

Program evaluation: a tale of lost opportunities

Peter Orpin, University Department of Rural Health, University of Tasmania

INTRODUCTION

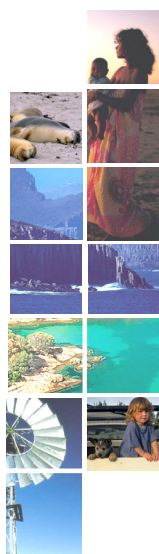
There is still much to do in building the evidence base for rural health. While this calls for a range of research approaches, the nature of the issues, the organisations likely to carry the major responsibility for building a research culture, and the current funding and policy mechanisms, all mean that many of the best opportunities will come, not through specifically funded “pure” research, but through program and policy evaluation activities. The grounded, bureaucratic and practice oriented environments in which evaluations are generally carried out complicate the search for academically credible research outputs, however, the research potential of evaluation has for too long been neglected, particularly in the health area. Of all the health areas, rural health has perhaps the most to gain from a reversal of this neglect. With appropriate care and rigour, evaluation offers rich opportunities for building the rural health evidence base.

The responsibility for this work is likely to fall almost exclusively on University Departments of Rural Health (UDRH) and Rural Clinical Schools (RCS). Since the formation of the first two UDRHs in 1996, both their focus and the external expectations placed on them, have expanded from the original brief to “... improve access for rural and remote communities to appropriate services” (1), to that of addressing the full spectrum of interlinking issues that underlie rural/remote/urban health differentials. This has entailed a mixture of strategies: professional education and support, program design and delivery, and knowledge generation. If UDRHs and RCSs are going to meet their own, and other’s expectations for building the rural health evidence base, they need to become better at exploiting the knowledge generating potential of the full range of their activities. Evaluation that is properly resourced, carefully designed, rigorously conducted and widely disseminated offers huge untapped, potential in this area.

RURAL HEALTH RESEARCH

The issue at the heart of rural health is quite straightforward and well established: on average, health status and health service access decline with distance from major metropolitan centres. Beyond this fact, our knowledge base concerning contributing factors and possible fixes is quite sparse and relatively shallow. Rural health presents a particular research challenge since the issues span the full spectrum from the purely biomedical to the social sciences, with a distinct balance towards the latter. Our understandings about the underlying factors, and their solutions, are more likely to be enhanced by studies on alternate models of service delivery, new ways of effecting and sustaining behaviour change and innovative policy developments, than by studies on disease mechanisms or new drug treatments.

This means that many of the best opportunities for knowledge generation lie outside traditional health research paradigms. While randomised control trials and major



epidemiological studies will have their roles to play, a lot of the most interesting, and potentially fruitful opportunities lie in the somewhat more “messy” areas of qualitative and evaluation research. UDRHs, in particular, operate in a no-man’s land between academia, the bureaucracies and the polity (government and non-government), and the day-to-day world of clinical practice. This means that they need to seize their research opportunities when and where they can and that, more often than not, will mean turning small “r” research opportunities into big “R” research outputs. Knowledge generated through evaluation-style research will rarely satisfy the “gold standards” used to judge biomedical research (2), but there is no reason that it cannot generate rural health knowledge that is valid, reliable and, most importantly, useful.

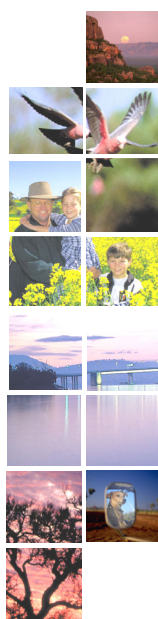
EVALUATION

Some form of evaluation is integral to almost every program, service delivery or support activity occurring in rural health. However, it is still often last in line when it comes to planning input and resource allocation and outputs are generally only narrowly disseminated. Its major processes and outputs are usually concentrated near, or even beyond, the end of programs and are primarily focused on a *post facto* measure of success or failure. This is both a cause and effect of its present marginal status. The evaluation activities are ramping up at the very point when everyone’s attention is beginning to disengage from the project, and there is an understandable reluctance in action-oriented organisations to commit too many resources to looking backwards and “navel gazing” and to asking questions which may reveal unwanted answers.

Among those who can bring a disinterested, critically reflective brief to evaluation, the academics, there is, at the very least, a sceptical view of its research worth, even, one suspects, within the “applied” environment of the UDRHs. There are sound reasons for this wariness. As Lomas (3) points out, there are few incentives for academic researchers to move into the area. It is unlikely to be a good career move: “the incentives of the promotional and status criteria in academe are a central barrier to much of the move to applied research ... [they] are often upside down” (1997:25). The same author also argues that there are considerable difficulties in making the connections necessary to translate even good research into practice change; policy makers and service deliverers speak a different language and work within very different agendas and timeframes.

There are encouraging signs of moves to address these issues at the highest level. Beginning with the Wills Report (4), the National Health and Medical Research Council (NHMRC) is revising its agenda and protocols to direct more funding and status to strategic, developmental and evaluation research. While this should eventually have some flow-down effect to all levels, there are and will continue to be, substantial research and learning opportunities presenting and missed in the day-to-day health service delivery and policy formulation activities of agencies. Nowhere is this loss more regrettable than in rural health where the issues and the environment are so conducive to an evaluation approach to research.

The following comments draw on material from the Evaluation Report on the first round of the Commonwealth’s Coordinated Care Trials (CCT) (5) and emerging experience from the Sharing Health Care Initiative (SHCI). While neither of these



programs is aimed specifically at rural health, both have considerable current presence and future promise in caring for the rural aged and chronically ill. More importantly, they highlight the dilemmas facing rural health academics seeking to mine the research potential of evaluation.

THE ISSUES

The issues to be resolved in realising the research potential in evaluation are those attending any research: study design and methodology, resourcing, and interpreting and communicating outcomes. Evaluation, however, presents particular problems in all these, especially the first. However, while evaluation research will rarely climb to the top of the Cochrane Collaboration “hit parade” this does not necessarily render the outputs invalid or useless. The paper explores thoughts around a number of design and methodology issues that particularly bedevil evaluation research and offers some suggestions on how these may be addressed.

Asking the right question(s)

All quality research rests on first and foremost on asking the right question. Most of the work of design and methodology is done at the point of framing the question(s) – or in alternate form, the hypothesis for testing. Among the many pitfalls, three stand out: asking the wrong question(s) altogether, asking a question that cannot be answered – either at all, or more commonly, within the design of the study – and starting with a question that has not been clarified and focused. All of these are particular problems for evaluation research but especially the first.

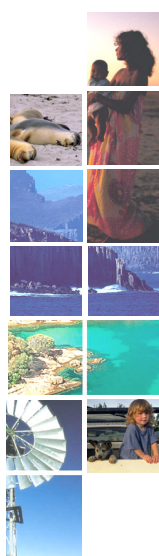
A major trap for evaluation research is seeking a simple answer to one big simple question: “Does it work?” Evaluation is most likely to produce conditional answers to small complex questions of the form: “Which ‘bits’ work for which people under which conditions?” To structure the evaluation around one or two simple hypothesis/questions is to set it up for failure, obscuring both successful elements and crucial learning.

The first round of the Australian CCTs, with a budget of over \$127 million, were the most ambitious trial of a health services model ever in this country. It consisted of nine trials based around variations on a base model of co-ordinated care for people with complex conditions. The Commonwealth as funders of the trial appear to have committed fully to a quality evaluation, philosophically and financially and to a policy of open dissemination of its findings.

The evaluation design was based around a single primary hypothesis:

That co-ordination of care of people with multiple service needs, where care is assessed through individual care plans and funds pooled from existing Commonwealth, State and joint programs, would result in **improved individual client health and well-being within existing resources** [bolding not in original] (6 :9)

The question, in effect, is twofold: will co-ordinated care models improve individual client well-being and can it be done without costing more money? The six secondary hypotheses are all designed to explore the conditions or limitations qualifying the answer to the primary question. With the benefit of hindsight it is easy to see that the



framing of the primary hypothesis virtually guaranteed failure. It posed a big, simple question to which it was never going to be possible to give a big, simple answer. The co-ordinated care model, while seemingly straightforward in conceptualisation, was always going to be exceedingly complex in any practical application, let alone nine different applications each with its own unique model details.

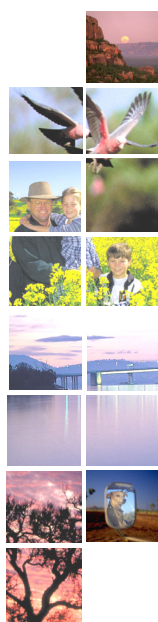
“Does it work?” is the quintessential experimental trial question based primarily on the deductive side of the “wheel of science” (7). It is usually only satisfactorily answerable in the strictly bounded conditions typical of the controlled trial, and then really only for those conditions. While such a question will always sit at the back of any evaluation, making this the primary explicit question/hypothesis is likely to be counterproductive. Evaluations almost always involve looking at complex contextual phenomena under conditions which include multiple uncontrolled, and largely uncontrollable variables. This makes them a very different prospect to, say, testing a drug in a controlled trial, or even a clinical protocol under controlled conditions. Evaluation is better suited to a more inductive model; that is, one which focuses more on building theoretical understanding from a series of empirical observations. Of course, in practice, deduction and induction are always linked in a circular process but the emphasis on one methodology or the other can have major implications. An evaluation which seeks to incrementally build knowledge through a series of questions around its multiple components is likely to yield better learning outcomes than one that treats the phenomenon under study as a single entity testable with one or two simple encompassing hypotheses. This is exactly the task we face at present in rural health.

A similar problem faces evaluators for the Commonwealth Sharing Health Care Initiative (SHCI), a chronic disease self-management program with seven demonstration projects, all driven by the same basic thesis but varying widely in geographical setting (well over half of the sites are rural or regional), target clientele and the intervention model. Again, the simple encompassing question of “Does self-management work?”, hides within it a multitude of unresolved issues and problematic questions: What exactly constitutes self-management?; Work for whom and in what ways?; Which strategies work for which people under which conditions? It is the accumulated body of knowledge building from partial answers to these multiple questions which holds the greatest promise for informing future policy and practice, not a simple “works” or “doesn’t work”.

This highlights one of the major dilemmas facing those seeking to practice evaluation as research. Program funders and policy makers, at least to some extent, require simple answers to big questions to satisfy political and fiduciary responsibilities. It is beyond the scope of the present paper to explore these issues except to note that finding a way to bridge that particular gulf is vital if we wish to move to “evidence-based” policy making and service design in rural health.

Statistics and power

Statistical competence and sophistication is seen as a primary marker of research competence, particularly in the health field. This is a highly specialised and esoteric field, although one that has become increasingly accessible to the barely competent through the proliferation of affordable statistical software. It is also one of the aspects of research that is most likely to scare away otherwise highly competent policy makers



and service designers/ delivers from engaging in high quality research oriented evaluation. It is perhaps the aspect most likely to discourage reflective rural practitioners from embarking on potentially valuable formal research.

The message that needs to get out to such people is that good quality simple statistics are not only accessible to most with basic mathematical competence, and sufficient for most evaluation purposes, but also highly preferable to poor quality or badly done complex statistics. Many, if not most of the important questions in program and policy evaluations can be answered with simple descriptive statistics like means and cross-tabulations. Where inferential statistics are used, power and significance issues will often be outweighed by size of effect and generalisability issues. In any case, the multiple variables involved will often mean that the subject numbers in any given cell in an analysis will be too small to reach the sorts of confidence limits usually sought for inferential statistics and the generalisability beyond the actual case will be problematic. The argument is for simple statistics not sloppy statistics.

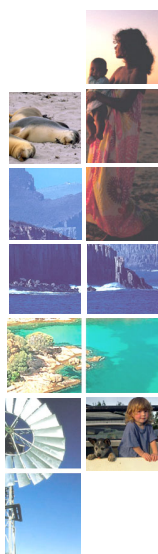
The instruments

Evaluation faces particular problems in the validation of the instruments used for measurement. Quality research requires quality instruments yet the process of psychometric validation of social science survey instruments is a highly specialised and resource hungry exercise. The financial and time constraints under which most evaluation functions, preclude the development and proper validation of instruments yet the chances are fairly slim of finding a fully validated “off the shelf” instrument that exactly suits the requirements of a given evaluation. It is little use having a valid instrument if it is not fully suited to capturing the phenomena under study. The co-ordinated care trial’s use of the Short Form-36 Health Survey appears to be in this category; although well validated and normalised for the Australian population, it appears to have been unable to capture important changes because of significant floor and ceiling effects.

The approach taken by the Sharing Health Care National Evaluation appears to represent one workable compromise: the careful adaptation of an existing validated instrument. While this may bring understandable criticisms about invalidation from the methodological purist it appears to represent an advance over the alternatives of an unsuitable but valid instrument or a purpose designed but completely un-validated instrument. For a large percentage of evaluations, there will be little choice but to use the latter strategy. There is no simple answer to the question of the implications of this for results, however, there is no more justification in summarily labelling any resultant such findings unscientific or invalid than there is for assuming a validated instrument guarantees scientific merit and validity. Careful and thoughtful design and some limited piloting can go a long way towards addressing validity concerns.

Controls

The randomised control trial (RCT) is the undisputed gold standard of health research, endorsed by both the NHMRC (2) and the Cochrane Collaboration. This form, however, is simply not feasible for most evaluation. Apart from the substantial resource implications, the effects of interventions can be difficult to fully define and quarantine, both practically and ethically. There is little doubt that controls, particularly randomised controls, provide increased internal and external validity,



however, this value is lost unless both controls and subjects are held under very tight experimental conditions. As shown in the CCTs, it can be difficult to sustain a valid control cohort over a long program. Also, the much quoted *levels of evidence* hierarchy (2) in which randomised control trials dominate, is just one part along with statistical precision, size of effect and relevance to practice of a broader schematic for assessing research evidence. Reliable and valid baseline measures – too often missing in evaluation – can provide a methodologically sound basis for testing interventions as long as the methodology and trial conditions are fully and openly documented. The SHCI evaluators have determined that the likely methodological gains simply do not justify the difficulties and likely costs of incorporating controls. The crucial issue is that the decision to have, or not have, controls be a methodologically considered one.

Interpretation issues – generalisability

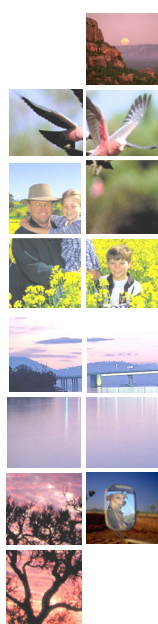
Research that produces no knowledge or insight generalisable beyond the specific study case has limited value. Likewise there can be little value in the evaluation that does not inform subsequent policy or programs. Too often, evaluation is conceived as a *post facto* justification of policy or ideas. This encourages a success/failure approach to the evaluation and sets up the conditions for many of the more common pitfalls; the wrong question posed for the wrong reasons – designed to confirm and justify rather than test and learn – inadequate resourcing, failure to integrate the evaluation into the program, biased analysis and interpretation and inadequate dissemination. A “failure” may well be a better generator of useful knowledge than a success, but considerably less likely to see that potential mined and disseminated.

The learning generated from even rigorously scientific evaluation will always have limited predictive utility. Because human behaviour is basically non-linear (8) the social sciences, unlike the physical sciences, can rarely elucidate relatively stable “laws” of behaviour. Evaluation research, therefore, must seek to build, piece by piece, some general propositions about a complex of actors and interactions that will allow at least a better understanding of other instances of a similar, but not identical, type. The crucial factors are not the sample numbers or the strict methodological control within a single carefully controlled example of the type, but developing and testing general propositions across a wide diversity of such types to find where and when the propositions hold or break down. A major advantage of structuring evaluation as research is that it is more likely to result in the formal theorisation and recording of the sorts of knowledge that reflective bureaucrats and practitioners already informally accumulate and apply in their work.

Trial conditions

The RCT draws most of its methodological rigour from the trial model in which all the variables not directly measured are either controlled or eliminated by strictly quarantining the phenomenon under study from external influences. An important part of the variable control process is being able to strictly limit the number of interventions tested at one time and to hold the intervention and the conditions constant for the duration of the study. Evaluation research generally presents difficulties in meeting these conditions

There is less opportunity in evaluation research to limit the number of intervention dealt with at one time. A program, even when defined as a single intervention, bears



little relationship to the sort of single intervention met within biomedical research. It inevitably consists of multiple elements in a complex of fluid relationships with borders that meld into the surrounding environment. This makes quarantining the phenomenon under study very difficult, if not impossible. When variables cannot be controlled or eliminated, rigour can be safeguarded to a large extent by, wherever possible, identifying and accounting for as many of the uncontrolled variable as possible in the analysis and interpretation. This means detailed measurement and documentation of the program or policy environment

Holding the intervention constant, while methodologically desirable, can also be in conflict with the purpose and philosophical underpinning of the program or policy being evaluated or even with the evaluation itself. If the purpose of good evaluation is, at least in part, to test programs or policies in order to refine them, then it can be counterproductive and perverse, to proceed with a given element in the face of clear evidence that it is flawed merely to satisfy methodological concerns. In this, as in most of the issues discussed above, the need is to find middle ground.

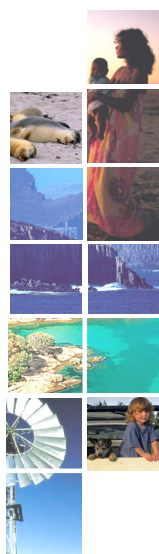
Dissemination issues

The first round CCT national evaluation report runs to 1558 densely printed pages bound into four volumes (5, 6, 9, 10). Similar volumes must fill the bookshelves of almost every government, and many non-government organisation offices across the country. It can be predicted that the SHCI will produce an equally impressive number of words and numbers. It must be said that with the advent of the World Wide Web this material is considerably more accessible than it has been in the past but there is a long way to go to match the accessibility of most academic research. For those who find their way to the two main CCT reports: the Final Technical National Evaluation Report and its associated appendices, there is a huge volume of data; although it is doubtful that careful study would yield much more insight than that already revealed in the 144 page summary document.

The Recollections of an Evaluation volume (9) provides richer insights, however, given the size and complexity of the program, this is a disappointing return in term of research knowledge. A search of *Medline* and *ProQuest Health and Medical Complete* databases using “Coordinated” “Care” “Trial” as keywords yielded just five responses. Much of the learning potential of these trials does not appear to have been realised, either because the data was not collected, the analysis not done or results not disseminated in forms and forums where it will be easily accessible to interested parties. It is difficult to establish just what percentage of evaluations make it into the peer-reviewed and professional literature and forums but it is undoubtedly quite small.

CONCLUDING COMMENTS

If the CCT reports represent the quality end of the evaluation spectrum, the wasted opportunity and resources represented by the myriad evaluations taking place around the country every day, must be huge. This loss is particularly regrettable in rural health where the nature and extent of the present activity bubble presents such a rich, but probably quite short, window of opportunity. The reasons behind these lost opportunities re-enforce each other: the failure by both evaluators and funders to



recognise the research potential, few incentives to exploit these even when they are recognised, the wrong questions asked for the wrong reasons, flawed methodologies, inadequate resourcing at all stages and restrictions on the dissemination of outcomes.

Building quality research into evaluation requires compromise at many levels. While the difficult task of balancing compromise and quality is an integral part of everyday life for those working in policy development and service delivery, it remains a dirty word in academe. The two concepts are not mutually exclusive but they do take some juggling.

The formulation of evaluation hypotheses/questions will always be driven to a large extent by those commissioning the evaluation but this does not necessarily preclude the framing of some questions or the addition of others to serve research and wider knowledge building aims – although this may require considerable ingenuity in working around resource restraints and commissioner anxieties about unexpected or unwelcome outcomes. Resources and design restraints will often preclude the use of sophisticated statistical analysis and yield less than ideal confidence intervals but simple statistics are not synonymous with sloppy statistics and statistical power is not the only criteria for utility and validity, especially when dealing with complex real world phenomena.

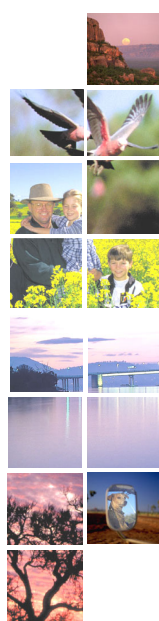
It would be difficult to sustain an argument against psychometric evaluation of instruments and the use of controls wherever possible, but when resource and design issues rule out both, thoughtful design, careful practice and cautious interpretation of results can go a long way towards maintaining rigour and validity. Caution is a sound guiding principle in making generalisability claims for any form of research, including RCTs, but is particularly apposite in any research in which human behaviour is a significant element. Better to present the case, float the possibilities, and allow time for the evidence to build from multiple sources.

In the end, the greatest safeguard of rigour and validity in evaluation research is transparency and peer review. Perhaps the surest guarantee of adequate testing of all aspects of the evaluation process and outcomes, is making it accessible to the widest possible peer audience.

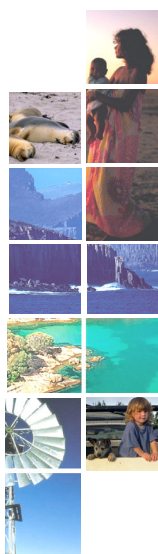
Building the evidence base for rural health, will certainly require increased dedicated research funding but it will also mean taking every opportunity to exploit the research potential of all we do as rural health organisations or practitioners and being bold and open in sharing our experience.

BIBLIOGRAPHY

1. University Department of Rural Health Tasmania. Plan for the First Triennium 1998/99-2000/01. Launceston: University of Tasmania, 1998.
2. National Health and Medical Research Council. How to review the evidence: systematic identification and review of scientific literature. Canberra: National Health and Medical Research Council, 1999.
3. Lomas J. Beyond the sound of one hand clapping: A discussion document on improving health research dissemination and uptake. Sydney: University of Sydney, 1997.



4. National Health and Medical Research Council. The Virtuous Cycle: Working together for health and medical research. Health and Medical Research Strategic Review Summary. Canberra: Commonwealth of Australia, 1998.
5. Commonwealth Department of Health and Aged Care. The Australian Coordinated Care Trials: Final Technical National Evaluation Report on the First Round of Trials. Canberra: Commonwealth Department of Health and Aged Care, 2001.
6. Commonwealth Department of Health and Aged Care. The Australian Coordinated Care Trials: Summary of the Final Technical National Evaluation Report on the First Round of Trials. Canberra: Commonwealth Department of Health and Aged Care, 2001.
7. Babbie E. The Practice of Social Research Sixth Edition. Belmont, California: Wadsworth Publishing Company, 1992.
8. Reisch G. Chaos, History and Narrative. History and Theory 1991;30:1-20.
9. Commonwealth Department of Health and Aged Care. The Australian Coordinated Care Trials: Recollections of an Evaluation. Canberra: Commonwealth Department of Health and Aged Care, 2001.
10. Commonwealth Department of Health and Aged Care. The Australian Coordinated Care Trials: Final Technical National Evaluation Report on the First Round of Trials – Appendices. Canberra: Commonwealth Department of Health and Aged Care, 2001.



PRESENTER

Peter Orpin is Senior Research Fellow and Research and Evaluation Program Area Co-ordinator in the University Department of Rural Health, Tasmania. Peter has extensive experience in both the biomedical and social sciences. After a career as a medical laboratory scientist he moved on to pursue studies in psychology and sociology and completed a PhD in sociology in 2001. This rather eclectic training has fuelled a personal passion for bridging the gulf that presently appears to exist between biomedical and qualitative approaches to health research.

