

AES 2005 International Conference keynote address¹

Reflections on evaluation practice and on the
2005 conference: some observations from a
grumpy old evaluator

a grumpy old evaluator

My thanks to the Conference Organising Committee for inviting me to give a keynote address that reflects on evaluation practice and on this year's conference. Such invitations are often expected to give rise to congratulatory statements and exhortations to go forth and do good evaluation work until we meet again. However, in this, the year of TV series such as Grumpy Old Men and Grumpy Old Women, I decided to take the opportunity to be a grumpy old evaluator, but a grumpy old evaluator that does hold much hope for the future of our practice (and just a few exhortations).

Before the conference I identified a few gripes about current practices in evaluation. During the conference I have been looking for counter examples that might show that I am being unnecessarily grumpy. I am pleased to say that I did in fact find some examples during the conference that showed that others shared my concerns and that in some cases they were actively addressing those concerns. I also saw some evidence that my concerns are in fact justified. Perhaps some of you will recognise examples (both positive and negative) from your own work environment and from sessions that you may have attended at the conference.

The list of gripes that I brought to the conference relate to the following four points:

- 'M&E'—the package deal
- the 'silver bullet' mentality of some programs
- evaluation frameworks that straight-jacket all program participants to achieve similar outcomes
- uncritical and indiscriminate use of data for evaluation purposes.

Sue Funnell



Sue Funnell, FAES, is a Past President of the AES and is Director of Performance Improvement Pty Ltd, Sydney. Email: <funn@bigpond.com>.

Monitoring and evaluation (or M&E)—the package deal

Over the last several years there has been increasing reference, in Australasia, to something called M&E (monitoring and evaluation). Notice how they come as a pair, a package deal? It is almost as if you say ‘monitoring and evaluation’ fast enough it will ease the pain of both. The conjoined use of M&E has been in place for many years in international aid programs and perhaps we are following that practice.

I see that conjoined use as an unfortunate trend. Why? Because there is a tendency to think that if you are doing one you are also doing the other. In my view, evaluation is the big loser in this partnership. What I am seeing is a lot of monitoring but not as much evaluation. There is a lot of counting and reporting against targets, targets that are often set quite arbitrarily. A preoccupation with monitoring and counting can lead to a focus on what’s easy to measure rather than what’s important to measure. It can also lead to a misplaced satisfaction that having done the counting the evaluation job has been finished. It has not! Counting is only one aspect among many.

Worse still, these counting exercises are often set within program logic models that appear to give the counts a legitimacy that they do not always deserve. It is not that I am against logic models. In fact I have been one of their greatest advocates. But I do have concerns about the way logic models are being used. More on that later in my discussion of ‘black box’ approaches.

I am exaggerating the situation a little for purposes of making a case and also because that’s what grumpy old evaluators do as they apply the lenses of hindsight. In fact, I don’t want to undervalue monitoring—it plays an important role in seeing whether we are on track, signalling areas that we might need to take a closer look at and it contributes to evaluation. We as evaluators have been wishing that monitoring data were both available and better for many years so we can hardly complain when people start collecting such data. But monitoring is not evaluation!

The more intense focus on monitoring than on evaluation represents a return, perhaps a retreat, to black box thinking. There was a period in the 80s-90s when we eschewed black box thinking and started to look a lot more inside the black box, to ask questions about attribution and to make value judgements about the performance we were measuring. Indeed logic models, especially some of those that emerged in Australia (Funnell 2000), were initially developed specifically for the purposes of ensuring that causal attribution and value judgements based on criteria and comparisons were not overlooked.

Current results frameworks (results and services, outputs and outcomes) that various Government agencies have put in place in Australia are portrayed like logic models but in practice sometimes incorporate a return to the black box thinking. They have adopted simplistic input–output–outcome (or similar terminology) pipeline approaches to program logic: approaches that over several decades we have found to be deficient. The current results frameworks provide a basis for mechanistically monitoring the various components in a logic model separately but make little attempt to assess the causal connections among the various program components (inputs, outputs, activities, outcomes—immediate, intermediate and ultimate, changes in level of need that gave rise to the program) or the relationship between those components and the complex wider context in which the program operates.

In addition, in many cases the results frameworks do not assist with the process of making value judgements about performance measured. They simply give rise to reports of quantitative measures often unaccompanied by value judgements about whether performance is good or poor, better or worse and so on, or even any analysis of meaning of the results. Targets can be used as one benchmark against which to monitor performance but often the way in which they are set is questionable for one or more of many reasons that I will not address in this paper.

In my view, current results frameworks while incorporating certain valuable features of program logic (e.g. logic modelling diagrams) have often lost sight of the need to monitor and interrelate other very important features of program logic that relate to factors (both internal program factors and external non-program factors) that influence the achievement of outcomes. It is on the basis of the analysis of the relationship between claimed outcomes on the one hand and program and non-program factors on the other hand that so-called outcomes can be truly said to be, at least to some degree, outcomes of the program(s) in question. Another concern about the focus on monitoring is that it sometimes leads to cynicism among staff—they know that monitoring data tells only part of the story. Under such circumstances, monitoring may come to be seen primarily as a reporting obligation and peripheral to the real business of managing programs. But we all know of the maxim *what gets measured (and, one might add, reported) gets done*. So what may start as a cynical gesture to meet reporting requirements, divorced from real work and the things that staff know are important but difficult to measure, gradually gets a life of its own that insidiously infiltrates the real work. So staff may well be committed to getting results and

they may understand the pitfalls of monitoring but the measurement and reporting frameworks and priorities, focused as they are on monitoring components distract them from progressing to evaluation processes to find out about real results and to explore issues of attribution. Monitoring on its own is not up to the job that evaluation can do in terms of addressing issues of causal attribution. However, taking an even longer term perspective, a positive development has been that while pre-80s monitoring used to focus on inputs and activities and occasionally outputs there is now a greater emphasis on measures of outcomes or what might at least potentially be outcomes if only we could demonstrate the causal relationship between those would-be outcomes and the program. But, reiterating an earlier point, until we can demonstrate that causal relationship, the so-called outcomes are just occurrences or trends that may or may not be a derivative wholly or in part of the program in question.

I have been encouraged by the fact that several papers at this conference while reporting on 'M&E' approaches have recognised the complexity of the measures they are dealing with and are expressing misgivings about simplistic measures. They are looking to develop evaluation techniques that will address such problems for evaluation as 'difficulties attributing cause and effect, long time frames over which interventions are likely to occur, multiple sets of activities and stakeholders and different levels at which change occurs' (Greenaway & Allen 2005).

On the basis of the papers I attended at this conference, I saw few examples of the shortcomings of the simplistic M&E models to which I have referred. In fact I would suggest that these shortcomings are well recognised by practitioners of monitoring and evaluation. They are not so readily accepted by politicians and senior management in agencies who want simple indicators of performance, even though privately they may acknowledge that such indicators tell only a pale version of the true story. Our challenge is to be responsive to the needs of politicians, senior management and the community while not being co-opted by them in our choice of evaluation approach simply for the sake of a quieter and easier life. We have an ongoing role in reminding and educating about the inadequacies of the very measures with which we as a profession tend to be identified.

The 'silver bullet' mentality of some programs

Causality is complex but a 'silver bullet' mentality is still reflected in the types of evaluation questions that we are often called upon to answer. For years evaluators have emphasised the importance of getting the evaluation questions right but I believe

we continue to fall short in this regard. Much education is required of evaluators and of those who commission evaluations alike.

Socio-ecological models have been around a long time and endeavour to represent the complexity of the contexts within which programs operate. Elaborate diagrams show everything as related to everything else and while that might be true it makes programs almost impenetrable and evaluation very difficult. At what point do evaluators enter the system? Evaluation has tended to focus on specific interventions within the system often seeing those interventions as 'silver bullets'. The environment that those bullets must penetrate is either seen as just so much noise that needs to be in some way controlled and held in suspended animation while we look at a program for a moment in time or so noisy that we can't decipher signal from noise and throw up our hands in despair, concluding that the intervention is un-evaluatable.

If only we could keep everything constant so that we could draw conclusions such as 'this program is or is not effective'! But we all know that sort of control is rarely possible and probably even undesirable. Moreover, why would we ever expect that the same set of conditions would or should be replicated in future? A focus on local adaptation and a healthy contempt for any type of authority that would straight-jacket program delivery staff into delivering what they 'know' will not work in their communities, will surely militate against replicability of program delivery. So, what policy implications could ever be drawn from assessing the impact of such a contrived program delivery scenario whose replicability in future real-life circumstances is questionable?

One approach to resolving this methodological dilemma of what to do in the event of complexity and variation (whether natural or encouraged and deliberate) is to think about asking our evaluation questions in different ways. I believe the realist approach to evaluation (Pawson & Tilley 1997) that has emerged over the last decade provides some useful insights for reshaping our evaluation questions. Some examples of its application in relation to real programs have been provided at this conference (see, for example, Alison Chetwin's paper entitled 'Realistic Evaluation of Police Practice in Reducing Burglary').

Instead of asking what works or doesn't work, a realist approach asks: 'What works for whom and under what circumstances?' It embraces the complexity and tries to understand it rather than treating it as noise. It encourages us to look at outliers and exceptions, not just for the purpose of disproving universal hypotheses but for the purpose of obtaining a better appreciation of the range of theoretical propositions that might apply—

building up conditional hypotheses and theories. So applying this approach we come to relish finding exceptions, seeing them as a source of new hypotheses and new learning, widening our horizons rather than narrowing them by using sophisticated methodological designs to eliminate the noise. We do not want to fall into the trap of narrowing the scope of our investigations to eliminate noise so much that we might fit within the somewhat uncharitable definition of an academic as ‘someone who knows more and more about less and less until she or he knows absolutely everything about nothing’ (source unknown).

One implication for evaluators is that we need to be educating those who commission or request us to conduct an evaluation about how to ask appropriate evaluation questions. We need to be encouraging them to ask questions that are posed in conditional terms, and to expect answers that are conditional, messy though they may be. The syntax of the typical terms of reference may need to change from ‘Is this program effective?’ to something like ‘Under what circumstances and with what types of people and in what ways is this program effective?’

Indeed realists such as Pawson (2002) would go further and encourage us to take our focus off the program per se and redirect our focus to the mechanisms that the program employs and perhaps to look at our findings in the light of other programs that use similar mechanisms, for example mechanisms that relate to motivating people to act by giving them an incentive – loosely referred to as carrot mechanisms; mechanisms that relate to changing people’s behaviour through deterrence—loosely referred to as sticks; or mechanisms that relate to changing people’s behaviour through providing information and education. Or, as Barry Leighton referred to in his conference presentation ‘Around the Moon and Back? Evaluation in the Canadian Federal Government’: carrots, sticks and sermons.

There are many different mechanisms that can be explored from many different theoretical positions and theories of change. Some of this has occurred in Australia. A paper I co-authored with Bryan Lenne (another past AES president) in 1990 (Funnell & Lenne 1990) looked at a typology of public sector programs and the different mechanisms underlying those types of programs. I and others have found that typology useful over the years though somewhat limited in its scope, especially as other types of programs (e.g. community capacity building) have come into prominence as part of public policy. The preconference workshop that I gave on generic program theories expands on the initial typology in the light of much work that has been undertaken over the last decade on theories of change and emerging policy directions.

Much work is still to be done on developing and testing theories of change and mechanisms and it is, I believe, an area in which we as evaluators can make valuable contributions. However, being funded to make that contribution is a challenge. Much funding of evaluation remains at the program level (in line with program budgeting, accountability and other considerations). A focus on mechanisms would require us to look across programs that share similar underlying mechanisms but which are ostensibly different in order to get a better understanding of how those mechanisms work. The title of a workshop that I gave at the AES conference in Wellington in 1996, ‘Performance-based Pay and Random Breath Testing—What do They Have in Common?’ illustrates the thinking behind a focus on mechanisms rather than programs. It probably also illustrates why it might be difficult to source funds for these efforts!

Pawson (2002), based at the Queen Mary University of London, has been able to do so. He looked at six applications of incentives in different policy contexts to see what could be learnt about how the mechanism of ‘incentives’ works under different conditions. For the purpose of the study he defined the mechanism through which incentives work in the following terms:

The incentive offers deprived subjects the wherewithal to partake in some activity beyond their normal means or outside their normal sphere of interest, which then prompts continued activity and this long term benefit to themselves or their community.

The policy contexts that he looked at included Health, Safety, Corrections, Transport, Housing, Education. The important thing to note was that he was looking for (and found) propositions that might arise from individual policy contexts but that could have wider application across other policy contexts. In the purest form, these propositions might be ‘policy-context-free’.

To expedite the sharing of learning across policy contexts, perhaps in future we could organise a conference program around mechanisms (e.g. carrots, sticks and sermons) rather than around policy contexts or evaluation issues.

Evaluation frameworks that straight-jacket all program participants to achieve similar outcomes

In measuring outcomes we often assume that the same set of outcomes will be appropriate for all participants. Performance monitoring systems are certainly that way inclined. So following on from my discussion of realist approaches, not only do

we need to be asking conditional questions about what works for whom under what circumstances but we also need to be asking questions that allow for the fact that very different outcomes may be both useful and achievable for different groups of people. This means that our outcome measures, attributes and expected levels of achievement may need to differ if we are to understand the nature and importance of the impact of a program.

The differences may be quantitative in terms of measuring self-referenced amounts of progress. Self-referenced measures assess the impact of interventions by looking at progress of recipients relative to their starting points. Some applications of goal attainment and global attainment scaling encourage this approach. Self-referenced measures of change can in principle be used to see whether programs add value. For some years there has been discussion of the use of measures of the value added by programs or interventions as an alternative to or adjunct to measures of absolute performance. Success in applying these approaches has been varied: some measures of change are fraught with difficulty and need to be applied in the light of sound methodological advice especially in relation to issues of reliability of measures and causal attribution. Conclusions need to be tempered accordingly.

In addition, many government programs these days explicitly encourage diversity of outcomes. John Owen in his keynote address referred to the devolution of authority from the centre of social systems as a key socio-political trend. Government-funded programs now often invite communities to identify outcomes that are important for *them* and to progressively develop solutions, to learn and adapt along the way. Their starting points, their needs and their solutions may vary enormously and standardised measures of outcomes would not only be difficult but quite possibly irrelevant and may be counterproductive. Yet such programs, despite their rhetoric about local solutions for local problems, often resort to simple and universally applied measures of common outcomes when reporting to government.

That is not to say that evaluators themselves necessarily adopt that approach. In fact in a session about research on attitudes and beliefs about evaluation practice it was reported that one factor that differentiated AES evaluators from AEA (American Evaluation Association) evaluators was a tendency for the former to be more inclined to participant and community-centred approaches to evaluation focusing on empowerment and cultural competence (Turner, Wolf & Toms 2005). Perhaps this preference stems from Australasian programs being less prescriptive and less standardised. The empowerment features of some of our programs

and of our approaches to evaluation may reflect broader cultural issues. Whatever the reasons, it would seem that we do have a context that should be receptive to the use of less standardised and more 'locally' responsive measures of performance. This has profound implications for our choice of methodologies, a choice that must be made in terms of fit for purpose. We should reject the notion that there are gold standard methods such as Randomised Control Trials (RCT) to which all evaluations should aspire. There is no need to apologise for not using RCTs if they are the wrong method given the situation. There is every need to apologise for trying to use RCTs when they are the wrong method simply in order to meet what we believe to be some type of context-free gold standard for evaluation methodology.

Let me give you a simple example of how standardised approaches to outcomes definition and measurement can be counterproductive. We often set up evaluation frameworks and logic model diagrams that have as the bottom rung of the ladder measures of outcomes that relate to numbers of participants and whether the numbers meet targets. We tend to see participation as a low-level outcome that we measure primarily because it is a precondition for achieving other higher level outcomes.

But what if for some individuals and communities, especially indigenous communities and disempowered marginal communities, the very act of participation represents a monumental achievement indicative of increasing trust and confidence in the wider community and a sense that they can make a difference? What if to develop this trust, manifested in a simple measure of increased participation, there had been an extended period of working with that marginalised community to develop trust? Might not participation rates, when portrayed as a low-level measure and the only measure of success, somehow devalue the developments that had occurred in the community and the work of program staff and others in that community to bring them to that point? Might not participation instead be portrayed as a relatively high-level outcome for these groups—an indicator that trust had developed? If so, what might be the lower level/initial and intermediate outcomes that should be sought and measured as leading up to increased participation?

So our program objectives, evaluation frameworks, measures of success and the things we celebrate as success need to be able to incorporate these wider and differing criteria for success for different individuals and communities in different circumstances. This is not about saying we will accept lower standards for different and, in particular, marginalised communities. This is

about better understanding of those communities, their needs and how they can be best addressed, about better appreciation of the kinds of progress that are valued by those communities and about recognising that different time frames may be required for different individuals and groups to achieve similar outcomes, given their different starting points and contexts. Case management processes have long taken this approach at the level of individuals. The challenge has been for them to aggregate data across individuals in order to draw conclusions about program outcomes and to inform program development. Chamberlain and Pressnell in their paper discussed some of these challenges and progress being made to address them.

Uncritical and indiscriminate use of data for evaluation purposes

It's good to triangulate using different types of data but we are sometimes too uncritical and lack discrimination in the way in which we use the data. We have a long tradition in Australia of using multiple methods born perhaps more out of pragmatism and necessity than out of conscious epistemological or methodological preferences. However, in the course of doing so we have perhaps not been as reflective as we might have been about the best ways to use those methods and different types of data in combination. Using multiple sources and methods interactively rather than simply additively can help with this process.

Different evaluation questions also require us to look at data in different ways. For example, realistic evaluation involves asking questions about what works for whom under what circumstances. It actively encourages us to look for what *doesn't* work under what circumstances. What bucks the general trend and what can we learn from that? Once we start to look at outliers and exceptions (or negative examples) we need to use different processes for pattern recognition and to be content with drawing different types of conclusions: conditional or contingent conclusions rather than universal conclusions.

There is currently a policy push in Australia and elsewhere for 'evidence-based practice'. On the face of it, it is difficult to disagree with such an approach. However I believe that in many cases this clarion call is more grounded in rhetoric than in really understanding the wide array of processes that need to be used, not just for collecting evidence but also for appraising and using evidence. Evidence-based practice may have had its origins in the field of public health and promoted such approaches as Randomised Control Trials, meta-analysis and narrative analysis but the concept of evidence-based practice now goes much wider

than those methodologies. Marianne Berry in her conference paper defined evidence-based approaches as follows:

Being *evidence based* means that in your practice or management you are either *using* techniques and policies that are grounded in positive tests of their effectiveness (from research, program evaluation and information about results) or that you are *gathering* information as you practise or manage in order to determine effectiveness.

She provided a useful list of 10 attributes of managers and practitioners who have evidence-based mindsets and practices.

The call for evidence-based practice is at least a tacit acknowledgement that evidence has not played as big a part in the past as some would have wished. The rhetoric around evidence-based practice does present us with an opportunity as evaluators to look more carefully at the types of evidence we use. Perhaps we can draw on other fields for insights about how to use evidence.

For many years evaluation practice tended to be dominated by the prevailing research methodologies in the fields of health and education. We can also learn some lessons from other fields. I have been heartened over the years to see increasing involvement in evaluation of people working in the natural resource management (NRM) field. Just look at our conference program this year—there are many papers relating to natural resource management, perhaps more than ever before.

People working in the field of NRM have been grappling with the fact that many of their traditions and methodologies come from the physical sciences, whereas they are working in complex ecosystems. They are looking for new ways of collecting and using evidence in valid, believable and defensible ways. Perhaps we can all learn from them, if only our different terminologies don't get in the way.

I am indebted to Helen Watts from the NSW Department of Natural Resources who gave a poster session at this conference, for introducing me to literature around Multiple Lines and Level of Evidence and Weight of Evidence approaches. Similarly the seminal papers of Claude Bennett in the 1970s from the US Department of Agriculture presented hierarchies of research evidence ranging from randomised control trials to field trials to use of anecdotes. As I have indicated earlier in this paper I am more inclined to judge methodologies in terms of whether they are fit for purpose rather than in terms of a preordained hierarchy of methods ranging from most desirable to least desirable. However, setting out these hierarchies has served the useful purpose of putting them side by side so that their relative usefulness for different situations can be appraised.

So I don't believe those approaches from other fields such as NRM have all the answers yet either but there are aspects of their work that will be applicable to various types of evaluations outside the natural resource management area. They too have partial data, paucity of data and confront difficulties in drawing conclusions about causal relationships. They too are working on processes for stepping through different types of data to reach causal conclusions. Perhaps those in other fields such as human service delivery can learn something from the logic of the processes that they apply, if not the detail of the processes.

Other fields from which we might draw include the evidentiary processes used in the judicial system. We might also learn from the way performance auditors use and weight criteria. Perhaps forensics holds some approaches for us. Given the popularity of various television series such as *CSI: Crime Scene Investigation* that relate to forensics, this may also be an approach that we can explain more readily to some of our audiences in order to demystify our methods.

I believe that if we are going to embrace evidence-based approaches we need to become much more analytical and transparent about the ways in which we use multiple sources of data to build up a picture of what we are evaluating. We need to develop and apply criteria and make judgements about the credibility of the data and how to use it (especially how to look for patterns and deviations from patterns) and about the plausibility of causal relationships.

At the same time we need to be aware that the required levels of certainty of evidence will vary depending on the purpose of our evaluation. There are many such purposes. John Owen, in his keynote address at this conference, described several different areas of evaluation practice with different purposes. The areas of practice to which he referred were: knowledge *about* impact, knowledge *for* program planning, knowledge *for* program consolidation, knowledge *for* quality control, and knowledge *for* participative action.

One of the reasons we need to become better at judging the quality of evidence is that we often don't have a lot of control over the data that we need to use. We inherit secondary data such as documents on file, end-of-project reports and so on, and often have limited capacity to collect primary data specifically for the purpose of evaluation.

Even when we design and deliver our own primary data collection instruments, such as questionnaires, we find that the quality of the responses varies enormously across respondents. For example, when we ask open-ended questions we find the answers vary from those that make ambit claims to those that both make claims and

substantiate them with evidence. We need to be able to distinguish between those answers of varying quality and sometimes to use the data in different ways depending upon its quality. Reports on the practical experiences of evaluators in appraising, using and synthesising different types of data and data of variable quality would be a useful subject for future conferences. Such reports would be about the real world of evaluation and how to come to terms with it.

We often temper our conclusions by reference to quality of the data, pointing to the limitations that it imposes on our conclusions. This is an area in which we as a profession could be more deliberate in our processes and develop some standards for doing so. The evaluation of data before using it for evaluation is, of course, one form of meta-evaluation, the subject of Valerie Caracelli's keynote address at this conference.

The evaluation of quality of evidence will be especially important as increasing focus is placed on relatively new evaluands such as whole-of-government programs, partnerships and community capacity building. New methods for evaluating success of these evaluands will evolve and we need to be able to judge the usefulness of the evidence. I have been encouraged to see several examples at this conference, not only of projects that are focused on these new evaluands but also of different methods for collecting information that is relevant to those evaluands (e.g. Keast and Brown in their paper 'The Network Approach to Evaluation: Uncovering Patterns, Possibilities and Pitfalls') rather than resorting to methods that have been developed for other types of evaluands.

I'll close on that note of encouraging us all to be engaged in a continuing process of meta-evaluation through being reflective about our practices and to do so in practical ways, as well as through engaging in the informed debate, the opportunity for which conferences such as this afford us.

Note

- 1 The paper was finalised following the conference, as it was a requirement that the paper reflect on the conference and therefore could not be finalised in advance. A few paragraphs in this paper were not included in the actual presentation.

References

- Bennett, C 1979, *Analyzing impacts of extension programs*, US Department of Agriculture, Washington DC.
- Berry, M 2005, 'Challenges in evaluation and performance measurement in children's services: experiences in the US and other countries', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Caracelli, V 2005, 'Evaluation in the eyes of the beholder: meta-evaluation as a tool of reflection', keynote address at the AES International Conference, Brisbane, 10–12 October.
- Chamberlain, A & Pressnell, M 2005, 'Evaluating outcomes: client audit and outcomes measurement', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Chetwin, A 2005, 'Realistic evaluation of police practice in reducing burglary', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Funnell, S 2000, Developing and using a program theory matrix for program evaluation and performance monitoring', in P Rogers, T Hacsí, A Petrosino & T Huebner (eds), *Program theory in evaluation: challenges and opportunities, New directions for evaluation*, No. 87, Jossey-Bass Publishers, San Francisco, California, pp. 91–101.
- Funnell, S & Lenne, B 1990, 'Clarifying program objectives for program evaluation', *Program Evaluation Bulletin*, Office of Public Management, NSW (out of print).
- Greenaway, A & Allen, W 2005, 'Evaluation and collaborative learning: critical practices for sustainable environmental management', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Keast, R & Brown, K 2005, 'The network approach to evaluation: uncovering patterns, possibilities and pitfalls', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Leighton, B 2005, 'Around the moon and back? Evaluation in the Canadian Federal Government', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Owen, J 2005, 'Turning current conceptions of evaluation inside out', keynote address at the AES International Conference, Brisbane, 10–12 October.
- Pawson, R 2002, 'Evidence based policy: the promise of 'realist synthesis'', *Evaluation*, vol. 8, no. 3, pp. 340–358.
- Pawson, R & Tilley, N 1997, *Realistic evaluation*, Sage, London.
- Turner, D, Wolf A & Toms, K 2005, 'The same only different: approaches to ethics in professional practice in Australasia and North America', paper presented at the AES International Conference, Brisbane, 10–12 October.
- Watts, H 2005, 'Evaluation design for natural resource management programs', poster session presented at the AES International Conference, Brisbane, 10–12 October.