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.

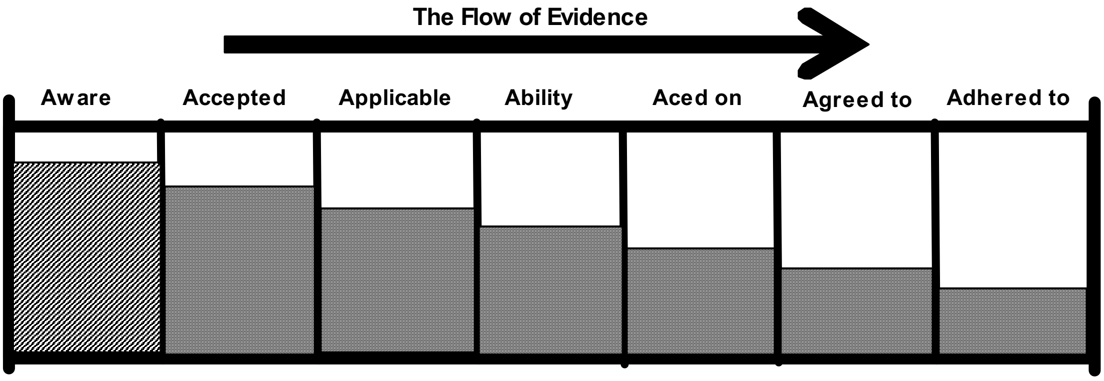
9.2 Processes of evidence-based policy need to be different for interventions with simple, complicated and complex aspects

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).

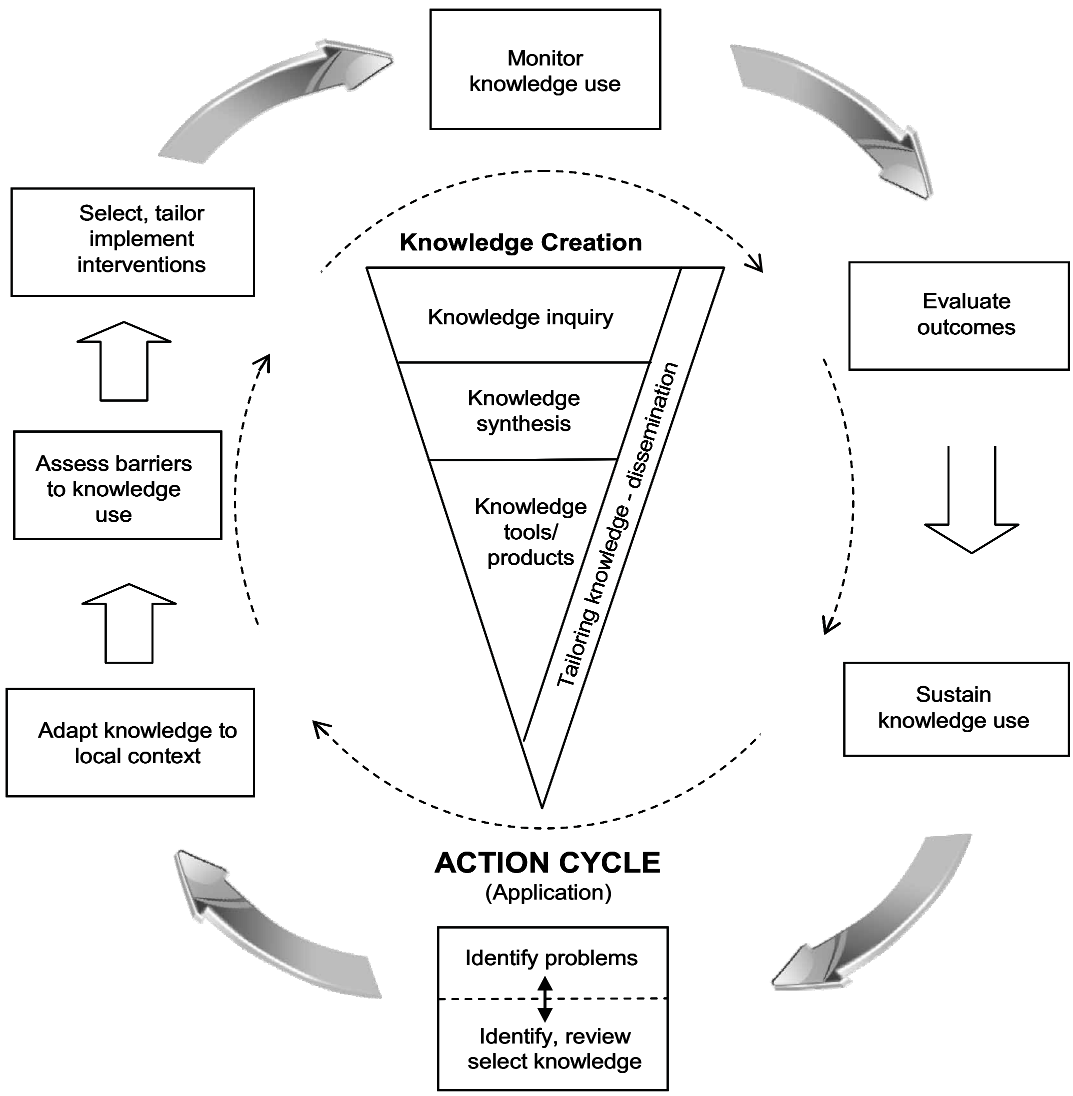
For complex interventions, which need to be constantly adaptive in response to emerging needs, challenges and opportunities, the classic linear approach to evidence-based policy presents even more challenges. There is no standardised intervention to study or replicate, but an ever-changing program. For 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 Best practices	Contingent Good practices in particular situations	Dynamic and emergent 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.