offering highly concentrated benefits but with very diffuse costs, are being vigorously advocated by powerful private interests.

On that basis, we examine the situation in telecommunications. In essence, neither the Howard Government (1996–2007) nor its successor placed any weight on systematic analysis of the costs and benefits of major telecommunications decisions. The most spectacular recent instance is, of course, the decision to build a 'National Broadband Network' (NBN) with significant taxpayer funding. As the Government has stated that no cost–benefit analysis of this decision has been, or will be, undertaken, we carried out such an assessment. Our results suggest that the incremental benefits of the NBN, when compared to the counterfactual scenarios, do not justify the incremental costs.

Given that evaluation of project decisionmaking in telecommunications, we turn to transport in the longer version of this paper. We outline some major trends in transport cost–benefit analysis in Australia, including those resulting from the creation of the Building Australia Fund and the establishment of Infrastructure Australia as a policy advisory body. To assess the quality of the evaluation processes, we undertake a detailed analysis of the east–west rail project in Victoria. Although that project involves several components, some of which are not now proceeding (or have been deferred), it remains extremely large and has now received very substantial funding from the Commonwealth. However, this is a

2 2008–09 Budget Paper No. 1, (Statement 4, pp. 14–15).

project which, even in its sponsor's cost–benefit analysis, had benefits that were not far above costs. Our examination of that cost–benefit analysis in the longer version of the paper raises a number of concerns, including double-counting of benefits and substantial difficulties with the approach the cost–benefit analysis adopts to the calculation of the project's 'wider economic impacts' (essentially, pecuniary externalities associated with the project).

6.1 The nature and role of cost–benefit analysis

In essence, cost–benefit analysis is a technique for evaluating collective decisions that hinges on the comparisons of the costs of a proposal to its benefits, where costs and benefits are valued in monetary terms. Cost–benefit analysis asks whether the sum of the amounts the individuals who comprise the community at issue would be willing to pay for the project to proceed exceeds the costs of that project. Generally, a project enhances wealth — in the sense of the aggregate monetary valuation of the community's resources — if it meets a properly specified cost–benefit test.

Cost–benefit analysis can be viewed from four complementary perspectives.

First, cost–benefit analysis is related to (though not identical with) the basic equi-marginal condition for overall efficiency in resource allocation. Thus, given a cardinally measurable objective function and perfect knowledge of the effect on welfare of any decision, it is a condition of an optimal set of decisions that the marginal dollar of public expenditure has a benefit equal to that of the marginal dollar of private expenditure (thus assuring that the overall level of public expenditure is optimal) and that the benefit of a marginal dollar of public expenditure is equalised across programs, projects and project elements. Because cost–benefit analysis aggregates willingness to pay across agents with different marginal valuations of income, it is not a perfect measure of underlying utility (and hence cannot be treated as an ideal social welfare function); nonetheless, taking that important caveat as given, one would at least question whether a set of public decisions was optimal if it did not maximise the aggregate benefits obtainable for given aggregate costs or minimise the aggregate costs required to obtain a given aggregate benefit, in each case measured using cost–benefit analysis.

Second, set against the backdrop of a given portfolio of projects, cost–benefit analysis can be used to evaluate whether one or more public projects should be added to or removed from that portfolio. In other words, cost–benefit analysis is a tool that can be used to assess whether wealth (the difference between the aggregate valuation of outcomes and the cost of obtaining those outcomes) would be increased by the decision to, say, proceed with a particular project, compared to the relevant

alternatives (which may involve doing nothing, deferring or otherwise varying the project, or proceeding with an alternative project).

Third, cost–benefit analysis is an instrument that the principals in public sector governance can use to improve the decisions taken by their agents, and to enhance their supervision of those agents (see, for example, Adler and Posner 2006, Posner 2001 and Spence and Cross 2000). As a result, the requirement to carefully assess and report the costs and benefits of decisions can improve the quality of decision-making and reduce the information asymmetry between principals and agents. In doing so, it can:

- help reduce the risk of 'capture', in which the agent's decisions, rather than reflecting the interests of the principal, come to be determined either by the agenda of self-interested third parties or by the agent's own interests and aspirations
- help correct 'policy bias', which is a situation in which those working in an agency have policy commitments that differ from (and may undermine) those of the public
- help overcome 'shirking', in which agents do not exercise as much diligence in taking decisions as would be warranted
- help disclose and correct the cognitive biases that affect decisionmaking
- increase consistency in decisionmaking, both by standardising the information base on which decisions are taken and by highlighting anomalies, such as differences between project appraisals in the valuation of common elements
- improve performance auditing and accountability by providing a standardised *ex ante* statement of key expected values for costs and benefits.

Ultimately, all of these effects mean that cost–benefit analysis is never merely an analytical tool; rather, as Aaron Wildavsky (1966) emphasised many years ago, it is inevitably an instrument in shaping bureaucratic structure and process, both within each public sector body and between that body, the other elements of the public sector with which it interacts, and the wider political system.

Fourth and last, cost–benefit analysis can be an anchoring device that reduces undesirable policy instability.

6.2 Telecommunications

We start by explaining the relevant context and then examine recent decisions in the light of cost–benefit analysis.

Context and background

Ergas (2008a) sets out the background to recent telecommunications decisions. Two trends dominated the period leading up to the 2007 change in government.

First, an impasse developed in relations between Telstra and the Australian Government over the issue of upgrading the Australian telecommunications network to higher broadband speeds.

Second, the Australian Government engaged in a wide range of spending programs (with appropriations totalling close to $4 billion, in 2008 prices) aimed at promoting service upgrading, usually in regional areas, and implemented an ever broader and more draconian range of quality of service regulations.

None of these spending initiatives or quality of service regulations were ever subjected to proper cost–benefit analysis (or if such analysis was undertaken, it was never disclosed). However, an analysis by one of the authors found that in 1999 the total benefits associated with addressing claimed service quality problems (including in terms of consumer gains and network-related cost savings) were between $644 million and $713 million in present value terms over the length of the project life. These benefits were outweighed by the costs which (again in present value terms) were estimated at $1387 million over the project life (Ergas and Hardin 1999).

The lack of attention to systematic evaluation of the costs and benefits of policy initiatives has continued under the Rudd Government. The Minister for Broadband, Communications and the Digital Economy, Senator Stephen Conroy, when asked by the opposition whether a cost–benefit study of the proposed expenditure had been carried out, said (according to a report in the *Communications Day* of 13 May 2009), that there was 'no need' for such a study, as 'Labor's commitment to build a high speed broadband network has been clear … A range of studies have been carried out all over the world that have investigated the economic impact of broadband.' (Bartholomeusz 2009)

Since then, one study, by Professor Joshua Gans (2009), has been submitted as evidence to a Senate inquiry into the NBN. Although its author notes that the

calculations are essentially back of the envelope, the submission suggests that the social benefits of the NBN will exceed the costs. However, these calculations are seriously flawed. These deficiencies are summarised in Appendix A of the longer version of this paper. Even more seriously, however, Professor Gans's submission uses the wrong test for assessing whether a project is worth while: it compares total costs and benefits, when the correct test is whether the incremental gains from the project (relative to network capabilities in the base case) exceed the associated incremental costs.

Before turning to examine the project's costs and benefits, it is useful to undertake a wider consideration of the relevant decision. In particular, it is uncontroversial that sensible policy evaluation requires a specification of the problem to be addressed and of the policy options for addressing it. As a result, it is reasonable to ask what precise problem the NBN is intended to resolve, and what other means might have been used to resolve it.

The Government's primary concerns appear to be with the availability of broadband access and its price. However, the data on availability that the Government has cited actually refers to take-up of broadband services, and hence might be more indicative of the demand for broadband than of its supply. This is all the more probable given that broadband availability appears to greatly exceed demand, with some 80 per cent of PSTN lines being connected to ADSL2+ enabled exchanges and close to 50 per cent of copper lines being short enough deliver very high speeds. Moreover, competing hybrid fibre-coaxial networks (which currently deliver up to 30 Mbps but which can, at relatively low cost, be upgraded to much higher speeds) either pass or run very close to some 60 per cent of premises.[3] Despite all of this, high-speed fixed services account for a relatively small share of total broadband services.[4] It is therefore not implausible that penetration levels simply reflect

[3] Low incremental costs for hybrid fibre-coaxial upgrade are discussed in the Soria and Hernández-Gil paper (2009), as well as in the Telstra documents (2008a, 2008b and 2008c). It is worth noting that, according to the *Communications Day* of 31 July 2009, Telstra will upgrade its hybrid fibre-coaxial network in New Zealand to 100 Mbpss for NZ$10 million. The cost of deploying the proposed FTTP network in those coverage areas is likely to be at least 10 to 20 times greater.

[4] According to Telstra's most recent annual results (released on 13 August 2009), Telstra's wireless broadband subscriptions doubled over the year to reach over 1 million (this figure includes data card subscribers only; it does not include customers with 3G handsets). In contrast, Telstra's high-speed services (20 Mbps plus) had 241,000 high-speed subscribers in June 2009, up from 160,000 the previous year. This represents about 10 per cent of Telstra's broadband customers. See http://www.telstra.com.au/abouttelstra/investor/docs/tls685-fyr2009esults announcement.pdf.

consumers' low valuations of the incremental benefits of higher speed fixed network access.

A similar picture emerges regarding business access to high-speed broadband. Competing, ubiquitous fibre networks cover all of the capital city central business districts (CBDs). Larger business premises outside the CBDs are almost always on direct fibre optic connections, even in non-metropolitan areas, as are premises such as hospitals and government offices. Smaller businesses have access to business parks, which are almost invariably on fibre access networks, and those smaller businesses that operate in activities where high-speed communications are an important element tend to locate in those business parks (where they can also benefit from other economies of agglomeration). Symmetric high-speed services over copper (such as BDSL) are available in virtually all urban locations and in many regional centres. There is, in short, no evidence of any absence of business access to high-speed broadband networks.[5]

There is also no evidence that suppliers of social services lack access to high-speed services — indeed, the opposite is the case. In other words, availability does not appear to be the constraint the NBN deployment assumes.

Australian broadband prices are in the upper half of OECD comparisons. However, prices in a number of countries are distorted by subsidies, and those subsidies would need to be added back, along with a mark-up to reflect the marginal social cost of funds, for a welfare comparison to be made. Additionally and importantly, there is significant competition in Australian broadband supply and key input prices are regulated. Service supply to CBDs and business parks is intensely competitive, as is the wiring of new residential estates. As for established premises in metropolitan areas, broadband is widely provided by Telstra's competitors using Telstra's Unconditioned Local Loop Service (ULLS), a regulated service that provides third-party access to the copper pair. As is shown in Appendix B of the extended version of this paper. Australian regulated ULLS charges are relatively low in urban areas, while take-up of ULLS has increased very rapidly.

As for non-metropolitan areas, the case that supply is failing to keep up with demand is also weak. Overall, these outcomes, like those above, suggest that the

5 Residential mobility and new household formation rates in Australia are relatively high. As a result, consumers who value high-speed access will tend to move to locations at which access is available and will incur low incremental costs from doing so. We are unaware of any evidence of a residential housing price premium associated with access to high-speed broadband. These elements suggest that latent demand, and welfare losses from lack of access, are likely to be low.

primary obstacles to take-up may lie in low customer demand, which implies low customer valuation of any new network.

This is not to say that there are no issues with respect to investment in, and upgrading of, Australia's telecommunications network — the opposite is true. As argued in Ergas (2008a, 2008b), the current telecommunications-specific access regime vests enormous and unwarranted discretion in the regulator. In this industry, as in others, such discretion creates a risk of time inconsistency; that is, of regulatory decisions which *ex post* expropriate the returns on socially worthwhile investments.[6] To that extent, an option for the Government would have been that of reforming the regulatory arrangements (along lines already adopted in the energy industries) so as to provide greater investor confidence, and then seeing whether socially desirable investment in network upgrading materialised.[7] As for areas where service is commercially unviable, these could have been dealt with at relatively low cost through a voucher scheme, which would have the merit of being technologically and competitively neutral (Ergas and Ralph 2008). There is, however, no evidence, at least in material disclosed to date, that the costs and benefits of those options were assessed relative to the option of simply building a new network.

The economics of the new network

What then can be said about the costs and benefits of the new network? To examine the underlying economics, we have used a cost model developed by Concept Economics.[8] The model describes the rollout of a fibre to the home (FTTH) network with a footprint covering 90 per cent of the Australian population by modelling the construction cost of new infrastructure.

---

6 Simply put, time inconsistency refers to situations where a policy that is optimal (from the point of view of the policymaker) *ex ante* turns out not to be the optimal policy *ex post*. If the policymaker cannot commit to a policy, it may then find itself wanting to change its policy *ex post* (say, after a regulated firm has made an irreversible investment decision), regardless of what it promised *ex ante*. Such an approach to policy is said to be time-inconsistent (Kyland and Prescott 1977). Specific applications of the concept to regulated industries can be found in the literature (Evans, Levine and Trillas 2008; Guthrie 2006; Levine, Stern and Trillas 2005). Ergas (2009b) provides a test of whether ACCC decisions in telecommunications are time-inconsistent (with the conclusion that they are).

7 Obviously, some care is required in the design of such an option. In particular, if there remains a material threat of the Government expropriating the returns on that investment, for example by subsequently building a network of its own, then socially desirable investment may be deterred. Jullien, Pouyet and Sand-Zantman (2009) systematically discuss the conceptual issues involved.

8 The model was developed by Dr Dieter Schadt, and we are grateful for his assistance in this respect. Obviously, he bears no responsibility for our use of the model's results.

As regards capital costs, we have assumed a weighted average cost of capital (WACC) in which the cost of equity is determined according to the capital asset pricing model. This reflects three considerations. First, this investment substitutes for private sector investment in competing infrastructure. Use of any other cost of capital than that for the private sector alternative will distort resource allocation between the public and the private sector (see, for example, Steiner 1974). Second, the Government has confirmed on a number of occasions that it intends the project to earn a commercial rate of return, suggesting that it values capital devoted to this project at that rate of return. Third, investing in a new broadband network has a high level of systematic risk. As a result, the Arrow–Lind conditions for use of the risk-free rate as the discount factor (which depend on the assumption that the benefits of the investment are independent of variations in overall incomes) do not hold in this instance, and the cost of the project to taxpayers must reflect the project's systematic risk.[9]

The model is designed to allow testing of the sensitivity of the results to a range of variables. Setting these variables to their base case levels (which involves a GPON-gigabit passive optical networking-architecture), we estimate a final retail cost per customer (on a nationally averaged basis) of just over $170 per month. This amount is the cost of the access network plus the cost of backhaul to the service provider's network, and an allocation for usage and other retail costs. It is, in other words, broadly comparable to the charge for a broadband service, minus the cost of any content.

While both the input assumptions and the outcomes are broadly consistent with studies undertaken in other countries (see, for example, Analysys Mason 2008), the cost estimates are sensitive to a range of assumptions, including, with respect to consumer take-up rates and cutover arrangements, the extent of aerial deployment, the project cost of capital, achievable operational efficiency improvements and the quality of service provided. Variations in those parameters lead to a possible range for unit per customer costs of between $125 per month and $225 per month. There is also very significant variation in costs between metropolitan and non-metropolitan areas. Thus, for the most likely estimate of $170 per month, unit costs in metropolitan areas are $133 per month, while those in non-metropolitan areas are just under $380.

9 In a classic article, Arrow and Lind (1970) showed that if a government project is 'small' (in relation to the total wealth of taxpayers) and 'the returns from a given public investment are independent of other components of national income', then the social cost of risk for project flows that accrue to taxpayers tends to zero as the number of taxpayers tends to infinity. The required assumption, in other words, is that the returns from the project are not related to (in the sense of being dependent on) income from other investments in the economy.

Given these sensitivities, we have run a variant which seeks to minimise unit costs, including by assuming that eventually all premises will subscribe to the service. This variant, which also sets initial service quality at relatively low, but perhaps not inappropriate, levels (in terms of the Committed Information Rate used to dimension backhaul) and somewhat reduces the WACC, only slightly reduces unit retail costs in metropolitan areas but could reduce unit retail costs in non-metropolitan areas to around $280 per month. Nonetheless, even these costs are high compared to current charges. They are about double the level of current non-content payments for telephony and broadband service (that is, the sum of the monthly rental and of the non-content component of DSL charges) in metropolitan areas and three or more times those in non-metropolitan areas.[10]

These costs need to be compared to alternatives. The most straightforward counterfactual involves continuation and some upgrading of the current copper-based network alongside progressive upgrading of the hybrid fibre-coaxial, with copper delivering speeds of some 20 Mbps to 40 Mbps and the hybrid fibre-coaxial delivering speeds of 50 Mbps to 100 Mbps. The costs of this scenario could be in the order of one-third those of the NBN in metropolitan and regional areas, up to around 80 per cent of the population. Remaining areas would primarily be served by wireless, at costs that would be around one-half those of the NBN, with speeds of 10 Mbps to 30 Mbps. Regulatory reform that increased investment certainty would make the progressive upgrading that took place in this counterfactual both quicker and more extensive.

Incremental cost-based retail network charges for broadband service per connectable premise under the counterfactual would therefore be in the order of $50–$70 per month in metropolitan areas, rising to around $80–$100 per month in regional areas, with a difference relative to the NBN scenario of around $75 per month in metropolitan areas and of $120 per month in regional areas (noting that the regional areas have less population coverage than is envisaged for the NBN, so that the like-for-like comparison involves assuming a regional cost-based rate in the NBN of around $210). Broadly speaking, the additional outlays (of $75 per month in metropolitan areas and of $120 per month in regional areas) allow speeds to rise to 100 Mbps in one step. However, this benefit is somewhat qualified by the fact that deployment of the new network may take 7 to 10 years (if not longer), but the

[10] They are even higher when compared to the access payments made by the average residential premise, remembering that about 30 per cent of households do not subscribe to any form of broadband service. Relative to those current average monthly payments, they are more than twice the current average monthly payments in metropolitan areas and about four times those in non-metropolitan areas.

prospect of that deployment may prevent the somewhat more limited, but sooner, upgrades that would otherwise have occurred from occurring.

The question then is whether the valuation of the incremental speed associated with the NBN outweighs the incremental costs. While there are some symmetric services (such as very high-quality videoconferencing) that could benefit from higher speeds, the difference in delay and overall service quality between (say) 30 Mbps and 60 Mbps would only rarely be discernible. This is all the more so as once the access network operates at reasonably high speeds; the relevant constraints on service quality are likely to come from performance in the core network (that is, the links between the first point of traffic aggregation and the global Internet), with further increases in access network speeds having little effect. Holding all else constant, it is therefore reasonable to expect the valuation of further reductions in download time to decline as average download times themselves decline (that is, as speeds increase). The median consumer's willingness to pay (WTP), taken as a function of service bit rate, would, in other words, increase more slowly for successive increases in speed.

This can be illustrated using the standard Becker (1965) time-allocation model. Naturally, the incremental benefits are higher for those earning higher wages (that is, who have a higher opportunity cost of time), but, all else being equal, the incremental benefits decline with the square of the speed.[11] For any given set of applications, the valuation of speed will therefore be significantly concave, though the location of the valuation curve will shift over time, as 'bandwidth-hungry' applications develop and as a greater number of consumers attain a utility level from access to broadband that induces them to obtain the service (that is, that exceeds the service's start-up costs). Appendix C of the extended version of this paper details the model.

Incremental willingness to pay and net benefits for the new network

Given these considerations, we have undertaken an assessment of the costs and benefits for the project. As with any such assessment, a substantial number of assumptions need to be made. In this section, we explain the approach we have adopted.

[11] Goolsbee and Klenow (2006) use Becker's framework to compute the consumer benefits of access to the Internet, but they do not examine the welfare effects of greater download speeds.

Computing incremental benefits of a project requires specification of a baseline scenario with which to compare the project scenario. We consider three such scenarios, which entail the following alternative comparisons.

Scenario A

- Baseline: The median consumer initially has speeds of 10 Mbps, which gradually increases to 60 Mbps by year 6, and remains on 60 Mbps.
- Project: The median consumer initially has speeds of 10 Mbps, which gradually increase (but at a slightly slower rate than the baseline) to 60 Mbps by year 9, and then has speeds of 100 Mbps from year 10 onwards.

Scenario B

- Baseline: Same as for scenario A.
- Project: The project is delayed by five years, during which time the median consumer is on the same path as the baseline. The median consumer then goes on to 100 Mbps at a later time than with the project Scenario A — from year 15 onwards.

Scenario C

- Baseline: Same as for scenario B.
- Project: Same as project for scenario B, but the project is targeted at consumers with a relatively high willingness to pay — those consumers in the top quintile. This is, in other words, a targeted version of the project, with the aim of serving only high WTP areas.

These speed adoption paths are plotted in figures 6.1 and 6.2.

Figure 6.1 **Time path of speeds: scenario A**

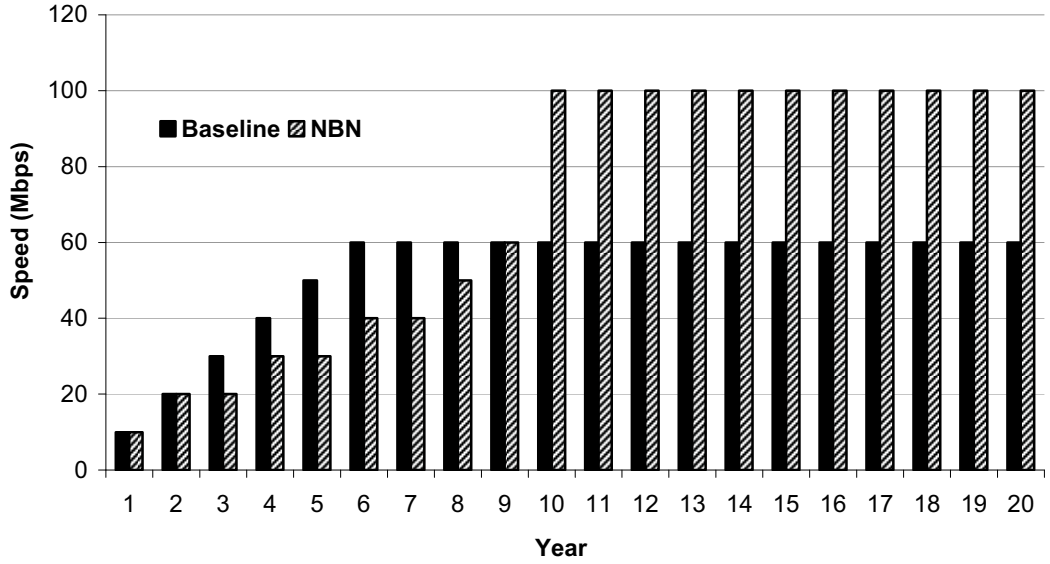
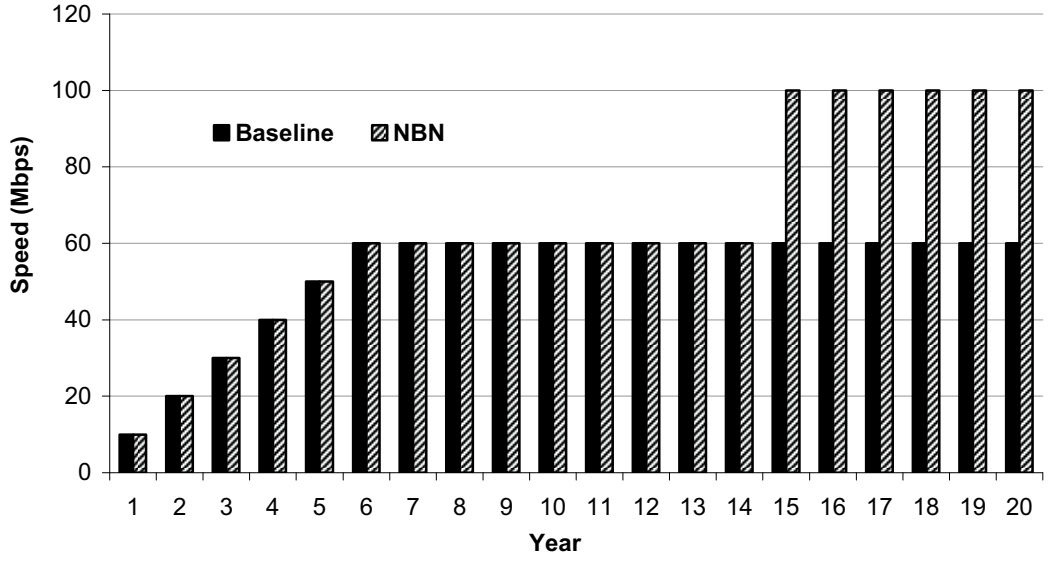


Figure 6.2 **Time path of speeds: scenarios B and C**



For scenario C, we assume the same WTP curves, except that the relevant consumer that is targeted when the project is built has a much higher WTP. By construction, the top 25 per cent of consumers are assumed to have initial valuations exceeding $100, and we take this consumer as the representative consumer that is targeted by the project under scenario C. We also assume that the growth rate of this consumer's WTP is 5 per cent per year (figures 6.3 and 6.4)

Figure 6.3 **Time path of willingness to pay curves: scenarios A and B**

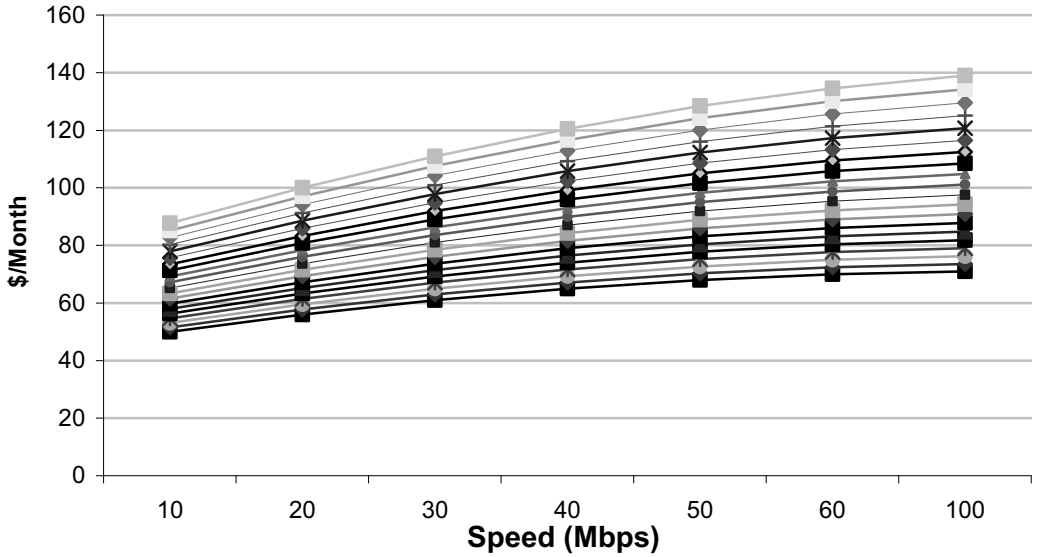
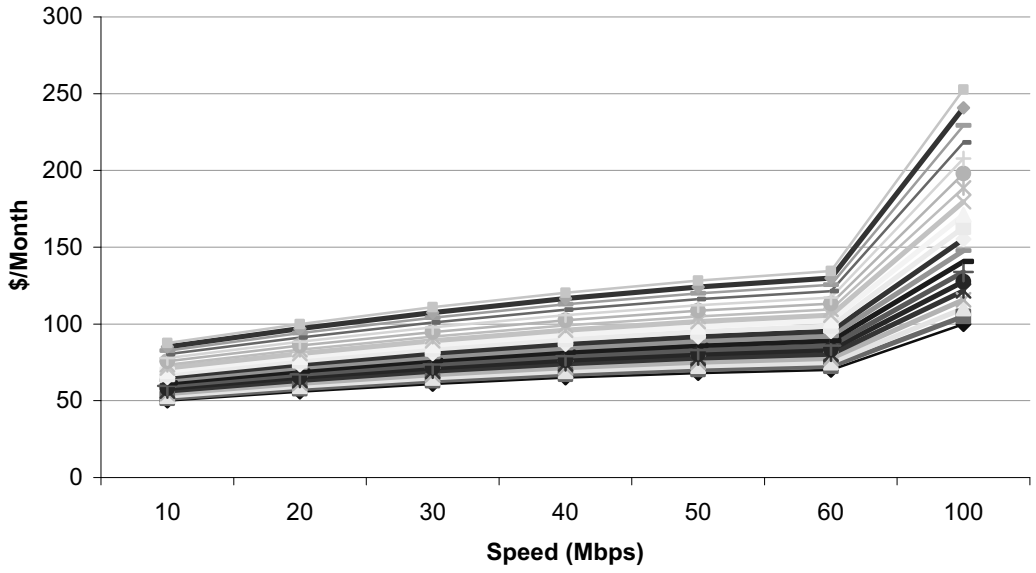


Figure 6.4 **Time path of willingness to pay: scenario C**



The next step is to combine the speed adoption path and the WTP curves to calculate a WTP curve for the baseline and the project under each scenario, and also compute the difference in the path of WTPs under each scenario. This gives us the incremental WTP curve — it is the path of benefits that the representative consumer would receive if the project went ahead, instead of the baseline.

These are plotted in figures 6.5 to 6.7.

Figure 6.5 **Path of willingness to pay under the baseline and the National Broadband Network: scenario A**

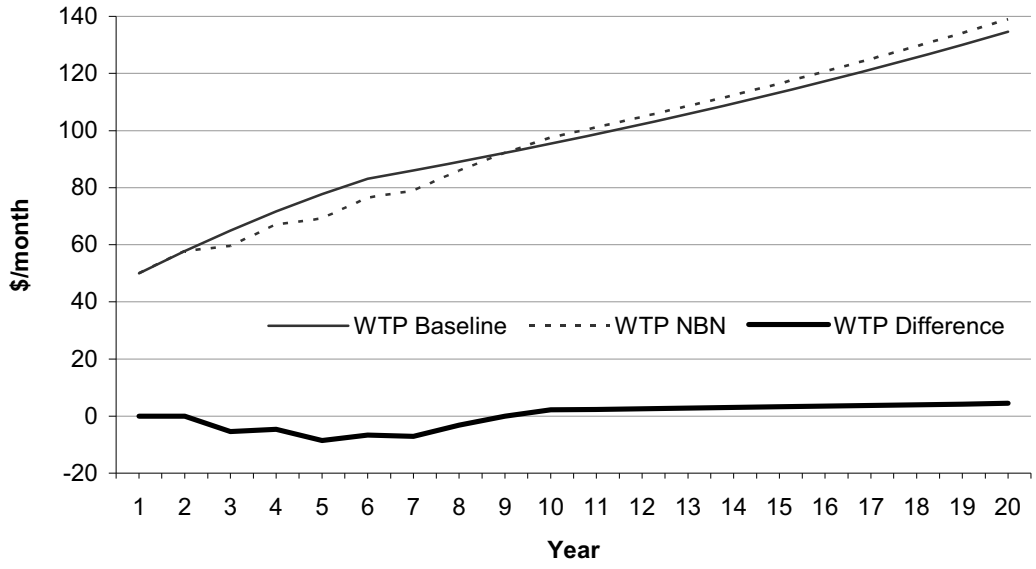


Figure 6.6 **Path of willingness to pay under the baseline and the National Broadband Network: scenario B**

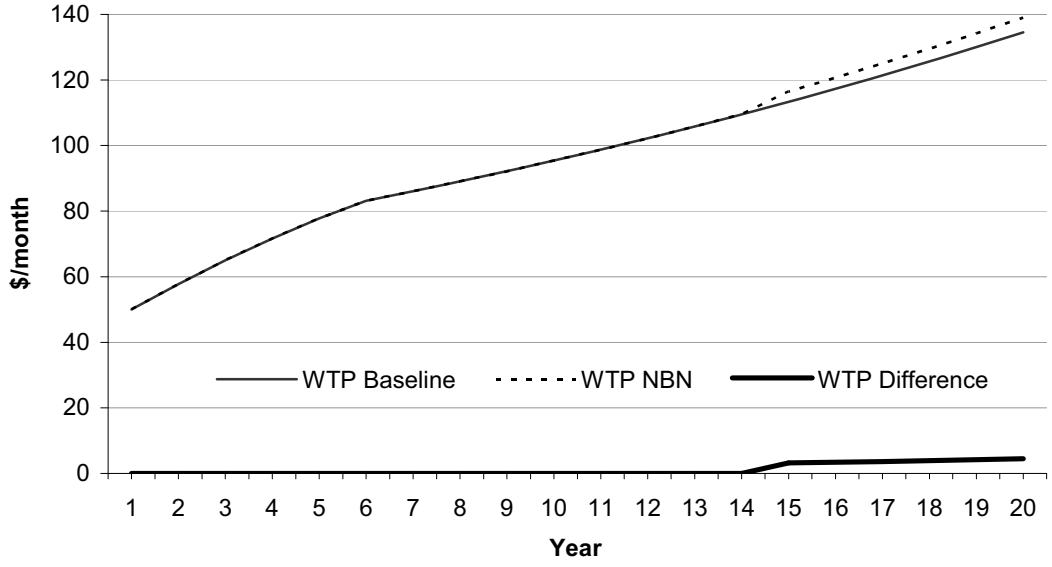
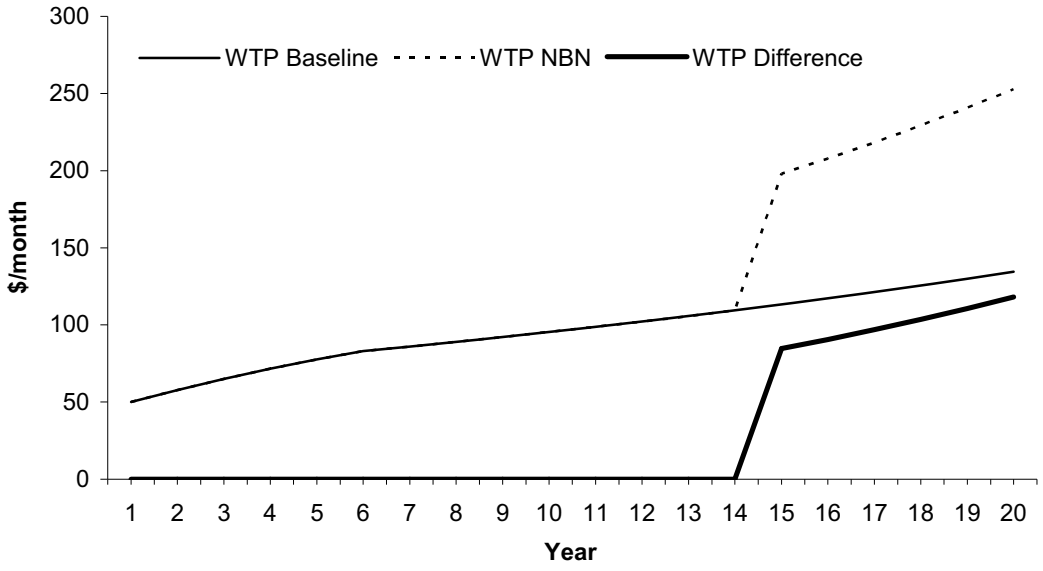


Figure 6.7 **Path of willingness to pay under the baseline and the National Broadband Network: scenario C**



We then compute the present value of the stream of benefits under each scenario, using a range of discount rates. The numbers in the tables are the present value of the consumer's WTP, expressed in dollars per month. Thus, the number in the first row of the first column ($1273) is the present value of the future stream of benefits that the consumer expects to receive.

The tables also compute the 'monthly constant equivalent', which is the constant amount that a consumer with the relevant discount rate would be willing to pay in each and every month over the next 20 years to receive the given stream of benefits. So, for example, under scenario A, a consumer with a 4 per cent discount rate would be willing to pay $0.52 every month (rounded up to $1 in the table) for the next 20 years to not have the NBN, and instead receive the benefits under the baseline.

To arrive at a final assessment of costs and benefits, we subtract the incremental costs computed earlier from these incremental benefits. Note that under scenario A the incremental benefits are negative, and so accounting for the incremental monthly costs that were computed earlier (of around $75 per month in metropolitan areas and of $120 per month in regional areas), the NBN has incremental net benefits that are negative. For all the other scenarios, the incremental benefits of the NBN are far below the incremental costs; indeed, it is difficult to conceive of credible scenarios for the NBN that would make its incremental costs fall below the incremental benefits (that is, result in the project yielding net benefits to Australia).

Indeed, in all of the scenarios, the incremental upgrading path is always the most socially beneficial.

Table 6.1 **Incremental benefits under various scenarios**

		NPV of per month benefits			*Monthly equivalent*		
	Discount rate	*Baseline*	*NBN*	*Increment*	*Baseline*	*NBN*	*Increment*
	%	$	$	$	$	$	$
Scenario A	4	1237	1228	–9	91	90	–1
	8	846	834	–13	86	85	–1
	12	612	599	–13	82	80	–2
Scenario B	4	1237	1249	11	91	92	1
	8	846	852	6	86	87	1
	12	612	615	3	82	82	0
Scenario C	4	1237	1540	303	91	113	22
	8	846	1002	156	86	102	16
	12	612	695	83	82	93	11

NPV = net present value

Sensitivity analysis of willingness to pay paths

To what extent do these results depend on the willingness to pay curves? To examine this question, we have conducted a sensitivity analysis on the WTP assessment, by examining 'enhanced' WTP curves in each of the three scenarios.

The results of the enhanced WTP analysis are very similar to the standard analysis. The ranking of the three scenarios remains unchanged, with the delayed project (scenario B) and the targeted project (scenario C) becoming slightly more attractive from an incremental benefit point of view. The incremental benefits under scenario A actually fall and become more negative under the enhanced WTP setting. In other words, increasing the willingness to pay for higher speed reduces the attractiveness of the NBN option, essentially because it also increases the density of demand in the midspeed tier (and hence increases the relative value of the options that involve incremental development of the access network).

Put slightly differently, the enhanced WTP curves have higher marginal WTP at lower speeds relative to the original analysis. Under the NBN the consumer misses out on those relatively high marginal gains in the early years, even though the consumer eventually receives high absolute benefits. This fact, combined with the logic of discounting, means that scenarios B and C become more attractive, while scenario A becomes less attractive.

Table 6.2 **Incremental benefits under various scenarios: enhanced WTP**

		NPV of per month benefits			*Monthly equivalent*		
	Discount rate	*Baseline*	*NBN*	*Increment*	*Baseline*	*NBN*	*Increment*
	%	$	$	$	$	$	$
Scenario A	4	1608	1609	1	118	118	0
	8	1087	1070	–17	111	109	–2
	12	776	753	–23	104	101	–3
Scenario B	4	1608	1648	41	118	121	3
	8	1087	1108	21	111	113	2
	12	776	787	11	104	105	1
Scenario C	4	1608	1918	310	118	141	23
	8	1087	1247	160	111	127	16
	12	776	861	85	104	115	11

NPV = net present value

Overall, the results are relatively robust because WTP is concave in speed, network coverage and in the rate at which upgrades are deployed, while costs are convex at a discontinuity (the upgrade to FTTP).[12] Moreover, the results reported above tend to understate the consequences of this fundamental feature of the situation, as we consider a median user, while there are substantial numbers of users — especially in non-metropolitan areas — who have low willingness but very high costs to serve.[13] In the counterfactual, the loss incurred on these users is limited by the more limited coverage of the upgrading; in the NBN, these costs are incurred in full and relatively soon.

Comparison of project costs and benefits

To examine the net benefits and costs of the NBN, we examine a scenario (scenario D) that is intentionally conservative as far as service quality is concerned, as it involves speeds under the base case rising to only 20 Mbps, which is less than the hybrid fibre-coaxial networks can currently provide.

[12] Costs are, in other words, concave in speed up to 30–60 Mbps and then leap at the discontinuity. Costs are always likely to be convex in the geographical breadth of deployment and in the speed of deployment, while the WTP gains in each of these dimensions are likely to be concave.

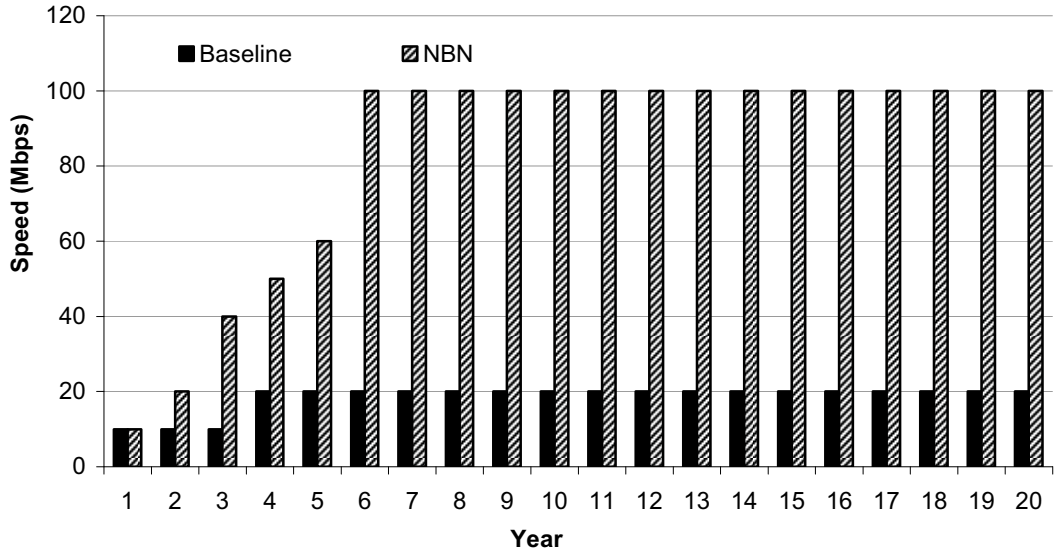
[13] This probably reflects the fact that WTP is correlated with human capital endowment, and human capital — especially that associated with 'information' activities — tends to be concentrated in metropolitan areas (see O'Flaherty 2005).

Scenario D

- Baseline: The median consumer initially has speeds of 10 Mbps, which increase to 20 Mbps in the fourth year and remain there.
- Project: The median consumer initially has speeds of 10 Mbps, which gradually increase to 100 Mbps by the sixth year of the NBN project, where they remain.

These speed adoption paths are plotted in figure 6.8.

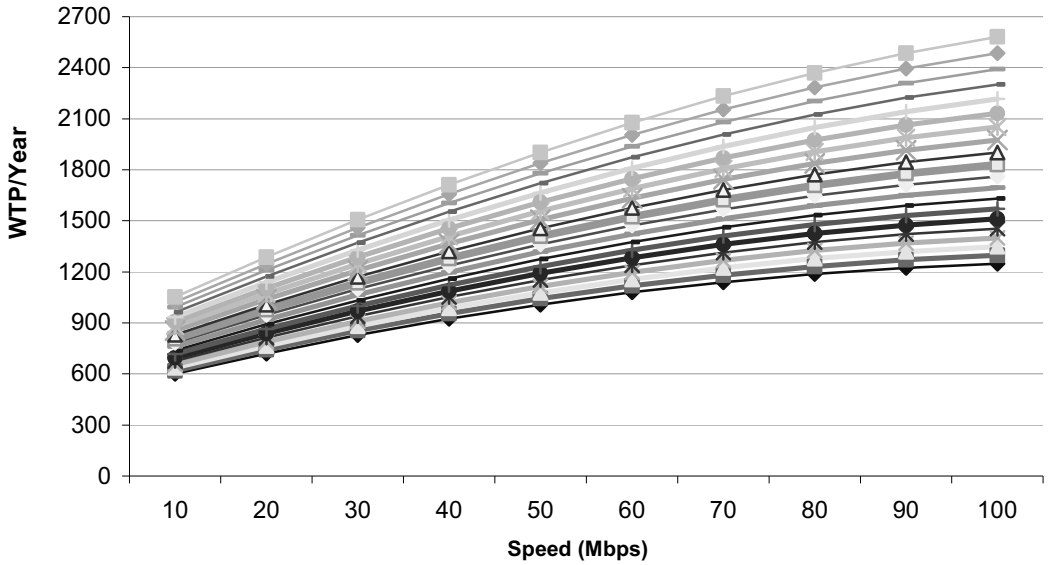
Figure 6.8 **Time path of speeds under scenario D**



For consumer willingness to pay, we assume a monthly WTP of $50 for 10 Mbps, increasing to $104 for 100 Mbps. To estimate aggregate willingness to pay, we assume that all consumers are alike. We also assume an annual growth rate of 3 per cent in WTP at the lowest speed, but assume that the growth rate increases as we move up the WTP curve. Thus, we assume an annual growth of 3 per cent for WTP for 10 Mbps, with the growth rate rising to 3.9 per cent for 100 Mbps. The initial annual WTP curve for scenario D and its growth rate over time is shown in figure 6.9.

Our next step is to combine the speed adoption path and the WTP curves to calculate a WTP curve over time for the baseline and the project, and also compute the difference in the path of WTPs under each scenario. This gives us the incremental WTP curve — it is the path of benefits that the representative consumer would receive if the project went ahead, instead of the baseline. These are plotted in figure 6.10.

Figure 6.9 **Time path of annual willingness to pay curves: scenario D**



Under the scenario D baseline, we assume that retail prices are $30 per month in metropolitan areas, and $50 per month in non-metro areas, which gives a national monthly cost recovery retail price of $32.90 (assuming an 85%–15% split between urban and non-urban areas).

For the NBN, under scenario D, and the assumption of a CIR (Committed Information Rate) of 1 Mbps, the engineering cost model provides estimates of break-even retail prices of $128 per month in metro areas, and $313 in non-metro areas, for a national average cost recovery price of $155 (again assuming a 85–15 per centsplit between metro and non-metro areas).

To compute aggregate costs and benefits, an assumption must be made about the path of demand. Under scenario D, the NBN engineering cost model assumes an S-shaped take-up pattern over time, with 50 per cent of the population taking up the service by year 6 and a saturation rate of 80 per cent. For the baseline case, we assume a slightly more rapid take-up rate, with the same starting percentage as under the NBN but with a final saturation rate of 90 per cent. These two demand profiles are shown in figure 6.11.

Figure 6.10 **Path of annual individual annual WTPs under the baseline and National Broadband Network: scenario D**

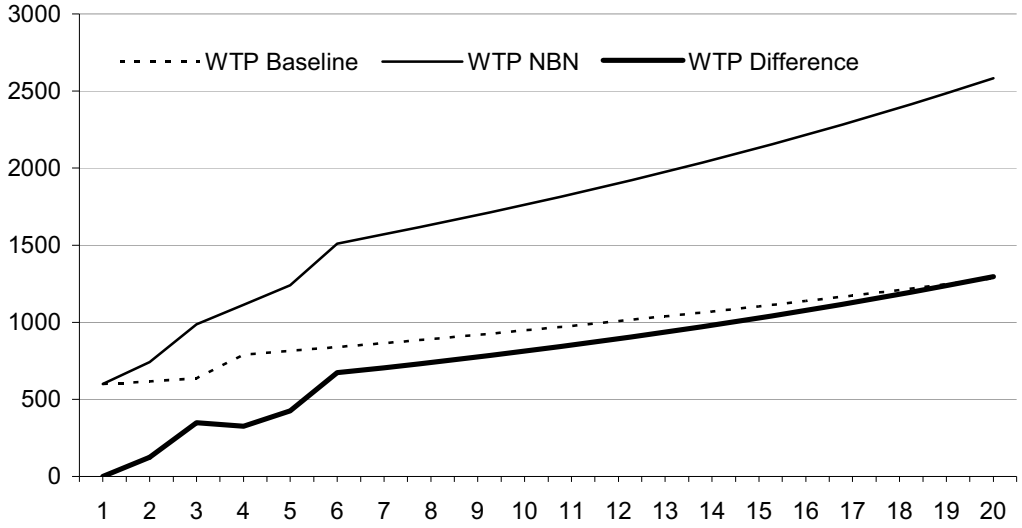
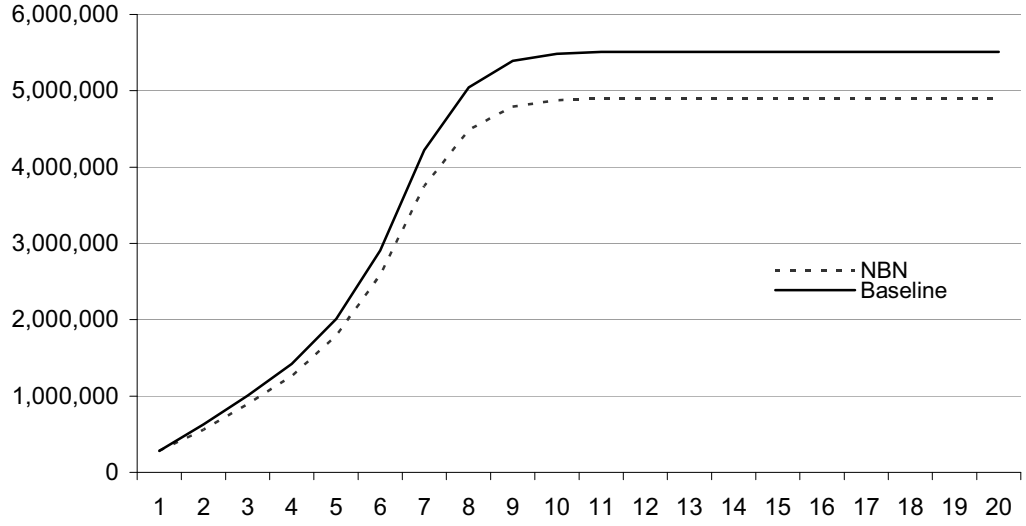


Figure 6.11 **Take-up paths: scenario D**



Finally, we can put all of this together and compute aggregate costs and benefits under the baseline and the NBN, and compute the present value of net incremental benefits of the NBN (table 6.3). The numbers in the table are the estimated present value of the net incremental benefit of the NBN, relative to the baseline. The estimates suggest that undertaking the project will result in a social loss in present

value terms of between $13.9 billion and $20.4 billion, depending on the discount rate chosen.

Table 6.3 **Present value of the net incremental benefits of the National Broadband Network: scenario D**

Discount rate	*Present value*
%	$(2009) billion
6	–20.4
7	–19.2
8	–18.1
9	–17.2
10	–16.2
11	–15.4
12	–14.6
13	–13.9

Since we have assumed that the willingness to pay for the NBN far exceeds that for the baseline, it is clear that the key drivers of the NBN's social losses are the large capital and operating costs of the project.

In fact, the central result of our modelling can be expressed in terms of the familiar condition for replacement investment. More specifically, it is economic to replace the existing network with a new network if the net present value of the ***total*** costs of the new network is less than the net present value of the ***incremental*** costs of the existing network, in each case adjusted for relative service quality (which we do through the willingness to pay calculation). It turns out that the NBN would satisfy this condition only if the present value of the additional cost of deploying and operating the NBN, compared to even the 20 Mbps scenario, were no more than $14 billion (evaluated at a discount rate of 13 per cent) to $24.7 billion (evaluated at a discount rate of 6 per cent).

Put slightly differently, assuming a midpoint discount rate of 10 per cent, it is irrational to spend more than $17 billion on the NBN, even if the alternative is a world in which the representative consumer cannot obtain service in excess of 20 Mbps and even if demand for high-speed service is rising relatively quickly. This amount of $17 billion is well below current estimates of the costs the NBN will involve, especially if it is to serve non-metropolitan areas. Alternatively and more realistically, if the base case (that is, the alternative to the NBN) is one in which the representative consumer is assumed to ultimately have access to 40 Mbps (rather than 20 Mbps as above), then it is inefficient to proceed with the NBN if the present value of its incremental costs of deployment and operation, evaluated at a 10 per

cent discount rate, exceed $10.6 billion, which is below the lowest bound of the estimates of these costs.[14]

Discussion of the results

It may be thought that these estimates understate the gains from the project because they do not take account of wider economic and social benefits. While it is likely that use of higher speed access lines will allow productivity gains, we would expect those gains to be reflected in consumers' and businesses' willingness to pay for that use. As a result, treating the productive efficiency gains as an added benefit amounts to double counting. As for wider social benefits, it is unclear what they consist of, and whether they are indeed greater under the project than under the counterfactual. Moreover, to the extent such social benefits exist, there must be the question of whether the project is the most efficient means of ensuring their delivery.[15] Without more precise specification of those benefits, it is not possible to assess whether they have any substance, although some that have been cited in the press seem dubious.[16]

Rather, it is our view that the estimates understate the likely project-related social costs. Thus, it seems probable that, evaluated at a rate of return that reflects the risks the project imposes on taxpayers, the project will incur losses.[17] While those losses themselves are a transfer, the distortions associated with financing them through taxation are not, and need to be added to the social costs of the project. In contrast,

---

14 Using 40 Mbps is especially realistic if sorting is allowed to occur — that is, if account is taken of the fact that suppliers will target those customers who place a high value on speed, and that those customers will have incentives to choose locations (for instance, at which to site offices) that offer such access.

15 If these benefits can be obtained at lower cost under some alternative option, then the cost increase from forgoing the use of that lower cost option (that is, from using the NBN to deliver those benefits, rather than the cheaper alternative) is a net cost to the project and should be treated as such in the analysis.

16 Claimed wider benefits such as the promotion of tele-medicine seem very difficult to credit. With respect to tele-medicine, it is not clear what residential medical applications require access to residential fibre optics, short of a future being projected in which individuals will have CAT scanners in their homes. As for GPs and medical centres, there is no evidence that network access costs and speeds have any effect on their use of tele-medicine; Paolucci et al. (2009) survey the literature on the effectiveness of tele-medicine and do not find such evidence. Finally, hospitals are generally already connected to high-speed access networks and would be so under the factual and counterfactual alike.

17 Of course, the project might be profitable were it given a monopoly or regulatory protection from competition (say, through an exemption from the merger laws that allowed it to acquire assets that would otherwise act as an effective competitive constraint). However, were that the case, then the efficiency costs of such a monopoly would need to be brought to account in the cost–benefit analysis.

under the counterfactual, taxpayer outlays would be limited to any vouchers used to subsidise demand by consumers in high-cost areas. Moreover, the prospect of taxpayer financing of the project's losses can lead to moral hazard, as well as to direct political interference in project decisions, diminishing the productive efficiency with which the project is pursued. Our estimates, however, do not gross up financing costs for the difference in the value of private and public income (that is, for the marginal social cost of funds) and assume the project is deployed and operated at least cost.

Additionally, the NBN project, whatever its merits, will create risks to the integrity of the regulatory system. First, the Australian Government will be both the primary investor in a major competitor and the industry policymaker and regulator, creating sovereign risk for private investors and introducing potential distortions to policy and regulatory decisions. Second, the NBN may involve some form of joint venture between entities that would otherwise have the scope to compete on a head to head basis, with the associated dangers of collusion. Third, there will be strong pressures for geographically uniform pricing, which can add distortions not only to resource allocation but also to competition (for example, if restrictions or taxes on bypass are used to protect the flow of cross-subsidies). These costs are not taken into account in our estimates.

At the same time, our estimates of the project benefits do not take account of offsetting equilibrating processes and therefore tend to overstate them. In particular, it is clear that, in the counterfactual, those consumers that place the greatest value on high-speed access will generally have such access, for two reasons: first, suppliers will have incentives to provide it, including through geographically targeted upgrades; and second, over 10 to 15 years , geographical mobility is relatively high, and consumers will sort themselves geographically in a way that, inter alia, reflects the valuations they place on different forms of broadband access. As a result, the population that gains access to very high-speed broadband in the NBN world relative to the counterfactual is likely to be that segment that places the lowest valuation on broadband access. To that extent, our estimates, which do not allow for this sorting process, exaggerate the gains from NBN deployment. This is all the more the case as our counterfactual scenario (scenario D) assumes relatively low speeds would be available should the NBN not proceed.

This overstatement of project gains is accentuated by our approach to estimating net benefits, which compares the willingness to pay for the incremental speed the project provides to the incremental cost of providing that speed. However, whether benefits are realised depends to a significant extent on the entity's future pricing policies. For example, if prices are set at average costs, then some potential utility gains will not be realised (as those consumers who value the project output at more

than incremental cost, but less than project average cost, will not consume its services).[18] This is equivalent to the issue that arises when toll roads are built: the cost–benefit analysis for the road link may be undertaken on the basis of potential social gains; however, the tolls may lead to some users whose valuations exceed marginal costs (and hence who are counted towards the cost–benefit analysis's estimate of benefits) not actually using the road, causing realised benefits to fall below assumed levels. Because we do not discount our estimated benefits for this effect of the entity's pricing policies, we probably overstate the likely benefits.[19]

Finally, we have not costed the most natural alternative — which is simply to delay the project and re-examine its economics every few years. This option to delay is likely to have high value, particularly if it is accompanied by regulatory reform that addresses the current disincentives to invest. Such an option would allow any public investment to be more narrowly targeted to areas of genuine and durable market failure and would reduce both the risk of asset stranding and of significant deadweight losses due to the tax financing of project losses.

In short, we believe our estimates overstate the likely gains and understate the likely costs from the NBN.

All that said, the notion of wider productivity benefits from broadband deployment is a popular one, with especially frequent reference being made[20] to an estimate by Access Economics that:

---

18 As noted above, the Australian Government's *National Baseline of School Broadband Connectivity 2008*, shows that while 'the majority of schools in metropolitan locations reported using fibre (51.6 per cent) and most schools in provincial locations also reported using fibre (46.5 per cent)', most schools 'use download speeds of up to 4 megabits per second, which is the lowest download speed range used in the FCS baseline survey. This disparity may be due to affordability of the service or the specific contractual arrangements negotiated, throttling and issues relating to the availability of suitable online curriculum resources and tools.'

19 Obviously, were perfect lump sum taxes and transfers available, then no such social costs would eventuate. Project charges to users would, in such a world, be set to marginal costs, and any fixed costs would be covered through public transfers. Unfortunately, such perfect lump sum taxes and transfers are not available, and hence it may be efficient to impose break-even constraints (or at least some degree of fixed cost recovery) on public suppliers. The welfare costs of any such constraints then need to be taken into account.

20 'Access Economics predicts that a national high-speed broadband network would mean economy-wide productivity growth 1.1 per cent higher after ten years compared to if the network was not built.' Senator the Hon Stephen Conroy, Minister for Broadband, Communications and the Digital Economy, speech to CeBIT Australia 2009 AusInnovate Conference, 12 May 2009. The Minister goes on to say: 'It is worth noting that Access Economics views this as a conservative estimate.' However, as discussed below, the comparison Access Economics makes is to a world in which only dial-up service is available (noting that as of the time of writing, 70 per cent of Australian households subscribe to some form of broadband).

> … economy-wide multifactor productivity levels would be around 1.1 per cent higher in an Australian economy with HSBB [high-speed broadband] available everywhere relative to an Australian economy without any HSBB after ten years. That is, the average annual growth rates in productivity would be around 0.1 percentage points a year higher in a complete HSBB world *compared with a situation where only, say, dial-up was available.* (Access Economics 2009, p. 20, emphasis added.)

However, as the Access Economics report plainly states, these productivity gains are relative to an economy in which only dial-up service, or similarly very low-speed access options, would otherwise be available. Moreover, it is also plain from the Access Economics report that the numbers cited are no more than assumptions, albeit ones Access Economics believes to be conservative for the comparison being made.

To take account of these differences, we believe that the Access Economics estimates of productivity gains should be set to one-third to one-half their initial levels, given that 70 per cent of households now have some form of broadband access. Additionally, account needs to be taken of the likely crowding-out effects of the public expenditure. We use a simple macroeconomic model with crowding out (> 0) to assess the likely impacts. The results, set out in table 6.4, are expressed as the present value of the cumulative change in GDP over a twelve year period, discounted to the present at a discount rate of 7 per cent (the rate used by Access Economics) and put in 2009 dollars. Broadly, the results suggest that cumulative GDP declines, despite an assumed increase in productivity.

Table 6.4 shows the present value, in 2009 dollars, of the cumulative 12-year change in GDP due to construction of the NBN, for a range of values of productivity increase and of extent of crowding out of other investment.

Table 6.4 Present value of the cumulative 12-year change in GDP due to construction of the National Broadband Network

Increase in productivity level	*Degree of crowding out*					
	0.5	*0.6*	*0.7*	*0.8*	*0.9*	*1.0*
0.3	–12	–17.1	–22.2	–27.3	–32.4	–37.5
0.4	–7.6	–12.7	–17.8	–22.9	–28	–33
0.5	–3.2	–8.3	–13.4	–18.5	–21	–23.6

Note: A discount rate of 7 per cent is used, for comparability with the Access Economics (2009) results.

This loss is not directly comparable to that derived from a comparison of incremental project costs and consumer valuations; however, some component of it — that part that reflects distortions due to the burden of taxation — could properly be added to the cost–benefit analysis loss (as that loss is calculated without regard to the difference between the private and public value of income). Unfortunately,

this component is not separately identifiable, being simply an element in the assumed crowding-out parameter.

Conclusions on telecommunications

In short, under both the Howard and Rudd governments, important telecommunications decisions have been made without formal, transparent assessment of costs and benefits. Our review — both of the quality of service regulations implemented by the previous government, and of the proposed NBN — suggests such an assessment would conclude that the policies at issue impose costs than exceed the relevant benefits.

6.3 Improving the evaluation process

The case studies set out above and in the long version of this paper suggest that at least some important infrastructure decisions are being taken on the basis of little evidence and in at least some instances, inadequate analysis. This is an obvious concern given the scope poor infrastructure decisions have to reduce capital productivity and hence lower living standards in the longer term. Mounting evidence of inefficiencies in the way our infrastructure is run — with the search for 'ribbon-cutting' opportunities displacing adequate investment in maintenance, causing a rapidly growing maintenance deficit that is well documented in Victoria and New South Wales (NSW Audit Office 2006, Victorian Auditor General 2008) — only adds to the concerns. What then can be done to strengthen the evaluation process?

Ultimately, the quality of evaluation depends on the value governments place upon it. Governments that view project evaluation as merely a nuisance that stands in the way of the decisions they want to take, and that believe they can get away with no evaluation or poor quality evaluation, will, over time, invariably succeed in devaluing the evaluation process. This has, we believe, occurred in Australia in recent years.

In part, this simply reflects a loosening of government budget constraints due, first, to sustained economic growth and, second, to a belief that the global financial crisis meant that high levels of government spending were not only feasible, but also desirable. As the threat of recession loomed, confused reasoning led to a belief that infrastructure investment could legitimately be claimed to be a tool of macroeconomic policy, even though, in an economy with monetary and aggregate fiscal policy instruments, infrastructure investment should play no role in

stabilisation policy and cyclical conditions should not affect the timing or extent of infrastructure outlays, other than through their effects on projected demand and on the shadow prices of inputs (effects which, properly analysed, can suggest that infrastructure projects should be deferred, rather than accelerated, during downturns); see, for example, Bureau 1985.[21]

There are, however, also longer term forces at work. These forces reduce the effectiveness of accountability and increase the attractiveness of infrastructure decisions as elements in rent-seeking bargains.

The first is the ever greater blurring of responsibility for infrastructure between the Commonwealth and the States, and the progressive loosening, by the Commonwealth, of budget constraints at a state level. This reduces the electoral accountability of, and electoral pressure on, State Governments, while reducing the opportunity cost that State Governments incur for poor investment decisions. To some extent, the Commonwealth has sought to offset the resulting moral hazard by imposing performance obligations on the States, such as the evaluation requirements built into Auslink. However, much as with foreign aid, these requirements typically bear only a very indirect link to the outcome being sought (which, in this case, is quality decisionmaking) and readily become (at best) 'tick-the-box' constraints that are often easily gamed (as the quality of compliance is rarely monitored, and, when monitored, even more rarely acted upon). Threats of conditionality have little credibility, especially when doing so would impose a significant political cost on the Commonwealth itself. Again, much as with foreign aid (see Azam, Devarajan and O'Connell 1999; Brautigam 2000; Knack 2001; Alesina and Weder 2002; Bardhan 2005; Easterly 2006; Moss, Pettersson and van de Walle 2006; and Janus 2009), the result is a degradation in institutional quality and in ultimate outcomes.

These issues associated with fiscal federalism have become even more complex with the creation of the Building Australia Fund and of Infrastructure Australia. Although there can be merit in coordinated approaches to infrastructure selection, there can be little doubt that the new mechanisms create significant incentive

21 Bureau develops a non-Walrasian model with an external constraint, a monetary policy instrument and fiscal policy. While no policy instrument should be thrown away, his main result is that macroeconomic considerations should enter into the evaluation of infrastructure investment only to the extent that the consequences of that investment are orthogonal to those of the macroeconomic instruments. As for the impacts of cyclical factors on the cost–benefit analysis, where public assets will compete with private assets (as in the case of the NBN), then the costs of those public assets will rise during recessions, even in the presence of Keynesian unemployment; see, for example, Johansson (1991, pp. 122–3). Additionally, to the extent demand expectations are reduced, this should lead to lower infrastructure investment.

problems. To the extent to which the projects they fund are worth while, that funding may simply displace funding of those projects by the States themselves, but with higher transactions costs and possibly poorer monitoring and other performance incentives in the process.[22] There may, in other words, be incentives for adverse selection, and then for moral hazard in project execution to boot.[23]

The second factor that has contributed to a decline in the quality of project evaluation is the growing involvement of the private sector in the design, construction, financing and operation of major infrastructure projects, both through the contracting out of almost all aspects of project implementation and perhaps especially, through public–private partnerships (PPPs). While these may have merits in terms of productive efficiency, the use of high-powered incentives[24] has complex, and often undesirable, impacts on the quality of public administration (see for example, Estache and Martimort 1999). In particular, because the incentives are high powered (that is, the private party secures substantial gains from reducing costs under the contract), these arrangements increase the returns to rent-seeking and to tainted deals between governments and private sector suppliers. Particularly with PPPs, the effects are then threefold: they concentrate the gains from the project (as some share of these is now captured by the private participant), and, by so doing, increase the payoffs from collusion between the public decisionmaker and the project's private beneficiaries; they allow crucial aspects of the project to be cloaked in commercial commerciality, thus reducing the transactions costs of collusion; and they relax (or, more properly, are widely but incorrectly claimed to relax) the public sector budget constraint. Each of these effects induces a deterioration in the efficiency of decisions and overall outcomes.

Ultimately, PPPs are only as good as the governments that make them; and if governments are intent on poor decisions, these partnerships can not only make

22 Obviously, if the Commonwealth funding were simply matching grants associated with the pure spillover effects of state infrastructure decisions — that is, a Pigouvian subsidy — the issue of displacement would not arise. Conversely, if the projects are so poor that they would never have been undertaken by the states then there will indeed be a 'flypaper' effect and aggregate infrastructure outlays will rise (on which see, generally, Brennan and Pincus 1990; as per Brennan and Pincus, this is a case where the grant pushes spending to the corner solution).

23 The question of how to design multi-level funding institutions and associated cost–benefit analysis processes so as to deal with these effects has received some attention in the EU, although with few readily implemented results to date; see Florio 2007.

24 The 'power' of an incentive structure is determined by the extent to which the agent to whom that incentive structure applies can secure for itself the gains from cost reductions (or other improvements in performance). Incentives are said to be 'high powered' when the agent secures a large share of the gains (as in a fixed price contract); conversely, they are 'low powered' when the agent's share of any gains is small (as in a cost-reimbursement contract).

those decisions more (privately) profitable but allow them to be locked in through long-term, judicially enforceable, contractual commitments.[25]

A third factor, which is yet to fully play itself out, is the recourse to hypothecated funding sources for long-term infrastructure finance, most notably the Building Australia Fund. While economic theory yields ambiguous results as to the effects of hypothecation on fiscal efficiency[26], it does identify a number of important ways in which earmarking it can reduce the quality of public expenditures.

First, earmarking implies inflexibility in the allocation of revenues among competing uses. If the earmarking is substantive, in the sense of being effectively constraining, social rates of return are unlikely to be equalised at the margin across uses. Tax rates, expenditure levels or more likely both, will be distorted as a consequence.

Second, reserving revenues to a program gives it a monopoly over those revenues, encouraging and potentially perpetuating technical inefficiency in its supply.

Third, earmarking can facilitate rent-seeking by allowing the interest groups that benefit from the hypothecated revenue stream to focus their activities more effectively. Rather than competing against other interest groups for a larger share of general revenues, the relevant groups can limit their efforts to seeking an increase in (or protecting from erosion) the hypothecated fund. At the same time, the political commitment they secure is potentially made more credible by the earmarking, increasing both the 'price' that the interest groups are willing to pay in exchange and the resources they are willing to dissipate in obtaining it. Rent-seeking coalitions therefore become easier to create and sustain, and the aggregate costs to the community from rent-seeking rise, as Kimenyi, Lee and Tollinson (1990) found in their study of the US Highway Trust Fund.

Fourth, these adverse consequences are made all the greater by the risk created by earmarking of fiscal illusion; that is, of the hypothecated revenues not being as visible as other forms of public revenue and expenditure.

---

25 This is similar to the 'Landes–Posner effect', whereby an independent judiciary increases the extent of rent-seeking by making it easier for legislators to lock in tainted deals (Landes and Posner 1975).

26 For example, earmarking may be a way of increasing the credibility of promises, reducing the inherent incompleteness of the implied contracts between government and the public. As well as any direct benefits arising from greater credibility of commitments, this may allow proponents of programs to signal the quality of the programs, of the proponents or both. Thus, in the model of Brett and Keen (2000), a commitment to dedicate revenues to a particular use, which is of value to the public but would not be of value to a 'poor-quality' politician, can support a separating equilibrium in which politicians signal their quality to the electorate.

All of these factors create risks that the new earmarked funds, though they may increase spending on infrastructure, could reduce the quality of that spending.

Set against these long-term forces, project evaluation is a relatively weak reed, and the effects of changes to evaluation processes alone may well be relatively small. Nonetheless, we would suggest three areas for reform.

The first is **greater transparency**. There is no reason why cost–benefit analyses should not be publicly disclosed as a matter of course. Instead, most cost–benefit analyses are never released, and those that are are often difficult to locate. Governments should also regularly publish, in readily accessed form, the cost–benefit analysis rankings of those projects they have decided to proceed with and those they have considered and rejected (as is done in Finland, for example). Were disclosure of cost–benefit analyses routine, the fact that a cost–benefit analysis had not been conducted on a particular project would become more obvious, as would the relative quality of the cost–benefit analyses that had been carried out.

The second is greatly enhanced **auditing**. Auditing plays an important role in improving the efficiency of principal–agent relations, both by allowing principals to better assess the outcomes of the efforts made by agents and by deterring collusion between agents and third parties (see Mookherjee and Png 1989). The introduction of an independent auditor, whose interests are separate from those of the party being audited, increases the likelihood of poor conduct being detected, including when that conduct takes the form of bias (for instance, associated with 'excess optimism' or with the strategic understatement of costs[27]).

The auditing we believe desirable would take two forms. To begin with, there is substantial merit in having independent reviews of all cost–benefit analyses for 'mega-projects' (say, projects with projected outlays in excess of $500 million). This could be done by an office answerable to Parliament, rather than forming part of the Executive. Such an office could be similar to the Congressional Budget Office in the United States. Were establishing such an institution considered too radical, at the very least adequate specialist resources should be provided to a parliamentary standing committee to engage the kind of forensic analysis required. This is not to cast doubt on the Australian National Audit Office, but rather to suggest that its competence, and standard form of operation, are not especially well suited to this task.

As well as this form of review, there is a pressing need for much more to be done in terms of post-completion review of projects. Although a few useful post-completion

[27] The pervasiveness of these forms of bias in transport assessments is amply documented by Flyvbjerg, Bruzelius and Rothengatter (2003).

reviews of cost–benefit analyses have been undertaken (BTE 2001; BTRE 2007a, 2007b; NSW Audit Office 2006; NSW Auditor General 2005; NSW Treasury 2008; Victorian Auditor General 2009), these are ad hoc, which limits their effectiveness both as instruments of accountability and as a means of learning from experience. The Auslink program mandated post-completion reviews; unfortunately, this requirement has not been rigorously enforced. We believe it should be.

Mandating systematic and transparent post-completion review could have far-reaching consequences. To begin with, it would force Commonwealth and state entities to more properly document and archive material related to the cost–benefit analyses and the cost–benefit analyses themselves. In contrast, as matters currently stand, cost–benefit analyses are typically undertaken before the final form of projects is determined, and then never updated. Additionally, little investment is made in documenting cost–benefit analyses and in ensuring the integrity of the documentation chain. A genuine system of post-completion reviews would require all of those deficiencies to be addressed. At the same time, such reviews could be used both to benchmark jurisdictions and to more effectively learn from mistakes.

In short, we would strongly endorse — and argue more should be done to implement — the conclusion Little and Mirrlees (1994, p. 206) reached in reviewing, after two decades, the impact of their great cost–benefit analysis manual:

> If good project appraisal warrants expenditure, as we argue, so does good appraisal of appraisal.

Third and last, there is a great deal that could be done both to increase the **quality of cost–benefit analyses** and to promote a greater sense of professionalism in the group of people engaged in project evaluation. There are still many complex technical issues to tackle in Australian project evaluation — including the selection of the criterion function (where, unfortunately, the use of Benefit–Cost Ratios is still widespread, despite its well-known deficiencies), the treatment of the marginal social cost of funds (which is usually ignored), the determination of the discount rate (often set in a manner that is somewhat arbitrary), the assessment of changes in service quality and reliability (which is particularly important in public transport, as well as in communications), the appropriateness or otherwise of corrections for 'optimism bias' (which, in the authors' opinion, are likely to be ineffective at best and distorting at worst), the role of 'wider economic benefits', and so on. While many of these issues are well traversed in the literature (if not in the practice) of project evaluation, there are other important issues that are relatively under-explored, such as the implementation of cost–benefit analysis in the context of

hypothecated funds (where congruence requirements should come into play[28]) or the design of incentive-compatible evaluation schemes for structures such as Infrastructure Australia.

There is consequently considerable potential for cooperative research across jurisdictions, and for using that research, and its dissemination, as an instrument of ongoing training for both practitioners and users of cost–benefit analysis. Moreover, that process could help give greater standing to the 'profession' of project evaluation and help define a community of those involved in project evaluation across different areas of infrastructure policy. There is an important role here for the Bureau of Transport and Regional Economics and also for the Productivity Commission. Thus, the Productivity Commission could, much as it did in regulation review, issue 'information notes' recommending particular approaches to the technical issues analysts face. While we do not believe there is one 'right' approach to all of these issues, and hence would not favour mandatory standardisation across the States, that should not impede the exchange of views and the fostering of comparability of analyses across jurisdictions (so that the effect of different approaches can be identified). Much has been done in this respect by the Australian Transport Council's 2006 *National Guidelines*, but the list of issues identified above highlights the task that remains.

6.4 Conclusions

Infrastructure investment is a cost, not a benefit; a means, not an end. This proposition, which is obvious to economists, is as utterly alien to contemporary Australian politicians as the notion of comparative advantage was to their predecessors.

That matters should be so is by no means a new phenomenon. Thus, in Hancock's magnificent *Australia* (1930), now sadly out of print, the great historian famously said that it was a failing of democracies, and especially of Australian democracy, to constantly confuse ends and means, and to show too much reluctance 'to refuse

[28] When decisions are delegated to agencies, and agencies are instructed to make optimal use of their budgets, the expected growth path of agency budgets on the one hand and of investment opportunities on the other becomes an important factor in determining the optimal pattern of outlays. When an agency regards both its current budget and its current set of investment opportunities as representative of future opportunities — either because these regenerate periodically or because they are linked — it is referred to as having congruent expectations. Agencies should, in defining the choice set for evaluation, choose a set of projects and time horizon that can reasonably be regarded as congruent. Where agencies are budget funded, it is not unreasonable to assume the current budget defines such a set; however, this assumption cannot simply be carried over to an agency whose budget is hypothecated.

favours, to count the costs, to discipline the policies they have launched'. '[The] policies therefore yield diminishing returns, until at last, they may become a positive danger to the national purpose that called them into existence.' Nowhere was this more marked, Hancock noted, than with public involvement in infrastructure ventures such as rail, where Australian government was 'particularly slow to confess it has got into a bad business, for its mere entry … has created vested interests which immediately express themselves in politics … So … it throws good money after bad, and hopes that something will turn up. In this way, losses accumulate in a lump, and the crisis, when it comes, is likely to be prolonged and severe.'

The costs and risks of this approach to infrastructure have also been known for many years. There are surely many echoes in current telecommunications decisions of the tendency, identified by Butlin, Barnard and Pincus (1982, p. 294) in their analysis of the development of the Post-Master General's Department, for Australian public enterprise to provide 'services that were too large, too quickly supplied and too cheap'. That so little should have changed is not encouraging.

Set against that background, how great a contribution can improved project appraisal make to securing better outcomes? Little and Mirrlees (1994, pp. 225–7) develop a simple model of the value of information in which good project appraisal yields benefits that, in expected value terms, are in the order of 10 per cent of project value.[29] For an economy investing over $10 billion per year on its transport and communications infrastructure, 10 per cent of project value would seem like a saving well worth seeking. That said, the Little–Mirrlees model assumes unbiased estimates and a decision-maker who, as a benevolent social planner, maximises social welfare; it is hardly contentious that those assumptions do not hold — if they did, central planning would be a far better system than it has ever proved to be.

To recognise this, however, is not to imply that no value should be placed on good appraisal; on the contrary, it is one of the protections taxpayers deserve to have. Testimonials of commitment to 'evidence-based policy' notwithstanding, shaping an environment in which project appraisal can effectively discharge this task remains as great a challenge as it has ever been.

Overall, our review suggests the following conclusions:

[29] The Little–Mirrlees formulation yields a value of appraisal that is at least 10 per cent of standard deviation of the errors removed by the appraisal, multiplied by the ratio of that standard deviation to the standard deviation of the errors not removed. This ratio should be about 1, though with competent appraisal it could be much more than that. As a result, a conservative estimate of the value of appraisal is 10 per cent of project value.

- Insufficient attention is paid in the evaluation process to options that would avoid investment, or, more broadly, that would focus on securing greater efficiency from the existing capital stock. Simply put, infrastructure investment appears to be viewed as a benefit, rather than a cost.
- The distortions arising from this undesirable narrowing of the range of options considered are then compounded by evaluations that are too vulnerable to 'fudge factors'. In a Gresham's law of evaluation, bad evaluations (often by consultants) can drive out good, given that they trade at equal values.

In our view, these outcomes are driven by governments that see little real value in major project evaluation. They may see merit in evaluation of essentially routine decisions (such as the decision to place a new roundabout or improve a road surface) or in cost-effectiveness analysis of the options available for meeting pre-determined goals (such as improving bus transit in a congested area) but not in the full analysis of objectives and options (including the option of not spending taxpayers' money). This, we argue, reflects the impact of a perception (initially due to strong economic growth, and then to a belief that the global financial crisis justifies greatly increased outlays) that public funds have a negligible opportunity cost. This perception has been accentuated by the growing blurring of accountability in the Australian federation, which reduces the budget disciplines on the States, and the blurring also of responsibility for financing infrastructure as between the public and private sectors (which, whatever its other merits, increases the return to rent-seeking deals between governments and private infrastructure developers). Together, these trends risk making cost–benefit analysis merely a box to be ticked, rather than an exercise that has real value, not least to government itself.

We are not optimistic that changes to cost–benefit analysis processes alone can counteract these powerful trends. Nonetheless, we think three changes would have merit:

- a requirement for all cost–benefit analyses to be disclosed that would also highlight which projects had not been subjected to economic project evaluation
- far greater and systematic auditing of cost–benefit analyses, both at the stage of the financing decision and post-project completion. In contrast, there is little or no such audit currently, and in many instances, cost–benefit analyses are not even updated, maintained or properly archived after the initial 'go/no go' decision is taken.
- the establishment of a centre of excellence or reference for cost–benefit analysis within the Australian Government, preferably in an independent entity, such as the Productivity Commission.

The Little–Mirrlees rule suggests that the value of proper project appraisal is at least 10 per cent of the value of projects. With Australia spending ever more on infrastructure, these are gains well worth seeking. Whether they can be achieved is obviously an open question.

References

Access Economics 2009, *Impacts of a National High-speed Broadband Network*, March.

Adler, M.D. and Posner, E.A. 2006, *New Foundations of Cost–Benefit Analysis*, Harvard University Press, Cambridge, Mass.; London, England.

Alesina, A. and Weder, B. 2002, 'Do corrupt governments receive less foreign aid?', *American Economic Review*, vol. 92, no. 4, pp. 1126–37.

Analysys Mason for BSG 2008, *The Costs of Deploying Fibre-based Next-generation Broadband Infrastructure*, Final report, 8 September.

Arrow, K.J. and Lind, R.C. 1970, 'Uncertainty and the evaluation of public investment decisions', *American Economic Review*, vol. 60, no. 3, pp. 364–78.

Australian Transport Council 2006, *National guidelines for transport system management in Australia*.

Azam, J.P., Devarajan, S. and O'Connell, S.A. 1999, 'Aid dependence reconsidered', Policy Research Working Paper No. 2144, The World Bank.

Bardhan, P. 2005, *Scarcity, Conflicts, and Cooperation: Essays in the Political and Institutional Economics of Development*, MIT Press, Cambridge, Mass.; London, England.

Bartholomeusz, S. 2009, 'Conroy's stab in the dark', *Business Spectator*, 15 May.

Becker, G.S. 1965, 'A theory of the allocation of time', *The Economic Journal*, vol. 75, no. 299, pp. 493–517.

Brautigam, D. 2000, *Aid Dependence and Government*, Almqvist and Wiksell, Stockholm, Sweden.

Brennan, G. and Pincus, J. 1990, 'An implicit contract theory of intergovernmental grants', *Publius*, vol. 20, no. 4, pp. 129–44.

Brett, C. and Keen, M. 2000, 'Political uncertainty and the earmarking of environmental taxes', *Journal of Public Economics*, vol. 75, no. 3, pp. 315–40.

BTE (Bureau of Transport Economics) 2001, *The Black Spot Program 1996–2002: An Evaluation of the First Three Years*.

BTRE (Bureau of Transport and Regional Economics) 2007a, *Ex-Post Economic Evaluation of National Highway Projects: Case Study 1: Wallaville Bridge*.

—— 2007b, *Ex-Post Economic Evaluation of National Highway Projects: Case Study 2: Northam Bypass*.

Bureau, D. 1985, 'Cohérence entre choix des projets et politique de régulation macroéconomique', *Annales d'Economie et de Statistique*, INSEE, Paris (English version in Champsaur, P. (ed) 1990, *Essays in Honor of Edmond Malinvaud,* vol. 2, *Macroeconomics*, The MIT Press, Cambridge, Mass.).

Butlin, N.G., Barnard, A. and Pincus, J.J. 1982, *Government and Capitalism: Public and Private Choice in Twentieth Century Australia*, George Allen & Unwin, Sydney.

Easterly, W. 2006, *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Oxford University Press, Oxford.

Ergas, H. 2008a, *Wrong Number: Resolving Australia's Telecommunications Impasse*, Allen & Unwin, Sydney.

—— 2008b, 'Setting access prices: a critique of the ACCC's approach in telecommunications', *Agenda: A Journal of Policy Analysis and Reform*, vol. 15, no. 4, pp. 35–58.

—— 2009a, 'Time consistency in regulatory price setting', *The Review of Network Economics*, vol. 8, no. 2, pp. 153–63.

—— 2009b, 'In defence of cost benefit analysis', *Agenda: A Journal of Policy Analysis and Reform*, vol. 16, no. 3, forthcoming.

—— and Hardin, A. 1999, 'Annuity and levelisation issues in forward looking cost models: the case of telecommunications', Paper presented at the ACCC Depreciation Forum, 30 September.

—— and Ralph, E. 2008, 'A policy framework for a new broadband network', in *Growth 60: Australia's Broadband Future: Four Doors to Greater Competition*, CEDA.

Estache, A. and Martimort, D. 1999, 'Politics, transaction costs, and the design of regulatory institutions', World Bank Policy Research Working Paper no. 2073, http://ssrn.com/abstract=620512

Evans, J., Levine, P. and Trillas, F. 2008, 'Lobbies, delegation and the under-investment problem in regulation', *International Journal of Industrial Organization*, vol. 26, no. 1, pp. 17–40.

Florio, M. 2007, 'Multi-government cost–benefit analysis, shadow prices and incentives', in Florio, M. (ed), *Cost–Benefit Analysis and Incentives in*

Evaluation: the Structure Funds of the European Union, Chelten, UK.; Northampton, Mass, pp. 1–46.

Flyvbjerg, G., Bruzelius, N. and Rothengatter, W. 2003, *Megaprojects and Risk: an Anatomy of Ambition*, Cambridge University Press.

Gans, J. 2009, 'The right policy for telecommunications and broadband', Submission to the Senate Select Committee on the National Broadband Network, 18th June.

Goolsbee, A. and Klenow, P.J. 2006, 'Valuing consumer products by the time spent using them: an application to the internet', *American Economic Review*, vol. 96, no. 2, pp. 108–13.

Guthrie, G. 2006, 'Regulating infrastructure: the impact on risk and investment', *Journal of Economic Literature*, vol. 44, pp. 925–72.

Hancock, W.K. 1930, *Australia*, Ernest Benn, London.

Janus, T. 2009, 'Aid and the soft budget constraint', *Review of Development Economics*, vol. 13, no. 2, pp. 264–75.

Johansson, P-O. 1991, *An Introduction to Modern Welfare Economics*, Cambridge University Press.

Jullien, B., Pouyet, J. and Sand-Zantman, W. 2009, 'Public and private investments in regulated network industries: coordination and competition issues', mimeo.

Kimenyi, M.S., Lee, D. and Tollinson, R.D 1990, Efficient Lobbying and Earmarked Taxes, *Public Finance Review,* Quartery, vol. 18, no. 1, pp. 104-113.

Knack, S. 2001, 'Aid dependence and the quality of governance: cross-country empirical tests', *Southern Economic Journal*, vol. 68, no. 2, pp. 310–29.

Kyland, F. and Prescott, E. 1977, 'Rules rather than discretion: the inconsistency of optimal plans', *Journal of Political Economy*, vol. 85, pp. 473–92.

Landes, W.M. and Posner, R.A. 1975, 'The independent judiciary in an interest-group perspective', *The Journal of Law and Economics,* vol. 18, pp. 875–901.

Levine, P., Stern, J. and Trillas, F. 2005, 'Utility price regulation and time inconsistency: comparisons with monetary policy', *Oxford Economic Papers*, vol. 57, no. 4, pp. 447–78.

Little, I.M.D. and Mirrlees, J.A. 1994, 'The costs and benefits of analysis: project appraisal and planning twenty years on', in Layard, R. and Glaister, S. (eds), *Cost Benefit Analysis*, 2nd edn, Cambridge University Press, Cambridge, England; New York, NY.

Moss, T., Pettersson, G. and van de Walle, N. 2006, 'An aid-institutions paradox? A review essay on aid dependency and state building in sub-Saharan Africa', Working Paper no. 74, Centre for Global Development.

Mookherjee, D. and Png, I. 1989, 'Optimal auditing, insurance, and redistribution', *The Quarterly Journal of Economics*, vol. 104, no. 2, pp. 399–415.

NSW Audit Office 2006, *Condition of State Roads: Roads and Traffic Authority of NSW*.

NSW Auditor General 2005, *Performance Audit: Liverpool to Parramatta Bus Transitway*.

NSW Treasury 2008, 'Circular: revised project size/risk thresholds for the submission of business cases and gateway reports', NSW TC 08/07.

O'Flaherty, B. 2005, *City Economics*, Harvard University Press, Cambridge, Mass; London, England.

Paolucci, F., Ergas, H. Hannan, T. and Aarts, J. 2009, The effectiveness of health informatics, in press.

Posner, E. 2001, 'Controlling agencies with cost–benefit analysis: a positive political theory perspective', *University of Chicago Law Review*, vol. 68, p. 1137.

Soria, B. and Hernández-Gil, F. 2009, Exploring potential natural monopoly properties of broadband access networks, Paper presented at the 19th European Regional Conference of the ITS, 20 September.

Spence, D.B. and Cross, F. 2000, 'A public choice case for the administrative state', *Georgetown Law Journal*, vol. 89, p. 97.

Steiner, P.O. 1974, 'Public expenditure budgeting', in Blinder, A.S., Break, G.F., Netzer, D., Solow, M. and Steiner, P.O. (eds), *The Economics of Public Finance*, The Brookings Institution, Washington D.C, pp. 241–361.

Telstra 2007a, Application to ACCC for exemption from standard access obligations in respect of the Singtel Optus HFC network, 17 December 2007.

—— 2007b, Application to ACCC for exemption from standard access obligations in respect of the SingTel Optus HFC Network, Schedule A, 17 December 2007.

—— 2007c, Application to ACCC for exemption from standard access obligations in respect of the SingTel Optus HFC Network, 17 December 2007, Schedule A, Annexure 2, Expert report of Michael G. Harris: Use of HFC to deliver broadband services, 12 December 2007.

Vanderbilt, T. 2008, *Traffic*, Allen Lane, London.

Victorian Auditor General 2008, *Maintaining the State's Regional Arterial Road Network,* Government Printer, Melbourne.

—— 2009, *Buy-back of the Regional Intrastate Rail Network*, Government Printer, Melbourne.

Von Hagen, J. 2006, 'Political economy of fiscal institutions', in Weingast, B.R. and Wittman, D.A. (eds), *Oxford Handbook of Political Economy*, Oxford University Press, N.Y, ch. 26.

Wildavsky, A. 1966, 'The political economy of efficiency: cost–benefit analysis systems analysis, and program budgeting', *Public Administration Review*, vol. 26, no. 4, pp. 292–310.

7 Evidence-based policy: reflections from New Zealand

Grant M. Scobie[1]
Principal Advisor, The Treasury, Wellington

Abstract

Major institutional reforms have often have proceeded on very imperfect evidence. A greater role for sound evidence in policy-making rests on three fundamental elements: data, models and institutional frameworks.

The creation of longitudinal databases and linked data sets has permitted important insights into policy design, and history shows creating such databases (and giving access to them) can help answer policy questions that were not anticipated at the time of the investment in the data. Such investment in data should be regarded as the creation of 'policy capital'.

Economic models can contribute to better policy selection, as evidenced in New Zealand by the evolution of thinking about aspects of income tax design.

Institutional frameworks can also be designed to enhance the quality of policy making. Examples include the regulation impact statement process. While the ideal of informing policy choice by open, peer-reviewed analysis by multiple think-tanks and interest groups is hard to sustain in small economies, the use of informal networks and patterns of collaboration among analysts may help.

[1] My email address is: grant.scobie@treasury.govt.nz. The guidance of a large number of my Treasury colleagues is acknowledged, as is the assistance I received from John Creedy, Gary Hawke, John Yeabsley, Richard Bedford, Richie Poulton, Tony Blakely, Philip Stevens, Arthur Grimes, Dean Hyslop and Dave Maré. The views, opinions, findings and conclusions or recommendations expressed in this paper are strictly those of the author. They do not necessarily reflect the views of the New Zealand Treasury or the New Zealand Government. The New Zealand Treasury and the New Zealand Government take no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

> There is nothing more horrible than the murder of beautiful theory by a brutal gang of facts. (Francois de la Rochefoucauld, 1613–80)

After the extended debate throughout 1890 and 1891 on the question of an Australian federation, New Zealand, which along with Fiji had had a seat at the table, finally withdrew. So, in view of the title of this roundtable, I wish to acknowledge the graciousness of my Australian hosts in including a speaker from New Zealand.

That graciousness is doubly notable as, while working for the Commonwealth Public Service in Canberra in the early 1960s I was informed that I was the top ranked finalist for a Harkness Fellowship to study for a PhD in the United States; however, I had the bare-faced impertinence to decline the offer of walking round the corner from my office in the Barton woolsheds to the Tariff Board and taking out Australian citizenship: a three minute operation for a Kiwi, I was assured, but a necessary condition for the award.

The decision by the New Zealand delegates not to pursue federation raises interesting questions about evidence-based policy. With the advent of refrigerated shipping New Zealand saw much broader markets for its produce, and did not want to be dominated by Australian interests (a recurring theme to this day). In addition, at that time New Zealand's economic performance was seen as superior (sadly, a non-recurring theme to this day). In short, based on the evidence at hand, a monumental policy decision was taken that arguably has had far reaching consequences for what will literally be centuries.

I will argue that there are two lessons we can draw from this snippet of history that are relevant to our deliberations at this roundtable. The first is that, when dealing with major institutional reforms, we typically have very imperfect evidence. And arguably those big institutional changes are what really matter in the broad sweep of history.

The second arises because, very often, the evidence we bring to bear as the basis for advising policy makers comes from the past. Furthermore, in some cases it may be overly influenced by recent events. We try to distil from the historical record some indications that if we recommend option A rather than B, the evidence suggests we could expect a 'better' outcome. Let us set aside for the moment what we mean by 'better'.

The point is a simple and well-known one: to what extent will evidence from the past be useful in predicting future outcomes? The matter assumes even greater importance when those future outcomes are spread out over many generations. We

might even postulate a new theorem of policy making: namely, the cost of being wrong rises with the square of the planning horizon. Retirement income policies would be a case in point from the contemporary policy debate in many countries.

I have chosen to organise my remarks around three broad headings; these can be summarised as data, models and structures. I will address each in turn, and will draw on some examples where evidence has made (or at least has the potential of making) a demonstrable contribution to policy formation in New Zealand. I conclude on a slightly provocative, if somewhat pessimistic, note.

7.1 Data

Clearly, without databases quantitative evidence cannot be forthcoming (Hanuschek 1999). I want to focus on two major developments in New Zealand in this regard: the growing importance of longitudinal databases, and the evolution of linked datasets. These are exciting developments and hold great promise for strengthening evidence-based policy. They are focused on detailed unit record data of firms, individuals, families and households. Increasingly it is recognised that deeper insights into the functioning of complex economies cannot come from pondering macroeconomic statistics alone. Rather, we need to better understand the micro-level behaviour of firms and individuals.

Longitudinal databases

Before turning to more recent developments, I want to highlight one of New Zealand's longest standing longitudinal studies: the Dunedin Multidisciplinary Health and Development Study, which has followed 1000 individuals born in Dunedin in 1972–73. The study has produced well over 1000 reports which have influenced policy and practice, both within New Zealand and beyond.

As an example, research findings have helped understand and respond to severe antisocial behaviour. It is now widely recognised that there are two distinct categories of antisocial behaviour (Odgers et al. 2008).

The first involves a relatively small group (mostly male) who exhibit signs of antisocial behaviour from a very young age, including an excess of neuro-cognitive deficits, hyperactivity and under-controlled temperament. Their antisocial behaviour persists as they grow up. Eventually, this small group accounts for about 50 per cent of the crime in society, including the most extreme criminal acts.

The members of the second group, who comprise some 20 per cent of the population (equally male and female), begin offending for the first time during adolescence. They have unremarkable early life histories. Their antisocial behaviour is largely driven by peers, and typically disappears by their mid-twenties, in part because they had sufficient human capital prior to adolescence.

The implications for intervention with these two groups are quite different. For the first group, very early intervention with both the child and the family is called for. For the second, group intervention (such as incarceration) is to be avoided, as this merely reinforces the strong influences of peers. Prisons are universities for criminals. These findings, originating from the Dunedin study, have underpinned new approaches to dealing with antisocial behaviour that have become mainstream in New Zealand and abroad.

However, the Dunedin study (and a Christchurch counterpart) are based on small samples from local areas. A more recent study is the Survey of Family Income and Employment, a national-level undertaking planned for eight waves, and based on an initial sample of 22 000 individuals. An additional feature is that in alternate waves there are substantial modules on health and on assets and liabilities. It is early days, but already new studies have emerged on topics such as the relation between mental and physical health and wealth accumulation, the influence of health status and chronic diseases on labour supply (Holt forthcoming), and estimates of saving behaviour based on changes in net wealth over time (Henderson and Scobie 2009).

Another development with enormous potential for providing evidence is the Longitudinal Business Database. This is built on the Longitudinal Business Frame, to which is added goods and services tax returns, financial accounts of businesses, pay-as-you-earn returns, and shipment-level export and import data from customs records. One example of the richness of this database is the ability to ask such questions as: do firms with international connections (either through trade or ownership) demonstrate higher productivity relative to purely domestic firms? Critically, one can test whether entry into exporting results in a 'learning effect' that leads to higher productivity or, alternatively, whether higher productivity firms self-select into exporting (Fabling et al. 2008). Such evidence informs policies aimed at promoting exports, for example.

Longitudinal studies can be targeted at specific sectors or population groups. The Longitudinal Immigration Survey (LisNZ) is designed to give detailed information on the settlement outcomes of migrants at 6, 18 and 36 months after taking up permanent residence. Such evidence has policy implications for the sorts of immigrants New Zealand should encourage and their support systems after arrival. The Health, Wealth and Retirement Survey, a national longitudinal study, is

providing detailed evidence on the wellbeing and retirement decisions of those aged 55–70 (Enright and Scobie forthcoming)

Why are longitudinal databases so important to strengthen evidence-based policy? Cross-sectional surveys of firms or individuals are plagued with the fact that so many of the things that matter are unobservable or at best captured by a weak proxy. Our ability to isolate the specific factors that lead to innovation by firms, that might lower recidivism by convicted criminals, or that might result in better educational outcomes, is severely handicapped by the multitude of things we cannot measure. By observing the same individual repeatedly through time, we can, under a weak assumption, control for the unobservables and have potentially more robust evidence.

Linked databases

In many countries there are large administrative databases, such as those held by tax offices, hospitals or social welfare agencies. Increasingly there are examples where these have been linked with other sources, such as surveys and census data from the national statistics office. The distinction between linked and longitudinal databases is not clear-cut. A linked database may well have a longitudinal dimension. New mechanisms have been sought to ensure the essential confidentiality is preserved.

An outstanding New Zealand case is the Linked Employer–Employee Database (LEED). Monthly returns by firms to the Inland Revenue Department list all paid employees, their earnings and their tax. This is linked to the firm data in the Longitudinal Business Frame, together with benefit data from the Ministry of Social Development. This has provided the basis for addressing such issues as the effect of minimum wages on teenage employment; employment rates of former benefit recipients; measuring labour productivity; implications of changes in the composition of the workforce over the business cycle for labour productivity, job mobility and earnings dynamics (Stillman and Hyslop 2006).

By linking health records with data on the condition of housing, an unequivocal relation has been established between cold, damp, poorly insulated houses and the health status of the occupants. The Healthy Housing program conducted by public health specialists at the Wellington School of Medicine has had a profound effect on government policy, culminating in a recently announced substantial program of public subsidies for insulating houses (Howden-Chapman et al. 2007 and 2008).

Sometimes the evidence is simply assumed rather than sought. I have lost count of the number of textbooks on public economics that cited lighthouses as a classic example of a public good. Of course, the immediate policy implication was that

they would need to be publicly provided. Then came Ronald Coase, who, in combing the evidence, found that every single lighthouse built in Britain between 1610 and 1675 was the result of private investment.

Before concluding these reflections on the role of data, I want to make the case for 'if we build it, they will come'. I would argue that, in designing a data collection effort, we cannot always foresee exactly what future questions will arise. But experience has shown that comprehensive databases can be drawn on to address a wide range of policy questions. The Australian survey HILDA (Household, Income and Labour Dynamics) is testimony to this proposition. SoFIE (Survey of Family Income and Employment), the parallel database in New Zealand, already has contributed to policy formation by addressing questions as diverse as housing affordability and deposit insurance, despite limited access.

I have been fortunate to work in an environment where some value is placed on assembling evidence on which to base the policy advice that The Treasury is required to give. I have come to refer to this process as 'building policy capital' — that is, the stock of knowledge on which we can draw to give the most informed judgments possible to the government of the day. Just as producing goods and services requires physical and human capital, the production of evidence for policy making requires 'policy capital' — and databases are an essential element of that capital.

7.2 Models

Rarely will we have a picture complete enough to provide 'evidence' on the full sweep of possible implications of a proposed policy. The use of models to arrange the evidence and to simulate the effect of policy proposals is an essential tool in the armoury of a policy analyst (Coleman and Scobie 2009).

In some cases insights are derived from a judicious blend of evidence and models. Non–point source pollution from agricultural runoff is raising the nutrient levels in a number of major lakes. Detailed land use and environmental data are being combined with economic modelling to design nutrient-trading schemes which are being adopted. In short, by working closely with farmers in affected catchments, local and regional policy makers are drawing on the evidence and the modelling to introduce innovative ways to deal with a complex problem (Kerr and Lock 2008).

The question of housing affordability has been highlighted by the sharp rise in property prices. What are the underlying drivers of house prices? One hypothesis relates to the extent of supply constraints arising from land use and building

regulations at the local level. A model of price determination combined with evidence across a large number of local jurisdictions has indeed confirmed that prices rise much more sharply in response to a demand shock in those localities with more restrictive policies. These findings have been one catalyst for a review of the regulatory environment for housing (Grimes and Aitken 2006).

Change to the income tax regime is one area where modelling becomes an essential ingredient. In 2008, a range of options was reviewed. The then Minister of Finance strongly favoured creating a tax-free threshold in order to target tax relief at lower income individuals. As a result of the tax modelling, he eventually rejected the option. Interestingly, this is one of the relatively rare cases where there is a documented account of the impact of evidence on policy making:

> … my initial preference was to create a tax-free income threshold. Intuitively, this seemed a very effective way to deliver relief to low-income workers, as it would make a significant proportion of their pay tax free. This was an idea that gained some currency in the wider community, the Labour Party, and the union movement. When I asked officials to report to me on the likely impact of such a move, however, it became clear that it would have only a minimal benefit for a very small number of low income earners. (The Hon. Dr. Michael Cullen, New Zealand Minister of Finance, speech to the New Zealand Institute of Chartered Accountants, 7 May 2008).

How much should the state invest in research and development (R&D)? In effect we are asking: what evidence do we have about the marginal rate of return to spending on R&D? If this rate is higher than the social opportunity cost of capital, then we might have evidence that society has underinvested in research (Shanks and Zheng 2006).

This has proved a challenging topic and one that, despite the use of creative models and extensive data, has failed to provide conclusive evidence. Long lead times are ubiquitous, complementarities between public and private R&D investments are difficult to untangle, and productivity growth is influenced by a host of other factors. In small, open economies such as New Zealand's, it is reasonable to suppose that a great deal of the knowledge on which we draw is generated offshore. The evidence has indeed highlighted the role of foreign knowledge in raising productivity in New Zealand, and thereby underscored the importance of policies which foster our international connectedness (Hall and Scobie 2006).

Even when there is an abundance of data, extracting meaningful evidence can present serious challenges to the modeller. Two problems stand out: self-selection bias and causality. Let me illustrate these with examples from research on savings behaviour. Consider the case of designing a workplace-based saving scheme. A robust relationship might be established showing that those who enrolled in such schemes have higher net wealth. However, the possibility that the enrolees were a

self-selected group, possessing a 'squirrel gene' that made them superior savers in any event, cannot be dismissed.

Separating out causality from mere correlation is a ubiquitous problem. In a number of countries there is renewed interest in greater financial literacy. Surveys have shown that the more financially literate also have higher net wealth, after duly correcting for the influence of a host of other variables (Lusardi and Mitchell 2006). But we are inevitably left with the nagging doubt that having higher net wealth may have been the catalyst for greater investment in financial literacy. Caution is called for, before enthusiastically endorsing a public program in high schools on the subtleties of derivatives and credit swaps.

The skilful analysis of longitudinal panel data employing the latest in statistical techniques can help unravel some of these dilemmas. This further strengthens the case I have made for greater use of longitudinal data.

At times there will be no real evidence or even prior experience. While in some instances there may perhaps be some overseas evidence, we may well be reduced to faith-based policy making. However, that faith need not always spring from divine inspiration alone. It can come from an accepted body of theory — that is, we must rely on our theoretical models of how economies behave. An example might well be the benefits of unilateral tariff reform. In large surveys, the vast majority of economists have been found to agree that, as a rule, lowering tariffs will indeed enhance national income.

7.3 Structure

In this third section I address the structure or institutional frameworks that can enhance the use of evidence-based policy. In the first instance I would argue for the free and unfettered entry into the evidence-producing industry. Different perspectives, value judgments and methodologies should all have an opportunity to compete in the marketplace for ideas.

Research will seldom produce clean and unequivocal results. As Alfred Marshall noted:

> Every statement in regard to economic affairs which is short is a misleading fragment, a fallacy or a truism. (Alfred Marshall, in a letter to L. Fry, 7 November 1914, cited in Pigou 1925, p. 484).

Robust evidence is more likely to emerge from a process involving peer review. Universities, think-tanks, research centres and units within government all have a role. There is no single institutional model that would have universal applicability.

This plurality of providers assumes greater importance when it is recognised that in many agencies there is an inherent tension between their operational responsibilities and strategic research and evaluation — the very essence of evidence building. Such activities are often seen as diverting effort from the short-run demands placed on the agency. In such an environment there are clear benefits from having other providers.

A number of countries have established institutional structures for assembling evidence for regulatory impact statements (RISs). Examples include the Office of Information and Regulatory Affairs in the United States, the Regulatory Impact Unit in the United Kingdom, the Office of the Coordinator of Regulatory Reform in Canada and the Office of Best Practice Regulation in Australia.

In New Zealand, the Cabinet currently requires that RISs accompany new proposals, but this is not a statutory requirement. The RIS is a product of a regulatory impact analysis aimed at assessing the economic, social, cultural and environmental impacts of any proposed regulation. The RIS summarises key information covering the problem, objectives, options, impacts, preferred option, implementation and review and consultation (The Treasury 2009).

While this structure falls short of a separate watchdog agency to monitor regulatory behaviour, it does provide, at least potentially, for the assembly of all relevant evidence which itself can help to identify the problem (if any), to elucidate options and to build the case for the recommended option (Wilkinson 2001). To help ensure that the regulatory process is open and transparent, RISs are published either at the time the relevant bill is introduced to parliament or the regulation is gazetted, or at the time of ministerial release.

In a small country such as New Zealand there can be a legitimate sense of frustration that efforts to build and incorporate evidence in policy making are fragmented across a host of groups spanning the government, academia and not-for-profit organisations. Many of these suffer from chronic shortages of funding and have difficulty in attracting top analysts — in short, they are institutionally fragile. The history of creating enduring institutions which each had a role to assemble evidence for the government is not encouraging. The Monetary and Economic Council, the Planning Council, the Economic Development Commission and the Institute for Social Research and Development have all merged, withered or failed. The New Zealand Institute of Economic Research would be an exception. Economies of scale are hard to achieve in such a setting. However, in my judgment, the answer lies in the operation of informal networks and collaborative efforts.

Let me return to the concept of 'better' that I touched on in my introduction. In assessing the evidence we will seek the option that will lead to a better outcome. The use of cost–benefit analysis has traditionally been one tool for reaching that judgment. However, there are well known conceptual and practical difficulties in applying cost–benefit analysis (Cowen 2000). Even absent such hurdles, cost–benefit analysis can only tell us whether A is more 'efficient' than B. Applied in this way it can discriminate between regulatory options; some will pass the test, others will be rejected.

The difficulty is that few decision makers would accept screening proposals solely on the grounds of economic efficiency. Other criteria are invariably taken into account. Furthermore, political decisions are typically taken on the basis of the distributional consequences — which groups will be advantaged — rather than the overall welfare of society. Costs are not necessarily assessed against the marginal benefits. A recent substantial increase in the subsidy to tertiary students was arguably driven more by advantaging a particular group than by enhancing the social return from public investment in education. So, rather than being a final arbiter, cost–benefit analysis can at least be an indicator of economic efficiency, identify 'lemons' that should proceed no further, and provide some sense of the trade-off implied when efficiency is sacrificed to attain other goals.

There is one dimension of having the appropriate institutional structures that may be overlooked — namely, that of communication. It is a fact that public perception is not always aligned with the evidence. Two examples will serve to illustrate this proposition. Parents (and, invariably, teacher unions) clamour for smaller classes. Not infrequently, policy makers respond, despite the mixed evidence from the Organisation from Economic Co-operation and Development that, at best, class size is weakly correlated with educational outcomes.

My second example concerns perceptions of risk. We often overestimate the probability of low-risk events and underestimate the probability of high-risk events (Viscusi 1983). Such misperceptions can be costly in so far as they lead to inappropriate policies, overinvesting in certain safety requirements (for example, oxygen on aircraft) and under-investing in other areas (for example, prostate cancer).

Communication, in its broadest sense, can operate through other channels. As researchers and business leaders establish credible track records, they may be invited onto public commissions (for example, on tax, infrastructure or innovation), thereby providing another institutional channel to influence the policy-making process.

7.4 Concluding remarks

As analysts, how are we to know when we have been successful — that is, when our evidence was, in fact, the basis for policy? Sometimes, perhaps too infrequently, we can point to a policy decision that was, in fact, evidence based; I have cited some examples from New Zealand. At other times we might have provided solid evidence but for political reasons certain options were not pursued. The 2001 McLeod Tax Review made a strong case, based on first principles, for the taxation of imputed rents on owner-occupied housing and for the taxation of capital gains (The Treasury 2001). However, the government of the day (doubtless after hearing the initial public clamouring) declined to pursue either option. But, presumably, we should not classify that as a failure.

More difficult is the case where the evidence strongly indicated that a certain policy would not be desirable and, heeding the evidence, policy makers opted not to pursue the matter. In other words: how many times has 'good evidence' been responsible for avoiding the implementation of 'bad policy'?

Presumably evidence gatherers would want this counted as a success — and it might well be so. But deciding on the counterfactual is always a messy business — and we should probably avoid creating incentives for bureaucracies to succumb to the temptation to claim credit for all sorts of bad things that might have happened but did not. Such a strategy would quickly see reports to ministers being filled with long lists of undesirable options each with a solid base of evidence as to why they should not be adopted.

Let me conclude by returning to the question posed for this session by the organisers, namely: how robust is our evidence-based policy making? You will forgive me if I adopt the position of the student who wrote on his examination paper: 'I don't know the answer to that question, but let me give you the answer to another question'.

My reason for adopting this devious approach is my conviction that we cannot generalise about evidence-based policy making. Each circumstance is different — the historical, economic and political context of each policy debate has its own characteristics and its own dynamic. Sometimes the process will get high marks for robustness — in other cases, the process will be fraught and fragile.

My alternative question is this: what steps can be taken to ensure evidence-based policy making is made more robust? In this regard I have argued that a number of steps can be taken. They relate in turn to each of the three key areas I have addressed. These include:

- attending to the unglamorous but basic task of data collection and management
- improving the conceptual frameworks and models we use to convert that data into information
- ensuring that the institutional arrangements we have for assembling, processing and communicating evidence in a manner that is useful for policy making are relevant and responsive.

It seems to me that if we can make advances, albeit in mincing steps, on all three of these fronts, the robustness of evidence-based policy making will be enhanced. The good news is that in many countries, including my own, we can point to progress.

The bad news is that for many of the really big, important questions the prospects for greater reliance on evidence are, I would argue, still rather slim. Many of these questions involve institutional change and, like the New Zealand delegation to the federation conventions, the ultimate decision is typically not informed by evidence. This not because of a lack of diligence or goodwill, but rather because the very nature of institutional change takes us into unchartered territory where evidence is a scarce commodity.

Let me illustrate this with a few examples from current policy debates:

- Should we revamp the tax system to place greater emphasis on taxing immobile rather than mobile factors, and place more reliance on consumption as distinct from income taxes?
- Should New Zealand enter into a free-trade agreement with country X?
- Should we institute a permanent retail deposit guarantee scheme?
- Should future obligations of the state (in, say, health, education and retirement incomes) be partly pre-funded?
- Should significant new prudential regulations be introduced for the finance sector?

While I do not wish to end on too pessimistic a note, I can only conclude that evidence-based policy making will at times remain a minor ingredient in the policy mix for the 'stuff that really matters'. True, historical episodes and the experience of others can and should be well canvassed for insights. And, at the very least, empirical evidence can help settle factual disputes. But, in the end, wise heads, cool judgments, the wisdom of crowds and ultimately political imperatives will determine how we respond to the tough 'wicked' questions. Evidence-based policy will remain a two-bit player in the big policy arenas.

Political leaders wax and wane in their demand for evidence. Abraham Lincoln summed up his approach in a speech to the Republican convention in Springfield, Illinois, when he said:

> If we could first know where we are, and whither we are tending, we could then better judge what to do and how to do it. (Abraham Lincoln, 16 June 1858)

It is difficult to imagine a more apt and succinct challenge as we seek to strengthen evidence-based policy.

References

Coleman, A. and Scobie, G.M. 2009, 'A Simple Model of Housing Rental and Ownership with Policy Simulations', Working Paper 09/05, The Treasury, New Zealand Government, Wellington.

Cowen, T. 2000, 'Cost-Benefit Analysis and its Policy Limitations', George Mason University.

Enright, J. and Scobie, G.M. 2001, 'Healthy, Wealthy and Working: Retirement Decisions of Older New Zealanders. Working Paper (forthcoming), The Treasury, New Zealand Government, Wellington.

Fabling, R., Grimes, A., Sanderson, L., and Stevens, P. 2008, 'Some Rise by Sin and Some by Virtue Fall: Firm Dynamics, Market Structure and Performance', Occasional Paper 08/01, Ministry of Economic Development, New Zealand Government, Wellington.

Grimes, A., and Aitken, A. 2006, 'Housing Supply and Price Adjustment', Working Paper 06-01, Motu Economic and Public Policy Research, Wellington.

Hall, J. and Scobie, G.M. 2006, 'The Role of R&D in Productivity Growth: The Case of Agriculture in New Zealand: 1927 to 2001', Working Paper 06/01, The Treasury, New Zealand Government, Wellington.

Hanushek, E. 1999, 'The Evidence on Class Size', in Mayer, S.E. and Peterson, P. (eds.) *Earning and Learning: How Schools Matter*, Brookings Institution Press, Washington D.C., pp. 131–68.

Henderson, K. and Scobie, G.M. 2009, 'Saving Rates of New Zealanders: A Net Wealth Approach'. Working Paper 09/04, The Treasury, New Zealand Government, Wellington.

Holt, H. 2009, 'Health and Labour Supply'. Working Paper (forthcoming), The Treasury, New Zealand Government, Wellington.

Howden-Chapman P., Matheson, A., Crane, J., Viggers, H., Cunningham, M., Blakely, T., Cunningham, C., Woodward, A., Saville-Smith, K., O'Dea, D., Kennedy, M., Baker, M., Waipara, N., Chapman, R. and Davie, G. 2007, 'Effect of Insulating Existing Houses on Health Inequality: Cluster Randomised Study in the Community', *BMJ*, March, vol. 334, no. 7591, p. 460.

Howden-Chapman P., Pierse, N., Nicholls, S., Gillespie-Bennett, J., Viggers, H., Cunningham, M., Phipps, R., Boulic, M., Fjällström, P., Free, S., Chapman, R., Lloyd, B., Wickens, K., Shields, D., Baker, M., Cunningham, C., Woodward, A., Bullen, C. and Crane, J. 2008, 'Effects of Improved Home Heating on Asthma in Community Dwelling Children: Randomised Controlled Trial', *BMJ*, September, vol. 337, no. 231, p. a1411.

Kerr, S., and Lock, K. 2008, 'Nutrient Trading in Lake Rotorua: Choosing the Scope of a Nutrient Trading System', Working Paper 08-05, Motu Economic and Public Policy Research, Wellington.

Lusardi, A., and Mitchell, O.S. 2006, 'Financial Literacy and Planning: Implications for Retirement Well-being', Working Paper 2006-01, Pension Research Council, Wharton School, University of Pennsylvania .

Odgers, C.L, Moffitt, T.E., Broadbent, J.M., Dickson, N., Hancox, R.J., Harrington, H., Poulton, R., Sears, M.R., Thomson, W.M. and Caspi, A. 2008, 'Female and Male Antisocial Trajectories: From Childhood Origins to Adult Outcomes', *Development and Psychopathology*, vol. 20, no. 2, pp. 673–716.

Pigou, A.C. (ed.) 1925, *Memorials of Alfred Marshall*, Macmillan, London.

Shanks, S. and Zheng, S. 2006, 'Econometric Modelling of R&D and Australia's Productivity', Staff Working Paper, Productivity Commission, Canberra.

Stillman, S. and Hyslop, D. 2006, *Examining Benefit-to-Work Transitions Using Statistics New Zealand's Linked Employer-Employee Data*, Statistics New Zealand, New Zealand Government, http://www.stats.govt.nz/publications/workknowledgeandskills/leed-reports/examining-benefit-to-work-transitions-using-leed.aspx (accessed 20 January 2010).

The Treasury 2001, *Tax Review 2001*, New Zealand Government, http://www.treasury.govt.nz/publications/reviews-consultation/taxreview2001 (accessed 20 January 2010).

—— 2009, *Regulatory Impact Analysis Handbook*, New Zealand Government, http://www.treasury.govt.nz/publications/guidance/regulatory/impactanalysis (accessed 20 January 2010).

Viscusi, W.K. 1983, *Risk by Choice: Regulating Health and Safety in the Workplace*, Harvard University Press, Cambridge, Massachusetts.

Wilkinson, B. 2001, 'Constraining Government Regulation', Discussion Paper, New Zealand Business Roundtable, Wellington.

General discussion

After Session 2's emphasis on the practice of applying evidence to policy analysis, discussion focussed on the fundamental role of data access in triggering wider and better analysis, the potential for better use of academics' skills, and technical issues in cost-benefit analysis.

Data access

Reflecting on Grant Scobie's examples of how access to new data sets had enabled better New Zealand policy analysis, several participants noted that an advantage of freer access to data (which nevertheless met proper privacy protections) was that it broadened the ranks of analysts beyond public servants, whose analysis was often confidential to governments and who could not directly contribute to public analytical debate. Additional transparency, more sources of analysis and better checking of findings though replication would provide appropriate 'checks and balances' for government as well as strengthening the quality of analysis. Better access to data for academics, not-for-profit organisations, research bodies and economic consultancies, would help governments to best capitalise on the existing resource base.

Bruce Chapman noted the breakthrough for Australian analysis which had come with better Australian Bureau of Statistics pricing structures for academic institutions. Before that reform, Australian academics had to buy individual access to confidentialised record files of Australian data, and could do better work on UK, US and Canadian policy (because of free data access) than they could do on Australian policy (because of the high cost of access to data). As Jeffrey Smith put it, by making 'research data sets available to academics you'll get a big pile of free policy-relevant research — because that's what academics do'.

Another participant sounded a cautionary note against general appeals for more data, given that data collection is not costless — governments, businesses, individuals and community groups all pay a price in some form for providing data. Rather, she argued that governments should focus on collecting the right type of data to answer policy questions and make better use of the data they already collect.

Academics' role and Public Service interaction

The discussion touched on the role of public servants in evidence-based policy and, in particular, the challenge of providing independent, public advice that may conflict with a government's position. One participant asked whether this could be addressed in the public service code of conduct, with public servants having an obligation or right to publicly state their findings. Others thought this would be inappropriate.

Bruce Chapman and another participant observed that academics sought relevance for their work, and prized dialogue with governments and officials that could help shape a useful analytical agenda. Better liaison between bureaucrats and academics could improve the use of evidence to better inform policy.

Issues in Cost-Benefit Analysis

Jeffrey Smith noted that many cost-benefit analyses in North America generally did not address the issue that publicly-funded projects involved a cost greater than their net fiscal cost because of the excess burden of raising finance. (The cost of raising a dollar of tax revenue exceeds a dollar because of the cost of operating the transfer system and the distortionary nature of taxes.) He asked how Australian evaluators handle that issue. Henry Ergas responded that usually there was no attention to adjusting the net cost for the marginal social cost of raising funds, which in his view leads to serious errors in cost-benefit appraisal. His impression was that Australian cost-benefit analysis practice was inferior to practice in France or Finland, and suggested that as well as actually performing more publicly available cost-benefit analysis for major projects, it would be useful for the Australian Government to develop some guidance notes covering these technical issues.

SESSION 3

FROM RHETORIC TO PRACTICE — HOW DO WE IMPROVE THE AVAILABILITY AND QUALITY OF EVIDENCE?

8 Facilitating better linkages between evidence and health policy: the role of the Cochrane Collaboration

Sally Green and Miranda Cumpston
Australasian Cochrane Centre

Abstract

Partly because of the large volume of scientific research available to today's policy makers, using research evidence in shaping policy is still difficult. The Cochrane Collaboration aims to increase the use of evidence in health care decisions by overcoming the barriers to research use, such as lack of contact with researchers or the need for skills in critical appraisal. The Cochrane Collaboration publishes systematic reviews of reliable research in The Cochrane Library, and is conducting research into strategies to effectively support policy makers in using high quality research evidence.

8.1 Introduction

> There is a gap between today's scientific advances and their application: between what we know and what is actually being done. (Lee Jong-wook, former Director-General, World Health Organization 2004)

This quote sums up the challenge of the work of the Cochrane Collaboration. This paper, and the presentation that accompanies it:

- introduces the Cochrane Collaboration, including the purpose and scope of Cochrane systematic reviews
- outlines strategies to support the use of research evidence in policy-making environments, including an overview of the Australasian Cochrane Centre's Policy Liaison Initiative.

8.2 What is a Cochrane systematic review?

A systematic review is a scientific tool used to summarise, appraise and communicate the results and implications of otherwise unmanageable quantities of research. Systematic reviews bring together all the separately conducted studies answering a particular healthcare question, sometimes with conflicting findings, and synthesise the results. Systematic reviews are an efficient way to access all of the available research to answer a question. In this way, systematic reviews recognise that science is cumulative and all research should be viewed in the context of what has been done before, aiming for decisions based on a body of evidence, rather than a single piece of research.

The Cochrane Collaboration is an international, not-for-profit organisation that aims to help people make well-informed decisions about health care by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions (www.cochrane.org). The Cochrane Collaboration looks at health decision making at the policy, individual practitioner and consumer levels, and endeavours to provide information that is evidence-based, easily accessible, internationally developed, quality controlled and useful. Cochrane reviews are updated every two to three years. The questions addressed by Cochrane reviews are predominantly those of clinical intervention, although we have many reviews relevant to effective practice and organisation of health care, public health and consumers and communication, as well as a new initiative for systematic reviews of diagnostic test accuracy.

Cochrane reviews are published electronically in *The Cochrane Library* (www.thecochranelibrary.com). In Australia we are privileged to have a national subscription to *The Cochrane Library*, funded by the Department of Health and Ageing.

While Cochrane reviews are predominantly of randomised control trials, there are several examples where Cochrane reviews extend beyond randomised control trials depending on the question (for example, reviews to inform the organisation of health systems or public health interventions are likely to synthesise research other than randomised controlled trials).

The Cochrane Collaboration has a sister organisation, the Campbell Collaboration, which is undertaking similar work preparing and maintaining systematic reviews relevant to social policy and education. While more recently established than Cochrane, Campbell is publishing an increasing number of reviews, which are available at its website (www.campbellcollaboration.org).

8.3 How do we strengthen the uptake of evidence from research into health policy?

There are barriers to using research evidence in both practice and policy environments, resulting in a 'gap' between the large amount of information available, much of which is based on very high quality scientific research, and how much of that information makes its way into policy and practice. The most frequently identified barriers to the use of research evidence are lack of time; lack of access to research; lack of skills to find, appraise and apply research; and lack of capacity within organisations to support research use. The Cochrane Collaboration aims to overcome these barriers by publishing systematic reviews of all available reliable research evaluating healthcare interventions in an accessible format.

In addition to synthesising and disseminating evidence, The Cochrane Collaboration is increasingly working to close the gap by complementing our reviews with strategies for evidence-based implementation. There is a research agenda evolving in this area to identify and evaluate effective strategies to integrate high-quality research evidence into health policy, fostering communication and exchange between researcher and policy makers, yielding better policy decisions and ultimately improved health outcomes.

There are many different models describing the complexity of health-care decision making and the role of research in supporting evidence-based policy. Figure 8.1 details one of those models, highlighting the pivotal importance of research, balanced with the many other factors at play when policy is developed. Within that complexity, however, policy makers have identified factors that can facilitate their use of evidence.

A systematic review of different strategies to increase the uptake of research-based information into health policy (Innvær et al. 2002) includes 24 interview-based studies with a total sample of 2000 policy makers around the world. The greatest barriers to research use were found to be lack of personal contact and trust between researchers and policy agencies; that research is not always timely or relevant to policy decisions; power and budget struggles; poor quality of available research; and political instability.

Another study has investigated factors that predict the use of systematic reviews by health policy makers. Key factors included demonstrating a culture that values the use of research for decision making within the organisation; provision of access to online database searching; and training in critical appraisal for teams. Also, if the user believed that systematic reviews reduce the time required to find and use evidence, help to overcome lack of critical appraisal skills, or were easy to use, the use of reviews increased (Dobbins et al. 2007).

Figure 8.1 **A healthcare decision-making model**
Factors influencing the role of research in policy decision-making

Source: Davies (2005).

A number of policy organisations internationally are now beginning to implement strategies to address these identified factors and remove barriers to the use of evidence in policy. Those strategies aim to enable policy makers to access and use evidence not only during large-scale evaluations and for new policy proposals, but as a routine part of their daily activities.

8.4 A case study: the Policy Liaison Initiative

An example of facilitating the use of evidence in policy making is the Australasian Cochrane Centre's Policy Liaison Initiative. This work, funded by the Department of Health and Ageing, is designed to address policy makers' needs and to help make Cochrane reviews more accessible to the policy-making environment.

The initiative was established in 2004, incorporating the available evidence on effective interventions of this kind, and informed by surveys of policy makers that identified their needs and perceptions in relation to research. The initiative is focused within Australia's National Health Priority Areas, but also includes reviews relevant to effective practice and organisation of care, consumers and communication, and public health.

Strategies aiming to facilitate greater access to Cochrane reviews within the Policy Liaison Initiative are those termed 'facilitating user pull' (Lavis 2006); that is, those aiming to increase research receptiveness and the usability of systematic reviews. We have a website, indexed by subject, which includes summaries of relevant Cochrane reviews as well as other information resources relevant to evidence-based policy. As one of the barriers to evidence use is lack of personal contact, a staff member at the Australasian Cochrane Centre serves as a policy liaison officer, available by email or phone to answer the questions of individual staff members. We prepare quarterly bulletins and have provided numerous seminars and training workshops on subjects such as asking answerable questions, how to search and find evidence, critical appraisal, implementation and e-health.

8.5 Conclusion

Many opportunities to use evidence from research in the development and evaluation of policy are currently missed. Policy makers describe many barriers to using research findings, one of which is difficulty accessing summaries and syntheses of research. The Cochrane Collaboration, through the preparation, publishing and maintenance of reliable systematic reviews, and through ongoing efforts to support their implementation in practice and policy, has an important role to play in facilitating better linkages between evidence and health policy.

References

Davies, P. 2005, Workforce development to support evidence-informed public health, presentation delivered at Cutting Edge Debates, Melbourne, on 27 October 2005

Dobbins, M., Rosenbaum, P., Plews, N., Law, M. and Fysh, A. 2007, 'Information transfer: what do decision makers want and need from researchers?', *Implementation Science*, vol. 2, no. 20.

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

Lavis, J. 2006, 'Perspective: ideas at the margin or marginalized ideas?', *Journal of Continuing Education in the Health Professions*, vol. 26, pp. 37–45.

WHO (World Health Organization) 2004, *World Report on Knowledge for Better Health*, World Health Organization, Geneva, November 2004.

9 Learning from the evidence about evidence-based policy

Patricia J. Rogers
CIRCLE (Collaboration for Interdisciplinary Research, Consulting and Learning in Evaluation), RMIT University

Abstract

From the long history of efforts to improve policy by drawing systematically on evidence about effectiveness, a number of recommendations can be made. The approach to evidence-based policy needs to be matched to each particular situation, especially in terms of whether the intervention has complicated or complex aspects. The quality of evidence about effectiveness should be judged not by whether it has used a particular methodology, but whether it has systematically checked internal and external validity, including paying attention to differential effects. The availability of evidence can be improved through supporting the different processes of knowledge transfer, knowledge translation and ongoing knowledge generation. Transparent processes of generating and using evidence are needed, including access to data to allow reviews of its quality and of the conclusions drawn.

9.1 Introduction

Important lessons should be drawn from the long history of efforts, in Australia and internationally, to improve public policy by drawing systematically on evidence. This history dates back at least to Lind's study of scurvy in the British Navy in the 1700s, Snow's investigation of cholera in London and Semmelweis's unsuccessful attempts to reduce maternal mortality from puerperal fever in the mid-1800s, and Rice's comparative assessment of approaches to teaching spelling in the United States in the 1890s. More recent efforts to base public policy on empirical evidence have used diverse methods and approaches, including experimental designs since the 1960s, action research since the 1970s, performance indicators since the 1980s, and more recently methods from epidemiology, statistics, philosophy and complexity science, including case-control designs, propensity scores, realist

synthesis, and systems dynamics. Some discussions of evidence-based policy, which ironically fail to draw on these experiences, risk repeating mistakes and having to rediscover what constitutes quality evidence.

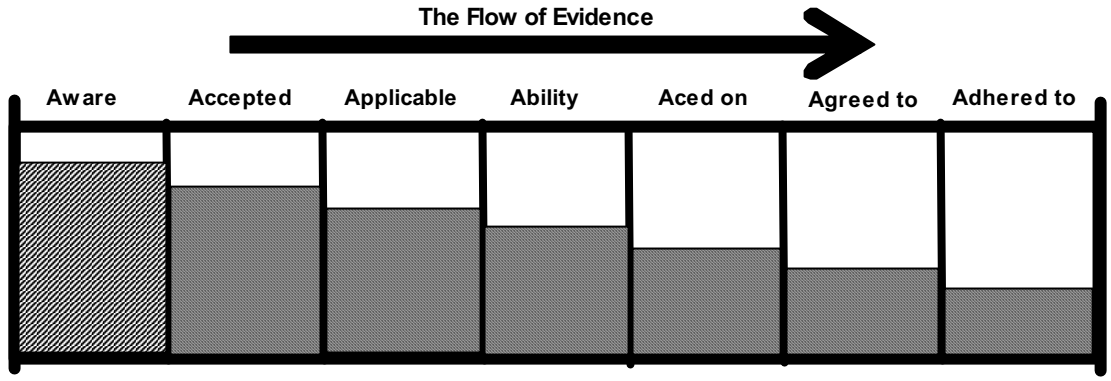
9.2 Processes of evidence-based policy need to be different for interventions with simple, complicated and complex aspects

Public policy interventions are diverse, and the processes of evidence-based policy need to match their varied features. In particular, it is important to distinguish between interventions that are essentially simple (consisting of a single, well-defined and predictable process); and those that have important aspects that are complicated (involving multiple components or processes that work differently in different situations); or complex (dynamic and emergent). This three-part distinction, drawn from complexity science (Glouberman and Zimmerman 2002; Kurtz and Snowden 2003; Stacey 1992), has been shown to be useful for planning evaluations (Patton 2008; Ramalingam et al. 2008; Rogers 2009). It can be applied to interventions of any scale — projects, programs, strategies or policies — and is most useful when it is applied to aspects of interventions rather than used to classify an entire intervention.

Processes for evidence-based policy about simple interventions

Some interventions can be characterised as essentially simple — that is, they are both necessary and sufficient to produce the intended result, and work in the same way in different settings and for different people. Some (but not all) vaccination programs might be usefully thought of in this way. In these programs, everyone who is vaccinated develops antibodies and immunity against the disease, and no-one develops immunity without the intervention. Therefore, simple with/without comparisons between treatment and control group are adequate. If the benefits of avoiding the illness outweigh the costs of administering the vaccination, then the policy decision is also simple — implement the program for everyone. Uptake of evidence is also simple — replicate the procedures used in the trial. Uptake of evidence about simple interventions focuses on compliance with the research evidence. For interventions of this type, it can be appropriate to think of the process of evidence-based policy as a ‘leaky pipeline’ where evidence uptake can only be compliant or something less then this (Glasziou 2006).

Figure 9.1 **Evidence uptake as a 'leaky pipeline'**



Source: Glasziou (2006).

Although few, if any, interventions are totally simple, it can sometimes be useful to think of them in this way, and to focus on the average effect of an intervention, identify 'what works', introduce it at all sites, and monitor compliance. However, not all interventions are like this.

Processes for evidence-based policy about important complicated or complex aspects

Where interventions have important complicated or complex aspects, it can be unhelpful or even dysfunctional to use this simple model of evidence-based policy, and to use research to make 'one-size-fits-all' policy recommendations.

Interventions often have important complicated aspects, where results differ in different situations — different implementation environments, different participant characteristics, or in conjunction with other interventions. These differential effects can be critically important. Sometimes it means that an intervention is only effective for some groups and less effective, useless or even harmful for others. In these circumstances, the average effect is a poor guide for policy and for practice.

For example, a review of early intervention programs for children in disadvantaged families found some programs which were effective on average but which were either ineffective or damaging for some of their participants (Westhorp, 2008). Those who did not benefit or who showed negative outcomes often had multiple and complex needs or were concentrated amongst the most disadvantaged families. The Early Head Start program, for example, was found to have unfavourable impacts on child development outcomes in families with multiple risk factors (Mathematica Policy Research Inc. 2002).

For interventions with important complicated aspects, research and evaluation need to go beyond 'What works (on average)?' to answer the question 'What works for whom, in what circumstances?' An effect that only occurs in particular situations can be invisible if results show only the average effect. For example, after the introduction of the British Road Safety Act, which introduced penalties for drink driving, time-series data of road fatalities showed no apparent effect until they were disaggregated to look particularly at data for Friday and Saturday nights (Glass 1997).

If an intervention works quite differently for different people or in different situations, how should policy address this? What are the risks in developing a policy based on the average effect? Should policy require an intervention that works best on average, or for the most people, or for the most disadvantaged? For example, since the chronological age at which children are ready to start school varies considerably, should the policy enforce a minimum to reduce the risk of children being sent too early, even though this can increase the risk of them having to wait too long, or allow differential practice — and, if so, should these decisions be made by those who fund services, or be delegated to service deliverers, or to service recipients?

Where an intervention only works in particular situations, the evidence needs to be disaggregated to show this, and then the practical significance of this selective effect needs to be assessed. Should an intervention be ignored if it is not a 'silver bullet' but only works in particular circumstances? Will it be possible to change the situation at other sites, or tailor the intervention itself, so it is effective in more places? Or is this more limited effect by itself worthwhile? Does policy need to specify conditions under which it is to be used?

Developing a complicated message from evidence, while it might represent the evidence well, raises additional challenges for those who will use that evidence in a particular situation For example, the bush safety policy of 'Stay and Defend Your Property or Go Early' (commonly abbreviated as 'Stay or Go') appears to have been too complicated for residents to understand and apply appropriately without assistance (ABC News 2009), as they needed to take into account the particular characteristics of their property, their household and weather conditions in order to choose the appropriate action (Bushfire CRC 2006).

The process of evidence uptake for interventions with complicated and complex aspects involves translation into new settings, including appropriate adaptation. Evidence-based policy needs to support this process of translation and to document and learn from it as well. In a recent seminar for the Australian Research Alliance for Children and Youth on the processes of scaling up successful pilots, Professor

Homel highlighted the need for evidence in terms of 'implementation science', including the factors that affect implementation quality; the factors that affect engagement and sustained participation; the effect of management, organisation, financing and training on outcomes; and the types of coordination needed (Homel et al. 2009).

A recent project, the Catholic Education Office Melbourne's Literacy Assessment Project, has demonstrated how, with appropriate support, service deliverers can customise interventions to meet the particular needs of recipients. The leader of the project explained: 'We weren't telling the teachers how to teach. We were helping them to make decisions based on data.' (Griffin 2009).

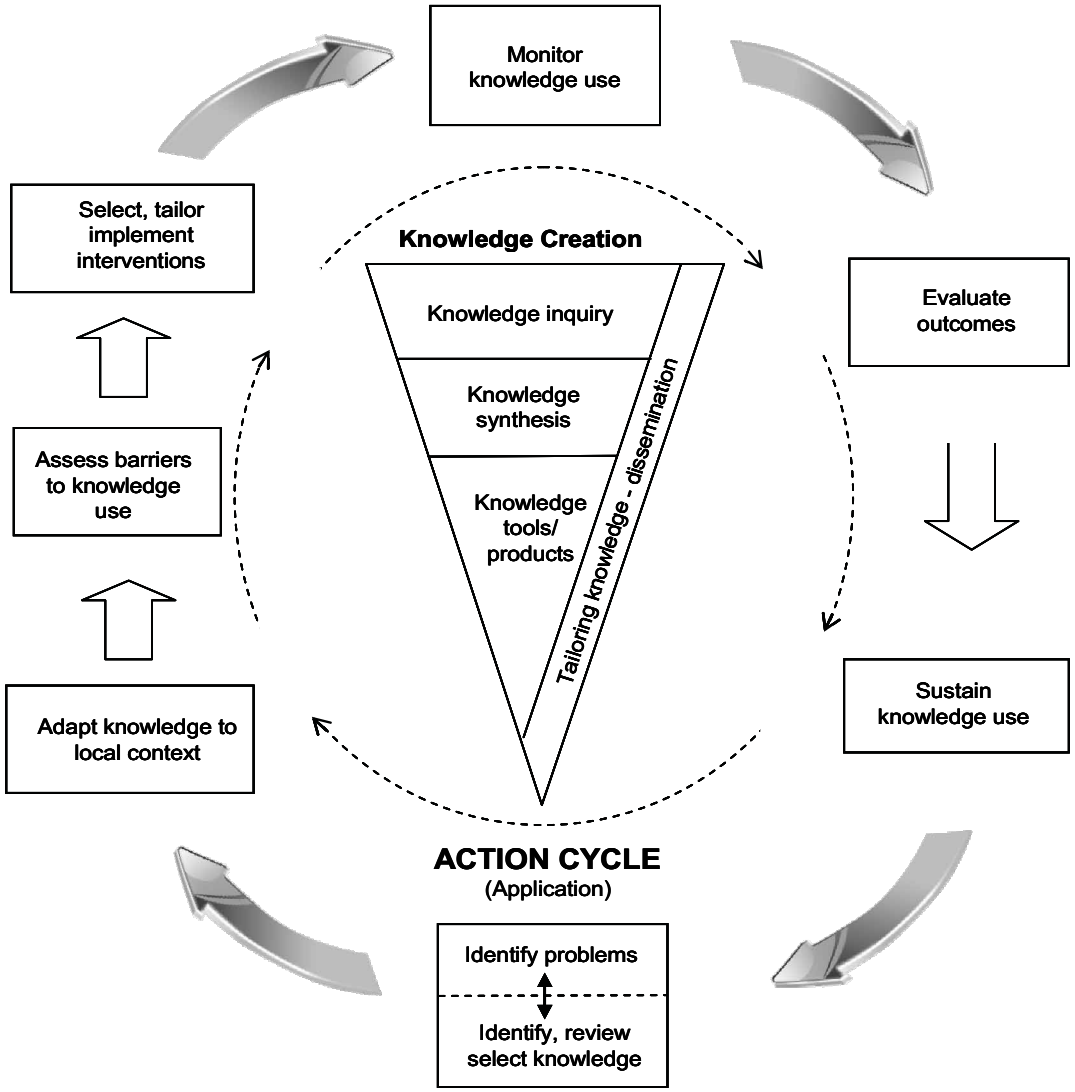
For complex interventions, which need to be constantly adaptive in response to emerging needs, challenges and opportunities, the classic linear approach to evidence-based policy presents even more challenges. There is no standardised intervention to study or replicate, but an ever-changing program. For interventions with important complex aspects, evidence-based policy involves developing broader principles to guide this adaptive practice, and for continuously learning and disseminating evidence, while recognising its limited generalisability and rapid obsolescence.

This process of learning can be more effective when those who are expected to use the learning are not just told the results, but are engaged in the process of generating it. 'Positive deviance' (Sternin and Pascale 2005) is a process for working with organisations or communities to solve seemingly 'intractable' problems that require behavioural and social change, such as child malnutrition in developing countries, female genital mutilation and methicillin-resistant *Staphylococcus aureus* (MRSA) (Morris 2009). Instead of an expert working to identify problems and suggesting ways to fix them, members of the community are supported to identify cases where exceptionally good results are being achieved, and to work together to see what can be learned from them and how they can be more widely implemented.

This different view of the evidence-based policy process is better represented by the iterative, spiral model shown in figure 9.2, with knowledge generation throughout the process and an ability to start from a piece of research or an identified success in practice.

Figure 9.2 **Evidence uptake as an ongoing, knowledge building process**

Knowledge to action process



Source: Graham et al. (2006).

These differences are summarised in table 9.1.

Table 9.1 **Approaches to evidence-based policy**
Simple, complicated and complex

	Simple	*Complicated*	*Complex*
What interventions look like	Discrete, standardised intervention	Interventions that are different in different situations, or that work only in conjunction with other components	Non-standardised and changing, adaptive, and emergent in response to changing needs, opportunities and understandings of what is working
How interventions work	Pretty much the same everywhere	Differently in different situations (different people or different implementation environments), which can be clearly identified	Generalisations rapidly decay, and results are sensitive to initial conditions as well as to context
Question needed for evidence-based policy	What works?	What works for whom in what contexts?	What is working and how?
Nature of advice given to policy	Single way to do it	Contingent	Dynamic and emergent
	Best practices	Good practices in particular situations	Principles
Process needed for evidence uptake	Knowledge transfer	Knowledge translation to new situations	Ongoing knowledge generation
Metaphor for evidence-based policy	Google directions	Transport map and timetable	Topographical map and compass

9.3 Credible comparative effectiveness evidence does not come only from RCTs (randomised controlled trials) nor do RCTs always provide it

With increasing explicit attention to evidence-based policy has come advocacy for particular methods — in particular for randomised controlled trials (RCTs) in which participants are randomly assigned either to a treatment group or to a control group. Some organisations, mostly based in the United States, such as the Coalition for Evidence-based Policy (2006) have advocated for the use of RCTs wherever possible, while other organisations, such as the Poverty Action Lab (2009), based at MIT, and the US Department of Education Institute of Education Sciences (2003), have defined strong evidence exclusively in terms of RCTs.

While evidence from RCTs can make a valuable contribution to policy, there are serious risks in judging the quality of the evidence by whether or not it uses RCTs. It is important to understand how evidence from RCTs can sometimes be misleading and how evidence from sources other than RCTs can sometimes be

credible. Uninformed advocacy for RCTs risks reducing the quality of evidence being used for policy by encouraging processes of evidence generation and synthesis and capacity building that include poor quality evidence from RCTs and exclude high quality evidence from other designs.

RCT data can provide poor quality evidence of effectiveness

It cannot be assumed that RCTs will always provide high-quality evidence of effectiveness, and care must be taken to avoid both false positives (where an intervention is incorrectly seen as effective) and false negatives (where an intervention is incorrectly seen as ineffective). While these issues have long been acknowledged in the methodological literature, they are not always evident in arguments advocating for the privileging of RCTs to build evidence for policy.

Potential quality issues in the conduct or interpretation of RCTs that can affect the validity of conclusions include poor measurement, poor adherence to randomisation, inadequate statistical power, unidentified differential effects, inappropriate comparisons, conducting numerous statistical analyses and only reporting statistically significant ones, differential attrition between control and treatment groups, unplanned crossover, and unacknowledged poor quality implementation of the intervention.

Even if these issues are addressed, other potential threats to validity remain: random error, treatment leakage, incomplete causal package, lack of blinding, limited effectiveness in real world practice, and questionable transferability to new situations.

Random error can occur when, due to the uncertainties of randomisation, treatment groups and control groups are not equivalent on all observable and unobservable variables. While a good RCT will include assessment of the comparability of treatment and control group on observable variables, it cannot assess comparability on unobservable variables, which creates a risk that differences in results may be due to unobserved differences between the groups (Worrall 2002). This is not simply a theoretical problem. In a study published in the *New England Journal of Medicine* (Concato et al. 2000), researchers compared findings about the effectiveness of five different clinical interventions produced from RCTs as compared to observational studies (using cohort or case-control designs). They found that, while the summary results from RCTs and observational studies were 'remarkably similar', findings from RCTs showed more variation between studies — to the extent that some of them produced findings at odds with results from the

other studies. This threat to validity means that no single RCT should be presented as providing a definitive answer.

Treatment leakage refers to ways in which the 'control' group actually receives access to the treatment. For example, in evaluations of *Sesame Street*, Comer schools, Head Start and drug and alcohol services for homeless men, results from experimental designs appeared to show that the interventions had no effects, until further investigation showed that the control group had accessed services from another source, or in other ways received something close to the treatment (Datta 2003).

Results from RCTs can be misleading for complicated interventions, when the intervention is effective only in particular circumstances. While it is possible for RCTs to examine differential effects if cell sizes are adequate and data are collected on the contextual variables, most examples of RCTs only report the average effect.

Clinical trials require double-blinding so that neither participants nor researchers know who has been allocated to the treatment and control groups. The difficulty, and sometimes impossibility, of achieving double-blinding raises more questions about interpreting results from RCTs, especially given the increasing recognition of the importance of the placebo effect.

Finally, there can be difficulties in extrapolating findings of efficacy in RCTs to likely effectiveness when treatments are scaled up or transferred to other contexts.

None of these issues are grounds for rejecting the use of RCTs, but they make it clear that a single RCT by itself will not provide definitive findings for most interventions, no matter how large or well implemented. Given the difficulties in adequately addressing these challenges for human services, there will be many situations where RCTs will not be suitable.

Non-RCT data can provide good quality evidence of effectiveness

Good quality evidence of effectiveness can also come from quasi-experimental approaches, which compare program participants to a comparison group rather than to a randomly assigned control group, and from non-experimental approaches, when such approaches systematically and rigorously test causal conclusions and combine evidence thoughtfully.

Sudden Infant Death Syndrome (SIDS) is one of two exemplars in the National Health and Medical Research Council guide *How to Put the Evidence into Practice: Implementation and Dissemination Strategies* (NHMRC 2000). It shows both the

value of drawing on a diverse set of evidence and how it is possible to develop effective policy even when the evidence is not definitive. Bringing together evidence from many studies, including retrospective and prospective epidemiological studies, pathological studies and case studies, a number of possible contributing factors were identified, and other possible causes (such as vaccinations) were ruled out. On the basis of this incomplete evidence, recommendations were developed — to put babies to sleep on their backs, avoid overheating and avoid cigarette smoke. No RCTs were used to test the effectiveness of these recommendations. The recommendations were communicated directly to parents and to health professionals working with parents, resulting in widespread change in the sleeping positions they used for infants. By 2005, the number of SIDS deaths had been reduced to fewer than 100, a decline of 83 per cent (ABS 2007).

This does not mean that any sort of anecdotal evidence should be considered adequate evidence of effectiveness. Other types of evidence should be rigorously analysed using general elimination methodology (GEM) (Scriven 2008), an approach to scientific inquiry that involves systematically identifying and ruling out alternative causal explanations for observed results, and multiple lines and levels of evidence (MLLE). MLLE involves bringing together different types of evidence, and systematically analysing the strength of the causal argument linking an intervention or a cause and its effects. This analysis might consider the strength of the observed relationship, specificity, temporality, coherence with other accepted evidence, plausibility, analogy with similar interventions, biological plausibility, dose and consistency of association. Given the specialist and often cross-disciplinary nature of the scientific evidence, the investigation is undertaken by a panel of credible experts, spanning a range of relevant disciplines, who are asked to judge the credibility of the evidence and the causal analysis (for example, Cottingham et al. 2005). MLLE has been used in human and ecological risk assessments and natural resource management (for example, Downes et al. 2002; Boyes 2006; NSW DECC 2009).

If only evidence from RCTs is included in syntheses, conclusions can be incorrect

Some of the limitations of RCTs, in particular random error and generalisability, can be addressed by synthesising multiple studies. However, limiting such syntheses to RCTs, as advocated by the Campbell Collaboration, can be problematic.

Where little RCT evidence is available, meta-analyses that consider only certain types of evidence can produce misleading or unhelpful conclusions. The limitations

of such an approach were demonstrated in a systematic review of the use of parachutes to prevent accidental death, published in the *British Medical Journal* (Smith and Pell 2003). The authors noted that, having found no RCT evidence of effectiveness, the usual recommendation would be to recommend against the use of this untested technology unless there was more evidence. Either this had to be accepted as a reasonable recommendation, or the process needed to be revised to include what they described as a 'commonsense' assessment of risks and benefits. While this has sometimes been dismissed as a ridiculous or even humorous example, it makes a serious point that is borne out in other examples.

What would have been the result if a systematic review had searched for evidence of effective interventions to prevent SIDS? What would have been the policy recommendation if no such evidence had been found? Should nothing be done until after there is evidence from one or more RCTs? Even if ethical issues had been satisfactorily addressed, there are practical difficulties in using RCT design for a condition with a low incidence, and huge sample sizes would have been required, making it difficult, if not impossible, to assess and ensure that the treatment and control groups were implemented as intended.

More recently, a Campbell Collaboration systematic review of the effectiveness of after-school programs in improving student outcomes (behavioural, social and emotional, and academic) using a similar protocol identified 88 studies, excluded all but five of them, and then concluded 'the collected evidence is not sufficient to make any policy or programming recommendations' (Zief et al. 2006, p. 25).

Even where systematic reviews enlarge selection criteria to include evidence from rigorous quasi-experimental studies, they leave out evidence from credible case studies and correlational studies, even where there is a credible argument of causal attribution.

An emerging alternative way to synthesise evidence is the use of realist synthesis, which was developed with support from the UK Economic and Social Research Council (Pawson 2006; Pawson et al. 2004). Realist synthesis includes any evidence where the conclusions are warranted on the basis of the data, including quality evidence from experimental, quasi-experimental and non-experimental research and evaluation. Rather than trying to produce a single answer of 'What works?' it seeks to answer the question 'What works for whom, in what circumstances and how?' by identifying and iteratively testing patterns of outcomes that are achieved through specific causal mechanisms in particular circumstances.

9.4 Transparent processes for generating and using evidence are needed

When the stakes are high, the quality of evidence and how it is used in policy can be misrepresented

Clinical trials are sometimes suggested as a model for the evaluation of human services programs. However in recent years, there has been increasing evidence of poor quality research about drug effectiveness being published and disseminated for commercial reasons. Recent reviews of clinical trials of new pharmaceuticals (House 2008) have revealed strategies that have misrepresented findings, including the choice of placebo as comparator (rather than a reasonable alternative), selection of subjects (Bodenheimer 2000), manipulation of doses (Angell 2004), method of drug administration (Bodenheimer 2000), manipulation of timescales (Pollack and Abelson 2006), suspect statistical analysis, deceptive publication (where the same results are published several times, inflating their weight in a meta-analysis), suppression of negative results (Mathews 2005), selective publishing (Armstrong 2006; Harris 2006; Mathews 2005; Zimmerman and Tomsho 2005), and opportunistic data analysis (where researchers tests all possible relationships for statistical significance) (Bodenheimer 2000). These problems have occurred despite peer review processes and conflict of interest disclosure requirements in journals, and the existence of regulators such as the Food and Drug Administration.

This does not bode well for evaluations of human service programs that are tied to commercial products, such as school textbooks and packaged intervention programs. Indeed, there is now an emerging body of research detailing similar problems in drug and violence prevention programs (for example, Gorman 2002; Weiss et al. 2008) and literacy programs such as the $US1 billion per annum Reading First program (Office of the Inspector General 2006).

For example, funding for projects under the Safe and Drug-Free Schools (SDFS) program, run by the US Department of Education, was conditional on schools implementing programs that have been proven to work. A list to help schools identify programs that would be eligible for funding identified nine prevention programs as ‘exemplary’ and 33 as ‘promising’ (programs that did not have sufficient evaluative data to justify the higher classification). A subsequent review of the evidence of effectiveness of the programs identified as ‘exemplary’, in this and similar lists used by other drug prevention agencies, revealed serious inadequacies in the quality of this evidence. For example, Project ALERT was included as an ‘exemplary’ program, as it had reported a statistically significant result on a relevant outcome measure. However, the evaluation had made 100

different comparisons between the program and control, using different substances, outcome measures, risk levels and two variations on the program, and calculated statistical significance on all of these. By chance, we would expect five to be significant at the .05 level, even if there were no real differences. The results showed that two were statistically significant — one of which showed that the program performed worse than the control (Weiss et al. 2008).

Data archives and documentation of evidence-based decision making can improve transparency

A recent report to the US National Academy of Sciences on ensuring the integrity, accessibility and stewardship of research data made a number of recommendations that would support these types of developments, including the following:

> All researchers should make research data, methods and other information integral to their publicly reported results publicly accessible in a timely manner to allow verification of published findings and to enable other researchers to build on published results, except in unusual cases where there are compelling reasons for not releasing data. In these cases, researchers should explain in a publicly accessible manner why the data are being withheld from release. (CEUIRDDA 2009).

The Australian Social Science Data Archive collects, preserves and makes available computer-readable data relating to social, political and economic affairs, and datasets are available without charge to organisations affiliated with Australian Consortium for Social and Political Research Incorporated (ACSPRI), which includes most universities and some Australian Government departments and agencies.

There is, however, currently no process for archiving the hundreds of evaluation reports produced in Australia to inform future policy, practice and research — and to permit review and validation of their conclusions. A national repository of evaluation reports, with suitable attention to matters of privacy and confidentiality, would improve the level of scrutiny and increase the range of evidence available.

9.5 Finding out 'what works' and implementing it will not necessarily improve results

For all the reasons discussed above, evidence-based policy is more than finding out 'what works' and implementing it. Finding a statistically significant difference between a treatment group and control group is not necessarily sufficient evidence to say that a policy will work when translated into wider practice. Interventions that

have been found to be effective might not be feasibly implemented in other settings due to a lack of skills, expertise or resources needed to properly implement the evidence-based intervention or adequate regulatory and supervisory processes to ensure adequate implementation. Even where they can be implemented well, there can be differential effects — what works on average can be ineffective or even harmful for some. Other unintended effects may only be evident over time, and some pilots cannot be scaled up effectively — for example, programs for the long-term unemployed may be effective on a small scale, but when scaled up end up just shuffling job queues unless additional employment opportunities are created.

Finally, it is important to note that different types of evidence are needed for different policy questions. Drawing on Davies' (2008) analysis, we can identify a range of questions that need different types of evidence to answer them, such as:

- What are the nature, size and dynamics of the problem? What are the risks of not addressing it?
- What resources are available?
- What are citizens' opinions, feelings, hopes and fears about this issue?
- How is the policy supposed to work? What are the risks of implementing it?
- What works? What works for whom, in what circumstances, how, and with what results (intended and unintended)?
- What are the cost–benefit ratio and comparative cost-effectiveness of interventions? What is the distribution of benefits and costs?
- What are the ethical implications of the policy?

These different questions remind us that, in addition to evidence of comparative effectiveness, evidence-based policy requires good descriptive quantitative and qualitative data about needs and factors producing problems; information about the availability of resources, including existing infrastructure and capital (including human and social capital) that can be leveraged; details of how previous interventions have been implemented; information about what different people value in terms of results and processes; and the identification of ethical issues . The evidence for policy making therefore needs to also encompass statistical databases; qualitative needs analyses; reports from previous projects, similar projects and pilot projects; opinion surveys; and expert reviews.

9.6 Conclusion

As Australia moves to embed an evidence-based approach to policy development and implementation, it is important to do so in a way that learns from previous attempts to use evidence to inform policy. This will include developing processes for generating and using evidence that suit the nature of policies and interventions, in particular whether they are essentially simple, or have complicated or complex aspects. The quality of evidence of effectiveness must be carefully assessed and not simply equated with use of any particular approach, such as the use of randomised controlled trials. Transparent processes for generating and using evidence will be needed, given the powerful incentives to misrepresent evidence. The process of evidence-based policy therefore needs to be understood not simply as a matter of finding out 'what works' and doing it.

References

ABC News 2009, 'Commission hears about dangers of stay or go confusion', 14 May, http://www.abc.net.au/news/stories/2009/05/14/2570753.htm (accessed 21 September 2009).

ABS (Australian Bureau of Statistics) 2007, *Australian Social Trends 2007*, Cat. no. 4021.0, ABS, Canberra.

Angell, M. 2004, *The Truth about the Drug Companies*, Random House, New York.

Armstrong, D. 2006, 'How the New England Journal missed warning signs on Vioxx, *Wall Street Journal*, 15 May, pp. A1–2.

Bodenheimer, T. 2000, 'Uneasy alliance: clinical investigators and the pharmaceutical industry', *New England Journal of Medicine*, vol. 342, pp. 1539–44.

Boyes, B. 2006, *Determining and Managing Environmental Flows for the Shoalhaven River, Report 1 — Environmental Flows Knowledge Review*, NSW Department of Natural Resources, http://www.dwe.nsw.gov.au/water/pdf/monitor_sholahaven_sh003.pdf (accessed 14 May 2009).

Bushfire CRC 2006, 'The stay and defend your property or go early policy: the AFCA position and the Bushfire CRC's current research, *Fire Note*, no. 7, http://www.bushfirecrc.com/publications/downloads/bcrcfirenote7staygo.pdf (accessed 11 March 2010).

CEUIRDDA (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, National Academy of Sciences) 2009, *Ensuring the Integrity,*

Accessibility, and Stewardship of Research Data in the Digital Age, The National Academies Press, Washington, DC.

Coalition for Evidence-Based Policy 2006, 'Which study designs can produce rigorous evidence of program effectiveness? A brief overview', Coalition for Evidence-based Policy Working Paper, http://www.evidencebasedpolicy.org /docs/RCTs_first_then_match_c-g_studies-FINAL.pdf (accessed 21 September 2009).

Concato J., Shah M.P.H. and Horwitz R.I. 2000, 'Randomized, controlled trials, observational studies, and the hierarchy of research designs', *New England Journal of Medicine*, vol. 342, no. 25, pp. 1887−92.

Cottingham, P., Quinn, G., Norris, R., King, A., Chessman, B. and Marshall, C. 2005, *Environmental Flows Monitoring and Assessment Framework,* technical report, CRC for Freshwater Ecology, Canberra.

Datta, L. 2003, 'Avoiding unwarranted death by evaluation', http://www.hfrp.org/ var/hfrp/storage/original/application/af7fd33cc8b440aba3b1b2cfe995493b.pdf (accessed 21 September 2009).

Davies, P. 2008, The role of impact evaluation in relation to other types of evaluation, presentation to World Bank Conference on Making Smart Policy, Washington, DC, 15–16 January.

Downes, B.J., Barmuta, L.A., Fairweather P.G., Faith, D.P, Keough, M.J., Lake, P.S., Mapstone, B.D., Quinn, G.P.,2002, *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*, Cambridge University Press, Cambridge.

Glass, G. 1997, 'Interrupted time series quasi-experiments', in Jaeger, R.M., *Complementary Methods for Research in Education*, 2nd edn, American Educational Research Association, Washington, DC, pp. 589–608.

Glasziou, P 2006, 'From research to practice: problems in the evidence pipeline' Centre for Evidence Based Medicine, University of Oxford, http://ebpg.nhri.org.tw/ImageUpload/File/Glasziou%20pipeline.pdf (accessed 11 March 2010).

Glouberman, S. and Zimmerman, B. 2002, 'Complicated and complex systems: what would successful reform of Medicare look like?', Commission on the Future of Health Care in Canada, Discussion Paper 8, http://www.change-ability.ca/Health_Care_Commission_DP8.pdf (accessed 11 March 2010).

Gorman, D.M. 2002, 'Defining and operationalizing "research-based" prevention: a critique (with case studies) of the US Department of Education's Safe,

Disciplined and Drug-Free Schools Exemplary Programs, *Evaluation and Program Planning*, vol. 25, pp. 295–302.

Graham I., Logan, J., Harrison, M., Straus, S., Tetroe, J., Caswell, W., Robinson, N. 2006, 'Lost in knowledge translation: time for a map, *Journal of Continuing Education in the Health Professions*, vol. 26, no. 1, pp. 13–24.

Griffin, P. 2009, 'Ambitious new project to raise literacy and numeracy levels in Victorian schools', http://newsroom.melbourne.edu/studio/ep-29 (accessed 11 March 2010).

Harris, G. 2006, 'FDA says Bayer failed to reveal drug risk study', *New York Times*, 29 September, pp. A1, A9.

Homel, R., Freiberg, K. and Branch, S. 2009, From macro to micro: identifying strategies to extend the reach of successful models of developmental prevention, ARACY Access Grid Presentation, 5 August, http://www.aracy.org.au/cmsdocuments/accessGrids/Microsoft%20PowerPoint%20-%20Micro%20to%20Macro%20-%20Access%20Grid%205-8-09%20%5BCompatibility%20Mode%5D.pdf (accessed 11 March 2010).

House, E. 2008, 'Blowback: the consequences of evaluation', *American Journal of Evaluation*, vol. 29, no. 4, 416–26.

Kurtz, G.F. and Snowden, D.J. 2003, 'The new dynamics of strategy: sense-making in a complex and complicated world', *IBM Systems Journal*, vol. 42, no. 3, pp. 462-83, http://xenia.media.mit.edu/~brooks/storybiz/kurtz.pdf (accessed 11 March 2010).

Mathematica Policy Research Inc 2002, *Making a Difference in the Lives of Infants and Toddlers and their Families: The Impacts of Early Head Start*, vol. 1, US Department of Health and Human Services.

Mathews, A.W. 2005, 'Worrisome ailment in medicine: misleading journal articles', *Wall Street Journal*, 10 May, pp. A1–2.

Morris, K. 2009, 'Positive deviants — role models for MRSA control, *The Lancet Infectious Diseases*, vol. 9, no. 5, pp. 275–275.

NHMRC (National Health and Medical Research Council) 2000, *How to Put the Evidence into Practice: Implementation and Dissemination Strategies*, handbook series on preparing clinical practice guidelines, NHMRC, Canberra.

NSW DECC (New South Wales Department of Environment and Climate Change) 2009, *Evaluation Framework for CMA Natural Resource Management*, Sydney, http://www.environment.nsw.gov.au/resources/4cmas/0982evalfworkCMAs.pdf (accessed 21 July 2009).

Office of the Inspector General 2006, *The Reading First Program's Grant Application Process: Final Inspection Report*, Office of the Inspector General, US Department of Education 2006, http://www.ed.gov/about/offices/list/oig/aireports/i13f0017.pdf (accessed 11 March 2010).

Patton, M.Q. 2008, *Utilization Focused Evaluation*, 4th edn, Sage Publications, Thousand Oaks, California.

Pawson, R. 2006, *Evidence-based Policy: A Realist Perspective*, Sage, London.

——, Greenhalgh, T., Harvey, G. and Walshe, K. 2004, 'Realist synthesis: an introduction', ESRC Research Methods Programme, University of Manchester, RMP Methods Paper 2/2004.

Pollack, A. and Abelson, R. 2006,'Why the data diverge on the dangers of Vioxx', *New York Times*, 22 May, pp. C1, C5.

Poverty Action Lab 2009, *Randomization*, http://www.povertyactionlab.org/research/rand.php (accessed 21 September 2009)

Ramalingam, B. and Jones, H. with Young, J. and Reba, T. 2008, 'Exploring the science of complexity: ideas and implications for development and humanitarian efforts', ODI Working Paper 285, ODI, London.

Rogers, P.J. 2009, 'Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation', *Journal of Development Effectiveness*, vol. 1, no. 3, pp. 1–10.

Scriven, M. 2008. A summative evaluation of RCT methodology & an alternative approach to causal research. Journal of MultiDisciplinary Evaluation, 5(9), pp. 11-24.

Smith, G., and Pell, J. 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trails', *British Medical Journal*, .vol. 327, no. 1459 – 1461

Stacey, R. 1992, *Managing the Unknowable*, Jossey-Bass, San Francisco.

Sternin, J. and Pascale, R.T. 2005, 'Your company's secret change agents', *Harvard Business Review*, May.

US Department of Education Institute of Education Sciences 2003, *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide*, http://ies.ed.gov/ncee/pubs/evidence_based/evaluation.asp (accessed 21 September 2009).

Weiss, C., Murphy-Graham, E., Gandhi, A. and Petrosino, A. 2008, 'Making the dream of evidence-based policy come true: the fairy godmother and her warts', *American Journal of Evaluation*, vol. 29, no. 1, pp. 29–47.

Westhorp, G. 2008, Development of realist evaluation methods for small scale community based settings, unpublished PhD thesis, Nottingham Trent University.

Worrall, J. 2002, 'What evidence in evidence-based medicine?', in *Causality: Metaphysics and Methods*, technical report 01/03, Centre for Philosophy of Natural and Social Science, London School of Economics.

Zief, S.G., Lauver, S., and Maynard, R.A. 2006, 'Impacts of after-school programs on student outcomes', *Campbell Systematic Reviews* 2006:3.

Zimmerman, R. and Tomsho, R. 2005, 'Medical editor turns activist on drug trials', *Wall Street Journal*, 26 May, pp. B1–2.

10 Evidence-based policy: summon the randomistas?

Andrew Leigh[1]
Research School of Economics, Australian National University

Abstract

Randomised experiments are a standard toolkit for the evaluation of new medical treatments, but are underutilised in Australian policy evaluation. Drawing on examples from Australia and overseas, I discuss the strengths, limits, ethics and politics of randomised trials. In a federation, it may be effective for the national government to establish a central fund for states and territories to conduct randomised policy trials.

10.1 Let's start with a vitamin tablet

For the past 10 years, I have taken a multivitamin pill once a day with my morning coffee. Like any good morning habit, it is a comfortable and familiar routine. But I've gotten a warm glow to know that each day begins with a simple act that helps take care of my body.

Earlier this year, a friend suggested that I read an article published in the *Journal of the American Medical Association* (Bjelakovic et al. 2007).[2] The authors set out to answer the question: do vitamin supplements make you live longer? To answer this,

1 Email: andrew.leigh@anu.edu.au. Web: http://andrewleigh.org Parts of this article draw on a keynote lecture delivered at the NSW Bureau of Crime Statistics and Research 40th Anniversary Symposium on 19 February 2009. I am grateful to participants at the Productivity Commission Roundtable, Terry O'Brien, Angela O'Brien-Malone and Nikki Rogers for valuable comments on earlier drafts. Jenny Chesters and Susanne Schmidt provided outstanding research assistance.

2 For an informal discussion of the issue, see Norman Swan's interview with one of the researchers on 5 March 2007, available at http://www.abc.net.au/rn/healthreport/stories/2007/1861068.htm. The researchers were at pains to point out that their findings should not be extrapolated to foods that are rich in vitamins, such as fresh fruit and vegetables.

they drew together all the best evidence in the world, which in this case meant randomised trials of vitamins A, C and E, beta carotene and selenium, and found no evidence that vitamins make you live longer. If anything, those who took vitamin supplements seemed to live shorter lives.

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

What makes me so sure about this decision? First, because the evidence comes from a study published in one of the world's leading medical journals. Second, because medicine has a well-established hierarchy of evidence. Grade I evidence in that hierarchy is 'well-conducted systematic reviews of randomised trials'. (Grade II is non-randomised controlled trials and uncontrolled experiments, while Grade III is descriptive studies and case reports.) There is a strong consensus in the medical profession that when it comes to questions like 'Are vitamins good for you?', the highest grade of evidence is a systematic review of randomised trials.

Yet suppose I were a policy maker, charged with deciding whether to scrap or expand a social program. In many cases, I would probably find that the only available evidence was based upon anecdotes and case studies — what medical researchers would regard as the lowest grade of evidence in their hierarchy. While the evidence base in social policy is steadily advancing, a lack of data and an unwillingness to experiment are two major factors that hamper our understanding of what works and what does not.

Uncertainty over the effectiveness of our social policies is even more striking when you realise that annual government social expenditure amounts to around $8000 per Australian.[3] Ask a handful of experts, and you will find no shortage of ideas for improving the system. More education, more work experience, conditional cash transfers, laptop rollouts, higher income support, wage subsidies, lower minimum wages, public works programs and prison reform are among the recent favourites. Knowing more about which of these policies work, and why, could help us to improve social outcomes, save money or both.

10.2 The counterfactual problem

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

[3] This calculation uses the OECD definition of 'social expenditure', which amounted to $165 billion in 2005.

In the social sciences, this turns out to be a much tougher problem to address.[4] If we simply use time series variation, we may find it tricky to separate the policy change from secular changes in incidence over time. For example, say that we wanted to estimate the impact of a training program on earnings. If we were to just track the earnings of participants, we might find it difficult to separate the effect of the overall economy from the impact of the program.

Another approach is to construct a counterfactual by using those who choose not to participate as the control group. For example, suppose that you wanted to see the impact of alcohol management plans on outcomes in Indigenous communities. One evaluation strategy might compare outcomes in communities that chose to establish an alcohol management plan with outcomes in those that did not. But if the two sets of communities are systematically different — say, because the treatment group has more social capital or stronger leadership than the control group — then such an approach would be likely to overestimate the impact of the intervention.

10.3 The strengths of randomised trials

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

In Australian social policy, a canonical example of a randomised policy trial is the New South Wales Drug Court trial, conducted in 1999–2000. Offenders were referred to the Drug Court from local or district courts, underwent a detoxification program and were then dealt with by the Drug Court instead of a traditional judicial process. At the time it was established, the number of places in detoxification was limited, so participants in the evaluation were randomly assigned either to the treatment group (313 persons) or the control group (201 persons). They were then matched to court records in order to compare reoffending rates over the next year or

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

5 On randomised policy trials, see for example, The Economist (2002), Farrelly (2008) and Leigh (2003).

more. The evaluation found that the Drug Court was effective in reducing the rate of recidivism, and that while it was more expensive than the traditional judicial process, it more than paid for itself (Lind et al. 2002). At a recent conference celebrating the tenth anniversary of the Drug Court, speakers broadly acknowledged the role that the court has played in improving the lives of drug offenders and the general community (Knox 2009).

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

> Q: 'Minister, is it true that your program spends more on drug offenders? Why should taxpayers fork out more money to put drug addicts through detox programs, when we could spend less and just throw them in jail?'
>
> A: 'You bet we're spending more on this program, and that's because we have gold-standard evidence that it cuts crime. A year after release, those who went through the Drug Court were half as likely to have committed a drug offence, and less likely to have stolen. It's probably the case that Drug Courts help addicts kick the habit. But even if you don't care a whit about their wellbeing, you should be in favour of Drug Courts because they keep you and your family safe.'

In the case of the Drug Court, many of us probably had an expectation that the policy would reduce crime. But high-quality evaluations do not always produce the expected result. Staying for a moment with criminal justice interventions, take the example of 'Scared Straight', a program in which delinquent youth visit jails to be taught by prison staff and prisoners about life behind bars. The idea of the program — originally inspired by the 1978 Academy Award winning documentary of the same name — is to use exposure to prison to frighten young people away from a life of crime. In the 1980s and 1990s, several US states adopted Scared Straight programs.

Low-quality evaluations of Scared Straight, which simply compared participants with a non-random control group, had concluded in the past that such programs worked, reducing crime by up to 50 percent. Yet, after a while, some US states began carrying out rigorous randomised evaluations of Scared Straight. The startling finding was that Scared Straight actually increased crime, perhaps because youths discovered jail was actually not as bad as they had thought. It was not until policy makers moved from silver-standard evidence to gold-standard evidence that they learned the program was harming the very people it was intended to help (Boruch and Rui 2008; Petrosino et al. 2002).

Being surprised by policy findings is perfectly healthy. Indeed, we should be deeply suspicious of anyone who claims that they know what works based only on theory or small-scale observation. As economist John Maynard Keynes once put it in a

different context, ‘When the facts change, I change my mind. What do you do, sir?’[6]

10.4 Are randomised trials ethical?

Although it would be practically possible to randomly trial many of our social policy interventions, there is a reluctance among policy makers to subject policy interventions to gold-standard evaluation. In most developed nations, it is impossible to get a new pharmaceutical licensed without a randomised trial. Yet new social policies — often costing considerably more — require no such evaluation.

One common explanation proffered for this is the ethical challenge: when you have a program that you think is effective, how can you toss a coin to decide who receives it? The simplest answer to this is that the reason we are doing the trial is precisely because we do not know whether the program works. The simplest exposition of this is ‘Rossi’s Law’ (named after sociologist Peter Rossi), which states: ‘The expected value for any measured effect of a social program is zero.’ If you believe Rossi’s Law, it mostly does not matter whether a given individual is allocated to the treatment group or the control group. Indeed, for some programs — such as Scared Straight — participants in the control group ended up better off than those in the treatment group.

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

We should not lightly dismiss ethical concerns about randomised trials, but they are often overplayed. Medical researchers, having used randomised trials for several decades longer than social scientists, have now grown relatively comfortable with the ethics of randomised trials. Certain medical protocols could be adapted by social scientists, such as the principle that a trial should be stopped early if there is clear

6 Reply to a criticism during the Great Depression of having changed his position on monetary policy, as quoted in Malabre (1994, p. 220).

evidence of harm, or the common practice of testing new drugs against the best available alternative.

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

To answer these questions, Garner's team is running a randomised trial. In effect, when a Sydney 000 operator receives a report of a serious head injury, a coin is tossed. Heads, you get an ambulance and a paramedic. Tails, you get a helicopter and a trauma physician. Once 500 head injury patients have gone through the study, the experiment will cease and the results will be analysed.

When writing an article about the trial last year, I spoke with Alan Garner, who told me that, although he has spent over a decade working on it, even he does not know what to expect from the results (Leigh 2008). 'We think this will work', he told me in a phone conversation, 'but so far, we've only got data from cohort studies.' Indeed, he even said, 'Like any medical intervention, there is even a possibility that sending a doctor will make things worse. I don't think that's the case, but [until HIRT ends] I don't have good evidence either way.'

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

10.5 The limits of randomised trials

While randomisation is an underused tool in the policy drawer, it is not effective in all cases. Writing tongue-in-cheek in the *British Medical Journal*, Gordon Smith and Jill Pell (2003) argued that the quality of the evidence on parachute effectiveness was severely limited by the absence of randomised controlled trials. They pointed out that the only evidence that parachutes prevent deaths when people jump out of planes was based on observation and expert opinion, the lowest rank of

evidence in the hierarchy. Their conclusion: we need randomised trials of parachutes!

As the previous section noted, randomised trials often need to undergo ethical scrutiny. Just as we would not allow a randomised trial of parachutes, we would not countenance a randomised trial that withdrew all income support from lone parents, experimented with taking children out of school or doubled hospital waiting lists. The *randomistas* — as economist Angus Deaton has called them — are not welcome everywhere.

But it is rare that policy makers are actually countenancing a policy contraction or expansion this radical. More often, the kinds of questions that need to be answered are much more modest, and therefore readily amenable to randomisation. Do long-term unemployed youth benefit more from job training or wage subsidies? Do schoolchildren benefit more from teacher merit pay or class size reductions? Would more generous post-release payments prevent ex-prisoners falling back into bad habits? What kinds of intensive early childhood programs would work best for Indigenous children?

Another limit to what we can learn from randomised trials comes from scale effects. As anyone who has eaten cafeteria food knows, what works well on a small scale does not necessarily work on a large scale (Currie 2001). One problem is that small-scale programs are often 'boutique programs', which are resourced to a level that is not feasible if implemented across an entire system. Another risk is that small-scale programs might fail to measure spillover and displacement effects. In economic jargon, randomised trials are a very precise way of measuring partial equilibrium effects, but often do not allow us to get at the general equilibrium effects.[7]

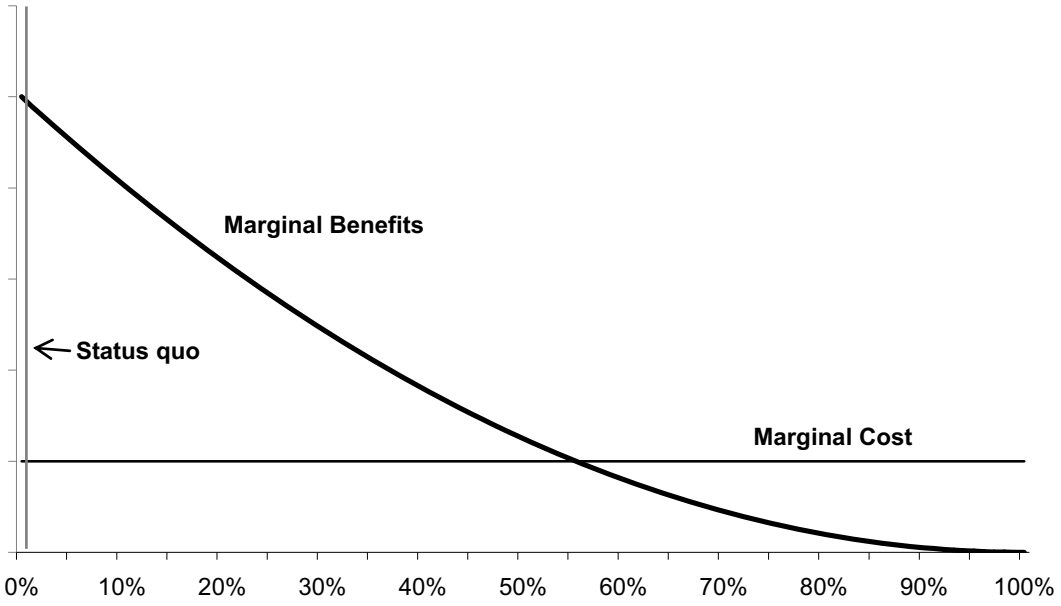
Because of these limitations, it is unlikely that we would ever want 100 per cent of government evaluations to be randomised trials. Most likely, the marginal benefit of each new randomised trial is a little lower than that of the previous one. At some point, it is indeed theoretically possible that we could end up doing more randomised trials than is socially optimal.

However, this is unlikely to ever occur, at least in my lifetime. Figure 10.1 presents my stylised sense of the state of play for randomised trials in Australia. My best

7 For a thoughtful discussion of these issues in the context of international development, see the papers presented at a Brookings Global Economy and Development Conference 'What Works in Development? Thinking Big and Thinking Small', held in Washington DC on 29–30 May 2008. Papers and presentations are available at http://www.brookings.edu/events/2008/0529_global_development.aspx. See also the recent exchange in Deaton (2009), Heckman and Urzua (2009) and Imbens (2009).

estimate is that less than 1 per cent of all government evaluations are randomised trials (excluding health and traffic evaluations, the proportion is probably less than 0.1 per cent).[8] Another way to put this is that, *to a first approximation, Australia currently does no randomised policy trials*. Policy makers could safely embark on a massive expansion of randomised policy trials in Australia before we come close to the point where the costs exceed the benefits.

Figure 10.1 **Costs and benefits of more randomised policy trials**



Note: All datapoints in this chart are assumptions.

8 I was unable to find a comprehensive database of all government evaluations in Australia, so opted instead to conduct a Google search on the gov.au domain. A search on 14 August 2009 for 'randomised AND evaluation' brought up fewer than 10,000 hits, while a search for just 'evaluation' brought up more than 1.7 million hits. It is likely that both these numbers overestimate the true numbers of randomised and non-randomised evaluations, since they are counts of hits rather than unique files. However, to the extent that the ratio of hits to unique files is the same for randomised and non-randomised evaluations (which seems a reasonable assumption), this suggests that about 0.5 percent of Australian government evaluations in recent years have used a randomised design. The assumption about most Australian randomised trials being health and traffic evaluations is based on the fact that the Australian trials contained within the Campbell Collaboration's randomised trials register (C2-SPECTR) are largely in those two categories (see Leigh 2003 for cross-national comparisons of the contents of this register).

There are also limits to the power of good evidence to change the minds of policy makers. In a recent speech, Gary Banks catalogued several instances — ranging from fertility policy to industry policy — in which the Productivity Commission showed that policies had failed to achieve their goal (Banks 2009). Despite the rigour of the econometrics, some readers may be surprised to learn that the Commission's recommendations have not been universally adopted by the Australian Government. Nonetheless, raising the evidence bar matters. High-quality evaluations are harder for policy makers to dismiss than low-quality evaluations.

10.6 Promoting randomised trials

How might randomised trials be promoted in Australia? One possibility would be for the federal government to systematically set aside resources for states to conduct rigorous randomised trials that would have a national benefit. For example, despite spending billions of dollars to reduce class sizes over the past few decades, Australia has never conducted a randomised trial of the impact of smaller classes on student performance. Part of the reason for this is political: state politicians are uncomfortable telling voters that a lottery will be used to determine which students get to sit in the smaller class. But if the program were federally funded, it is possible that a state government might be willing to administer the experiment. In cases where federal programs are being channelled through the states, small amounts set aside for 'random assignment evaluation' can have a large knowledge payoff. Box 10.1 sets out three examples from recent pieces of US federal legislation that explicitly set aside funds for randomised evaluation.[9]

[9] These examples are drawn from the Coalition for Evidence-Based Policy, which is part of the Council for Excellence in Government (http://www.excelgov.org/), and a presentation by Adam Gamoran, 'Measuring impact in science education: challenges and possibilities of experimental design', NYU Abu Dhabi Conference, January 2009.

Box 10.1 US federal legislation that specifically funds randomised evaluation

1. The Second Chance Act, dealing with strategies to facilitate prisoner re-entry into the community, sets aside 2% of program funds for evaluations that 'include, to the maximum extent possible, random assignment … and generate evidence on which re-entry approaches and strategies are most effective'.
2. The No Child Left Behind Act calls for evaluation 'using rigorous methodological designs and techniques, including control groups and random assignment, to the extent feasible, to produce reliable evidence of effectiveness.'
3. Legislation to improve child development via home visits directs the Department of Health and Human Services to 'ensure that States use the funds to support models that have been shown in well-designed randomized controlled trials, to produce sizeable, sustained effects on important child outcomes such as abuse and neglect'.

Another way that randomised trials might be promoted is through the regular use of an 'evidence hierarchy' by social policy makers. Such hierarchies — common in the medical literature — are steadily gaining currency among policy makers. As an example, Box 10.2 depicts the hierarchy proposed in Leigh (2009). Although evidence hierarchies are invariably imperfect, they can help to focus attention on the quality of the available knowledge base. If a politician is told 'we think you should implement Option A, but you should also know that the state of knowledge is very poor', he or she may be more inclined to sow the seeds of a few new randomised trials.

Box 10.2 A possible 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 they are more similar to the policy under consideration.

Source: Leigh (2009).

10.7 Conclusion

In Australian policy debates, the term 'evidence-based policy making' has now become so meaningless that it should probably be jettisoned altogether. The problem in many domains is not that decision makers do not read the available literature — it is that they do not set up policies in such a way that we can learn clear lessons from them.[10] In employment policy, the early 1990s recession saw Australia spend vast amounts on active labour market programs, without producing a skerrick of gold-standard evidence on what works and what does not. In Indigenous policy, there are as many theories as advocates but precious few randomised experiments that provide hard evidence about what really works.

Sometimes randomised trials will justify the expansion of a politically difficult intervention. But we should never forget the social benefit of evaluations whose results show us that a program does not work. Thanks to randomised evaluations of multivitamin tablets, my household now has $30 a year to spend on other things. The same is true of evaluations that find government programs to be ineffective. A randomised trial that conclusively shows a program had no impact is a valuable piece of knowledge in improving Australian public policy.

References

Banks, G. 2009, 'Evidence-based policy-making: What is it? How do we get it?', lecture given at the *ANZSOG/ANU Public Lecture Series*, Canberra, 4 February 2009, http://www.pc.gov.au/__data/assets/pdf_file/0003/85836/cs20090204.pdf (accessed 3 August 2009).

Bjelakovic, G., Nikolova, D., Gluud, L.L., Simonetti, R.G. and Gluud, C. 2007, 'Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis', *Journal of the American Medical Association*, vol. 297, no. 8, pp. 842–857.

Boruch, R. and Rui, N. 2008, 'From randomized controlled trials to evidence grading schemes: current state of evidence-based practice in social sciences', *Journal of Evidence-Based Medicine*, vol. 1, no. 1, pp. 41–49.

Currie, J. 2001, 'Early childhood education programs', *Journal of Economic Perspectives*, vol. 15, no. 2, pp. 213–238.

10 Harvard economist Roland Fryer is particularly scathing about the quality of evidence in education. In a recent interview with the *New York Times*, he said: 'If the doctor said to you, "You have a cold; here are three pills my buddy in Charlotte uses and he says they work," you would run out and find another doctor. Somehow, in education, that approach is O.K.' (quoted in Hernandez 2008).

Deaton, A.S. 2009, 'Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development', NBER Working Paper no. 14690, http://www.nber.org/papers/w14690 (accessed 3 August 2009).

Farrelly, R. 2008, 'Policy on trial', *Policy*, vol. 24, no. 3, pp. 7–12.

Heckman, J. and Urzua, S. 2009, 'Comparing IV with structural models: what simple IV can and cannot identify', NBER Working Paper no. 14706, http://www.nber.org/papers/w14706 (accessed 3 August 2009).

Hernandez, J.C. 2008, 'New effort aims to test theories of education', *New York Times*, 25 September.

Imbens, G.W. 2009, 'Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009)', NBER Working Paper no. 14896, http://www.nber.org/papers/w14896 (accessed 3 August 2009)

Knox, M. 2009, 'Applause for former drug users who turn their lives around', *Sydney Morning Herald*, 7 February.

Leigh, A. 2003, 'Randomised policy trials', *Agenda: A Journal of Policy Analysis and Reform*, vol. 10, no. 4, pp. 341–354.

—— 2008, 'A good test of public policy', *Australian Financial Review*, 8 April.

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

Lind, B., Weatherburn, D., Chen, S., Shanahan, M., Lancsar, E., Haas, M. and De Abreu Lourenco, R. 2002, *NSW Drug Court evaluation: cost-effectiveness*, NSW Bureau of Crime Statistics and Research, Sydney, www.courtwise.nsw.gov.au /lawlink/bocsar/ll_bocsar.nsf/vwFiles/L15.pdf/$file/L15.pdf (accessed 3 August 2009).

Malabre, A.L. 1994, *Lost Prophets: An Insider's History of the Modern Economists*, Harvard Business Press, Cambridge, Massachusetts.

Petrosino, A., Turpin-Petrosino, C. and Buehler, J. 2002, *'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency* (Updated C2 Review), Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE), available at www.campbellcollaboration.org.

Smith, G.C.S. and Pell, J.P. 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *British Medical Journal*, vol. 327, pp. 1459–1461.

The Economist 2002, 'Try it and see', 2 March, pp.73–74.

General discussion

The discussion period focused on the merits of randomised controlled trials (RCTs) in generating robust evidence for policy development, and examples where other forms of evidence might be more useful.

RCTs as a 'gold standard'?

A number of roundtable participants questioned how widely randomised trials could be used, and asked whether they deserved their 'gold standard' status or superiority in the evidence hierarchy, arguing that RCTs:

- are not useful for answering some types of policy questions (for example, monetary policy or climate change policies are not amenable to randomisation)
- can be costly and politically challenging
- may not necessarily provide an unbiased estimate of a policy's effects — the design of the trial might have been poor, subjects might drop out or original random assignment could be compromised by subjects' behaviour
- the trial may not necessarily replicate the conditions of a full-scale rollout of the policy — for example a high-quality administrative team might be used for the trial.

Andrew Leigh, Jeffrey Smith and Sally Green acknowledged that randomised trials are not a panacea ('trials are not a substitute for thinking'), but they argued that they can be a powerful tool with unique potential to avoid selection bias. Andrew Leigh argued they had particular potential in illuminating a large range of social policy questions, particularly in areas such as education, crime and income support.

Patricia Rogers reminded the roundtable that randomised trials share many of the same potential problems that other empirical studies have (such as attrition bias), and some RCTs have been poorly designed or implemented. Sally Green noted that one of the improvements secured through the CONSORT Statement, was better journal reporting of the common biases within trials.

Andrew Leigh argued that Australia is in 'no danger of over-relying on randomised trials', or blindly accepting them as the only method of evaluation. To date, Australian governments have undertaken only a handful of randomised policy trials.

On balance, speakers concluded Australian governments should do more randomised trials, but in conjunction with other forms of evidence. Even where RCTs could make a powerful contribution, there would often be a need for other forms of evidence to address interest in scaling effects, or general equilibrium effects.

Getting relevant, quality evidence at the right time

A participant asked how research organisations could best meet the challenge of engaging with policymakers to ensure they are producing the right type of evidence at the right time. Sally Green indicated that initially the Cochrane Collaboration did not really engage policymakers to help set the research agenda, as it was a volunteer organisation that relied on volunteers undertaking systematic reviews, usually on topics of their interest and expertise. The result was that a lot of reviews were undertaken at a little financial cost, but there were some big gaps in terms of coverage. For example, many volunteers did reviews on pregnancy and childbirth, but far fewer on heart disease. As the organisation has matured, the Collaboration has devoted a significant effort to engaging policymakers and practitioners to help set the research agenda and prioritise reviews by asking policymakers what sorts of questions they would like to see answered.

One participant noted that although there are good examples of research evidence developing into a compelling narrative for policymakers, there was an equal number of cases where it had not. This was especially so in an environment where a 'single bottom line' is important, with research often complex and cloaked in qualifications and caveats. Sally Green argued that it was particularly important for the research community to ensure their results are correctly communicated to policymakers and the public. A number of research bodies are currently studying the best ways to communicate results, including by looking at how people can best understand uncertainty and complex 'decision trees'.

SESSION 4

INSTITUTIONALISING AN EVIDENCE BASED APPROACH — HOW CAN AN EVALUATION CULTURE BE EMBEDDED INTO POLICY-MAKING?

11 Institutionalising an evidence-based approach to policy making: the case of the human capital reform agenda

Peter Dawkins
Secretary, Department of Education and Early Childhood Development, Victoria. The author is also a Professorial Fellow of the Melbourne Institute at the University of Melbourne.

Abstract

This paper is based on a case-study of evidence-based policy — that is, the development and implementation of the human capital reform agenda in Victoria and Australia. It is argued that this is an outstanding example of an evidence base generating a major reform agenda. It is also concluded that an outcomes framework which can be linked with progress measures and targets, and an associated evaluation framework, provides strong incentives for governments to adopt evidence-based policies that can be expected to have a desirable impact on the agreed outcomes. Third, it is argued that different kinds of evidence are useful in different circumstances and that often multiple sources of evidence are ideal. There are important differences between evidence needed for strategic policy design and specific policy initiatives. It is also suggested that the Council of Australian Governments' National Productivity Agenda is a very good illustration of how an evidence-based policy framework can support a federal–state reform agenda in the context of vertical fiscal imbalance. Finally, the way in which an evidence-based approach to policy development and advice to ministers is embedded in the modus operandi of the Department of Education and Early Childhood Development in Victoria is outlined.

11.1 Introduction

Policy decisions will be influenced by much more than objective evidence, or rational analysis. Values, interests, personalities, timing, circumstance and happenstance — in short, democracy — determine what actually happens. But evidence and analysis can

nevertheless play a useful, even decisive, role in determining policy-makers' judgments. Importantly, they can also condition the political environment in which those judgments can be made (Banks 2009).

In this paper I discuss the institutionalising of an evidence-based approach to policy making, in the context of a case study of the human capital reform agenda, initiated by the Bracks–Brumby Victorian State Government's *Third Wave of Reform* (DPC and DTF 2005) followed by the Rudd Federal Government's 'Education Revolution'.

I will also cover the way in which the Victorian Department of Education and Early Childhood Development has embedded an evidence-based approach to its policy advice to ministers and its evaluation of progress against objectives.

11.2 The human capital reform agenda

The third wave of reform

In August 2005, the then Victorian Premier, Steve Bracks, in association with the then Victorian Treasurer, John Brumby, launched *Governments Working Together: A Third Wave of National Reform* (DPC and DTF 2005). The first wave of the reform in the 1980s involved the floating of the dollar, the deregulation of financial markets and the effective end of tariff barriers designed to protect Australian industry. National competition policy was the centrepiece of the second wave in the 1990s. The Victorian Government was now calling for a third wave, in which a major focus would be a human capital reform agenda.

Victoria proposed this new National Reform Agenda to the Council of Australian Governments (COAG). In progressing the case for the agenda, extensive evidence was collected about the effects of early childhood development, schooling and vocational education on literacy and numeracy, labour force participation and productivity. This evidence was used in tandem with some computable general equilibrium (CGE) modelling, using the Monash Model, to simulate the potential economic effects of the reform agenda on gross domestic product (GDP) and tax revenues (DTF 2006).

Later, the Productivity Commission was asked by the Australian Government to undertake modelling of a similar kind, which resulted in an important report which also found substantial economic benefits for Australia from such a reform agenda, *Potential Benefits of the National Reform Agenda* (PC 2006).

In February 2006, COAG agreed to progress the human capital reform agenda. In the COAG communiqué, the focus of the reform and the framework for implementing it was outlined (COAG 2006). This included a list of 'indicative outcomes', such as the proportion of young people meeting basic literacy and numeracy standards, and the proportion of young people making a successful transition form school to work or further study.

To 'hold jurisdictions accountable for achieving these outcomes', COAG agreed that the progress of jurisdictions would be independently assessed and transparently reported. This led to the establishment of the COAG Reform Council.

Although this did not result in major federal–state investment in a human capital reform agenda before the 2007 election, the COAG Reform Council was still in place after the election of the Rudd Labor Government, which had committed to an Education Revolution.

Meanwhile, the Victorian Government produced a number of policy papers about the way in which it proposed to implement the human capital reform agenda in collaboration with the Australian Government, on the basis of joint Commonwealth–State investment, and a range of targets which an analysis of the evidence suggested were reasonable to aim for, and against which the progress of the policy implementation would be judged (DPC, DoE and DTF 2007; DPC, DHS, DTF and DoE 2007; DPC 2008).

The Education Revolution

In January 2007, Kevin Rudd, then Leader of the Opposition, announced with Stephen Smith, then Shadow Minister for Education and Training, the federal ALP's commitment to an Education Revolution:

> human capital development is at the heart of a third wave of economic reform that will position Australia as a competitive, innovative, knowledge based economy that can compete and win in global markets … (ALP 2007, p. 3).

They went on to quote international evidence of the effect of education on economic growth:

> OECD research estimates that a one year increase in the workforce's average number of years of education can add 3–6 per cent to GDP and increase annual growth by as much as 1 per cent …
>
> International research has shown a close relationship between high literacy standards and economic growth, with a 1 per cent premium on average literacy scores linked to a 1.5 per cent higher level of per capita GDP. (ALP 2007, p. 11).

Econometric research from the United States by Erik Hanushek is another example of the evidence base of the effect of education on economic growth (Hanushek 2009).

The ALP's paper also documented evidence of Australia being a laggard in human capital investment from early childhood development to university education, concluding that raising this investment and promoting higher quality educational outcomes would be one of three priorities for a federal Labor Government.

The Victorian Blueprint for Education and Early Childhood Development

In parallel with its negotiation through COAG for a national human capital reform agenda, the Victorian Government proceeded to develop its own strategy. This included a skills reform policy that involved moving to a more demand-driven system for vocational education and training (DIIRD 2008). The Minister for Education, Bronwyn Pike, and the Minister for Children and Early Childhood Development, Maxine Morand, also released their strategy for young Victorians from birth to 18 in the *Blueprint for Education and Early Childhood Development* (DEECD 2008). This involved four priorities, six goals, three broad strategies and 20 areas of action. The core mission was to 'ensure a high quality and coherent birth-to-adulthood learning and development system to build the capacity of every young Victorian'. It committed to basing the strategies and actions on an international evidence base:

> Directions emerging from international research and successful improvement strategies provide guidance on how we can make further improvements (DEECDa 2008, p. 13).

It also committed to an outcomes and evaluation framework for monitoring the success of the blueprint implementation that would be based on outcomes and progress measures simultaneously being developed for the COAG Productivity Agenda.

The COAG Productivity Agenda

With the election of the Rudd Labor Government, committed to its Education Revolution, momentum for a national human capital reform agenda was regained. At a COAG meeting in December 2007, a commitment was made to a National Productivity Agenda, in which human capital reform was to be central, and a Productivity Agenda Working Group was established, chaired by Deputy Prime Minister Julia Gillard, to develop this agenda.

In July 2008, COAG adopted an outcomes framework for the National Productivity Agenda, and associated performance measures (COAG 2008a). This is attached as Appendix 1 to this paper.

In November 2008, this resulted in a major multi-billion dollar investment in early childhood, school improvement and vocational education and training (COAG 2008b).

As well as a National Education Agreement being reached between the Commonwealth and the States and Territories about untied funding for state school systems, along with an agreed outcomes framework, and an agreed delineation of responsibilities between levels of government, a number of national partnership agreements were signed. In one, States and Territories committed to work with the Commonwealth to raise the quality of teaching. In another, agreement was reached to invest in low socioeconomic school communities because of the evidence about the association between low socioeconomic status of students and schools and educational outcomes.

11.3 Evidence-based policy: some important distinctions

Overall policy strategy and specific policy initiatives

In thinking about evidence-based policy, there is an important distinction between overall policy design and specific policy initiatives. To illustrate this, I give two examples of strategic policy design and two examples of specific policy initiatives.

Strategic policy design (1): COAG National Productivity Agenda

As outlined above, the COAG National Productivity Agenda was built primarily on a broad evidence base about the impact of human capital investment and human capital reform on economic growth (DTF 2006; Productivity Commission 2006).

Strategic policy design (2): Teacher quality

A major priority of the COAG productivity agenda is to raise the quality of teaching. This is based on a large number of empirical studies of the determinants of education outcomes. In September 2007, McKinsey & Co., under the leadership of Sir Michael Barber, produced a report titled *How the World's Best Performing School Systems Come Out on Top* (McKinsey 2007). It concluded that three things matter most:

> 1) getting the right people to become teachers, 2) developing them into effective instructors, 3) ensuring that the system is able to deliver the best possible instruction for every child. (McKinsey 2007, p. 5).

The most quoted US psychometric and econometric research on this subject is by Sanders with various other authors, based on research in Tennessee (for example, Sanders and Rivers 1996) and by Erik Hanushek with others (for example, Rivkin et al. 2005). In Australia, Andrew Leigh has undertaken similar research on the causes and effects of teacher quality (Leigh 2009).

Specific policy initiative (1): Performance and Development Culture in Victorian Schools

In 2003 the Victorian Education Minister announced to introduction of a process to be called the *Performance and Development Culture* (DE&T 2003). There was significant latitude for each school in the way it implemented the P&D culture, but five criteria were established for accreditation: effective induction and mentoring for new teachers; use of multiple sources of feedback on an individual teacher's effectiveness; customised teacher-development plans; individualised professional development; and endorsement of the presence of the P&D culture by the teaching staff. By the end of 2008, 94 per cent of schools had been accredited by a third party.

The department has evaluated the effect of the accreditation process on schools and found that during the process of accreditation a range of measures of school performance improve significantly (The Nous Group 2007). Banerjee and Kamener of Boston Consulting Group have also undertaken a review, which also found a positive impact of this initiative (Boston Consulting Group 2008).

Specific policy initiative (2): Performance pay for teachers

One policy idea that has been under discussion in Australia, to promote the quality of teaching and learning, is performance pay for teachers. This has been tried in some places, but there are a limited number of cases from which to draw evidence. Some research suggests that it can have positive effects (for example, Angrist and Lavy 2004; CTAC 2004; Figlio and Kenny 2006; Muralidharan and Sundararaman 2009; Podgursky and Springer 2006; Winters et al. 2008). In the Victorian *Blueprint for Education and Early Childhood Development* it was announced that Victoria would investigate rewards and incentives for effective teaching, and the Minister for Education, Bronwyn Pike, has recently announced that some trials will be conducted to evaluate two alternative approaches to performance pay (DEECD 2009b).

Different types of evidence

Evidence ranges from econometric studies of the contribution of education to economic growth, psychometric studies of the effect of teacher quality on student achievement, and evaluations of individual policy interventions, including (but not often) randomised trials, to reviews of the evidence from large numbers of different studies, including meta-studies. Some studies focus on identifying examples of success (such as successful school systems or successful school improvement agendas) and identifying the common factors associated with success.

The data collected, the statistical methods used, and the evaluation methods adopted vary in their degree of sophistication. The type of evidence required depends on the nature of the policy decisions to be taken. Strategic policy design, such as the human capital reform agenda, requires a range of evidence to support the thrust of the policy. Studies at a high level of aggregation, such as cross-country studies of economic growth, are highly relevant. In other circumstances detailed micro studies, involving pilot or trial programs, may be what is needed, such as in deciding whether and how to proceed with a performance pay system. Evidence from trials in other places may provide useful background research, but care has to be taken when trying to generalise from such specific experiments (Heckman and Smith 1995).

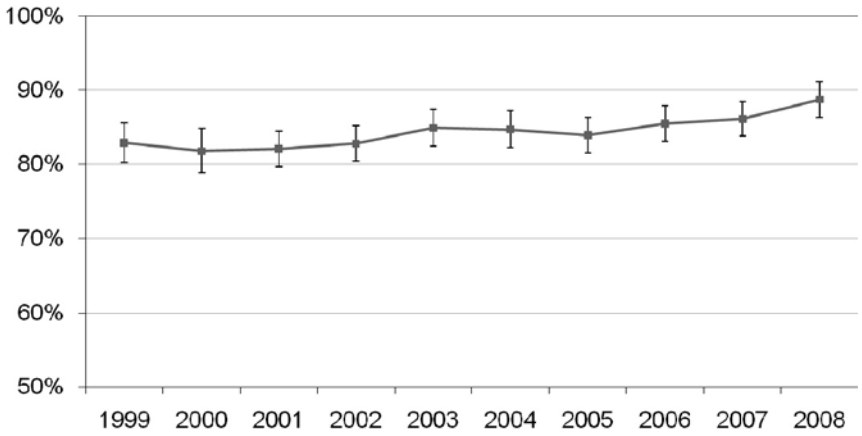
11.4 Institutionalising an evidence-based approach to policy making

Outcomes and evaluation frameworks and building reward mechanisms into policy design

A precursor to this COAG outcomes-based policy process is *Growing Victoria Together*, an outcomes framework established in 2000 by the Victorian Government, in which, for example, the literacy and numeracy of school students and Year 12 or equivalent completion were two key outcomes in education policy. A 90 per cent target was set for 2010, for Year 12 or equivalent completions for 20- to 24-year-olds. Also in 2000, a review was completed of post-compulsory education and training pathways, which led to the development and implementation of a policy agenda to improve post-compulsory pathways and amongst other things increase the Year 12 or equivalent rate to the 90 per cent target (DEET 2000 — the Kirby Report). Progress against this target appears in the Department Secretary's annual performance plan and annual performance review, and a process of constant monitoring and policy evaluation has been undertaken as Victoria's Year 12

completion rate progresses towards the 90 per cent target. Figure 11.1 demonstrates the substantial progress that has been made.

Figure 11.1 **Progress towards Year 12 completion rate (Victoria)**
Persons aged 20–24 years with Year 12 or equivalent (AQF 2) or above



Data source: *Australian Bureau of Statistics (2008)*.

This idea of motivating evidence-based policy through an outcomes and evaluation framework and creating incentives and rewards to encourage an evidence-based approach to achieving the agreed outcomes is embodied in the current COAG National Productivity Agenda (see Appendix 1 to this paper for the outcomes framework) and the implementation of the human capital reform agenda at the State level.

Examples include the following.

1. In the National Education Agreement, States and Territories and the Commonwealth commit to working together to promote these outcomes and monitor the progress measures.
2. The COAG Reform Council is a federal–state body that has been established to report to COAG on the progress nationally and by jurisdictions in relation to these outcomes and progress measures.
3. In the *Melbourne Declaration on the Educational Goals for Young Australians* (MCEETYA 2008), signed by all education ministers in parallel with the National Education Agreement, all Australian governments committed to sharing evidence on best practice in the pursuit of their jointly agreed educational goals, for example through a biennial national forum. Other mechanisms have subsequently been agreed though the Education and Early Childhood Ministerial Council and Senior Officials Committee.

4. The establishment of national literacy and numeracy tests, which commenced in 2008, and the associated Australian Curriculum, Assessment and Reporting Authority, which was also charged with developing national curricula, was also a step forward in promoting evidence-based policy. The publication of results by jurisdiction and in due course school by school, including the use of like school groups to take into account the school context (especially the socioeconomic status of a school's students), represents a further stimulus to evidence-based policy. The national database that will result will enable national research and evaluation of what works in promoting literacy and numeracy outcomes, which was only possible to do with state-level data previously.
5. In the national partnership agreements on teacher quality, literacy and numeracy and low socioeconomic status school communities, there are facilitation and reward payments to promote the use of evidence-based policy to improve the agreed educational outcomes. Facilitation payments are to support policy initiatives that are built on pre-existing evidence about what works and reward payments are to reward the achievement of outcomes that the reforms are seeking to achieve, and will be based on the use of progress measures.
6. In the implementation of the *Blueprint for Education and Early Childhood Development* (DEECD 2008) in Victoria, the Victorian Government committed to pursuing stretch targets in relation to the outcomes framework. Within the government school system, this in turn leads to targets for each region and each network and each school, and a process for monitoring progress against targets. There is an associated evaluation and research program to determine the factors that are driving success or failure in making progress towards these outcomes, and funding to support interventions in schools where insufficient progress is made.

The modus operandi of a government department

In this section I provide an overview of how my department, the Victorian Department of Education and Early Childhood Development, seeks to follow a systematic approach to an evidence-based approach to policy advice, implementation and evaluation.

In accordance with the outcomes framework in the *Blueprint for Education and Early Childhood Development* (DEECD 2008), the department operates an outcomes and evaluation framework and an associated research strategy (figure 11.2).

Figure 11.2 **Outcomes and evaluation framework structure**

Data source: DEECD Corporate Plan 2009–2011.

To oversee this strategic evidence-based approach to policy development and review, the department has a Portfolio Strategy Board as its peak governance committee. The board meets quarterly and receives a quarterly report on progress against all the measures in the outcomes framework. It also reviews the progress of the various strategies that have been adopted to affect the agreed outcomes. Its work is supported especially by two divisions of the department. These are the Data and Evaluation Division and the Policy and Research Division. The Portfolio Strategy Board approves any amendments to the evaluation strategy, the research strategy and appropriation of the research budget. It recommends any proposed changes to the progress measures or targets though the Secretary to the Portfolio Ministers and on to relevant whole-of-government and cabinet processes.

Outcomes, progress measures and targets are in turn incorporated in the department's business planning process through the relevant offices and regions and through the government school system to school networks and individual schools. Within the government school system there is an accountability and improvement framework, within which each school has an annual improvement plan and a regular review cycle.

In the area of schools policy, the Secretary also chairs a cross-sectoral committee, which oversees a process of dialogue between the government, Catholic and independent school sectors, about how we can work together in the best interest of all young Victorians.

This cross-sectoral committee has overseen the process whereby the national partnership agreements between Victoria and the Commonwealth, involving investment in schools in all sectors and a process for evaluating the success of the partnership agreements, have been negotiated.

In the area of early childhood, the department has a partnership agreement with the Municipal Association of Victoria (MAV), under which the department and the MAV encourage all local authorities to develop Municipal Early Years Plans and to share evidence about the development of children on an area basis.

A research committee of the department makes recommendations and reports to the Portfolio Strategy Board about the research agenda to support the evidence-based policy work. The Secretary also has a group of external experts from universities and research bodies, which meets quarterly as a think tank to support the department's strategic thinking about evidence-based policy. Two members of the group are also coopted onto the Portfolio Strategy Board — one an expert on education and one an expert on early childhood development. The department commissions extensive research from external university-based and other relevant experts. We have recently been developing formal advice to university researchers about how to connect with the department's research agenda.

11.5 Conclusions

This paper has focused on the human capital reform agenda as a case study of evidence-based policy in Australia. A first conclusion is that the case for a major human capital reform agenda, put forward by the Victorian Government and followed by the Australian Government's Education Revolution, was itself motivated by a strong evidence base about the impact of human capital, especially the quantity and quality of education, on participation, productivity and economic growth. This was supported by econometric modelling and CGE simulations. This is an outstanding example of an evidence base generating a reform agenda.

The second conclusion is that an outcomes framework, such as *Growing Victoria Together,* and the COAG National Productivity Agenda outcomes framework, also adopted in the Victorian Government's *Blueprint for Education and Early Childhood Development* (DEECD 2008), can be a strong stimulus for evidence-

based policy. An outcomes framework can be linked with progress measures and sometimes targets, and an associated evaluation framework, which provides strong incentives for government to adopt policies which the evidence suggest can have a desirable impact on the agreed outcomes. The development of policies in Victoria to achieve the 90 percent Year 12 or equivalent target for 20- to 24-year-olds is a good example. The current development of evidence-based policies to improve teacher quality, increase literacy and numeracy and improve outcomes for students in low socioeconomic school communities, in the COAG Productivity Agenda, are further examples.

Third, it is clear that different kinds of evidence are useful in different circumstances, and that often multiple sources of evidence are ideal. In motivating the human capital reform agenda, aggregate econometric modelling and CGE simulations were very helpful. So was psychometric and econometric evidence about the links between teacher quality, literacy and numeracy, Year 12 completions and labour force participation. When focusing on specific policy interventions within the human capital reform agenda, evidence about the effects of policies adopted in other jurisdictions in Australia and around the world is useful. It is helpful in this context to have the benefit of sometimes randomised trials, for example in the consideration of performance pay for teachers, although care has to be taken about generalising from the specific findings of a particular trial.

Fourth, the COAG National Productivity Agenda is an illustration of how an evidence-based policy framework can support a federal–state reform agenda in the context of vertical fiscal imbalance. The use of an outcomes framework, progress measures, targets and facilitation and reward payments can provide the Australian Government with confidence about getting a return on its increased investment in education, and provides a framework for State and Territory governments to pursue an evidence-based policy agenda supported by facilitation and reward payments, the latter where improvements are achieved in the progress measures.

Fifth, in this paper I have described the way that the Victorian Department of Education and Early Childhood Development has embedded an evidence-based approach to policy development and evaluation and advice to ministers. It involves a Portfolio Strategy Board overseeing an outcomes framework with progress measures and an evaluation framework for assessing the impact of strategies and policies on the desired outcomes. In the government school system, this scrutiny of evidence goes right down through an accountability framework to the classroom level. This is all supported by a Data, Outcomes and Evaluation Division, a Policy and Research Division, and a Secretary's think tank of expert advisers, and extensive use of external researchers from universities and elsewhere.

Appendix 1 **COAG outcomes for the National Productivity Agenda**

	Early Childhood Development	*Schooling*	*Skills and Workforce Development*
Aspirations	That children are born healthy and have access to the support, care and education throughout early childhood that equips them for life and learning, delivered in a way that actively engages parents, and meets the workforce participation needs of parents	That all Australian school students acquire the knowledge and skills to participate effectively in society and employment in a globalised economy	All working aged Australians have the opportunity to develop the skills and qualifications needed, including through a responsive training system, to enable them to be effective participants in and contributors to the modern labour market. Individuals are assisted to overcome barriers to education, training and employment, and are motivated to acquire and utilise new skills. Australian industry and business develop, harness and utilise the skills and abilities of the workforce
Outcomes	Children are born healthy Children acquire the basic skills for life and learning Children will benefit from better social inclusion and reduced disadvantage, especially Indigenous children All children have access to affordable, quality early childhood education in the year before formal schooling Quality early childhood education and care supports the workforce participation choices of parents with children in the years before formal schooling	All children are engaged in and benefiting from schooling Young people are meeting basic literacy and numeracy standards, and overall levels of literacy and numeracy achievement are improving Schooling promotes social inclusion and reduces the educational disadvantage of children, especially Indigenous children Australian students excel by international standards Young people make a successful transition from school to work and further study	The working age population have gaps in foundation skills levels reduced to enable effective educational, labour market and social participation. The working age population has the depth and breadth of skills and capabilities required for the 21st century labour market. The supply of skills provided by the national training system responds to meet changing labour market demand. Skills are used effectively to increase labour market efficiency, productivity and innovation, and ensure increased utilisation of human capital.

Appendix 1 **(continued)**

	Early Childhood Development	*Schooling*	*Skills and Workforce Development*
Indicative Progress Measures	Proportion of children born of low birth weight Proportion of children with basic skills for life and learning, and who are vulnerable, as identified by the Australian Early Development Index Proportion of disadvantaged 3-year-olds in early childhood education Further performance measures need to be identified for children aged 18 months to 3 years Proportion of 4-year-olds accessing quality early childhood education Proportion of parents who can access the quality early childhood education and care services required for their preferred labour force participation	Proportion of children enrolled in and attending school Literacy and numeracy achievement of Year 3, 5, 7 and 9 students in national testing Proportion of students in the bottom and top levels of performance in international testing (e.g. PISA, TIMMS) Proportion of the 19-year-old population having attained at least a Year 12 or equivalent or AQF Certificate II Proportion of young people participating in post-school education or training six months after school Proportion of 18–24-year-olds engaged in full-time employment, education or training at or above Certificate III	Literacy and numeracy achievement of working age people in national and international testing. Proportion of 20–64-year-olds with or working towards the post-school qualifications in: • Cert III and Cert IV • diplomas and advanced diplomas. Level and proportion of total investment in structured (including nationally recognised) training by industry, individuals, businesses and government. Proportion of graduates employed after completing training. Extent of skills shortages, recruitment difficulties and labour market vacancies. Proportion of people employed at or above the level of their qualification.
COAG Targets	Universal access to early learning for all 4-year-olds by 2013 Halving the gap in mortality rates for Indigenous children under five years old within a decade In five years all Indigenous 4-year-olds in remote Indigenous communities will have access to a quality early childhood education program	Lift the Year 12 or equivalent attainment rate to 90 per cent by 2020 Halve the gap for Indigenous students in reading, writing and numeracy within a decade At least halve the gap for Indigenous students in Year 12 or equivalent attainment rates by 2020	Halve the proportion of Australians aged 20 to 64 without qualifications at Certificate III level and above by 50% between 2009 and 2020. Double the number of higher qualification completions (diploma and advanced diploma) between 2009 and 2020.

References

ALP (Australian Labor Party) 2007, *The Australian Economy Needs an Education Revolution*, Australian Labor Party, Canberra.

Angrist, J.D. and Lavy, V. 2004*, The Effects of High Stakes High School Achievement Awards: Evidence from a Group Randomized Trial*, IZA Discussion Paper 1146*,* Institute for the Study of Labor, Bonn, Germany.

Australian Bureau of Statistics 2008, *ABS Survey of Education and Work, Australia,* Cat no. 6227.0, Canberra.

Banks, G. 2009, *Challenges of Evidence Based Policy Making*, report to the Productivity Commission, Australian Public Service Commission, Canberra.

Boston Consulting Group 2008, *Organisational Learnings from the Roll-out of Performance and Development Culture in Victorian Government Schools: Background Paper*, Boston Consulting Group, Melbourne.

COAG (Council of Australian Governments) 2006, *COAG Communiqué, 10 February 2006*.

—— 2008a, *COAG Communiqué,* 3 July.

—— 2008b, *COAG Communiqué,* 20 November.

CTAC (Community Training and Assistance Centre) 2004, *Catalyst for Change: Pay for Performance in Denver Final Report*, CTAC, Boston, Massachusetts.

DE&T (Department of Education and Training) 2003, *Performance and Development Culture,* Melbourne, Victoria, http://www.education.vic.gov.au/management/schoolimprovement/panddc/default.htm (accessed 10 September 2009).

DEECD (Department of Education and Early Childhood Development) 2008a, *Blueprint for Education and Early Childhood Development*, DEECD, Melbourne, Victoria.

—— 2009b, *Rewarding Teaching Excellence: Blueprint Implementation Paper*, DEECD, Melbourne, Victoria.

DEET (Department of Education, Employment and Training) 2000, *Ministerial Review of Post Compulsory Education and Training Pathways in Victoria* (Peter Kirby, Chair), DEET, Melbourne, Victoria.

DIIRD (Department of Innovation, Industry and Regional Development) 2008, *Securing Jobs for Your Future*, DIIRD, Melbourne, Victoria.

DPC (Department of Premier and Cabinet) 2008, *Next Steps in Australian Health Reform: The proposals of the Victorian Premier, Brumby J.*, DPC, Melbourne, Victoria.

DPC and DTF (Department of Premier and Cabinet and Department of Treasury and Finance) 2005, *Governments Working Together: A Third Wave of National Reform — A New National Reform Initiative for COAG: The Proposals of the Victorian Premier*, DPC, Melbourne, Victoria.

DPC, DHS, DTF and DoE (Department of Premier and Cabinet, Department of Human Services, Department of Treasury and Finance and Department of Education) 2007, *Council of Australian Governments' National Reform Agenda: Victoria's Plan to Improve Outcomes in Early Childhood*, DPC, Melbourne, Victoria.

DPC, DoE and DTF (Department of Premier and Cabinet, Department of Education and Department of Treasury and Finance) 2007, *Council of Australian Governments' National Reform Agenda: Victoria's Plan to Improve Literacy and Numeracy Outcomes*, DPC, Melbourne, Victoria.

DTF (Department of Treasury and Finance) 2006, *The Economic and Fiscal Dividends of a New National Reform Agenda*, Working Paper, DTF, Melbourne, Victoria.

Figlio, D.N. and Kenny, L. 2006, *Individual Teacher Incentives and Student Performance,* National Bureau of Economic Research Working Paper 12627, National Bureau of Economic Research, Cambridge, Massachusetts.

Hanushek, E.A. 2009, 'The economic value of education and cognitive skills', in Sykes, G., Schneider, B. and Plank, D., *Handbook of Education Policy Research*, Routledge, New York, pp. 39–56.

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

Leigh, A. 2009, *Estimating Teacher Effectiveness from Two-Year Changes in Students' Test Scores*, Research School of Social Sciences, Australian National University, Canberra.

MCEETYA (Ministerial Council for Education, Employment, Training and Youth Affairs) 2008, *The Melbourne Declaration on the Educational Goals for Young Australians,* MCEETYA, Melbourne.

McKinsey & Co. 2007, How the World's Best Performing School Systems Come Out on Top, http://www.mckinsey.com/clientservice/socialsector/resources/pdf/Worlds_School_Systems_Final.pdf

Muralidharan, K. and Sundararaman, V. 2009, *Teacher Incentives in Developing Countries: Experimental Evidence from India,* National Bureau of Economic Research, Working paper No. 15323, Cambridge, Massachusetts.

Podgursky, M.J. and Springer, M.G. 2006, 'Teachers, schools and academic achievement', *Econometrica*, vol. 73, no. 2, pp. 417–58.

PC (Productivity Commission) 2006, *Potential Benefits of the National Reform Agenda — Report to the Council of Australian Governments*, Productivity Commission Research Paper, Productivity Commission, Canberra.

Rivkin S.G., Hanushek, E.A. and Kain, J.F. 1995, 'Teachers, students and academic achievement', *Econometrica*, vol. 73, no. 2, pp. 417–58.

Sanders, W. and Rivers, J. 1996, *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*, University of Tennessee Value Added Research and Assessment Centre, Knoxville, Tennessee.

The Nous Group 2007, *Evaluation of Flagship Strategy 4: Creating and Supporting a Performance & Development Culture in Schools (Year 2)*, report for the Department of Education, Victoria.

Winters, M.A., Ritter, G.W., Barnett, J.H. and Greene, J.P. 2008, 'An evaluation of teacher performance pay in Arkansas', working paper submitted to the *Journal of Public Economics*, Department of Education Reform, University of Arkansas, Arkansas.

12 Drawing on powerful practitioner-based knowledge to drive policy development, implementation and evaluation

Robert Griew
Associate Secretary, Department of Education, Employment and Workplace Relations

Abstract

Policy makers often have difficulty implementing the principles of evidence-based policy in the real world. It is important to construct policies that are not only consistent with evidence but also draw on practitioners' experience and are well integrated with the structure of the communities they serve. To determine the foundations of successful policies we must engage with the successful practitioners operating in the field. In this paper I describe four examples of the use of evidence to support practitioner-led action.

12.1 Introduction

In this paper I want to come at the topic of institutionalising an evidence-based approach to policy development and implementation from a slightly different angle than most of the other papers. I am not going to disagree with the importance of evidence, or with the importance of attention to the standards of evidence used.

If I have a straw person I would like to attack, however, it is not the waste and damage poor evidence and poorer motives can inflict on an unsuspecting public. Others have made that case very well. Rather, it is the opposition too glibly assumed between good evidence and the real world of professionals and the communities in which their services (and our evidence) either work or not.

I would like to discuss this by doing a couple of things. First, I want to discuss a couple of case examples, drawn both from previous lives and from my current role

at the Australian Government Department of Education, Employment and Workplace Relations (DEEWR).

In all of these I will argue that a commitment, not just to evidence, but to professional engagement — that is, to a direct relationship with the professionals whose life is the evidence — marks out these examples. It is not enough to assert the importance of evidence, as if the nature of evidence itself is unproblematic. It is also important to be clear about whose evidence, what evidence and evidence for what purpose.

Finally, I would like to consider this message in relation to the new structures of our federalism and to the Australian Government's ambitious Indigenous policy agenda, to 'close the gap'.

My key message is that a good evidence base alone does not guarantee good policy. A convincing evidence base will not redeem policies that are poorly integrated with the contexts of the communities they are developed to serve. To replicate successful reforms, we must engage with successful practitioners to isolate the specifics that underpin success and to harness them to the task.

Over years working across Commonwealth, State and Territory roles in health, welfare and education, I have observed an impressive commitment to evidence among my colleagues. And I have seen this both empower and inhibit action. I have even seen a commitment to evidence posited as a strategy to contain or even exclude 'professional interests'. We have all witnessed instances of an obsession with questions of technical evidence paralysing action or maybe even providing cover for an unwillingness to act.

12.2 Some reflections on evidence-based knowledge and evaluation

The Blair government is sometimes credited with reviving the idea that policy making 'should be better informed by research and evidence over the … conviction approaches of the Thatcher and Reagan eras' (Sanderson, quoted in Althaus et al. 2007). The *Modernising Government* white paper in 1999 focused on the capacity of the public service to 'to ensure that policies are strategic, outcome focused, … and robust' (United Kingdom Cabinet Office 1999).

However, nearly a decade later, the 2007 report *Analysis for Policy: Evidence-based Policy in Practice* asserts that many policy makers find evidence-based policy difficult to practise, in the context of expectations for quick decision making, poor

understanding of the relative merits of different forms of evidence and difficulty in mapping evidence to the 'real world' of policy implementation (GSR 2007). In Australia we face our own, as well as these common, challenges.

First, even when we know what works, it is difficult to put this knowledge into practice on a wider scale. The difficulties with replicating the successes in improving education outcomes for Aboriginal kids at Cherbourg State School in Queensland are related not only to the question of 'What works?' but also to the related questions — 'How does it work?' and 'Why does it work?' Sustainability and scalability are always key questions and are not answered simply by knowing that a specific intervention works. In the federal context, building this knowledge can be difficult, as the government with the greatest scale and reach is largely not responsible for direct service delivery.

Second, while commitment to evaluation has improved, evaluation is still often left to the end of the policy-making process, and we too rarely engage practitioners in the development and assessment of evaluation outcomes.

Third, this exclusion is exacerbated by the time pressures and the largely confidential reality of Cabinet government.

12.3 Utilising a practitioner-based knowledge to inform policy

I want to turn now briefly to outline four examples of evidence being used either to draw on or support practitioner-led action. The first two are from community health, and both are related to Aboriginal health and wellbeing.

The first is in improving the performance of the primary health sector in the management of adult metabolic and cardiac disease in remote communities. These largely syndromic diseases account for a huge proportion of the excess mortality suffered in those communities.

In this first case, a continuous quality improvement model known as ABCD (standing for Audit of Best Practice in Chronic Disease) was built on the basis of the 1999 Northern Territory Preventable Chronic Diseases Strategy (Australia's first). It involved:

- gathering data on performance across a number of annual cycles, eventually in each of more than 40 remote clinics
- analysing the performance of each clinic individually against a very specific set of performance standards

- working with practitioners at a clinic-by-clinic level to make key adjustments to service provision to improve each clinic's performance against these measures
- returning to each clinic over a number of annual cycles.

A 2007 article in the *Medical Journal of Australia* reported success: improvements in primary health care provision and closer adherence to best practice and small improvements to overall health outcomes in the same year that the Northern Territory first recorded an improvement in adult Aboriginal (female) life expectancy and a decline in the mortality from three of six major causes, sufficient to contribute to a closing of the gap (Bailie et al. 2007; NT DHCS 2007; Thomas et al. 2007).

The value of the CQI (continuous quality improvement) approach is that it offers more than functional evaluation, as it actually integrates with both the values and the day-to-day reality faced by staff.

In my second example, I worked with a group of colleagues to consider a cross-sectoral model for primary health care for Indigenous children, focusing not just on health but also on social outcomes. There is substantial evidence that antenatal, family and social environments all contribute to children's social, developmental and cognitive capacities, for good and bad, and to their long-term health outlook.

Accordingly, the report tries to 'envisage a service system where the boundaries between the service silos and professions are substantially dissolved and new service forms for Australian Indigenous families can be developed.' The rich knowledge base is, however, not enough. It is now relatively well known but does not generate change. So we spent the last third of our report on the question of change.

We identified two options for cross-sectoral reform: a pragmatic approach to reform for the three professions, based on linking up services by upskilling workers in cross-sectoral skills; and a more radical approach, based on integrated early childhood service centres. We wanted to look at how our proposed reform process could be developed 'across the different jurisdictions and settings where Indigenous families live and in all services, not just those led by an exceptional clinician, manager or community leader'. It was clear that professional buy-in was the key rate limiter. Government silos are a problem and we did not underestimate them. But getting the professions to loosen up would require more than a Head Office instruction.

12.4 The impact of practitioner-based knowledge on the DEEWR productivity agenda

I want to turn now to two examples from DEEWR. DEEWR's agenda impacts on all Australians across the life course. The broadness of this agenda ensures that DEEWR's work relates to a significant range of practitioners across the states and territories, in the community sector and industry.

I will give two examples.

Most commentators would identify the Government's macro response to the global recession as an example of well-based policy — indeed, of Australia's policy advisers being leaders in the clarity of their reading of the evidence and the impact of their advice. However, a similar effort to explore the complex data on local labour markets has also informed Government's approach to the impact of the global recession on regional labour markets.

Early on, DEEWR was asked to analyse current labour market indicators and evidence from previous downturns. The department identified 20 priority regions hardest hit already, badly affected in previous downturns or with exposure factors predicting further deterioration (that is, those most likely to experience high unemployment and long-term unemployment). 'Local employment coordinators' have been engaged in these regions to help drive local responses to unemployment — supported by the $650 million Jobs Fund.

For this paper, the interesting point is the energy with which both administrative and local knowledge was able to be added to the standard Australian Bureau of Statistics data sources to address the need. The eventual database constructed by DEEWR analysts ranged from statistical data with strong validation among forecasters to local intelligence from DEEWR state and regional network staff and has proven very robust as it is continually updated and reviewed.

The department's role in the development and implementation of the Council of Australian Governments' early childhood agenda has also involved close collaboration and partnership with experienced, domain-specific experts. The international evidence demonstrates that government investment in early childhood interventions for disadvantaged young children and for young children in general is more effective than interventions that come later in life (Heckman and Masterov 2007; McCain and Mustard 1999).

In response to this evidence, the Government has committed to both a quality agenda, in collaboration with the States and Territories, and to a service extension/integration agenda, including early learning and care centres and child and family centres — different forms of integrated models. These reforms are based on strong international evidence (Wise et al. 2005). We are also, however, taking steps to improve our own Australian evidence base. The national rollout of the Australian Early Development Index (AEDI) was announced in the 2008-09 Budget.

This is a population-based measure of child development, which enables local communities and schools to assess how children in their catchments are developing by the time they reach school age. As Peter Dawkins's paper outlines, the Victorian Best Start program provides another good example of Government intervening by providing local data to empower local alliances of professionals and community leaders to take their own action to improve outcomes for children in their own localities.

I want to finish with a brief reflection on the current coincidence of reform imperatives: reform of our federal–state architecture, and a number of specific and challenging social policy objectives — closing the gap with Indigenous Australians and mitigating the impacts of the global recession on vulnerable local labour markets, among others.

How will these fit together? Is there a collision of priorities here? Will we default back to the old ways of input control, as have previous governments? I want to make three comments.

First, let us not kid ourselves about the power of input control. The levers we all grew up with were rarely, if ever, used and tended to focus us on the wrong things. Certainly, there is no sign that they did a lot to close the gap.

Second, no-one should underestimate the resolve of the current Australian Government regarding the measures that have been agreed in the new style of partnership agreements between the Commonwealth and the States and Territories.

This is elsewhere referred to as the 'transparency agenda'. The Government is unashamed about pursuing transparency agendas in a number of policy areas, consistent with its focus on outcomes, while also respecting the room for the States and Territories to move in implementation. Terry Moran's presentation yesterday to this forum made that clear.

And, of course, transparency is for all of us — for professionals in the field but also for all of us in government. From this perspective, the question is not about reform commitment. It is about whether we are having the right discussions, informed by the powerful evidence and with evidence gathered in a way which itself prompts practice improvement.

My contention is that the debate about closing the gap will not be about which level of government talks to the professionals, the practitioners and the communities. It will be about the substance of what they have to tell us and how, together, we respond to the challenges and opportunities richly embedded in their evidence.

References

Althaus, C. Bridgman, P. and Davis, G. 2007, Australian Policy Handbook, 4th ed., Allen & Unwin, Crows Nest.

Australian Government 2009, *Budget 2008-09 Ministerial Statements: Closing the Gap Between Indigenous and Non-Indigenous Australians*, http://www.aph.gov.au/Budget/2009-10/content/ministerial_statements/indigenous/html/ms_indigenous-02.htm (accessed 8 July 2009).

Australian Public Service Commission 2008, Address to Heads of Agencies and members of Senior Executive Service, the Hon. Kevin Rudd MP, Prime Minister of Australia, http://www.apsc.gov.au/media/rudd300408.htm (accessed 9 July 2009).

Bailie, R., Si, D., O'Donoghue, L. and Dowden, M. 2007. 'Indigenous health: effective and sustainable health services through continuous quality improvement', *Medical Journal of Australia*, vol. 186, no. 10, pp. 525–27.

Banks, G., *Challenges of Evidence-Based Policy-Making*, Contemporary Government Challenges Series, Productivity Commission, Canberra, http://www.apsc.gov.au/publications09/evidencebasedpolicy.htm (accessed 8 July 2009).

Briggs, L., 2006, New directions for implementation, ANZSOG–PM&C Conference: Project Management and Organisational Change, 21 February 2006, Darwin, http://www.apsc.gov.au/media/briggs210206.htm (accessed 14 July 2009).

Briggs, L., 2008. State of the Service Report 2007-08, presentation in Darwin, 23 March 2009, Australian Public Service Commission, Canberra, http://www.apsc.gov.au/media/briggs230309.htm (accessed 15 July 2009).

Civil Service 2008, *The Professional Skills for Government (PSG) Competency Framework*, http://www.civilservice.gov.uk/people/psg/index.aspx (accessed 15 July 2009).

COAG (Council of Australian Governments) 2008, *Intergovernmental Agreement on Federal Financial Relations*, Australian Government, Canberra, http://www.coag.gov.au/intergov_agreements/federal_financial_relations/docs/national_partnership/naional_partnership_on_early_childhood_education.pdf (accessed 10 February 2009).

DEECD (Department of Education and Early Childhood Development), 2009, *What is Best Start*, DEECD, Melbourne, http://www.education.vic.gov.au/ecsmanagement/beststart/what.htm (accessed 13 July 2009).

Farrelly, R., 2008, 'Policy on trial', *Policy*, vol. 24, no. 3, pp. 7–12.

GSR (HM Treasury, Government Social Research Unit), 2007. *Analysis for Policy: Evidence-based Policy in Practice*, http://www.gsr.gov.uk/downloads/resources/pu256_160407.pdf (accessed 10 July 2009).

Heckman, J. and Masterov, D., 2007. *The Productivity Argument for Investing in Young Children*, NBER Working Paper #13016, University of Chicago, Chicago, http://healthcare-economist.com/2007/05/01/the-productivity-argument-for-investing-in-young-children/ (accessed 11 July 2009).

Karoly, L.A., Greenwood, P.W., Everingham, S.S., Hoube, J., Kilburn, M.R., Rydell, C.P., Sanders, M., Chiesa, J. 1998. *Investing in Our Children: What We Know and Don't Know about the Costs and Benefits of Early Childhood Interventions*, RAND Corporation, Santa Monica, California.

McCain, M.N. and Mustard, F. 1999. *Early Years Study: Reversing the Real Brain Drain: Final Report*, Ontario Children's Secretariat, Government of Ontario.

Robert Griew Consulting, Eades, S.; Lea, T.; Peltola, C., JTA International, 2007, *Family Centred Primary Health Care: Review of Evidence and Models Funded by the Office for Aboriginal and Torres Strait Islander Health*, Department of Health and Ageing, Canberra, http://www.health.gov.au/internet/main/publishing.nsf/Content/F9B9C09A8203D507CA25751F007EA423/$File/FINAL%20FCPHC%20fr%20Robert%20Griew19.11.pdf (accessed 9 July 2009).

Sanderson, I. 2006. 'Complexity, "practical rationality" and evidence-based policy making', *Policy and Politics*, vol. 34, no. 1, pp. 115–32.

United Kingdom Cabinet Office 1999, *Professional Policy Making for the Twenty First Century*, http://www.nationalschool.gov.uk/policyhub/docs/profpolicymaking.pdf (accessed 10 July 2009).

University of Melbourne 2006, *Statewide Evaluation of the Best Start Program: Final Report*, commissioned by the Victorian Department of Human Services, Melbourne, http://www.dhs.vic.gov.au/beststart/docs/2007/bs_eval_report_Sept2006.pdf (accessed 13 July 2009).

Wise, S. 2005, *The Efficacy of Early Childhood Interventions*, report commissioned by the Australian Government Department of Family and Community Services, Australian Institute of Family Studies (AIFS) Research Report No. 14, AIFS, Melbourne. Available at: http://www.aifs.gov.au/institute/pubs/resreport 14/main.html [Accessed July 11 2009]

13 Intelligent federalism: accountability arrangements under COAG's reform of federal financial relations

Mary Ann O'Loughlin
Executive Councillor and Head of Secretariat, COAG Reform Council

Abstract

A key objective of COAG's reform of federal financial relations is to strengthen accountability for the quality and efficiency of the services delivered and the outcomes achieved. National Agreements in education, skills and workforce development, healthcare, disability services, affordable housing, and Indigenous reform set out agreed objectives, outcomes and performance indicators. The COAG Reform Council — an independent agency — assesses and publicly reports on the performance of the Commonwealth, State and Territory governments under the National Agreements. The accountability arrangements set standards to hold governments to account and, through a mix of incentives, encourage improved performance. This paper examines the COAG reforms and concludes that the new arrangements have the potential to improve government performance — and outcomes for Australians — through building the evidence base and fostering learning, both across and within governments.

At this conference, we mainly heard two kinds of accounts. The first were histories of past successes or failures to influence policy through the application of evidence.

In *The Intelligence of Democracy*, Charles Lindblom argues that pluralist democracy is superior to other political systems because of the greater number of incentives it contains to encourage intelligence and learning in the process of policy making (Bovens 2006, p. 26; Lindblom 1965). If this is so, in the process of policy making, federations such as Australia — with a built-in basis for comparing and learning across central and sub-national governments — have an advantage over unitary democracies.

The advantages of federalism are often touted (Twomey and Withers 2007, chapter 2). Federalism provides the opportunity — and often the pressure — to be innovative and to experiment in order to compete with other jurisdictions. In a federation, ideas can be tested by a jurisdiction and copied by others. Where experiments fail, federalism 'cushions the nation as a whole from the full impact of government blunders' (de Q Walker 2001, p. 38).

This paper looks at the accountability arrangements that can encourage 'intelligent federalism' under recent reforms of the Council of Australian Governments (COAG) — the peak intergovernmental forum in Australia, comprising the Prime Minister, State Premiers, Territory Chief Ministers and the President of the Australian Local Government Association.

The paper begins with an outline of COAG's reform of federal financial relations. It then looks at the associated accountability arrangements, focusing on the role of the COAG Reform Council in assessing performance under the new National Agreements between the Commonwealth, State and Territory governments. The paper argues that the new accountability arrangements — founded on and with a commitment to evidence-based policy — have the potential to improve governments' performance through fostering and strengthening learning both across and within governments.

13.1 Reform of federal financial relations

In March 2008, COAG endorsed a new reform agenda for Australia, agreeing to work together to:

> … boost productivity, workforce participation and geographic mobility, and support wider objectives of better services for the community, social inclusion, closing the gap on Indigenous disadvantage and environmental sustainability. (COAG 2008a, p. 2)

Reform of the architecture of Commonwealth–State financial relations is an essential part of this reform agenda, with COAG agreeing to implement a new framework for federal financial relations — 'the most significant reform of Australia's federal relations in decades' (Commonwealth of Australia 2009, p. 3). The intent of the new framework is to improve the wellbeing of all Australians through improvements in the quality, efficiency and effectiveness of government services (COAG 2008b, p. 4).

The Intergovernmental Agreement on Federal Financial Relations (the IGA) provides the overarching framework for the Commonwealth's financial relations with the States and Territories. It establishes a foundation for governments to

collaborate on policy development and service delivery, and to facilitate the implementation of economic and social reforms. All policy and financial relations between the Commonwealth and the States and Territories are now governed under the provisions of the IGA (Commonwealth of Australia 2009, p. 9).

There are three main elements of the new financial arrangements.

- National Specific Purpose Payments (SPPs) supported by new National Agreements.
- National Partnership payments associated with National Partnership Agreements.
- a performance and assessment framework to support public reporting and accountability.

Under the new framework for federal financial relations, the previous more than 90 different payments from the Commonwealth to the States and Territories for specific purposes — many containing prescriptive conditions on how the funding should be spent — have been combined into five new *National SPPs* (Commonwealth of Australia 2009, p. 24). National SPPs are ongoing financial contributions from the Commonwealth to the States and Territories to be spent in the key service delivery sectors of schools, skills and workforce development, health care, affordable housing, and disability services. The States and Territories are required to spend each National SPP in the service sector relevant to the SPP but they have full budget flexibility to allocate funds within that sector as they see fit to achieve the agreed objectives for that sector (COAG 2008b, p. D-2).

National SPPs are associated with *National Agreements* between the Commonwealth and State and Territory governments. National Agreements establish the policy objectives in the service sectors of education, skills and workforce development, health care, affordable housing, and disability services. There is also a National Agreement on Indigenous Reform which does not have an associated SPP, although it links to other National Agreements and National Partnerships which have associated funding.

National Agreements set out the objectives, outcomes, outputs and performance indicators for each sector, which are agreed between all jurisdictions. The agreements also clarify the roles and responsibilities of the Commonwealth, States and Territories in the delivery of services and the achievement of outcomes. They do not include financial or other input controls imposed on service delivery by the States and Territories, and there is no provision for National SPPs to be withheld in the case of a jurisdiction not meeting a performance benchmark specified in a National Agreement.

National Partnership Agreements outline agreed policy objectives in areas of nationally significant reform or for service delivery improvements, and define the outputs and performance benchmarks. The Commonwealth provides National Partnership payments to support the delivery of specified projects, to facilitate reforms, or to reward those jurisdictions that deliver on national reforms (Commonwealth of Australia 2009, p. 26).

The extent of the change in federal financial arrangements under the IGA is shown in table 13.1. There has been a major shift away from the previous form of Commonwealth payments to the States and Territories for specific purposes, which often involved prescriptions on service delivery in the form of financial or other input controls.

- In 2007–08, 42.7 per cent of Commonwealth payments were in the previous form of payments for specific purposes, compared to only 4.1 per cent in 2009-10.
- There has been a shift to the new National SPPs; 28.5 per cent of Commonwealth funding is in this form in 2009–10.
- There has also been a shift to funding under National Partnership payments; 8.3 per cent of Commonwealth payments is in this form in 2009–10.

Table 13.1 **Commonwealth payments to the States and Territories**
2007-08 and 2009-10

Payments	*2007-08*	*2009-10*
	%	%
Existing payments for specific purposes	42.7	4.1
National Specific Purpose Payments (supported by National Agreements)	–	28.5
National Partnership payments	0.3	8.3
GST	56.9	58.0
Other general revenue assistance	0.2	1.1
Total	**100.0** ($74 960m)	**100.0** ($83 200m)

Source: Commonwealth of Australia (2009, chapter 2).

The third main element of the new federal financial relations arrangements is a *performance and assessment framework* to support public reporting and accountability. Under the IGA, the Commonwealth, States and Territories have agreed to greater accountability through simpler, standardised and more transparent performance reporting, and 'a rigorous focus on the achievement of outcomes — that is, mutual agreement on what objectives, outcomes and outputs improve the

wellbeing of Australians' (COAG 2008b, p. 5). The IGA gives the COAG Reform Council significant responsibilities for assessment and reporting of the performance of governments under National Agreements and National Partnerships.

13.2 Role of the COAG Reform Council

The COAG Reform Council assists COAG to drive its national reform agenda by strengthening accountability for the achievement of results through independent and evidence-based monitoring, assessment and reporting on the performance of governments. The council is independent of individual governments and reports directly to COAG. (An overview of the role of the COAG Reform Council is at http://www.coag.gov.au/crc/index.cfm.)

As set out in the IGA (COAG 2008b, p. A-4), the role of the COAG Reform Council is to:

- monitor, assess and publicly report on the performance of the Commonwealth, States and Territories in achieving the outcomes and performance benchmarks specified in National Agreements
- independently assess whether predetermined performance benchmarks have been achieved before an incentive payment is made to reward nationally significant reforms under National Partnerships.

More specifically, for National Agreements the COAG Reform Council provides annual reports to COAG containing the performance data and a comparative analysis of the performance of governments in meeting the objectives of the agreements. The reports are made public. The reports also:

- highlight examples of good practice and performance so that, over time, innovative reforms or methods of service delivery may be adopted by other jurisdictions
- highlight contextual differences between jurisdictions which are relevant to interpreting the data
- reflect COAG's intention to outline transparently the contribution of both levels of government to achieving performance benchmarks and to achieving continuous improvement against the outcomes, outputs and performance indicators (COAG 2008b, p. C-2).

13.3 Improving performance through accountability arrangements

The IGA is clear that improved accountability is a key objective of the new framework for federal financial relations. The framework aims to ensure that the appropriate government is accountable to its community 'not just for its expenditure in delivering services, but more importantly for the quality and efficiency of the services it delivers and the outcomes it achieves' (COAG 2008b, p. 5).

Public accountability has three main functions (Bovens 2006, pp. 25–6; Bovens 2007, pp. 192–3):

- to monitor and control the conduct of governments
- to enhance the integrity of public governance
- to improve the performance of government by strengthening the learning capacity and effectiveness of public administration.

Within the framework of the IGA, the main function of public accountability is to improve performance. Performance is improved through accountability arrangements that are:

- preventive — by setting standards to hold institutions to account
- remedial — by encouraging responsibility to fix problems and to prevent their recurrence
- educative — by tracing connections between past, present and future policies (Bovens 2006, p. 26).

The next sections discuss how the public accountability role of the COAG Reform Council under the IGA has the potential to improve governments' performance through arrangements that are preventive, remedial and educative, focusing on the council's role under the six National Agreements.

13.4 Setting standards

The six National Agreements — in education, skills and workforce development, health care, affordable housing, disability services and Indigenous reform — have a similar structure. The National Agreements clearly identify the outcomes, performance indicators and targets — the standards by which governments are held to account. Setting standards has a preventive function, steering reform and actions towards the achievement of outcomes:

> Accountability is not only about control, it is also about prevention. Norms are (re)produced, internalized, and, where necessary, adjusted through accountability. The manager who is held to account is told about the standards he must hold to and about the fact that in the future he may again (and, in some cases, more strictly) be called to account. (Bovens 2007, p. 193)

As an example of the setting of standards, figure 13.1 summarises the structure of the National Education Agreement. All National Agreements begin with the objective(s) of the agreement — the overall aim. The objective of the National Education Agreement is:

> … that all Australian school students acquire the knowledge and skills to participate effectively in society and employment in a globalised economy. (COAG 2008c, p. 1)

Each agreement also has a set of outcomes agreed by governments. As shown in figure 13.1, the National Education Agreement has five outcomes.

For each outcome there is a set of performance indicators which measure progress towards the outcomes. For example, under the National Education Agreement, for the outcome of 'young people meet basic literacy and numeracy standards and that levels of achievement are improving', the performance indicator is literacy and numeracy achievement of Year 3, 5, 7 and 9 students in annual national testing under the National Assessment Program — Literacy and Numeracy (NAPLAN).

Most National Agreements also include identified targets to achieve. The National Education Agreement has three (listed in figure 13.1).

Each year, the COAG Reform Council reports the performance information for all jurisdictions against National Agreement outcomes and performance benchmarks (COAG 2008b, p. A-4), reflecting the preventive nature of the accountability arrangements. This preventive function is given weight by the council's responsibility to publicly release its reports on governments' performance under the National Agreements.

The effectiveness of the outcomes, performance indicators and targets of the National Agreements in steering governments towards reform is dependent on the robustness of the evidence upon which they are founded. In his chapter in this volume, Peter Dawkins describes the rigorous evidence base to build the case for a national human capital reform agenda under COAG, which led to the development of the outcomes framework and associated performance indicators of the National Education Agreement.

Figure 13.1 **Structure of the National Education Agreement**

Objectives, outcomes, performance indicators and targets

Objective	**All Australian school students acquire the knowledge and skills to participate effectively in society and employment in a globalised economy.**			
Outcomes	All children are engaged in and benefiting from schooling	Young people are meeting basic literacy and numeracy standards, and overall levels of literacy and numeracy are improving	Australian students excel by international standards	Young people make a successful transition from school to work and further study
	Schooling promotes the social inclusion and reduces the educational disadvantage of children, especially Indigenous children			
Performance Indicators	The proportion of children enrolled in and attending school (by Indigenous and low socio-economic status)	Literacy and numeracy achievement of Year 3, 5, 7 and 9 students in national testing (by Indigenous and low socio-economic status)	The proportion of students in the bottom and top levels of performance in international testing (e.g. PISA, TIMSS).	The proportion of the 20–24-year-old population having attained at least a Year 12 or equivalent or AQF Certificate II (by Indigenous and low socio-economic status)
	The proportion of Indigenous students completing Year 10			The proportion of young people participating in post-school education or training six months after school
				The proportion of 18- to 24-year-olds engaged in full-time employment, education or training at or above Certificate III
Targets	Lift the Year 12 or equivalent attainment rate to 90 per cent by 2015 Halve the gap for Indigenous students in reading, writing and numeracy within a decade At least halve the gap for Indigenous students in Year 12 or equivalent attainment rates by 2020			

Source: COAG (2008c).

13.5 Encouraging responsibility for performance

Each year the COAG Reform Council also undertakes a *comparative* analysis of the performance of all the jurisdictions — Commonwealth, States and Territories — towards the outcomes, as measured by the performance indicators and targets (COAG 2008b, p. C-2). The comparative analysis compares the performance of jurisdictions against each other and also against their own year-on-year performance, reflecting the importance of achieving continuous improvement against the outcomes, outputs and performance indicators. The comparative analysis supports remedial action — by encouraging responsibility to fix problems and to prevent their recurrence:

> An administrator who is called to account is confronted with his policy failures and he is aware that, in the future, he can be called upon again, even more pitilessly, to render account. (Bovens 2006, p. 26)

To take an example of comparative analysis, figure 13.2 presents illustrative data for the States and Territories against the National Education Agreement's 'young people meet basic literacy and numeracy standards' performance indicator. The agreed measure of the indicator is the proportion of students achieving at or above the national minimum standard. Achievement of the minimum standard indicates that the student has demonstrated the basic elements of literacy and numeracy for the year level.

The data against this indicator are shown for Year 5 Reading. Year 5 Reading is a good indicator of performance, as reading is a foundation skill for writing and numeracy and by Year 5 the impact of jurisdictional differences in school starting age on the acquisition of skills should be diminishing.

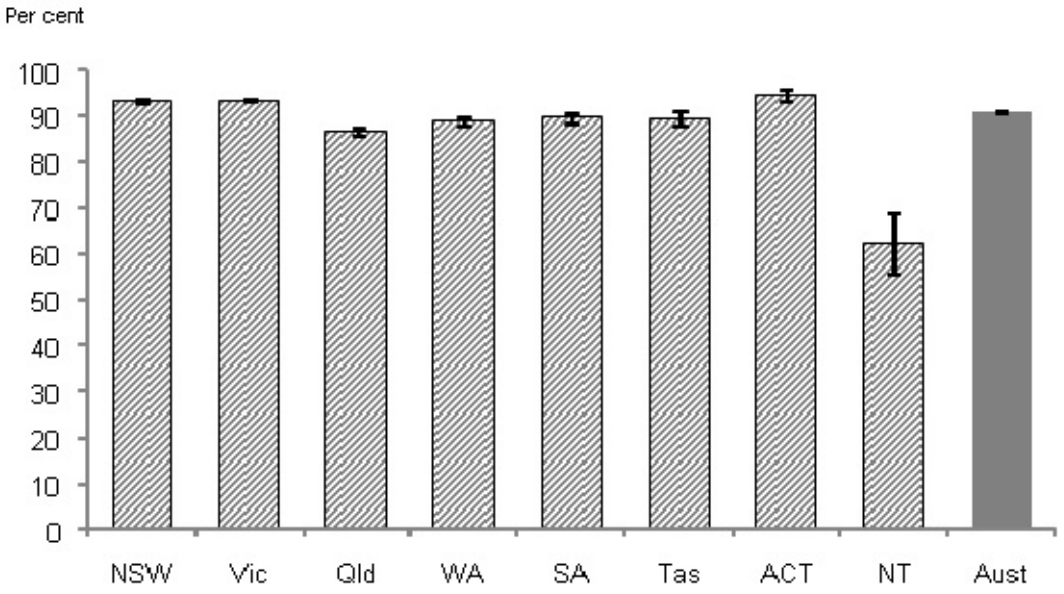
Nationally, a high proportion of all students — 91 per cent — achieve at or above the national minimum standard in assessments of reading at Year 5 (COAG Reform Council 2009, pp. 61–2). Comparing the performances of the States and Territories:

- three jurisdictions achieve higher levels than the national average — New South Wales, Victoria and the Australian Capital Territory — all with results of 94 per cent or above of students meeting the national minimum standard in Year 5 Reading;
- four jurisdictions cluster below the national average, with the proportion of students meeting the national minimum standard ranging from 90 per cent in South Australia and Tasmania, to 89 per cent in Western Australia, and to 87 per cent in Queensland;

- the Northern Territory differs markedly from other States and Territories, with 63 per cent of students meeting the national minimum standard for Year 5 Reading.

Figure 13.2 **Proportion of Year 5 students achieving at or above the national minimum standard for reading**

By State and Territory, 2008



Notes: 1. The achievement percentages shown in this graph include 95 per cent confidence intervals indicated by error bars. 2. Exempt students were not assessed and are deemed not to have met the national minimum standard. 3. Absent and withdrawn students did not sit the test and are not included in these data.

Data source: MCEETYA (2008, p. 56).

To help understand performance, the council is also required to highlight relevant contextual differences between the jurisdictions — such as differences in populations — that are associated with high or low performance (COAG 2008b, p. C-2). For example, the Northern Territory's results for Year 5 Reading reflect its high proportion of Indigenous students (41 per cent) and the high proportion of students living in remote areas (46 per cent) (COAG Reform Council 2009, p. 145).

The council's approach to the task of highlighting contextual differences reflects its general approach to the assessment of governments' performance (COAG Reform Council 2009, pp. 6–7). In particular, the council's approach is *dynamic*, emphasising changes in performance from year to year. The first reporting year — 2008 for the National Education Agreement — establishes the baseline data and benchmarks against which improvements in performance can be measured in subsequent years.

Given this approach, the contextual differences that are highlighted in assessing performance under the National Agreements are high level and small in number. They are focused on differences that help understand the data by giving the broad context, in particular student characteristics such as Indigenous and socioeconomic status. This is particularly relevant for first-year reports, as they present the baseline data for the comparative assessment of performance. However, these contextual factors are likely to be less relevant to understanding changes in performance, and hence are less likely to be relevant in subsequent years' reports as the focus shifts to assessing the performance of the jurisdictions over the years compared with their baseline data (COAG Reform Council 2009, p. 26).

Furthermore, while the council is required to highlight contextual differences between jurisdictions which are relevant to interpreting the data, COAG is clear that the council does not have a policy advising role — the council does not analyse the effectiveness of the governments' policies and programs behind the results of performance (COAG 2008b, p. A-4). This is consistent with the focus of the National Agreements and the associated National SPPs on outcomes, and with the principle of budget flexibility for the States and Territories to allocate funds as they see fit to achieve the outcomes. (The COAG Reform Council does have a role in highlighting good practice, which is discussed below.)

Thus, the comparative analysis does not explain *why* there are differences between the jurisdictions. Rather than providing an explanatory analysis, the comparative analysis is better thought of as providing *catalyst* data (Ekholm 2004, p. 1). The comparative analysis of performance — highlighting differences among the jurisdictions — leads one to search for reasons to explain the differences.

Catalyst data can be powerful, as pointed out by an editorial in *The Australian* on Queensland's results in NAPLAN testing:

> Following the woeful performance of Queensland primary school children in national testing last year, the Bligh Government turned to an expert for help. The state's children need it, after being ranked second-last in the nation. Only the Northern Territory where absenteeism and social disadvantage are more prevalent fared worse. (*The Australian*, 4 May 2009, editorial)

In response to Queensland's performance in NAPLAN in 2008, Premier Bligh took remedial action, seeking an independent review by Professor Geoff Masters from the Australian Council for Educational Research of the literacy and numeracy standards in Queensland primary schools and advice on how to improve students' skills (Masters 2009).

The COAG Reform Council's comparative analysis of performance must also reflect 'COAG's intention to outline transparently the contribution of both levels of

government to achieving performance benchmarks and to achieving continuous improvement against the outcomes, outputs and performance indicators' (COAG 2008b, p. C-2).

Each National Agreement identifies the roles and responsibilities of the Commonwealth, State and Territory governments in achieving the agreement's outcomes. The council must assess the performance of all governments against the indicators and benchmarks, given their respective responsibilities. To do this, the council will seek information from jurisdictions on their work towards the achievement of the nationally agreed outcomes, consistent with their roles under each agreement. This will be done for the second and subsequent years' reports in light of the comparative analysis of the performance data in the first year report.

13.6 Tracing connections between policies and outcomes

Accountability arrangements can be educative — by tracing connections between past, present and future policies:

> Public accountability … offers a regular mechanism to confront administrators with information about their own functioning and forces them to reflect on the successes and failures of their past policy. (Bovens 2006, p. 26)

The COAG Reform Council has two main 'educative levers' under the IGA. First, in reporting on performance under the National Agreements, the council is required to highlight examples of good practice and performance 'so that, over time, innovative reforms or methods of service delivery may be adopted by other jurisdictions' (COAG 2008b, p. C-3). The Council will do this in the second and subsequent years' reports on the National Agreements.

Under the agreed process, the Council will select, in consultation with the jurisdictions, one or more areas for good practice and performance analysis, based on the comparative analysis of the performance data for the previous year. This provides the opportunity for an in-depth analysis of the performance data and additional related data, and of relevant Australian and international research. The Council will liaise with jurisdictions and draw upon subject experts as needed.

The second 'educative lever' is the council's role in relation to National Partnerships. As noted earlier, the Council is required to independently assess whether performance benchmarks have been achieved before an incentive payment is made under National Partnerships. In addition, in the reports on the National Agreements, the Council is required to include an analysis of the performance

information for National Partnerships to the extent that they support the objectives in the National Agreement (COAG 2008b, p. A-4).

The National Partnerships vary in their intent, but at the core of many is the implementation of reforms to improve the outcomes in a sector. For example, table 13.2 summarises the National Partnerships related to the objectives of the National Education Agreement. To take one example, the National Partnership Agreement on Literacy and Numeracy aims to deliver sustained improvement in literacy and numeracy outcomes for all students, especially those who are falling behind, by focusing on the key areas of teaching, leadership and the effective use of student performance information (COAG 2008d, p. 1).

Consistent with the educative function of the accountability arrangements, in the second and subsequent years' reports on National Agreements, the COAG Reform Council will report the performance data of National Partnerships related to National Agreements, linking progress towards outcomes and objectives with the reform actions of governments to improve performance.

Table 13.2 **National Partnerships supporting the National Education Agreement**

National Partnership	*Objective*
Improving Teacher Quality	To improve teacher and school leader quality to sustain a quality teaching workforce by targeting critical points in the teacher 'lifecycle' to attract, train, place, develop and retain quality teachers and leaders in Australian schools.
Literacy and Numeracy	To deliver sustained improvement in literacy and numeracy outcomes for all students, especially those who are falling behind, by focusing on the key areas of teaching, leadership and the effective use of student performance information.
Low Socio-Economic Status School Communities	To support reforms to transform the way schooling takes place in participating schools and address the challenges facing students in disadvantaged communities.
Early Childhood Education	To provide universal access to quality early childhood education for all children in the year before full-time school by 2013.
Youth Attainment and Transitions	To improve access to increase educational attainment and the engagement of young people aged 15 to 24 with education, training and employment.

Source: National Partnerships are available at http://www.coag.gov.au/intergov_agreements/federal_financial_relations/index.cfm.

13.7 Conclusions

COAG's reform of federal financial relations — particularly under the new National SPPs and National Agreements — is an example of the 'devolution and transparency' paradigm of large-scale public sector reform. This paradigm is based on the devolution of responsibility to the organisational units delivering the relevant service and then the use of transparency to drive performance, by making public the results of differing units in a way that allows comparisons to be made (Barber 2007, p. 4).

The new accountability arrangements of the National Agreements clearly identify the roles and responsibilities of all governments — Commonwealth, State and Territory — in improving the quality, efficiency and effectiveness of government services. They also have a mix of incentives — preventive, remedial and educative — to encourage governments to improve performance. Further, they give an independent agency — the COAG Reform Council — responsibility for assessing and publicly reporting on the performance of governments.

The new arrangements take advantage of Australia's federal system with its built-in basis for benchmarking and comparing across Commonwealth, State and Territory governments. They are aimed squarely at improving performance through fostering and strengthening learning.

At the heart of this tradition of accountability is the question of the extent to which governments deal adequately with — learn constructively from — feedback about their own performance (Bovens 2006, p. 26).

At the heart of the success of the new arrangements is the extent to which Australian governments can embrace intelligent, *gutsy* federalism.

References

Barber, M. 2007, 'Three paradigms of public-sector reform', McKinsey & Company, http://www.cabinetoffice.gov.uk/media/cabinetoffice/strategy/assets/paradigm.pdf (accessed 26 October 2009).

Bovens, M. 2006, 'Analysing and assessing public accountability. A conceptual framework', *European Governance Papers*, No. C-06-02, http://www.connex-network.org/eurogov/pdf/egp-connex-C-06-01.pdf (accessed 26 October 2009).

—— 2007, 'Public accountability', in Ferlie, E., Lynn, L. and Pollitt, C. (eds), *The Oxford Handbook of Public Management*, Oxford University Press, Oxford.

COAG (Council of Australian Governments) 2008a, *Council of Australian Governments' Meeting, 26 March 2008, Communiqué*, http://www.coag.gov.au/coag_meeting_outcomes/2008-03-26/docs/communique20080326.pdf (accessed 26 October 2009).

—— 2008b, *Intergovernmental Agreement on Federal Financial Relations*, http://www.coag.gov.au/intergov_agreements/federal_financial_relations/index.cfm (accessed 26 October 2009).

—— 2008c, *National Education Agreement*, http://www.coag.gov.au/intergov_agreements/federal_financial_relations/docs/IGA_ScheduleF_national_education_agreement.pdf (accessed 26 October 2009).

—— 2008d, *National Partnership Agreement on Literacy and Numeracy*, http://www.coag.gov.au/intergov_agreements/federal_financial_relations/docs/national_partnership/national_partnership_on_literacy_and_numeracy.pdf (accessed 26 October 2009).

COAG Reform Council 2009, *National Education Agreement: Baseline Performance Report for 2008*, COAG Reform Council, Sydney.

Commonwealth of Australia 2009, *Australia's Federal Relations*, Budget Paper No. 3 2009-10, Canberra.

de Q Walker, G. 2001, 'The advantages of a federal constitution', *Policy*, Summer 2000-01, pp. 35–41, http://www.cis.org.au/policy/summer00-01/polsumm0001-8.pdf (accessed 26 October 2009).

Ekholm, M. 2004, 'Evidence-based policy research — some Swedish lessons', paper presented at the First OECD Conference on Evidence Based Policy Research, 19–20 April, Washington, http://prod.ceg.rd.net/usermedia/images/uploads/PDFs/OECD-Ekholm.pdf (accessed 2 June 2009).

Lindblom, C. 1965, *The Intelligence of Democracy*, Free Press, New York.

Masters, G. 2009, *A Shared Challenge: Improving Literacy, Numeracy and Science Learning in Queensland Primary Schools*, Australian Council for Educational Research, http://education.qld.gov.au/mastersreview/pdfs/final-report-masters.pdf (accessed 26 October 2009).

MCEETYA (Ministerial Council for Education, Employment, Training and Youth Affairs) 2008, *2008 National Assessment Program — Literacy and Numeracy: Achievement in Reading, Writing, Language Conventions and Numeracy*, http://www.naplan.edu.au/verve/_resources/2ndStageNationalReport_18Dec_v2.pdf (accessed 27 October 2009)

The Australian 2009, 'Testing time in Queensland', editorial, 4 May, p. 9.

Twomey, A. and Withers, G. 2007, *Federalist Paper 1 Australia's Federal Future: Delivering Growth and Prosperity*, a report for the Council for the Australian Federation, http://www.caf.gov.au/Documents/AustraliasFederalFuture.pdf (accessed 26 October 2009).

General discussion

The discussion provoked by session 4 papers centred on: institutional initiatives that might support better use of evidence; debate over using indicators for performance management; and the risks of inappropriate media uses of data and analysis.

Institutional initiatives for better use of evidence

Patricia Rogers, Robert Griew, and Peter Dawkins each noted the importance of building a greater public sector culture of transparency, and openness to evaluation and learning. Participants who had had confidential access to departmental archives of evaluations noted that methodological quality was often very poor; that many seemed to have been done to meet formal requirements rather than to promote organisational learning; and that staff are 'often terrified of being punished' as a result of any evaluation that identified ways a program could have been made more effective. Some argued the importance of ensuring that existing evaluations, even if of limited quality, were made publicly available for analysis and to spur improvements.

Speakers noted the value of institutional independence for quality evaluation, and some pointed to the Productivity Commission as an exemplar of institutional design to defend independence. Another approach to creating a degree of independence and expertise in assessment was the Office of Development Effectiveness within AusAID.

Another institutional path to strengthening evaluation was to ensure easier and better-directed access to others' evaluation work. Several participants noted they had been impressed by the example of the Cochrane Collaboration outlined by Sally Green (Chapter 8 above).

Departmental contributors noted a perennial problem of limited expertise in evaluation, with the intra-departmental evaluation function frequently hindered by rapid rotation of generalist staff. They saw a need to train and give experience to specialist evaluators.

Mary Ann O'Loughlin and some other participants also pointed to the potential of 'evaluation clubs'. Some mentioned practice in the area of evaluations of the effectiveness of foreign aid. Such 'clubs' brought together evaluators to share lessons, to contribute peer support and peer review, and in some cases, to fund high-

quality evaluations out of a pool of funds contributed by club members. One such club was the International Initiative for Impact Evaluation or "3ie", whereas in the area of domestic policy, Ron Haskins mentioned the work of the US Coalition for Evidence-Based Policy (of which he is a board member).

Using indicators for performance management.

Jeffrey Smith noted that, at best, well-selected indicators typically reflected outcomes, but outcomes were not the same as the impacts of policies at work in the area. An indicator that suggested a good outcome in one jurisdiction might not have identified a better-functioning policy, but might just result from that jurisdiction having a privileged subset of the broader population.

A strong improvement in an indicator in one jurisdiction might also follow from extravagant resourcing at the cost of other objectives, whereas lesser improvement in the same indicator in another jurisdiction could follow from much better targeted use of fewer resources. In such a case, perhaps the 'weaker performing' jurisdiction deserved the greater plaudits and funding, for achieving more with less.

He pointed out that some indicators (such as retention rates for secondary school completion in Australia, or eligibility for college admission in the US) could be met by policies that wasted human and financial resources. Better, he argued, to carefully evaluate the impacts of policies and enumerate their benefits and costs, than to risk creating incentives to meet inappropriate indicators.

Other speakers acknowledged that such problems could arise with some indicators, but argued that in the context of Australia and the COAG process, jurisdictions were well placed to stress the context surrounding their performance as captured in indicators, and that the information added by the indicators was a step towards better understanding of successful policies.

Misuse of data and analysis

Some discussion turned on the problem of misuse or misrepresentation of data and analysis. One example cited was the misuse of the Productivity Commission's 'outer envelope' estimates in its 2007 report to the Council of Australian Governments on the *Potential Benefits of the National Reform Agenda*. Speakers noted that these estimates of the long term, gross benefits, after adjustment of the economy, to the assumed full implementation of the National Reform Agenda, came with very clear caveats. Yet those caveats were often overlooked in reporting of the benefits.

Participants' comments on this episode included the view that little harm was done by such oversimplification; that there were some reasons in the early stages of knowledge about the NRA why elements in the estimates might have been conservative; and that even somewhat misguided use of estimates was part of a better-informed debate than if estimates had not been ventured.

SESSION 5

WHAT HAVE WE LEARNED AND WHERE TO FROM HERE?

14 Rapporteur's comments

Jonathan Pincus
University of Adelaide

Abstract

Economics has always relied on data and evidence, and on ways of thinking logically about them. So the plea for evidence-based policy is directed elsewhere, at politicians and their advisors. Undoubtedly, evidence has influenced policy, often for the good. However, practical policy choice is determined by interests, political preferences and power, as well as by evidence. Strident calls for more 'evidence-based policy' can reflect a political naiveté; or can hide a claim that politics should be run by 'experts'; or can be a cover for the role of interests.

At this conference, we mainly heard two kinds of accounts. The first were histories of past successes or failures to influence policy through the application of evidence. The second were programmatic contributions, about how to bring more evidence to bear, or better evidence; or how better to interpret evidence. We also had some discussion of a third kind, about the broader frameworks within which policy does operate or should operate. I will direct some of my remarks there, using the standard approach of 'public choice'.

In summary, to whom is the plea for more evidence-based policy being directed, and with what expectation of reaction? I argue that the prime target cannot be policy-oriented economists, whose practices do not fall foul of the criticisms that Archie Cochrane directed at the National Health Service. Rather, the pleas for evidence-based policy are primarily directed at politicians and their advisors. To make such pleas implies the expectation of an effect; an expectation based on some theory of public decision-making processes, or of politics, in which there is an important role of experts. Political decisions create winners and losers, and economists are not expert in assessing the values of various alternative distributions of economic welfare. That fact may lead to a degree of disappointment: whereas top lawyers in Australia can get to make actual public policy decisions (as judges), top economists rarely do more than give advice to decision makers.

14.1 Economists do use evidence

If you are confronted by a person who says 'I support evidence-based policy: what about you?', what response is there except: 'Of course I support evidence-based policy'? It is an adequate reply, even though you may also think but not say 'However, I need to know exactly what you mean by "evidence", and what you mean by "based"'.

Unlike the common practice of medicine before Archie Cochrane, policy-oriented economics was not a craft occupation, and certainly not a craft that neglected to heed the evidence which could be made available for interrogation as to the efficacy and efficiency of craft practice.

Henry Ergas kindly provided me with the following from Cochrane's 1972 book:

> I once asked a worker at a crematorium, who had a curiously contented look on his face, what he found so satisfying about his work. He replied that what fascinated him was the way in which so much went in and so little came out. I thought of advising him to get a job in the NHS, it might increase his job satisfaction, but decided against it. He probably gets his kicks from the visual demonstration of the gap between input and output. A statistical demonstration might not have worked so well. (Cochrane 1972, p. 12)

From the fact that this conference had a splendid address about the Cochrane Collaboration in Australia, it would be wrong to infer that economists and other social scientists suffer from envy of doctors. Yes, doctors do decide; but rarely on matters of public policy. (I will return to this issue later.)

In contrast with what Archie Cameron criticised in medicine, economics has had a very long tradition of searching for confirmatory non-theoretical evidence, at least, and, at best, for crucial or decisive evidence.

Although I am much more familiar with economics than with other social sciences, what I know of sociology, social psychology, geography and anthropology suggests that the same is true: they have all had long traditions of looking at the factual evidence through some theoretical filter or other. Even political science has long had an empirical stream, and political philosophy is gathering and using evidence these days.

Thus, the plea for evidence-based policy is not aimed at policy-oriented economists in universities and the like. Economics has always relied on data and evidence, and a set of ways of thinking logically about them. Nonetheless, there can be virtue in preaching to the converted: it hones arguments to be presented later to the un-converted; and it helps to keep up morale (Harries 1991).

As Gary Banks and Brian Head both stress in their conference contributions, policy decisions depend on more issues than those of pressing importance to evidence based policy–style economists. Economists and other social scientists, acting in their professional capacity, rarely get to decide anything of public importance. Top economists get the Nobel Prize; become secretary of the US Treasury, chairman of the US Federal Reserve, or governor of the Australian Reserve Bank, or enter other senior public offices; have rewarding careers on statutory bodies; become consultants; or hold prestigious university chairs.[1] But in the Anglosphere there are far fewer public decision-making positions for an economist-as-economist than there are for lawyers, who can aim at becoming judges, of which there are plenty. So maybe the conference needed a paper from the field — the law — that has given much thought to questions of admissibility, relevance and weight of evidence.

14.2 Interests matter

At the 2009 Australian Conference of Economists, John Siegfried made the Australian launch of a book that he edited for the American Economic Association, called *Better Living through Economics*. The book has 12 case studies in which economic research, according to Siegfried, 'had a profound effect on our way of life and material wellbeing affecting income and wealth, as well as mortality, health, happiness, and welfare'.

The 12 cases are:

- tradable pollution permits
- price index reform
- antitrust reform
- matching physicians and students
- spectrum auction design
- airline deregulation
- welfare-to-work reform
- Earned Income Tax Credit
- trade liberalisation

[1] Economists do find employment on various Australian statutory entities such as the Commonwealth Grants Commission (an advisory body, but effectively a decision-making one); the Australian Consumer and Competition Commission; the Australian Competition Tribunal; and wage-fixing bodies. Often, however, they generally share decision-making powers with lawyers. See Hughes (1980).

- saving for retirement
- monetary policy targeting
- voluntary military force.

In many instances, the original impetus was largely theoretical. For example, Milton Friedman advocated a volunteer army in his 1962 book *Capitalism and Freedom*, using at best casual empiricism, but mostly through powerfully argued theoretical and moral cases. For tradable permits, the foundational theory was provided by Ronald Coase and amplified by William Baumol and Wallace Oates. For the policy of reducing barriers to international trade, the original theoretical ideas date back at least to David Ricardo.

Theory was important, but not trumps. In all but maybe one case, some kind of empirical evidence was essential for 'selling' the idea to policy advisers and to politicians, and even to the general public; for example, some trials or experiments were run before the matching algorithm was adopted widely in medical and other schools. At the other end of the spectrum, the change in default enrolment in pension funds in 2006, the impetus was, to a great extent, empirical not theoretical.

I want to make two points about these and similar 'war stories'. The first is to reiterate that economics has long used evidence as the basis of policy recommendations.

My second point was signalled by my calling these 'war stories'. During the conference we were treated to similar tales about the influence of evidence on policy making, but I found them largely unconvincing. I tend to view such stories through the 'interest' approach to politics, associated with Arthur Bentley and E.E. Schattschneider in American political science; and with the 'public choice school' in American economics. The basic public choice idea is that diffused interests, and especially those standing to gain little per person, will not be as effective politically as concentrated interests that stand to gain (or lose) much per head. However, all other things being equal, inefficient policies are less attractive politically than are efficient policies.

Siegfried's selection is unabashedly of instances in which economic advice was taken—what could be called 'merited victories'. However, we need to look at an unbiased sample of all cases in which evidence was tendered, which include 'unmerited defeats' (good ideas not adopted) as well as 'unmerited victories' (bad ideas adopted: living worse through economics?).

For 'merited victories', we need to entertain the hypothesis that interests played a major role even when evidence was allegedly decisive in a public policy decision.

This does not mean completely ignoring the role of good evidence, or the roles of values, ideology and morality in public decision-making processes, but maybe giving them a weaker part than is played by interests.

To take an example of failure of evidence-based policy — the background paper for this roundtable details the three worthy goals of the National Drought Policy:

- encourage primary producers and other sections of rural Australia to adopt self-reliant approaches to managing the risks stemming from 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. (DAFF 2007)

These goals were supported by a considerable quantity of relevant evidence. Yet the money went elsewhere: expenditure has predominantly been directed as emergency payments to a minority of farmers in perceived hardship, and to farm businesses meeting eligibility criteria.

One interpretation is that there is a political value in listing worthy objectives of this kind in legislation, to give some 'cover' for what the program is really about, which is 'emergency' relief. Did the minister keep a straight face when introducing the legislation? Did parliament really expect that traditional drought relief would be replaced by payments according to the objectives of the National Drought Policy? (Rural electorates — because of the relative homogeneity and greater salience of their economic interests — have disproportionate power in the Australian political system, compared to city electorates.)

If we entertain a model of public policy that is largely interest driven, what is the role for evidence about what is good for 'community welfare'? Does the provision of such evidence explain, for instance, the great neoliberal experiment of reducing Australian rates of border protection against imported goods? Similar policy changes were taking place in other countries, so an Australian explanation should largely be directed at the timing of the change, not the fact of it. Various kinds of Australian evidence were adduced against the protectionist policy. The first was that it was expensive to national economic welfare, and increasingly so: this fact counted against its continuation in Australia and internationally — the balance of interests was being upset. Secondly, evidence about the dynamics of the Australian labour market calmed some fears about mass unemployment following a tariff cut, and the government put in place some special forms of adjustment assistance. But maybe at least as important was the fact that one rural-based politician, Bert Kelly, convinced farmers that the system was largely against their interests.

In his conference presentation, Bruce Chapman recounted the role of evidence in the introduction of university fees and the attendant income-contingent loan scheme in Australia. The evidence he cited mainly related to estimates of the average private and public rates of return from university study, and to the socioeconomic classification of university students. I can well understand how evidence of average rates of return can be used to justify university fees: Adam Smith would have approved. But we should look elsewhere for evidence to support a system of unconditional loans for student fees, with income-contingent repayments: presumably, evidence that promising potential students from low socioeconomic backgrounds had been dissuaded from enrolling at Australian universities for cashflow reasons, or would be, if they had to pay fees up front. (Forgone earnings were a much larger cost to the average student than were the fees being contemplated.) As far as I know, no such Australian (or other) evidence was adduced. Would it have mattered to the decision, if the evidence had been presented *against* the hypothesis of cash or credit constraints?

My point is that there is an alternative hypothesis about the use of evidence in this instance, namely, that Minister Dawkins wanted his educational revolution of a massification of the universities; that he knew that taxpayers would be very unhappy if they were told that they had to bear the full cost of the expansion of tertiary education; and that the Higher Education Contribution Scheme achieved the first goal without the second consequence.

14.3 Some win, some lose

What is being sought, through the campaign for evidence-based policy, is a stronger or wider role in policy decisions for experts in policy-relevant evidence. There seem to be two main targets of the campaign. First are politicians and their close advisors. Second are moralists, sometimes called ideologues. Of course, these two categories can overlap.

Maybe it is a good thing to preach to politicians and their advisors.[2] Presumably the purpose is to influence policy decisions. But policy decisions take account of matters upon which economists are not expert. Gary Banks wrote of the Australian experience that:

> It also reveals the sterility of academic debates about whether evidence can or should play a 'deterministic' role in policy outcomes. It will be clear to all at this gathering in Canberra that policy decisions will typically be influenced by much more than

[2] James M Buchanan argued that a sensible society 'pays the preacher' (including, possibly, bodies such as the Productivity Commission) (Buchanan 1994).

> objective evidence, or rational analysis. Values, interests, personalities, timing, circumstance and happenstance — in short, democracy — determine what actually happens. (Banks 2009, p. 4)

Although this quotation can be interpreted simply as stating a fact — namely, that values, interests and so on, do influence policy outcomes — I suspect that Gary Banks also had a normative intent: that some of these things should influence policy. (However, I also note that values and interests seem to be set in opposition to or cut across 'objective evidence' and 'rational analysis'.)

It is useful to list three questions about public policy, the last of which I have already briefly discussed:

1. What works?
2. What is worth doing?
3. What is done?

Early in the conference, we were reminded of Tony Blair's statement about being interested only in 'what works'. This is a question of efficacy, a technical question to be answered by specialists or experts. We do need to know, however, what it is that 'what works' works on. In particular, a factual matter of great interest to politicians and their advisers is who gains and who loses, and how much: what works to improve the wellbeing and affiliation of a set of voters?

Jeff Smith memorably reminded the conference to not forget the 'subscript *i*': or, put otherwise, that it is unknown for a policy to improve the outcomes for everyone who is affected, more than some alternative. Therefore, someone or some group has to make an overall judgment: Is this policy worth doing?[3]

Typically, the economist's solution is to employ practical utilitarianism — for example, cost–benefit studies, or computable general-equilibrium modelling — to decide what works to improve economic efficiency or community wellbeing. This presumes an answer to the second, normative question: namely, that improving 'community wellbeing' is worth doing. As a moral stance, utilitarianism is far better than many others that I can think of. But it is not the only respectable framework of normative evaluation, of a means of making an overall judgment about a policy that harms some interests or values, and helps others.

Utilitarianism is a consequentialist approach. In contrast, there is an important moral claim underlying Milton Friedman's *Capitalism and Freedom* — namely,

[3] For the use of the distinction between 'reason for belief' and 'reason for action', see Furedi (2009).

that freedom is a value in itself, one that exists apart from the consequential benefits or costs that freedom brings and entails.

I suspect that there is no completely non-consequentialist defence of freedom (except as a matter of religious faith, as in 'God wants us to be free'). But the main arguments about freedom — those that led to the change in the way that people were thought to relate to their rulers and state, and hence which led to democratic forms of governance and the rule of law — were moral, not factual or consequential: they were about the moral standing of individuals.

To take an extreme example, slavery was abolished in the British Empire not because it did not work, but because it came to be considered morally abhorrent.[4]

The moral constraint imposed by utilitarianism is that it approves only those changes that generate more gains than losses — all in terms of unobservable stuff called 'utility'.[5] Australian democracy is thankfully not damaged by a tradition of 'winner takes all' after a general election. But that does not mean that our politics are based, or should be based, on a philosophy of 'winner takes nothing unless it is more than is taken from the losers'. Utilitarianism involves a much greater constraint than imposed by the prohibition against 'winner takes all'. Therefore, it is not the dominant ideology of politics.

14.4 Conclusion

Modesty, not stridency, befits economic policy advisers. I vividly recall the speech given by the then Treasurer, Peter Costello, on the occasion of the thirtieth anniversary celebrations of the Industry Commission/Productivity Commission. He reminded the audience that it was the politicians, not their advisors and academics, who took the bold, even heroic, decisions on microeconomic reform. Good ideas do not prevail in policy through a magical process of shining the light of truth into darkness. Rather, they have influence within a system that defines problems and agendas, selects from competing approaches, then crafts, implements and seeks to justify responses — and involved in all these steps are interests, political preferences and power. Strident calls for more 'evidence-based policy' at best

4 Can we base moral judgments solely on factual evidence? There is a group of moral philosophers who use laboratory experiments, surveys and natural and field experiments to investigate what some have called the universal grammar of morality. But the relationship between these empirical investigations and moral argument has yet to be agreed.

5 In practice, economists selectively apply tests for economic efficiency: specifically, very rarely do economists take seriously the inefficiencies that the usual methods indicate arise from substantial redistribution. An exception is Mark Harrison (Harrison 2007).

reflect a political naiveté; or, worse, can hide a claim that politics should be run by 'experts'[6]; or, worst, can provide a smokescreen to disguise the role that interest plays, while simultaneously marginalising and delegitimising opposing viewpoints.

14.5 Coda: randomised controlled trials

This conference has ranged widely over types of evidence, from random assignments into treatment categories within a controlled environment, through natural experiments, through statistical analysis of panel data sets and other forms of econometrics, through computable general-equilibrium modelling, to theoretical propositions. A comment on those who hunger after medical-style evidence for the social sciences: despite Andrew Leigh's paper, I remain convinced that Australian randomised controlled trials (RCTs) are destined to provide major inputs into relatively minor social and economic policy decisions.

Say we had conducted an experiment through some kind of RCT in 2007 to find out how people react to an unexpected gift of cash from the government: what determines what they spend and what they save and, if they save, for how long; also, what are the effects on taxpayers, if the gift is debt financed? I strongly doubt that this would have told us what Treasury policy makers needed to know, when they advised the Rudd Government on the stimulus package.

Or take performance pay for teachers in state schools. Say that we were concerned that lessons from US studies would not transfer well to Australia, in view of the US tradition of locally controlled public schools, so we wanted to test the effects of performance pay in Australia before it was considered for widespread implementation. We clearly could not run double-blind assignments of teachers to one form of pay or another. Therefore, some teachers would be selected randomly and informed that they would be eligible for performance pay; and others, not. The first hurdle is that any experiment would require the rewriting of the contracts of individual teachers or, at least, a renegotiation of their conditions of employment. Assuming that that hurdle were surmounted, I would still expect that union opposition to performance pay would greatly influence the outcomes, in unknowable ways (whether the random assignments were made within each school or across a state) (Donohoe 2009).

---

6 The most famous expression of a yearning for economists to be regarded as experts is J.M. Keynes's 'If economists could manage to get themselves thought of as humble, competent people on a level with dentists, that would be splendid' (Keynes 1931, p. 373).

References

Banks, G. 2009, 'Evidence-based policy-making: What is it? How do we get it?', ANZSOG Public Lecture, 4 February, http://www.pc.gov.au/speeches/cs20090204. Also reprinted as 'Challenges of Evidence-based Policy', Australian Public Service Commission.

Buchanan, J.M. 1994, *Ethics and Economic Progress*, University of Oklahoma, Norman.

Cochrane, A.L. 1972, *Effectiveness and Efficiency: Random Reflections on Health Services*, Nuffield Provincial Hospitals Trust, London.

DAFF (Department of Agriculture, Fisheries and Forestry) 2007, *Drought and Exceptional Circumstances*, http://www.daff.gov.au/brs/climate-impact/drought (accessed 20 January 2010).

Donohoe, P. 2009, 'Performance Pay has no Place in Education', *Green Left Online*, 9 May, http://www.greenleft.org.au/2009/794/40875 (accessed 20 January 2010).

Furedi, F. 2009, 'Specialist pleading', *The Australian Literary Review*, 2 September, pp. 14–5.

Harries, O. 1991, 'A Primer for Polemicists', *Commentary*, http://www.libertarian.co.uk/lapubs/tactn/tactn010.pdf (accessed 20 January 2010).

Harrison, M. 2007, *The Outcomes of Income Transfers*, New Zealand Business Roundtable, Wellington.

Hughes, C.A. 1980, 'Government Action and the Judicial Model,' in Tay, A.E. and Kamenka, E. (eds.), *Law-making in Australia*, Edward Arnold, Melbourne, pp. 268–70.

Keynes, J.M. 1931, *Essays in Persuasion*, Macmillan, London.

Siegfried, J. (ed.) 2009, *Better Living through Economics*, Harvard University Press, Cambridge, Massachusetts.

General discussion

Following Jonathan Pincus's reflections on the first four sessions (chapter 14), the Roundtable concluded with comments from a panel made up of David Tune, Ron Haskins, Jeffrey Smith and Mary Ann O'Loughlin, and then a general discussion involving other participants.

Panel discussion

David Tune suggested the subject matter of the Roundtable would be valuable to both policy advisers and the politicians that they served. He noted that in long experience as a policy adviser, he had learnt there were often conflicting objectives in policy making (for example between excellence in design of a spending program, and fiscal cost).

A good policy idea could be proposed 5 or 6 times over 15 years, without ever gaining support, and then be accepted and implemented because various political and other forces had moved into alignment. One possible example was the reform of COAG arrangements, where there was a widening realisation among politicians and public servants at all tiers of government that existing processes weren't working well, and that continued strong economic performance would require redoubled reform efforts. Similarly, crises could also facilitate initiatives that couldn't otherwise have been undertaken.

Evidence could be influential when such opportunities arose, and public servants should try to anticipate the evidence requirements that would facilitate policy change. While this could be difficult, one successful example was precautionary thinking within the Commonwealth Treasury some 4 or 5 years earlier, about how governments should respond when faced with the threat of another recession. That exploration of the evidence had paid dividends in advice permitting a speedy response to the global financial crisis.

Finally, David Tune suggested that emphasis on evaluation may have faded over the last decade. Policy proposals used to require inclusion of an evaluation strategy, and funding was allocated for approved evaluation plans. But the quality of execution had sometimes fallen short, and it was desirable to build a stronger evaluation culture.

Ron Haskins noted that while the use of evidence in the political process was often 'ugly', there were avenues of influence that could be used to strengthen evidence-based thinking: individual evidence-minded legislators; transparency institutions such as the Congressional Budget Office in the USA and the Productivity Commission in Australia; journalists with an interest in good use of evidence; and (in the US context) Congressional Hearings.

Making good policy was a medium-term commitment: most major US reforms had been built over 4 to 8 years, and 'the truth will out in the long run', even if poorly-informed thinking held sway at particular stages of the debate.

US experience that had been useful in strengthening the quality, quantity and influence of evidence included shaping reform legislation that included a budget for high quality evaluation, using random assignment where appropriate, and charging the relevant cabinet secretary with responsibility for using that funding most effectively.

Jeffrey Smith cited claims from some Australian colleagues who thought Australia would stand towards the bottom of a league table of advanced economies in its use of evidence to inform policy. But reflecting on some of the creditable Australian examples presented at the roundtable, he doubted that assessment would be true. He was, however, less optimistic than Ron Haskins about the beneficial impact of quality reporting in advancing the cause of evidence-based policy. In the US, he found even specialist journalists in quality publications generally made a poor job of explaining how a particular body of evidence had been established, and what its strengths and limitations were.

A key issue in strengthening evidence was the training and employment of numerate and methodologically-skilled evaluators. The US experience was that anti-poverty programs since the 1960s had seen the gradual emergence and dominance of economists over previous generations of sociological researchers in this field, strengthening evaluation. But a professional monoculture was not desirable, and Professor Smith praised the work of the US Institute for Education Sciences, that had transformed the quality of education policy evaluation, including by providing doctoral training grants for students who studied rigorous evaluation of education policies – with the key being the methodological rigour of the study.

Professor Smith also outlined the case for 'routinization' of evaluation in a particular area (such as labour market programs, or agricultural programs). Routinization sought to standardize the processes of relevant evaluations, in effect to a template that could be applied to many projects by less skilled evaluators. This would hopefully produce acceptable quality evaluations in a greater range of cases

than could have been studied from first principles by a very limited pool of highly skilled evaluators.

Mary Ann O'Loughlin highlighted what for her were the most important messages from the Roundtable discussion: good evidence matters; good use of evidence has to be alert to differential impacts of policies on different groups, not just the average impact; randomised evaluation methodologies could be powerful, but were not a cure-all; and that the lags between gaining and analysing evidence, and its influencing a constituency for reform, were large.

On this last point, transparency and communication were important parts of the task. For policy makers and advisers to make sense of a large variety of rapidly-growing evidence, there were important roles for networks of experts, and for dissemination and processing institutions such as the Cochrane Collaboration.

General discussion

In brief closing discussion, speakers with policy experience in New South Wales, Victoria, Queensland and the Commonwealth all argued it was an opportune time to carry forward innovations to support stronger evaluation and better use of evidence in policy formation.

Data sets (for example on health and education) that had previously been closely held within jurisdictions were now beginning to become more widely available, and there was scope for facilitating access to data and protocols for data sharing that could perhaps be helped by agencies such as the Australian Bureau of Statistics, the CoAG Reform Council, or the Productivity Commission.

Speakers felt the Roundtable had highlighted some ideas that were ripe for practical application in Australia, and some from other countries that might be useful, such as the roles of the Economic and Social Research Council, the Government Social Research Service, and the Chief Government Social Researcher in the UK. Professor Brian Head noted the opportunity to draw on such ideas in development of the Australian Research Council's work.

One speaker observed that while there were obviously opportunities for government departments to do better evaluations, there would always be conflicting pressures on such departments, and there were virtues in transparency and independence in sponsoring greater evaluative contributions from institutions outside the public service.

A Roundtable program

Day 1 — Monday 17 August	
3.30 – 4.00	***Registration***
4.00 – 4.15	***Welcome***
	Gary Banks, Chairman, Productivity Commission
Session 1	***Evidence Based Policy: Its principles and development***
	Chair: Gary Banks
4.20 – 4.40	**Brian Head,** Institute for Social Science Research, University of Queensland
4.45 – 5.25	*Keynote speaker:* **Ron Haskins,** Senior Fellow, Brookings Institution
5.30 – 6.10	*Keynote speaker:* **Jeffrey Smith,** Department of Economics, University of Michigan
6.10 – 6.40	Roundtable discussion
6.45 – 7.15	*Pre-dinner drinks*
7.15 – 9.45	***Dinner***
	Guest speaker: ***Terry Moran,*** Secretary, Department of Prime Minister and Cabinet
Day 2 — Tuesday 18 August	
Session 2	***How robust is our evidence-based policy making?***
	Chair: Mike Woods, Deputy Chairman, Productivity Commission
8.45 – 9.05	**Bruce Chapman,** Crawford School of Economics and Government, Australian National University
9.05 – 9.25	**Henry Ergas,** Chairman, Concept Economics
9.25 – 9.45	**Grant Scobie,** Principal Adviser, Policy Coordination and Development, NZ Treasury
9.45 – 10.15	Roundtable discussion
10.15 – 10.30	***Morning tea***

Session 3	***From rhetoric to practice – how do we improve the availability and quality of evidence?*** *Chair:* Bernie Wonder, Head of Office, Productivity Commission
10.30 – 10.50	**Sally Green,** Australasian Cochrane Centre
10.50 – 11.10	**Patricia Rogers,** Public Sector Evaluation, RMIT
11.10 – 11.30	**Andrew Leigh,** Research School of Social Sciences, ANU
11.30 – 12.00	Roundtable discussion
12.00 – 1.00	***Lunch***
\Session 4	***Institutionalising an evidence-based approach – how can an evaluation culture be embedded into policy-making?*** *Chair:* Wendy Craik, Commissioner, Productivity Commission
1.00 – 1.20	**Peter Dawkins,** Secretary, Victorian Department of Education and Early Childhood Development
1.20 – 1.40	**Robert Griew,** Associate Secretary, Department of Education, Employment and Workplace Relations
1.40 – 2.00	**Mary Ann O'Loughlin,** Executive Councilor and Head of the Secretariat of the COAG Reform Council
2.00 – 2.30	Roundtable discussion
2.30 – 2.35	***Afternoon tea –*** to the table
Session 5	***What have we learned and where to from here?*** *Chair:* Gary Banks
2.35 – 2.55	*Rapporteur:* **Professor Jonathan Pincus**, University of Adelaide
2.55 – 3.25	*Panel discussion*: **David Tune,** Associate Secretary, Department of Prime Minister and Cabinet, **Mary Ann O'Loughlin, Jeffrey Smith, Ron Haskins**
3.25 – 3.55	Roundtable discussion
3.55 – 4.00	***Closing: Gary Banks***

B Roundtable participants

Gary Banks	Chairman, Productivity Commission
Dr Ron Ben-David	Chairman, Essential Services Commission, Victoria
Professor Laurie Brown	Research Director, National Centre for Social and Economic Modelling, University of Canberra
Dr Matthew Butlin	Chairman, Victorian Competition and Efficiency Commission
Professor Bruce Chapman	Public Policy, Australian National University
Blair Comley	Deputy Secretary, Department of Climate Change
Dr Wendy Craik	Commissioner, Productivity Commission
John Davidson	Assistant Director General, Australian Agency for International Development
Pam Davoren	Deputy Chief Executive, ACT Chief Minister's Department
Professor Peter Dawkins	Secretary, Department of Education and Early Childhood Development
Professor Meredith Edwards	Faculty of Business and Government, University of Canberra
Dr Henry Ergas	Chairman, Concept Economics
Susan Garner	Chair of Canberra Chapter, Australasian Evaluation Society
Dr Jenny Gordon	Principal Adviser Research, Productivity Commission
Professor Sally Green	Centre Director, Australasian Cochrane Collaboration
Professor Bob Gregory	Research School of Social Sciences, Australian National University
Professor Robert Griew	Associate Secretary, Department of Education, Employment and Workplace Relations
Lisa Gropp	Principal Adviser Research, Productivity Commission

Dr Ron Haskins	Senior Fellow, Brookings Institution
Professor Brian Head	Institute for Social Science Research, University of Queensland
David Kalisch	Commissioner, Productivity Commission
Larry Kamener	Consultant, Boston Consulting Group
Professor Paul Kerin	Melbourne Business School, University of Melbourne
Professor Andrew Leigh	Research School of Social Sciences, Australian National University
Angela MacRae	Commissioner, Productivity Commission
John McCormick	Director, Department of Premier and Cabinet — Tasmania
Terry Moran	Secretary, Department of Prime Minister and Cabinet
Terry O'Brien	First Assistant Commissioner, Productivity Commission
Mary Ann O'Loughlin	Executive Councillor and Head of Secretariat, COAG Reform Council
Professor Jonathan Pincus	Visiting Professor, University of Adelaide
Brian Pink	Australian Statistician, Australian Bureau of Statistics
Professor Patricia Rogers	Professor in Public Sector Evaluation, Royal Melbourne Institute of Technology
Dr Grant Scobie	Principal Adviser, New Zealand Treasury
Steve Sedgwick	Consultant
Judith Sloan	Commissioner, Productivity Commission
Professor Jeffrey Smith	Economics Department, University of Michigan
Andrew Stoeckel	Founding Chairman, Centre for International Economics
Dr Dahle Suggett	Deputy Director, General Policy and Strategy, Department of Premier and Cabinet — NSW
Louise Sylvan	Commissioner, Productivity Commission
David Tune	Associate Secretary, Department of Prime Minister and Cabinet
Philip Weickhardt	Commissioner, Productivity Commission

Serena Wilson	Deputy Secretary, Department of Family, Housing, Community Services and Indigenous Affairs
Dr Ian Winter	Executive Director, Australian Housing and Urban Research Institute
Bernie Wonder	Head of Office, Productivity Commission
Mike Woods	Deputy Chairman, Productivity Commission

Observers

Jonathan Ayto,	Principal Advisor, New Zealand Treasury
Kristy Bogaards	Productivity Commission
Miranda Cumpston	Research Officer, Australian Cochrane Collaboration
Sue Holmes	Productivity Commission
Alan Johnston	Productivity Commission
Paul Lindwall	Productivity Commission
Lawrence McDonald	Productivity Commission
Margaret Mead	Productivity Commission
Gary Samuels	Productivity Commission