Learning Communities

INTERNATIONAL JOURNAL OF LEARNING IN SOCIAL CONTEXTS

SPECIAL ISSUE: EVALUATION

Number 14
SEPTEMBER 2014

Editor

Ruth Wallace

Guest Editor

Emma Williams

Publisher

www.cdu.edu.au/the-northern-institute/lcj
Northern Institute
Faculty of Law, Education, Business and Arts
CHARLES DARWIN UNIVERSITY
Darwin, Northern Territory 0909 AUSTRALIA
CRICOS Provider No. 00300K | RTO Provider No. 0373
ISSN 1329-1440

Design and Print

The Exhibitionist
Phone: 08 8981 3166

Artwork Credit

Artist: Marina Strocchi
Title: Mystery Train 2007
Etching (workshop proof edn 60)
24.5 x 20cm [image]
Collaborators: Leon Stainer & Dian Darmansjah
Printer: Dian Darmansjah
Gifted by the artist & Northern Editions, 2007
Charles Darwin University Art Collection
CDU 1585

Learning Communities

INTERNATIONAL JOURNAL OF LEARNING IN SOCIAL CONTEXTS

SPECIAL ISSUE: EVALUATION

CHARLES DARWIN UNIVERSITY | Northern Institute

Learning Communities

INTERNATIONAL JOURNAL OF LEARNING IN SOCIAL CONTEXTS

SPECIAL ISSUE: EVALUATION

Number 14 – September 2014

CONTENTS

Foreword

Learning Communities has a tradition of special issues that reflect the unique contexts in which we work at the Northern Institute. I am pleased to publish this special ‘Evaluation’ issue of the Learning Communities journal, released in conjunction with the Australasian Evaluation Society holding its 2014 International Evaluation Conference at Charles Darwin University.

The focus on northern Australia emphasises the need for sound policy development, investment and evaluation to ensure we learn from the past and optimum return on investment – including the investment of time energy and money. The Northern Institute recognises the importance of evaluation to understand the impact of programs over time. Our ‘Evaluation for Northern Contexts’ area, has a mandate to work on theory development, stakeholder engagement and capacity building to support policy makers, agencies and communities working in social policy in the region. In addition the team undertake evaluation projects, linking theory and practice. The Northern Australia agenda, with other major regional policies and strategies, reinforce the need for ongoing work in all of these areas whether the topic is education, employment, community safety or Indigenous empowerment.

Just as the Northern Institute has committed to working with local stakeholders to increase evaluative capacity in the Northern Territory, we have also committed to growing our own capacity through linkages with other research institutions and experts. Just in this issue, contributors come from many universities and research agencies: the Tangentyere Research Hub, Griffith University, the University of Southern Queensland, Flinders University, the University College in London, the Australian Institute of Criminology, James Cook University and the University of Melbourne, as well as the Northern Institute at Charles Darwin University.

I am aware that in other countries there are strong cross-institutional evaluation relationships; Canada, for example, has a consortium of universities involved in evaluation education. I would like to think that this issue of the Learning Communities journal could serve as a step in developing similar, improved links between Australasian universities engaged in evaluation research – that is, not just conducting evaluation projects, but engaging in work on evaluative theories and approaches to improve the ability of evaluation to inform policy and practice.

Ruth Wallace
Director, Northern Institute

Introduction: ‘Evaluation for Northern Contexts’

This special ‘Evaluation’ issue of the Learning Communities journal was written to mark both the Australian Evaluation Society (AES) 2014 International Evaluation Conference being held on the Charles Darwin University campus, and the new ‘Evaluation for Northern Contexts’ area of the Northern Institute. Development of the issue offered the Institute an opportunity to strengthen existing relationships and build new ones in key areas of evaluation practice and theory.

Aboriginal issues are central to much of the work at the Institute, including evaluation. Ensuring that Aboriginal community members are co-investigators rather than ‘subjects’ has been a long term focus of the Institute. That theme is reflected here, particularly in the offering from Tangentyere Hub researchers and evaluators.

A more recent development is the Institute’s commitment to realist evaluation. In a region where new interventions are regularly rolled out and/or scaled up, a context responsive method of evaluation that does not simply ask ‘what works?’ but rather ‘what works in which conditions for whom (and how)?’ is well suited to the Institute’s focus on people, policy and place.

Community safety is another policy focus at the Institute. In 2013, the Institute signed a Memorandum of Understanding with the Australian Institute of Criminology (AIC), building capacity to address community safety, crime and justice issues through policy-informing research and evaluation. A number of joint projects have already been initiated, and the articles in this issue further demonstrate the value of this relationship.

Finally, the Institute has an ongoing commitment to build regional evaluative capacity. As well as developing internal capacity through the development of a network of adjunct Fellows, the Institute has offered workshops to local evaluation practitioners and stakeholders in Darwin and Alice Springs. New partnerships have been formed, such as one with the Centre for Program Evaluation (CPE) in the Melbourne Graduate School of Education within the University of Melbourne. We are pleased to include a contribution from CPE authors.

For the journal, invitations were issued to existing partners such as Australian Institute of Criminology researchers and members of the Australasian Evaluation Society’s Special Interest Group in Realist Evaluation and Realist Synthesis, as well as to Institute staff, adjuncts and partners. We made it clear that we did not want accounts of evaluation projects; those can be presented in reports or in other venues such as criminology or education journals. Instead, we asked potential authors to reflect on their recent work and identify findings that could inform evaluation methodology or theory. After some weeks of consultation, a number of potential papers were identified. We supplemented these with invitations to evaluators working in areas that seemed to be of special interest and addressing issues not often seen in the literature, such as realist design.

For the publication to be ready in time for the AES conference in September, the timeline for paper development and review had to be compressed. A number of authors were unfortunately unable to meet the tight deadlines; their manuscripts may be published in a future issue and/or appear in the planned 'Short Paper' series in the 'Evaluation for Northern Contexts' section of the new Institute website, due to be launched shortly. However, we are very pleased to provide the range of papers presented in this special 'Evaluation' issue.

The first paper, by Nick Tilley and other authors from Griffith University, University College in London, the Lucy Faithfull Foundation and the Australian Government, combines three strands that run through the journal: evaluation in Indigenous contexts, evaluation on a community safety topic, and a realist approach. It is also exciting to be able to provide a paper dealing with a realist project at the design stage.

The next three papers also focus on realist approaches, each from a different perspective.

The paper by Hannah Jolly and Lesley Jolly discusses how they used curriculum theory to distinguish context from mechanism in a cascade of contexts and mechanisms in a curriculum change evaluation. Providing fine-grained analysis of some aspects of their evaluation to demonstrate their point, they demonstrate how evaluators can employ theory to distinguish context from mechanism in a principled way.

The Hawkins paper examines the potential for the inclusion of experimentation in realist evaluation approaches, arguing for the adoption of experimental approaches to test realist theory as well as to estimate effect sizes. Hawkin's approach requires that program theory, not the program, is the unit of analysis and requires context to be brought into the effect size equation.

The Pointing paper is also in the realist vein, but focuses on the rapid realist synthesis approach rather than realist evaluation. The paper recounts how the author first encountered realist approaches, and details his experience in using realist synthesis methods to identify theoretical bases of the impact of closed-circuit television (CCTV) to reduce alcohol-related assault, and to design an evaluation of homelessness and alcohol harm reduction in a northern Australian city.

The focus of Pointing's paper on CCTV and alcohol-related harm reduction leads naturally into the next set of papers, which all stemmed from projects dealing with crime, justice and safety issues. However, for this journal, authors reflected on their criminological evaluation experience to identify issues relevant to evaluators in many disciplines.

Brown draws out lessons from his Systematic Social Observation (SSO) experience, discussing cases in the United Kingdom where he used this method in evaluating the impact of changes in alcohol licensing laws and in evaluating the impact of environmental clean-up campaigns. Brown discusses SSO design issues and notes some problems experienced by fieldworkers using the method, but also indicates the benefits of SSO in providing data that may not be secured through other research methods.

Boxall describes 'pragmatic' evaluation, and techniques for keeping an evaluation on track in the face of the many difficulties which often arise. Using an evaluation of the Family Group Conferencing pilot project (NSW) to illustrate her points, Boxall demonstrates how reflexive, adaptive and pragmatic approaches to evaluation that involve project stakeholders in the development of evaluation designs and research methods can help to keep evaluations on track when original plans become unfeasible for reasons beyond the control of the evaluators.

Morgan's paper focuses on the issue of collaboration with internal stakeholders when conducting independent evaluations using rigorous scientific methods, such as quasi-experimental designs. Using evaluations of multiple programs designed to prevent and reduce crime, and to respond to the needs of vulnerable populations in court settings, Morgan highlights the benefits but also the challenges in working collaboratively with program managers, staff and participants.

Willis and Tomison use an evaluation of a multi-faceted juvenile justice project in Thailand to demonstrate how applying a Participatory Action Research approach to program logic development can ensure shared understandings of evaluation in a cross-cultural, cross-language context. The authors describe how they used the approach when working with the Thailand Department of Juvenile Justice and Observation in the evaluation of a complex project aimed at improving outcomes for young offenders, and reflect on the potential application of the approach to other cross-cultural situations.

In the Northern Institute context, the most frequent cross-cultural encounters occur with Aboriginal community members. The Institute is privileged to be able to present the reflections of researchers/evaluators at the Tangentyere Research Hub on what evaluation means to them. The authors have also prepared two diagrams, one showing how they see external parties conducting evaluations and the other, in contrast, showing how they conduct evaluations within their community in a culturally grounded way. The authors express their hope that they might assist evaluators working with Aboriginal people to think differently and perhaps approach future evaluation in cultural communities differently.

Christie and Campbell, with substantial experience in Northern Territory Aboriginal communities, reflect upon Aboriginal contributions to the theory and practice of evaluation. They provide an example from an evaluation of Housing Reference Groups where Aboriginal voices did appear to influence government policy and practice, although this did not appear to occur until long after the evaluation report was submitted.

Cairns and McLaren remind us that there are other populations that require evaluators to reflect on their practice and ethics. They use a case study of an evaluation with Deaf and hard of hearing students to show how an inclusive, participatory evaluation methodological framework was implemented. The authors point out that adherence to ethical standards and principles was a primary consideration in the evaluation.

Williams also looks at evaluation ethics, but this time from the perspective of institutional ethics review. With a special emphasis on informed consent, Williams contrasts the issues of ethics in evaluation with those in bio-medical and social science research. She identifies potential opportunities for improved practice in response to a recent document issued by the Australian National Health and Medical Research Council on ethical considerations in evaluation.

Finally, Guenther examines the annual reports of a number of Australian state and territory education departments to identify how the performance measures in them, typically labelled as targets and outcomes, reflect stated and/or actual education priorities. He uses two case studies to demonstrate how broader aims for education reflected in foundational documents – sometimes considered too impractical or costly to measure – can be addressed through evaluation.

We hope that you enjoy these articles. If you would like to comment on any of them or provide feedback, please feel free to address comments to emma.williams@cdu.edu.au. We are also looking to create a venue within the Northern Institute website where more public dialogue on the points raised in the articles can be presented.

One final point — we thank the authors and the many dedicated reviewers who put in so many hours and such care into this issue — if you find the occasional rough edge in the journal, please attribute it to the ambitious timeframe rather than lack of authorial or editorial rigour.

Emma Williams
Guest Editor

On being realistic about reducing the prevalence and impacts of youth sexual violence and abuse in two Australian Indigenous communities

Nick Tilley

University College, London

n.tilley@ucl.ac.uk

Susan Rayment-McHugh

Griffith University

Stephen Smallbone

Griffith University

Martina Wardell

Department of the Prime Minister and Cabinet

Dimity Smith

Griffith University

Troy Allard

Griffith University

Richard Wortley

University College, London

Donald Findlater

Lucy Faithfull Foundation

Anna Stewart

Griffith University

Ross Homel

Griffith University

Keywords: realist evaluation, realistic evaluation, realist design, youth sexual violence and abuse, community safety evaluation, evaluation in Australian Aboriginal and Torres Strait Islander communities, Griffith Youth Forensic Service Neighbourhoods Project

Abstract

Social interventions, like medical ones, can produce negative as well as positive outcomes. It is important for policy and practice to learn what works, what doesn't work, and what produces unintended effects, for whom and in what contexts. This is the task of realist evaluation. The formulation and evaluation of programs aiming to deal with problems in Australian Aboriginal and Torres Strait Islander communities face a number of practical, conceptual and methodological problems. Here, realist methods for the design and evaluation of promising programs from which transferable lessons can be derived are discussed in the context of an initiative aiming to reduce the prevalence and impacts of youth sexual violence and abuse. Tentative conclusions are drawn for what this might mean for programs targeting similar problems elsewhere.

Introduction and background

Griffith Youth Forensic Service Neighbourhoods Project (GYFS-NP) aims to reduce the prevalence and impacts of youth sexual violence and abuse (YSVA) in two Australian communities – a remote Aboriginal community, and a culturally diverse suburban precinct within a regional city. It aims to do so by engaging closely with the local communities, measuring various aspects of the problem in a variety of ways, and developing, implementing and evaluating a suite of locally-tailored interventions. Because of the context-specific nature of the targeted problems, these particular interventions may not be directly transferrable to other sites or to other related social problems. At the same time, the project aims to develop and test an over-arching prevention model that is transferable to a wide variety of places, problems, and contexts.

YSVA has emerged as a hitherto only partially recognised problem in some Aboriginal and Torres Strait Islander (Indigenous)[1] communities. It is experienced in remote communities as well as within communities embedded in towns and cities. It takes the form of rape, prostitution, 'rough sex' in which girls appear to be resigned to being treated as objects of sexual satisfaction to boys, inculcation of children into highly sexualised peer groups, sexual teasing, self-abuse, and inappropriate touching.

The project whose evaluation is the concern of this paper began with an exploration of a range of data on the extent of YSVA in the project's target areas, and particularly amongst the Indigenous people residing in these areas. Data were assembled going back for as long as a decade. Rates of pregnancy amongst young girls, sexually transmitted infections, and recorded sex offences were many times higher within the project's target areas and amongst Indigenous residents than amongst the general population. Systematic observations were also made at key public locations within the project areas where it was thought that antisocial behaviour might be taking place, when it was deemed safe enough, not in the expectation that YSVA would be seen directly but to gauge the number and basic attributes of those present, what they were doing, and the presence or otherwise of formal and informal guardians. The project is led by psychologists based at Griffith University whose work has specialised in the assessment and treatment of court-referred youth sexual offenders. Work with these youth strongly suggested that the incidents coming to the attention of the authorities, particularly in the two target communities, represent only a small fraction of the total number of incidents occurring. Indeed it suggested that some forms of YSVA might be endemic in the two locations of concern (Smallbone & Rayment-McHugh, 2013).

Following an earlier study that aimed to empirically examine the scope, dimensions and dynamics of YSVA in these two communities (Smallbone, Rayment-McHugh, & Smith, 2013a), Australian Government funding was secured to support a three-year program that would devise, deliver and evaluate strategies to reduce the extent and

[1] We are mindful of sensitivities about terms used to refer to Australian Aboriginal and Torres Strait Islander peoples. We have used the term 'Indigenous' in this paper because it accords with references to policies and other documents.

impacts of the problem. It was agreed that a realist approach would be taken to the evaluation. Project funding provides for two full-time clinicians, a full-time researcher, an international advisory group, travel and associated expenses. It does not provide for the costs of most of the interventions that are likely to be formulated during the project. Though some of the interventions will be provided directly by the two full-time clinicians, others will depend on third party agreement and action.

Following public health principles, the project adopts a broadly preventative approach This includes primary prevention – measures aimed at the general population and background conditions that enable or foster YSVA; secondary prevention – measures targeting the types of setting that are known to foster YSVA and the groups known to be at high risk of YSVA offending or victimisation; and tertiary prevention – measures targeting existing sites for YSVA and known perpetrators or victims, to try to prevent further incidents. Prior work of the clinicians leading the project had focused on the tertiary prevention of continued sexual violence and abuse by known youth offenders (Smallbone & Rayment-McHugh, 2013; Smallbone, Rayment-McHugh, Crissman, & Shumack, 2008; Smallbone, Rayment-McHugh, & Smith, 2013b). However it soon became clear that even if these clinical interventions successfully prevented further offending among individual referred youth, this would have a very limited overall effect on the prevalence and impacts of YSVA in these communities. Serious concerns remained that other young people, exposed to the same situational, family, peer, organisational and neighbourhood risk factors, would go on to offend or be victimised unless their exposure to these risk factors was reduced or prevented.

A ‘Haddon matrix’ is being used to capture the different types of prevention and targets of intervention. Table 1 shows the matrix and provides some illustrative examples of the types of measures that might be considered (see Smallbone, Marshall, & Wortley, 2008, for a fuller discussion of the matrix).

Table 1: **Haddon matrix for YSVA with examples of measures that might be considered**

	Primary Prevention	***Secondary Prevention***	***Tertiary Prevention***
Offenders/ Potential Offenders	• Reduce exposure to known developmental risk factors • Introduce school-based sexual ethics programs	• Re-engage school-disengaged youth • Therapeutic services, particularly for boys exposed to known risk factors	• Incapacitate most prolific/ serious/ influential youth offenders • Expand offender rehabilitation services (YSVA specific + serious delinquency)
Victims/ Potential Victims	• Reduce exposure to known developmental risk factors • School-based resilience-building programs	• Interventions with at-risk girls (personal safety, guardianship, sex education)	• Improve reach and effectiveness of victim support and treatment services • ‘Cocoon’ the most vulnerable victims
Situations	• Create safe, attractive places for children and youth • Increase legitimate use of public spaces	• Improve natural surveillance in ‘at risk’ places • Increase planned/ legitimate/ supervised activities in ‘at risk’	• Improve targeting of police patrols (hot spots; hot times) • Disrupt problem youth group activities/ movements
Community	• Mobilise and focus community concerns about YSVA • Parenting programs tailored for the local context	• Responsible bystander training (youth and adults) • Problem-solving with community leaders to reduce barriers to community guardianship	• Mobilise and focus community concerns about YSVA • Community engagement focused on improving extended guardianship

The broad methodology for the development of GYFS-NP has been one of problem-solving. This involves breaking problems down into their constituent parts and then working out what might most plausibly be put in place to deal with them. The process is sometimes described in SARA terms (Eck & Spelman, 1987). The 'S' refers to 'scanning', which comprises a broad effort to assemble evidence on the nature and extent of the problem. The pre-project work involved scanning. The first 'A' refers to 'analysis', where hypotheses relating to the conditions giving rise to the problem are formulated and tested. The 'R' refers to 'response', where strategies and tactics to address the problem based on the analysis are devised and plans made for their implementation. The responses will fit into the Haddon matrix. The second 'A' refers to 'assessment', where the processes involved in practice and the outcomes ultimately achieved are identified by systematic research and evaluation so that transferable lessons can be learned. In practice, although the S→A→R→A sequence looks linear the processes are iterative where work at later stages lead to examination and reworking of earlier ones (Sidebottom & Tilley, 2011).

Local Implementation Groups (LIGs) comprising local community representatives and key local service providers have been established in the project communities to advise on the appropriateness and implementation of the suggested responses that emerge from the scanning and analysis. Negotiations are also being undertaken with third party partner agencies and organisations, whose participation will be necessary to deliver many of the measures that are likely to be suggested.

As this paper is being written scanning has been undertaken and analysis with a view to the formulation of responses is underway. The first interventions have been agreed and are in the early stages of implementation. No evaluation has yet been undertaken. Our focus in the present article is on the planned provision for realistic and realist evaluation. We set the scene with some brief comments about the expectations of government agencies concerning the evaluation of funded Indigenous community safety projects. We then set out a number of challenges to the evaluation of crime prevention projects generally, and to the present project specifically. In the main section of the paper we work through an example of how we are going about the task of applying the realist method to specific sub-problems associated with youth sexual violence and abuse.

Government interest in evaluation: Why evaluate?

'Value for Money' is a key principle that underpins all decision-making regarding the expenditure of public money. In Australia, the Commonwealth Grant Guidelines set out a number of key principles for grants administration, including 'Outcomes Orientation'. Put simply, the Guidelines articulate that agency staff should determine what change is expected as a result of a granting activity and then measure the actual outcome.

A strong focus on evaluation in the Indigenous community safety sector is deemed imperative not just in terms of accountability for public expenditure. It is also impelled by the severity of the issues to be addressed and the paucity of existing evidence about the outcomes of related initiatives. Indigenous people are consistently over-represented across all Australian criminal justice systems, both as victims and offenders, with direct ramifications for individuals, families and communities (Allard, 2010, 2011; Weatherburn, 2014). This over-representation is also a barrier to improving outcomes in a range of other areas including health, education and employment. The situation is compounded by significant gaps in the evidence base to guide policy and programs. The *"What Works to Overcome Indigenous Disadvantage – Key Learnings and Gaps in the Evidence"* report produced by the Closing the Gap Clearinghouse (2013) identified a large number of studies relevant to community safety. However, it also noted that the vast majority of these studies were descriptive only, with just two projects demonstrating positive outcomes.

Funding was allocated to GYFS-NP primarily to address a significant problem that has historically been insufficiently recognised, acknowledged or addressed, namely youth sexual violence and abuse. This problem undoubtedly exists in all communities, but seems to be especially prevalent in some Indigenous communities. Whilst value for money clearly matters, the issue is also important in its own right.

Community safety evaluation problems and how they apply to GYFS-NP

While knowing the overall value for money from different interventions is an admirable objective, it is important to be realistic about what is possible. The evaluation of criminal justice interventions faces major challenges in all settings. These problems are especially acute when some crime types, such as YSVA, are being addressed. In Indigenous community settings there are also some special difficulties.

This section sets out thirteen major challenges for evaluating crime prevention projects in general, and those focusing on YSVA in particular. At first glance these challenges may appear to present insurmountable barriers to the task evaluating outcomes of such a project. In the later sections of the paper we hope to show that progress, albeit sometimes slow and unsteady, can be made through a systematic process of theorising and empirical testing.

1. *Rare events.* Some crimes, including some forms of YSVA in Indigenous communities (e.g. stranger rapes), are high impact but rare events (Laycock, 2013; Tilley, 2009a). Obtaining meaningful numbers for statistical analysis aimed at measuring short-term outcome effects may be unrealistic.

2. *Multiple interventions.* High crime neighbourhoods typically face multiple problems in education, health and social welfare. For this reason they are often sites for many social programs. A program aiming to address one problem is liable to have an effect on others also. Teasing out the independent net effects of individual initiatives, where many are operating simultaneously and where new ones are being added and old ones fold, is often not possible using standard before and after, experimental and quasi-experimental methods. The YSVA initiative focused on here is a case in point. Both GYFS-NP sites have multiple on-going and new initiatives focused on health and social welfare as well as initiatives focused on other crime problems.

3. *Multiple components in single programs* (Tilley, 1996). Individual crime prevention programs tend to involve a range of components. They often include one or more interventions aimed at awareness-raising, opportunity-reduction, lessening crime motivation, and community capacity building. These composite measures are typically put in place by different agencies. Working out which of the measures are producing positive, negative and nil effects is highly challenging. Net impacts, if they can be assessed at all, may mask varying types of impact in different sub-groups. GYFS-NP will include multiple interventions.

4. *Dodgy and confidential data.* Counting crimes is always difficult. Not all crimes are reported and of those reported not all are recorded (Tilley, 1995). Moreover, rates of crime reporting and recording change. The reasons for change may or may not be related to the content of the program. Raising awareness of YSVA amongst the community and amongst those to whom it is reported is liable to affect the rate at which it is reported and recorded (Farrell & Buckley, 1999). This has largely been the experience of efforts to take domestic violence more seriously, for example. There are alternatives to recorded crime, but these present their own problems. Victimisation surveys, which are an obvious possible choice, are expensive, technically difficult to mount, risk responses that are not always candid, and have to be very large to capture relatively rare events. Where administrative records may be usable, for example from police, health or education, for data protection reasons they are often difficult to access at a level of detail which make them usable for evaluation purposes. In the case of YSVA this is an acute problem, not only for evaluation but also for analysis, where the details relating to subsets of young people are crucial.

5. *Adaptation*. Human beings are active agents who behave intentionally and knowingly in response to their situations and the resources they have to hand. They are liable to adapt to a change in response to new resources offered by a program or new understandings prompted by it. This, alongside multiple interventions and multiple components of single interventions, brings a characteristic complexity to the ways in which interventions produce changing and often diverse outcomes (Pawson, 2013; Byrne & Callaghan, 2013). An 'arms race' has been described for some situational measures, for example, whereby over time the offender community innovates in response to measures introduced to make offending more difficult or risky (Ekblom, 1997). Likewise those attempting to reduce crime will then try to find further means of making the crime more risky or less rewarding. With YSVA it is possible to envisage measures that will make it more difficult to find an opportunity for a sexual assault in a crime hotspot, but over time some offenders may find a way round them. The short and long term effects are liable to change so that what was effective at Time One may no longer be effective at Time Two.

6. *Rubbernecking*. Agencies in crime prevention, as in other fields, continually and commendably look for ways of improving their services to the community (Tilley, 2004). There is often also a strong oral culture where new and promising ideas spread rapidly (even poor ones can spread quickly if they have sufficient surface plausibility). This means that an initiative designed for implementation in one area is apt to be appropriated fully or in part in others also. This is liable to undermine comparisons

between experimental and control areas. Even where similar initiatives are not applied some compensatory alternative may be introduced for 'controls', again undermining the validity of any control/experimental group comparison. It is unclear yet whether or how this might apply to GYFS-NP.

7. *Unique conditions*. It is trite but true that each individual and each context is unique. Moreover each individual and each context undergoes continual change. Thus, what is in place for one individual at one moment is not precisely the same at the next (Cartwright & Hardie, 2012). Yet what is of interest in the evaluation of programs is not primarily what worked in the past at a specific time for specific groups but what will work now and in the future for the groups whose problems are being addressed. If community safety evaluations are to be useful, they have to be pitched at a useful level, one that transcends the specific. But the evaluation has also to avoid the banalities of the very general, for example that reduction in opportunity can reduce crime. Even if GYFS-NP has a net effect amongst those to whom it is applied, unless there is a reason to believe that similar effects can be achieved elsewhere little of use will have been learned. The inductivist's error is to assume that what went in the past will go just the same in the future. Evaluations need to achieve an appropriate middle-range level of generality. Few do. It is not yet clear that 'Indigenous community' comprises a unitary category in relation to which middle-range generalisations can readily be made.

8. *Defensive agencies*. Evaluations are frequently highly 'political' in the sense that there are vested interests in programs being found to be successful (Read & Tilley, 2000; Tilley, 2000). In the UK, police officers sometimes say of their crime prevention initiatives that they are 'doomed to succeed', by which they mean that findings of failure are unacceptable. Texts of evaluation reports can be and are massaged to suppress or occlude negative findings and to accentuate positive ones. At worst reports with negative findings are suppressed. The problems are most acute where those involved in the design of a program are also responsible for its evaluation and where further funding turns on positive outcomes. GYFS-NP potentially risks this in that the evaluation is not fully independent of the program.

9. *Implementation failure*. It is one thing to have a plan for an intervention. It is another for the plans to be properly delivered. Apparent negative findings may follow from inadequate implementation of an initiative that could have been effective had it been implemented as intended. In crime prevention implementation failures are common. This is partly because responsibility for community safety is generally assumed to fall to criminal justice agencies, whilst *competency* for much that could control crime lies elsewhere (Laycock, 1996). That is, those generally assumed to be accountable for the prevention of crime, notably the police, are unable directly to influence the conditions giving rise to it. For example, informal control in domestic settings can only be effected by families, the security of car parks by car park owners and managers, and the vulnerability of stores to shop theft by those running the stores. For YSVA in Indigenous communities, it is unlikely that the police have the requisite capability to deliver many of the most promising preventive measures, even if they are able and willing to deliver some. Even where project personnel can deliver the measures

themselves they will need the agreement and collaboration of others. Third parties will, however, need to be mobilised for implementation of most of the measures being designed. This reduces control of implementation and increases the risk of partial or non-implementation of the measures being developed.

10. *Monetisation mysteries*. As already indicated, value for money is a major principle behind rational government decision-making. Yet there are tricky conundrums in the calculation of net effects and accurate monetisation of inputs, outputs and outcomes in community safety initiatives. Net overall effects are important in calculating the value of an initiative. Inadvertent harms may be produced, for example by displacement of offences from one time, place, method, target, offence or offender. Inadvertent benefits may be produced, for example by diffusions of crime prevention benefits by time, place, target, offence or offender. The overall crime effect comprises the direct effect plus the diffusion of benefits effect minus the displacement effects, all calculated and monetised. In practice, tracing and quantifying indirect effect sizes is hugely problematic and so far the only types that have been measured with any sophistication are spatial displacement and diffusion of benefits in relation to programs aiming to reduce domestic burglary. Monetisation of short and long-term non-financial costs and benefits of the effects produced, assuming for a moment that they can be adequately quantified, have tended to use two methods, neither of which is satisfactory. Willingness to pay comprises one: how much would individuals be prepared to pay to reduce or avert a given crime? Willingness to accept comprises the other. How much would individuals accept to suffer a given crime? The first is unsatisfactory in that the ability to pay shapes the potentiality for payment and this will vary by time and community, where acknowledgment of the latter would be objectionable to many on grounds of social justice. The second is unsatisfactory as for some crimes, such as homicide, the sums are likely to be infinite, making calculations of differing returns on spend impossible (Adams, 1995). These difficulties face programs aiming to prevent YSVA in Indigenous communities in a quite acute form. The offences at issue are ones where the primary impacts are non-financial. Indirect diffusion of benefits and displacement effects alongside side-effects that have nothing to do with crime are all possible, albeit that we lack established standard methods for their adequate identification, measurement and monetisation.

11. *Slow burn outcomes*. Many community safety initiatives aiming to reduce risk factors are introduced relatively early in children's lives but cannot be expected to have an impact on the target problem until several years later. As time passes controlling conditions for the purpose of evaluation becomes more and more difficult. Causal attribution is thereby challenging. Yet funding bodies understandably often want robust results much more quickly because of their budgetary cycles. The precise measures to be implemented in GYFS-NP have not yet been fully determined but may relate, for example, to attempting to alter prevailing assumptions about how girls and boys relate to one another, a strategy that will take some years to deliver its benefits.

12. *Initiatives in remote locations*. There are special problems in undertaking fieldwork in remote locations where many Indigenous people live, including one of the sites for GYFS-NP. Some of the problems are logistical; it is expensive and time-consuming to travel to undertake fieldwork. Some of the problems are linguistic; English is the

first language of the research team, whereas a specific Aboriginal language is the first language of the residents. Some of the problems are cultural; non- Indigenous outsiders operate with somewhat different forms, norms and conventions relating to discourse from those prevalent in many Indigenous communities.

13. *Other stuff and secular trends.* A range of well-established sources of error in establishing the responsibility of programs for change face initiatives such as GYFS-NP (Shadish, Cook & Campbell, 2002; Tilley, 2009b). There may be secular trends that explain any observed change. There may be regression towards the mean following a peak that provokes the introduction of the program. There may be selection effects, where program participants comprise an atypical group that will either change anyway or will never change. Communities are not laboratories where conditions can be controlled. They are 'open systems' subject to external events and internal disruption that are liable to interfere with the delivery and relevance of any given program. This is the case for many local programs, where communities are susceptible to disruption from external events and where there are already trends reflecting pre-existing internal dynamics, and where programs may be introduced in response to peaks in rates that fluctuate in a pseudo-random fashion.

Realist evaluation

Realist evaluation takes its inspiration from science. Indeed, it aims to bring scientific method to program evaluation. The Nobel Prize winner for medicine, Peter Medawar, wrote a book on scientific method with the telling title, *The Art of the Soluble* (Medawar, 1967). Being realistic in evaluation means, among other things, focusing on what can be done and setting aside what may seem desirable but is not feasible. The rather obvious precepts that follow are: 'Don't try to measure the immeasurable!' and 'Concentrate your evaluation efforts where robust and useful findings are possible'. In practice this means that no evaluation could, and therefore no evaluation should, aim to measure and monetise all the intended and unintended outcomes that may be produced. It is essential, thus, to be selective and to select from what can be measured and attributed to the program.

Realist evaluation has a particular definition of what comprises science (Pawson & Tilley, 1997, 2009a; Pawson, 2006, 2013; Bhaskar, 1978). Science is concerned with identifying causal potential and causal mechanisms. However science also recognises that few causal mechanisms operate unconditionally. Their activation is contingent on conditions that are conducive to the release of their causal potential. The trick in applied science is to understand where and how to activate the causal potential of interventions to produce positive outcomes that will outweigh any negative ones. Theories are tested and refined over time, using a variety of methods, ultimately producing improved outcomes.

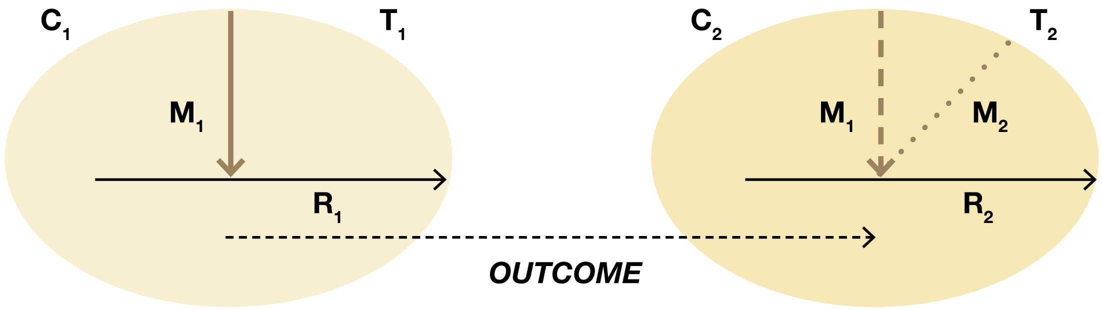
A recent book on the history of cancer treatments neatly illustrates the point being made here. As Mukherjee (2011) shows, cancer treatments have a very long history, rooted in changing theories and producing evolving and contrasting outcome patterns. Progress has consisted in gradually developing better and better understanding of common and contrasting features of different cancers together with preventive strategies and treatment regimes that attack different causal pathways and conditions for the cancers to grow and/or change in form and/or spread to further organs. Different theories have been tested using diverse methods, including randomised controlled trials, case comparisons, natural experiments, animal experiments and so on. The methods used have been those that are available and relevant to the theory at hand. Moreover, they have not been conclusive in their findings. Adherents of pet theories have been able to hold on to them! In practice, over time weights of evidence have led to changes in understanding and treatment, leading to lower age-related rates of cancer and better survival times for those who suffer its different forms.

A particular feature of the production of improved cancer prevention and treatment outcomes has been a program of research that has led to a more and more detailed understanding of the causal mechanisms producing cancer and leading to differing patterns of cancer growth amongst those with it. It is this detailed understanding that has paved the way for interventions aiming to pre-empt the production of or inhibit the growth and development of cancerous cells. The story continues, with much of the focus now on understanding the genetics of cancer and the formulation of ways to intervene at that level.

The realist and realistic evaluation planned for GYFS-NP accords with the approach that has been taken in improving the treatment and prevention of cancer. Realistic evaluation emphasises analysis by relevant subgroups furnishing salient variations in context where treatments/interventions activate or deactivate causal mechanisms to generate intended and unintended positive and negative outcome patterns. Figure 1 shows this in a formal way. C describes context. The intervention changes it from Time One (T1) to Time Two (T2). The initial problem pattern (state of affairs such as mortality rate or behaviour pattern, such as YSVA) is referred to as Regularity One (R1). This is generated by the mechanisms (M1) that are activated in the context (C1) furnished at T1. The intervention changes the context and hence the pattern of mechanisms activated. Some will be deactivated (broken M1 in C2) and others will be activated (M2). The new regularities, be they positive or negative or a mix of both, are described as Regularity Two (R2). The Outcome comprises the change in regularities (R2 minus R1) that occur from T1 to T2 as a result of change in context (C2 to C1) from T1 to T2.

There are three important complications not captured in Figure 1. The first is that the processes of change in crime prevention are never fully insulated from unplanned external events that may impinge on the activation or deactivation of relevant causal mechanisms. The second is that there are endogenous sources of instability in community safety; human agents act intentionally and adaptively in the face of change. These mark something of a difference from cancer. Cells do adapt but not intentionally. Moreover, external events do occur but on the whole more slowly. The third is that in the field conditions for community safety multiple mechanisms are activated in complex, overlapping and interacting ways. This is true also for cancer. Thus, Figure 1 is indeed figurative. It represents formal processes but does so by abstracting from much higher levels of complexity on the ground.

Figure 1: **The realist evaluation formal framework**



Realist evaluation and youth sexual violence and abuse

So what, realistically, will be possible for evaluation in GYFS-NP? The starting point is the theory of initiative and the problem-solving approach, involving some detailed theory-building to inform the data that will be collected and the methods planned for their collection.

The project team has begun by identifying sub-sets of problems, where the participants and generative mechanisms vary quite substantially, although with some overlaps. Sub-problems include, for example:

a. Peer to peer, casual 'rough sex', which is mostly unreported, to which many of the boys appear to feel entitled, and to which many of the girls appear resigned. Girls submit to the rough sex as a condition for group membership, albeit that taking part in it redefines them as morally reprehensible and thereby unacceptable as long-term partners. The boys take the girls' willingness for granted. The sex generally takes place in secluded open spaces, at night, where boys and girls congregate for recreational purposes. The boys also assume that if the girls socialise with their male peers in those places they know and accept what is expected of them.

b. Stranger rape offences against adult females, which are rarely committed, normally reported and widely publicised.

c. Opportunistic youth prostitution for small rewards, mostly provided by adult non-Aboriginal males. The incidents are rarely if ever reported.

d. Domestic sexual assaults and rapes committed by family members, some of whom are non-resident visitors. This may occur in overcrowded households where boys, girls and adults sleep in the same spaces and where children routinely see underage and adult sexual behaviour, which thereby becomes normalised.

e. Child sexual assaults and inappropriate proto-sexual behaviour in school, where those involved often cannot be seen because the layout of the school creates many suitable spaces. The behaviour follows from insufficiently recognised over-sexualised behaviour of children to which no systematic preventive response is currently in place.

There is some evidence in relation to these problems collectively and individually, but it is neither extensive nor direct. Much material comes from community interviews and file reviews, rather than from systematic primary research or the interrogation of detailed administrative records. With regard to the latter, police recorded crime data have been looked at in aggregate, but details of individual incidents have not been available. Rates of sexually transmitted infections have been examined, but again individual case files have not been interrogated. These varying sources complement one another and point in the same direction: that there is a substantial YSVA problem amongst the Indigenous youth in the project's settings and that it takes a variety of forms. The dimensions of the problem and various sub-problems appear to vary considerably between the two communities, though peer to peer sexual abuse seems very common in both. But the direct evidence on the nature and extent of the behaviours is rather limited and the quantitative evidence of their prevalence and incidence is unreliable. Rates of reporting and recording will be low for obvious reasons: some of the actions are at least partly consensual; in some cases the expected personal and familial costs of reporting incidents will likely be unacceptable; and in some cases victims may be too embarrassed to report incidents and may also fear the discomfiture of formal investigation. In addition systematic observations have been made at suspected hot spots, but for obvious reasons these have not specifically concerned sexual behaviour so much as general incivilities and antisocial behaviours. The best evidence suggests that formally available statistical data represent the proverbial tip of a YSVA iceberg, but the actual size or shape of the submerged part of the iceberg cannot be determined with any precision.

In their problem-solving efforts so far the project team has been generating accounts of the nature and causes of the problems, drawing on the available evidence and field visits to sites where the problem behaviours are believed to take place most commonly, including schools, parks and backtracks. Some possible interventions rooted in these accounts have been worked through. The team is also planning additional data gathering to supplement key gaps with what has been available, in order to test emerging hypotheses.

Take peer-to-peer rough sex as an example. Four broad initial prevention activities are proposed, with further analysis needed for three:

1. Enhance formal surveillance and community guardianship at locations found to be at high risk of peer-to-peer sexual assault through a) targeted police patrols, b) reintroduction of community night patrols, and c) installation of CCTV and increased lighting.

2. Design and implement youth based interventions to challenge concerning sexual attitudes and beliefs.

3. Undertake an analysis of youth social behaviour, networks and interactions, and

develop interventions to a) interrupt risky youth social networks and behaviour and b) increase active peer guardianship.

4. Understand and enhance familial supervision and address barriers to guardianship within immediate and extended familial systems.

Here we take just the first of these preventive activities, 'Enhance formal surveillance and community guardianship', focused on injecting police and community patrols. The evidence, theoretical rationale, proposed activity and related provision for evaluation at one of the two project sites are explained.

Evidence

Much of the concerning behaviour involving local youth occurs in public places that are hidden from view, difficult to access for police or other guardians, and provide ready escape routes if anyone tried to intervene. Locations of concern were identified through community interviews, file reviews and direct site observations where the detritus from covert youth behaviour could be observed. Community members reported many specific places to be unsafe, particularly at night, and described community-wide reluctance to intervene in problem incidents at these locations. "Hot spots" were identified on the basis of observed high risk youth activities in combination with low natural guardianship and low formal surveillance.

Rationale

The present design of public spaces presents numerous opportunities for risky and abusive behaviour, as well as significant barriers to formal or informal surveillance and subsequent responses. Given these public spaces have been identified as locations of concern for negative youth activity, general crime, and YSVA specifically, situational strategies to reduce opportunity, increase effort, and increase risk of detection in these locations are warranted.

Increasing formal surveillance and community-level guardianship of these public spaces is aimed generally at altering the existing use of these spaces, and specifically at reducing peer to peer sexual assaults. A range of situational crime prevention activities will be applied to increase effort and risk, in particular (initially) through increasing formal and informal surveillance.

Initial Proposed Activities

a. Undertake semi-structured community (adult) interviews focused on understanding barriers to, and opportunities for community guardianship.

b. Engage with key police staff regarding current patrolling practices, and the purpose and design of targeted police patrols.

c. Engage with community interest groups, former community patrol volunteers, relevant community service providers, the Regional Council and local church groups regarding the reestablishment of community night patrols.

d. Engage with Regional Council regarding potential locations for CCTV and increased lighting.

Action Plan

- Consult with local implementation group about proposed activities
- Clarify / review current police practices
- Consult the police regarding targeted patrols
- Engage with relevant stakeholders regarding the establishment of community night patrols
- Consult the cleaning authorities regarding the documentation of rubbish at identified public locations
- Develop and secure formal partnership agreements
- Implement additional research activities
- Analyse results and review intervention plans as necessary
- Implement intervention activities

It is clear that even if the underlying account of the rough sex in public places problem is correct (remembering that much of the evidence for it is anecdotal) and even if the proposed strategy of increasing guardianship is one which could inhibit the unwanted behaviour (by activating mechanisms to do with reducing opportunity by increasing the perceived risk and effort involved in it), enhancing guardianship faces substantial implementation challenges. There is an unstated body of fallible but plausible theory behind the efforts to secure increases in targeted patrols by the police and community members (Jones & Tilley 2004; Ratcliffe,Taniguchi, Groff & Wood, 2011). Moreover, it is possible that any increase in patrols would produce either a) displacement of the youth rough sex to different times and locations or b) diffusions of benefit to other forms of crime and antisocial behaviour taking place in the areas where patrols are planned (for example reductions in robbery, noise, littering, and public drunkenness which are also facilitated by and attracted to secluded places in nearby residential areas known to likely offenders and prospective offenders) or c) diffusions of benefit to times when the patrols were not actually taking place.

In realist terms the theory behind increased patrols is relatively simple:

Context One, with its mechanisms and regularities: The sites comprise easily accessible nearby open spaces for youth, in which covert activities can easily take place, especially at night. These activities include sexual behaviours that boys desire and the girls at least tolerate as the price of group acceptance. The activities involve low effort, low risk and high reward for the boys in particular but also for the girls who are unlikely to make crime reports. There is also no informal social control from significant others in the local community who could show disapprobation. Instead, the group norms of sexual behaviour prevail and it is expected that boys will want sex and girls will be willing to satisfy them. The result is a seldom reported under-age sexual regularity in the particular places. The conditions present also attract other forms of criminal and antisocial behaviour which are also low risk, low effort and high reward. Hence there are non-sexual as well as sexual crime regularities.

Context Two, producing an expected change in mechanisms and hence regularities: Police and community patrols are introduced at those times when illicit sexual behaviour appears to be most common. The change in context means a change in activated causal mechanisms. Real and perceived risks to boys and girls increase, prospective rewards for boys decrease, and effort is increased if the preferred times and places for the sexual behaviour cease to be so readily available. The reduced conduciveness of the context for YSVA also reduces it for other types of crime and antisocial behaviour which also decline as a side-effect, a type of diffusion of benefits. Furthermore because perpetrators are unaware of the patrol times there is a temporal diffusion of benefits beyond the times the areas are patrolled, affecting both sexual and non-sexual crimes, in relation to which prospective offenders overestimate the risks they face. Finally, spatial and temporal displacement may eventually take place as offenders switch the times and places of their (sexual and non-sexual) offences to ones where the risk and effort is lower.

The costs and benefits of the intervention would need to monetise the costs of the patrols and the organisation of the patrols. Against this, monetised benefits from prevented sexual crimes would need to be calculated. Furthermore the monetised costs of any displaced sexual crimes to other times and places would need to be added to the debit side of the equation and the monetised benefits from prevented sexual crimes beyond the time and place of the patrols and of other offences averted as a diffusion of benefits from the targeted patrol would need to be added to the credit side.

The data problems for the evaluation of this small segment of this proposed element of GYFS-NP (one of four types of activity, relating to one of five sub-problem types, in one of two targeted communities) are legion. The patrols are not the only intervention planned by the project team, and there are other initiatives constantly coming and going in the neighbourhood, for example from newly appointed community beat officers and initiatives aimed at other crimes that may impact YSVA. There is little formal reporting and recording of the behaviours, and although the rate of the behaviours is unknown it is likely to be quite small, making tracking real change as against normal fluctuations difficult to distinguish. It is not possible in principle to know where to look to catch all displacement and diffusion of benefits, and possible candidates will again likely have pseudo-random fluctuating levels. Calculation of the non-financial short and long term benefits and costs involve inexorably contestable assumptions. Producing a precise measure of impact size, costs and benefits is not only likely impossible; it would also be of dubious value if it were too orientated to the specifics of the local setting. A level of abstraction beyond the overly particular is needed if findings are to be more widely usable.

What we have shown here is a tentative theory for one part of GYFS-NP, formulated in realist terms. Its test would comprise a realist evaluation. How, if at all, can it be tested in practice, given all the challenges that face community safety evaluations in general and those relating to YSVA in Indigenous communities in particular? The answer is that those data that can be collected that speak to the theory, imperfect though they are, will be assembled and used. Note that they will not test the whole theory, in particular they will not test those parts of the overall program theory that speak to the implementation of the proposed measures.

Five data sources for this component of the project are being prepared.

a. Refuse analysis: The suspected high crime areas have regular rubbish collection. The rubbish, of course, comprises litter. It also provides unobtrusive measures of different types of crime and antisocial behaviour: drug paraphernalia indicates drug taking, spent condoms sexual behaviours and empty beer cans alcohol consumption, while the total amount of rubbish comprises an indicator of the total volume of antisocial behaviour. Rubbish collection takes place monthly. In the months leading up to the start of the patrols the collected rubbish will be analysed to obtain a background before measurement. Patrol conditions (police, community, police and community and no patrol) can then be alternated following each clean up and the rubbish next collected then acts as an indicator of changes in illicit usage of the patrolled area. These data will provide an indicator as to whether the types of YSVA and related behaviours are changed. It will also provide some data on possible diffusions of benefit within the patrol target area. Similar observations in nearby areas that are propitious for YSVA will also be made to gauge whether there is displacement or diffusions of benefit to them.

b. Observations: Controlled observations are planned in the period running up to the interventions. These will involve a checklist indicating who is seen and what they are doing. The observations will be repeated during the different interventions. They will also be undertaken in the hour before and hour after the times patrols are scheduled. Changes and variations in the observed populations and their behaviours comprise a second indicator. This indicator will also provide some evidence of temporal displacement and diffusion of benefits from the patrols.

c. Interviews: An opportunity sample of young people will be interviewed (using realist methods where the respondent is enjoined to feed back on the theory – see Pawson & Tilley, 1997) as the intervention unfolds to determine whether they have noticed the patrols and if so whether they or their friends have changed their behaviours as a result and if so how. Likewise those involved in the patrols will be interviewed to elicit their emerging theories about positive and negative consequences of their presence and how these are produced, with a view to testing these practitioner theories if the data that are collected (or can be readily assembled) are suitable.

d. Police calls for service: These will furnish data on the subset of incidents that come to police attention in the patrol areas and in possible displacement and diffusion of benefits areas. Some may relate to YSVA but only small numbers are expected. The calls for service data will enable comparisons to be made on changes in incident numbers in patrol and non-patrol areas and within patrol areas between patrol and non-patrol times, distinguishing between different patrol modalities.

e. Patrol records: These will check when patrols are undertaken, what the patrols experienced and what they did.

None of the data sources, a, b, c, d or e, is perfect, but collectively they speak to the theory that has been framed in realist terms. They would not enable a precise effect size to be estimated. They would, however, provide evidence on the direction of the direct and indirect, intended and unintended, outcomes from this segment of the program. Table 2 summarises the broad uses of the data in relation to the realist approach being adopted, and the hypotheses that have been formulated so far.

The detailed analysis, theory-building and ad hoc data collection plans, shown here for targeted patrols, are needed also for all other components of the project if they are to be evaluated realistically. It is a tall order but it marks a way for evidence-based improvement. Where similar types of intervention, rooted in similar theory, are tried in differing sites across the project it may be possible to synthesise findings at a higher level of abstraction to inform policy and practice more widely.

Table 2: **Summary of proposed data uses for realist theory test and elicitation**

	Refuse analysis	***Observations***	***Interviews***	***Police calls for service***	***Patrol records***
Patrol modalities context outcome variations	*	*	*	*	
Short, medium and long term diffusion of benefits/displacement	*			*	
Implemented measures		*			*
Mechanisms activated		*	*		
Alternative conjectures for possible test with other data collected		*	*		*

Any claim, however, that robust quantified estimated net effects of community safety programs, such those addressing YSVA, including intended and unintended positive and negative effects, monetising both financial and non-financial short and long term costs, would, we think, entail a form of alchemy even though it is clear that they could have potential value to policy-makers and practitioners in allocating their resources. At this stage of development at any rate, as with similar stages in developing treatments for cancer, more modest but nevertheless realistic and important results, emanating from ad hoc opportunist and bespoke data collection and analysis, are needed.

At a much higher level of abstraction some evaluation of the success of the underlying problem-solving methodology is planned. The context is one of 'wicked issues' (Rittell & Webber, 1973): one of unknown extent and severity, that cannot adequately be dealt with using existing standard practices, that has no known simple solution and that falls between or across the responsibilities of several agencies and organisations. YSVA, especially in Indigenous communities, is just such a 'wicked issue'. Wicked issues, the theory goes, can be addressed effectively by a) systematically unpacking the problem attributes in detail and working through possible points of intervention to prevent or ameliorate them, and then b) mobilising relevant agencies and organisations to co-operate and collaborate in delivering changed policies and practices in accordance with that analysis. For YSVA specifically, the public health framework is useful in shaping analysis of the problem and identifying potential points of intervention. Once made aware of the problem and what might be done more effectively to deal with it, well-meaning agencies, organisations and community members can be expected to co-operate in delivering changes in their practices to identify and respond to it. If this overall approach is successful in relation to YSVA, we would expect two measurable outcome patterns. The first would be an initial increase in formally reported YSVA as the problem becomes more widely acknowledged by victims, families and formal agencies and as agencies learn to respond to it more sensitively and effectively, followed by a fall in reports as the overall real rate goes down following effective interventions. The second would be a fall in clinic attendance for sexually transmitted infections amongst young people, given that one source is YSVA, and clinics will, we assume, be attended before, during and after the initiative. These are the broad, expected and measurable outcome patterns.

Conclusions

GYFS-NP provides a unique opportunity to learn about the strengths and challenges associated with the application of Realist Evaluation in the context of an Indigenous community safety project. Realist Evaluation has garnered broad, international respect for its clear focus on building understanding about 'what works for whom in what circumstances'. This combination of questions is appealing. It is essential that administrators of public funds are able to advise government if funded projects achieve intended outcomes. However, for positive results to be maintained or replicated elsewhere, teasing out the contextual elements and the mechanisms that make the outcomes possible is equally essential. Realist Evaluation presents a neat solution to the seemingly incompatible requirements of contextual specificity and generalisable knowledge. It is hoped that GYFS-NP will both achieve strong safety outcomes for the communities involved, and generate much needed evidence to inform safety initiatives elsewhere. The degree of difficulty of the questions Realist Evaluation seeks to address suggests that successful application will not be easy.

References

Adams, J. (1995) *Risk*. London: UCL Press.

Allard, T. (2010). *Understanding and preventing Indigenous offending* (Brief 9, December 2010). Retrieved from Standing Committee of Attorneys-General, Australian Institute of Criminology Indigenous Justice Clearinghouse website http://www.indigenousjustice.gov.au/briefs/brief009.pdf

Allard, T. (2011). Indigenous young people and the justice system: Establishing an evidence base. In A. Stewart, T. Allard & S. Dennison (Eds.) *Evidence Based Policy and Practice in Youth Justice*. Sydney: Federation Press.

Bhaskar, R. (1978). *A Realist Theory of Science*. London: Verso.

Byrne, D., & Callaghan, G. (2013). *Complexity Theory and the Social Sciences: The State of the Art*. London: Routledge.

Cartwright, N., & Hardie, J. (2012). *Evidence-based Policy*. Oxford: Oxford University Press.

Closing the Gap Clearinghouse (2013). *What works to overcome Indigenous disadvantage: key learnings and gaps in the evidence 2011-12*. Produced for the Closing the Gap Clearinghouse. Canberra: Australian Institute of Health & Welfare. Melbourne: Australian Institute of Family Studies.

Eck, J., & Spelman, W. (1987). *Problem-Solving: Problem-Oriented Policing in Newport News*. Washington DC: Police Executive Research Forum.

Ekblom, P. (1997). Gearing up against crime: A dynamic framework to help designers keep up with the adaptive criminal in a changing world. *International Journal of Risk, Security and Crime Prevention*, 2, 249-65.

Farrell, G., & Buckley, A. (1999). Evaluation of a UK police domestic violence unit using repeat victimisation as a performance indicator. *The Howard Journal of Criminal Justice and Crime Prevention*, 38, 42-53.

Jones, B., & Tilley, N. (2004) *The Impact of High Visibility Patrol on Personal Robbery Hot Spots* (Research Findings 201), London: Home Office.

Laycock, G. (1996). Rights, roles and responsibilities in the prevention of crime. In T. Bennett (Ed.), *Preventing Crime and Disorder: Targeting Strategies and Responsibilities*. Cambridge, UK: University of Cambridge.

Laycock, G. (2013). Happy Birthday? *Policing*, 6, 101-107.

Medawar, P. (1967). *The Art of the Soluble*. London: Methuen.

Mukherjee, S. (2011). *The Emperor of All Maladies: A Biography of Cancer*. London: Fourth Estate.

Pawson, R. (2006). *Evidence-Based Policy*. London: Sage.

Pawson, R. (2013). *The Science of Evaluation:* London: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Pawson, R., & Tilley, N. (2009) Realist evaluation. In H. Uwe, A. Polutta & H. Ziegler (Eds) *Evidence-based Practice: Modernising the Knowledge Base of Social Work?* Opladen & Farmington Hills: MI, pp. 151-180.

Ratcliffe, J., Taniguchi, T., Groff, E. & Wood, J. (2011). The Philadelphia Foot Patrol Experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology, 49,* 795–831

Read, T., & Tilley, N. (2000). *Not Rocket Science?* Crime Reduction Research Series Paper 6. London: Home Office.

Rittell, H., & Webber , M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*, 155-169.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Sidebottom, A., & Tilley, N. (2011) Improving problem-oriented policing: The need for a new model? *Crime Prevention and Community Safety, 13,* 79-101.

Smallbone, S., Marshall, W.L., & Wortley, R. (2008). *Preventing child sexual abuse: Evidence, policy and practice.* Cullompton, Devon: Willan Publishing.

Smallbone, S., & Rayment-McHugh, S. (2013). Youth sexual violence and abuse: Problems and solutions in the Australian context. *Australian Psychologist, 48,* 3-13.

Smallbone, S., Rayment-McHugh, S., Crissman, B., & Shumack, D. (2008). Treatment with youth who have committed sexual offences: Extending the reach of systemic interventions through collaborative partner*ships. Clinical Psychologist, 12,* 109-116.

Smallbone, S., Rayment-McHugh, S., & Smith, D. (2013a). *Preventing youth sexual violence and abuse: Scope, dimensions, and dynamics of the problem at (two sites).* Unpublished Report.

Smallbone, S., Rayment-McHugh, S., & Smith, D. (2013b). Youth sexual offending: Context, good-enough lives, and engaging with a wider prevention agenda. *International Journal of Behavioral Consultation and Therapy, 8,* 54-60.

Tilley, N. (1995). *Thinking about Crime Prevention Performance Indicators,* Crime Prevention and Detection Series Paper 57, London: Home Office.

Tilley, N. (1996). Demonstration, exemplification, duplication and replication in evaluation research. *Evaluation: The International Journal of Theory, Research and Practice, 2,* 35-50.

Tilley, N. (2000). The evaluation jungle. In K. Pease & V. McLaren (Eds.), *Crime Prevention: What Works?* London: Institute for Public Policy Research: pp. 115-130.

Tilley, N. (2004). Applying theory-driven evaluation to the British Crime Reduction Programme: The theories of the programme and of its evaluations, *Criminal Justice, 4,* 255-276.

Tilley, N. (2009a). What's the 'what' in "what works?" Health, policing and crime prevention.' In J. Knutsson and N. Tilley (Eds.). *Evaluating Crime Reduction Initiatives.* Crime Prevention Studies Vol 24. Monsey, NY: Criminal Justice Press and Cullompton, Devon: Willan, pp. 121-145.

Tilley, N. (2009b). *Crime Prevention.* London: Routledge.

Weatherburn, D. (2014). Arresting Incarceration. Canberra: Aboriginal Studies Press.

Telling Context from Mechanism in Realist Evaluation: The role for theory

Hannah Jolly

University of Southern Queensland

hannah.jolly@usq.edu.au

Lesley Jolly

Strategic Partnerships Research Consultants

ljolly@bigpond.net.au

Keywords: Realist evaluation, context and mechanisms, curriculum evaluation, curriculum model, realist program theory, formal theory in realist evaluation, curriculum model

Abstract

Realist evaluation is based on the premise that aspects of context trigger particular mechanisms in response to an intervention, which result in observable outcomes. This is often expressed in the formula C+M=O. Contexts are defined as the conditions that an intervention operates in (often but not exclusively sociocultural), while mechanisms are understood to be the future action that people take in response to the intervention. There is much debate, however, about the definitions and because distinctions are not clear-cut it can be difficult to decide which is which, particularly when the intervention concerns some program of curricular intervention. In this paper we discuss how we resolved this dilemma in an evaluation of a curriculum change in 13 universities in Australia and New Zealand. In that case we found a cascade of contexts and mechanisms, whereby what was a mechanism from one point of view (such as the decisions involved in course design) became a context triggering later mechanisms (such as teacher and student behaviours). The scholarly literature defining curriculum helped us to organise our thinking and subsequent analysis in a rational way, but in many evaluations there may not be a handy body of work that discusses how to understand the topic of the intervention in this way, nor do many consultant evaluators have the luxury of long hours in the library. We consider some ways in which evaluators might decide on defining contexts and mechanisms in principled ways and some of the consequences of those decisions.

Introduction

The contribution of Pawson and Tilley's (1997) realist approach to program evaluation has constituted a significant shift from available methods. It is most simply understood as a method for evaluating "what works for whom in what circumstances" (Pawson & Tilley, 1997). Rather than focus on global judgements about the worth of a program, it seeks to identify the varieties of success and failure that any program experiences and the factors that contribute to all of the eventual outcomes. The basic premise is that there will be a range of conditions, often sociocultural, that affect the outcomes of any program. These are referred to as Contexts (C). In addition the ways in which people respond – their reasoning about what they should do and the resources they can bring to bear (Pawson & Tilley, 1997, p.67) – will also vary. In the realist approach this is referred to as the Mechanism (M). Hypotheses about how the program results in observed outcomes (O) is often expressed in the formula C + M = O (CMO). The attraction of this approach lies in the fact that it notes real life programs are rarely entirely successful or entirely unsuccessful, but have patches of success and failure. Also, it is common to find that a program judged to have worked well in one place fails in another or in subsequent years. Realist Evaluation (RE) not only focusses on underlying factors behind outcomes but the various ways in which they can combine and recombine to cause outcomes.

Since its publication, the approach has been widely taken up and applied with varying methodological success (Pawson & Manzano-Santaella, 2012), suggesting that the application of the method is not so simple. Pawson and Manzano-Santaella (2012, p. 176) have now published a discussion of some of the challenges of the "practice on the ground," including the oft expressed problem of "I am finding it hard to distinguish Cs from Ms and Os, what is the secret?" (Pawson & Manzano-Santaella, 2012, p. 188). Whilst their paper discusses this issue in some detail, we will also address the subtleties of this challenge, and attempt to explore their recommendation that "which property falls under which category is determined by its explanatory role" (Pawson & Manzano-Santaella, 2012, p. 187).

The challenges of applying the realist evaluation approach

Whilst much of the discussion of the difficulties of applying the realist approach is given over to understanding the differences in function between Contexts and Mechanisms, this may be premature if a suitable understanding of the *function* of a CMO configuration *as a whole* is not applied to the process of evaluation. In their 2012 "workshop" on the method, Pawson and Manzano-Santaella (2012, p. 188) emphasise that "the function of CMO configurations…is that they are rather narrow and limited hypotheses, which attempt to tease out specific causal pathways, as pre-specified mechanisms, acting in pre-specified contexts spill out into pre-specified and testable outcome patterns." That is to say, these configurations are sensitive to the actual moment in the intervention process being considered. They need to be used at appropriate times and in appropriate ways during the data analysis if they are to help us to make meaningful evaluations.

In our case, we had an idea of what the intervention was meant to achieve and how it was meant to achieve it, and we began analysis by trying to define contexts and mechanisms directly from the data. When we took this approach we found that it led us in circles. This is because the *function of variables* in a moment of analysis (that is, whether a variable acts as a C (Context) or as an M (Mechanism) is very much dependent on the *focus of explanation* at a given point in the analysis. Something which is a mechanism at one stage of an intervention, such as the reasoning leading to particular decisions about how to design and implement a program, may then produce a fresh context for a later stage, such as the way subjects strategise in response to the program design.

This situation was complicated in the example evaluation by the fact that the program of intervention was taking place in multiple sites, and with differing purposes and methods of implementation in each site. We knew that the focus of explanation needed to vary from site to site, but had not yet pinned down how. Add to this that the program in question concerned a curricular innovation (the notion of curriculum being notoriously slippery), and we quickly discovered that analysis of the data we had collected was creating more questions than answers.

As Pawson and Manzana-Santaella (2012, p. 178) reiterate, “realist evaluation is [or should be] avowedly theory-driven; it searches for and refines explanations of program effectiveness.” While it can be daunting to be told that more theory is needed, in our case it turned out that the theory that helped us to define the specific causal pathways to be investigated was a quite practical one about the nature of curriculum. While this is a highly debated topic, once we had settled on an understanding of what “curriculum” encompasses and how the various elements interact, the evaluation task became much easier.

The example evaluation

The evaluation in question was of a program of curricular innovation that had taken place at a variety of universities across Australia and New Zealand. The program involved the introduction of the Engineers Without Borders (EWB) Challenge into the first year engineering curriculum. The EWB Challenge was conceived as a means of exposing students to the principles of engineering design and problem solving, by providing a design challenge based on the requirements of a real, third-world community who have worked with EWB on sustainable development projects. This program of innovation constituted a “widespread curriculum renewal in engineering education”, because:

> The first year in engineering had traditionally focussed on basic science and maths and the introduction of the Challenge and its associated team-based project work allowed for development of the so-called “soft skills” amongst the graduate attributes: communication and teamwork and an understanding of the need for sustainable development. The Challenge has been in operation since 2008 and every engineering school in Australia has made some use of them at one time or another. This [evaluation] project was carried out with the co-operation of 13

> universities from Australia and New Zealand who have maintained their use of the projects, albeit in widely divergent types of student cohort and courses. (Jolly, 2014, p. 3)

Thus, the evaluation was seeking to understand both how the program had been applied differently in different sites, and for different purposes, and what contributed to local success and failures. As such, the evaluation was focused on both process and outcome, in that it sought to discover both how the intervention worked, and with what effect. Realist Evaluation (RE) is ideal for this kind of multi-site, multi-context situation where correlations between variables are unlikely to apply in all cases and an understanding of the range of generative causation that can apply is required. In CMO terms, the ideal, desired operation of the intervention could be expressed in a highly compressed form (Table 1).

It needs to be noted that there are dangers in such shorthand representations of CMO configurations (Pawson & Manzana-Santaella, 2012), which we will discuss further below. For now we acknowledge that this hypothesis about how the program should work includes many finer grained levels of CMO configuration. In fact it was the task of the evaluation to find out just what those finer-grained configurations were.

Table 1: **The ideal CMO configuration for the program (based on Jolly, 2014)**

Context (C)	+	***Mechanism (M)***	=	***Outcome (O)***
• First year engineering curricula emphasise technical and theoretical subjects and pay little attention to practical “real-world” engineering. • Need to develop so-called “soft skills” such as communication and teamwork in engineering students. • Need to respond to widespread concern for environmental issues, especially sustainable development.		• The use of EWB projects with real-world clients will expose students to project-based design work in engineering and this exposure will influence the reasoning of students about how they develop as students and as engineers.		• Students will develop the targeted teamwork and communication skills; start to become familiar with engineering project and design methods; and will learn to incorporate sustainability considerations into engineering design. • These developments will be maintained and built on throughout their time at University.

With a high volume of data and differing focus and desired detailed outcomes across sites, the task of evaluation was difficult and complex. Fortunately, the data collection process had already been planned and executed on an explicit theoretical basis; that is, on a principled notion of "curriculum":

> Since the topic [of the evaluation] was curriculum change, [we had begun the process] with a review of what could be understood to be included in the term "curriculum" and this resulted in taking a broad scope, including aspects of institutional and course context as well as what happened in the classroom. Program logic analysis was carried out in each site to gain insight into how staff understood what they were doing and how outcomes should be achieved. Data was collected in a broadly ethnographic manner with the help of research assistants recruited and trained at each site. Data sources included documents produced by institutions, staff and students, observations of classroom activities, interviews and focus groups and an online survey for students. (Jolly, 2014, p. 3)

Jolly (2014) discusses the process of basing the data collection in program logic analysis in more detail. Here we are interested in how a theory of curriculum helped us to refine our understanding of the effectiveness of this program by identifying appropriate focusses of explanation within which to define Contexts and Mechanisms.

Defining and dealing with the notion of "curriculum"

Dillon (2009, p. 3) points out that review of the literature on curriculum shows more than 120 definitions of the term, "presumably because authors are concerned about either delimiting what the term means or establishing new meanings that have become associated with it…we need to be watchful, therefore, about any definitions that capture only a few of the various characteristics of curriculum". He approves the classic definition of Schwab, which is also amongst the more comprehensive in the literature:

> Curriculum is what is successfully conveyed to differing degrees to different students, by committed teachers using appropriate materials and actions, of legitimated bodies of knowledge, skill, taste, and propensity to act and react, which are chosen for instruction after serious reflection and communal decision by representatives of those involved in the teaching of a specified group of students who are known to the decision makers. (Schwab, 1983, quoted in Dillon, 2009, p. 343)

This definition begins by saying that curriculum is about outcome, 'what is successfully conveyed to differing degrees'. Curriculum in this definition also includes the more usual elements of materials, actions, knowledge and skills, but the phrase 'chosen for instruction' draws our attention to the fact that the activities that will develop knowledge and skills have to be designed. People with authority and those who 'represent those involved' need to make choices about what and how they teach. In order to make those choices they need knowledge of the prospective students.

To recast this statement into a realist form, we could assume that all curricula operate in contexts that begin at the institutional level. Institutions such as schools, training organisations, workplaces and industries set the conditions under which required knowledge and acceptable practice can be defined. As a result of those conditions, people who are designing syllabi and going into classrooms make certain decisions about how to develop the required knowledge and practice in their students. At this point in the curriculum process, mechanisms within the 'course design' process – the reasoning about how to use resources – have the power to explain outcomes in relation to the nature of the program. The result of those decisions and resource allocations is a set of learning environments that are the context in which students decide what they will do with the learning opportunities on offer. At this point "Course Design" has the power to explain learning outcomes when considered as a context in which a range of mechanisms are triggered. The outcomes of those mechanisms constitute 'what is successfully conveyed' and sometimes that can include things that educators had no intention of conveying. Thus an element of the underlying theory or model of curriculum, such as Course Design, cannot be labelled either as Context or as Mechanism until we know exactly what we are attempting to explain.

Defining and dealing with Contexts and Mechanisms

In the example evaluation discussed here, in attempting to sort out what the focus of evaluation at each site should be, we recognised the need to take a broad theory of curriculum but to organise it in line with Realist thinking. Table 2 shows a map of the curricular landscape that was devised accordingly. Each of the boxes in this map details the relevant focus of explanation at a given point in the curricular landscape. These points could then be examined independently for different sites, according to the data we were dealing with from those sites. While the headings used in the table were derived from the wider literature on curriculum, the descriptors were derived from our data, and the structure of the table reflects our belief in the "generative causation" that is at the basis of all Realist Evaluation (Pawson & Tilley, 1997, p. 56).

This curricular landscape map became an analytical device allowing the data to be unpacked into a single explanatory proposition about how the program of intervention was working in any given site or at any given level (macro to micro). This explanatory proposition is as follows:

> A high level curriculum context pertaining to institutions and programs [triggers] choices people [instructors and administrators] made about how and when to use the EWB challenge projects. The actions that resulted from those choices, such as what pedagogy would be used in the classroom or how assessment would be undertaken, then set a new context in which students made choices about how they would respond to the projects and what was being asked of them. (Jolly, 2014, p. 14)

Table 2: **The Curricular Landscape**

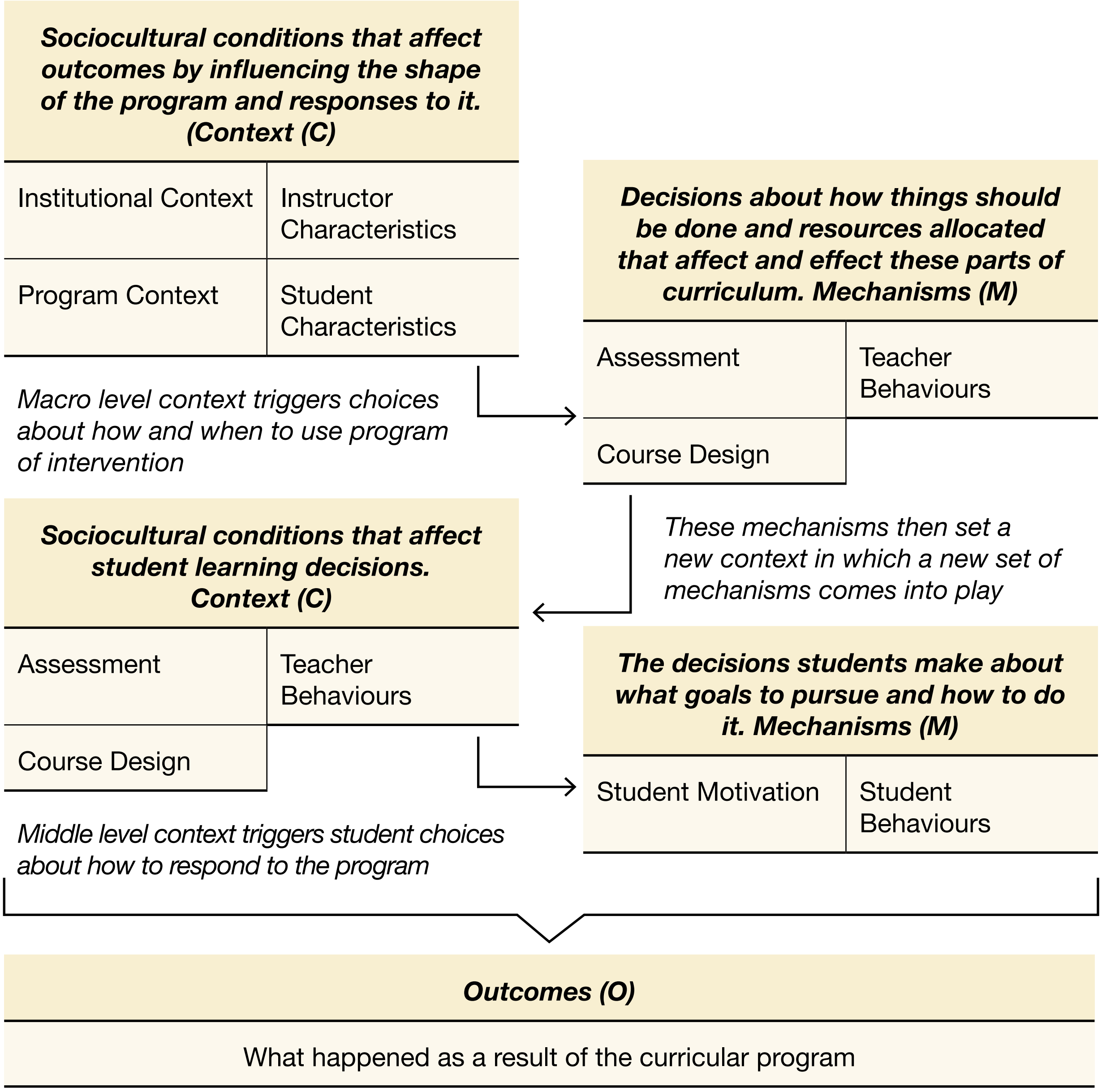
← *Potentially context* →

← *Potentially mechanism* →

Institutional Context The institutional factors affecting the way in which the course is implemented	**Instructor Characteristics** The experience, beliefs and attitudes of instructors before responding to this instance of the course	**Teacher behaviours** How teachers respond to this instance of the course, including decisions, attitudes, interactions		**Outcomes** What happened or changed as a result of teacher and student behaviour
Program Context The nature of the course within the program Factors affecting status, purpose and perceptions of the course within its program context	**Student Characteristics** The experiences, beliefs and attitudes of students prior to responding to the course	**Student Motivation** The factors affecting the kinds and level of effort/ interest, etc., that students put in to the course	**Student behaviours** The nature and amount of student participation in the course, including how and how much they focus on topics, processes and products	
	Assessment The nature of assessment tasks, the nature of criteria, the weightings of criteria			
	Course Design The nature, amount and sequence of organisation of the course, the topic focus, the resources and learning processes it incorporates and the level and nature of prior knowledge it assumes.			

This means that the curriculum map was able to be further unpacked into an evolving set of Cs and Ms, as seen in Figure 1.

Figure 1: **The CMO Cascade**



Sociocultural conditions that affect outcomes by influencing the shape of the program and responses to it. (Context (C)

Institutional Context | Instructor Characteristics

Program Context | Student Characteristics

Macro level context triggers choices about how and when to use program of intervention

Decisions about how things should be done and resources allocated that affect and effect these parts of curriculum. Mechanisms (M)

Assessment | Teacher Behaviours

Course Design

These mechanisms then set a new context in which a new set of mechanisms comes into play

Sociocultural conditions that affect student learning decisions. Context (C)

Assessment | Teacher Behaviours

Course Design

Middle level context triggers student choices about how to respond to the program

The decisions students make about what goals to pursue and how to do it. Mechanisms (M)

Student Motivation | Student Behaviours

Outcomes (O)

What happened as a result of the curricular program

In one way, this cascade complicates the task of deciding what counts as context and what as mechanism since many aspects of curriculum might be both. Pawson and Manzana-Santaella (2012, p. 187) remind us that ultimately the choice about "which property falls under which category is determined by its explanatory role". As already noted, any ingredient of the model may operate as either context or mechanism depending on what it is we are trying to explain about the curriculum and how it works. If, for instance, we want to explain the variations in the ways the 13 universities designed their courses, we need to see course design as a set of reasoning, decisions and allocation of resources (Mechanisms). But if we are examining why some students managed group work well and others did not, we might need to consider the course design as a relevant part of the context. In other words, it depends on whether we aim to explain the next point in the curriculum cascade (in this case the production of a course design) or the ultimate outcomes of the whole program (here a set of learning outcomes). In line with the requirement that realist evaluations need to produce many CMO configurations instead of a catalogue of characteristics (Pawson & Manzana-Santaella, 2012), we need to treat the curriculum model as tool for identifying potential CMO configurations. The model identifies potentially relevant moments in the working of curriculum, but the direction and detail of relevant CMO configurations need to be established empirically. While we cannot go into the whole analysis here (for the final report of the project, see Jolly, 2014) we will take one example to illustrate what we mean.

A detailed example

Identifying what had to be explained

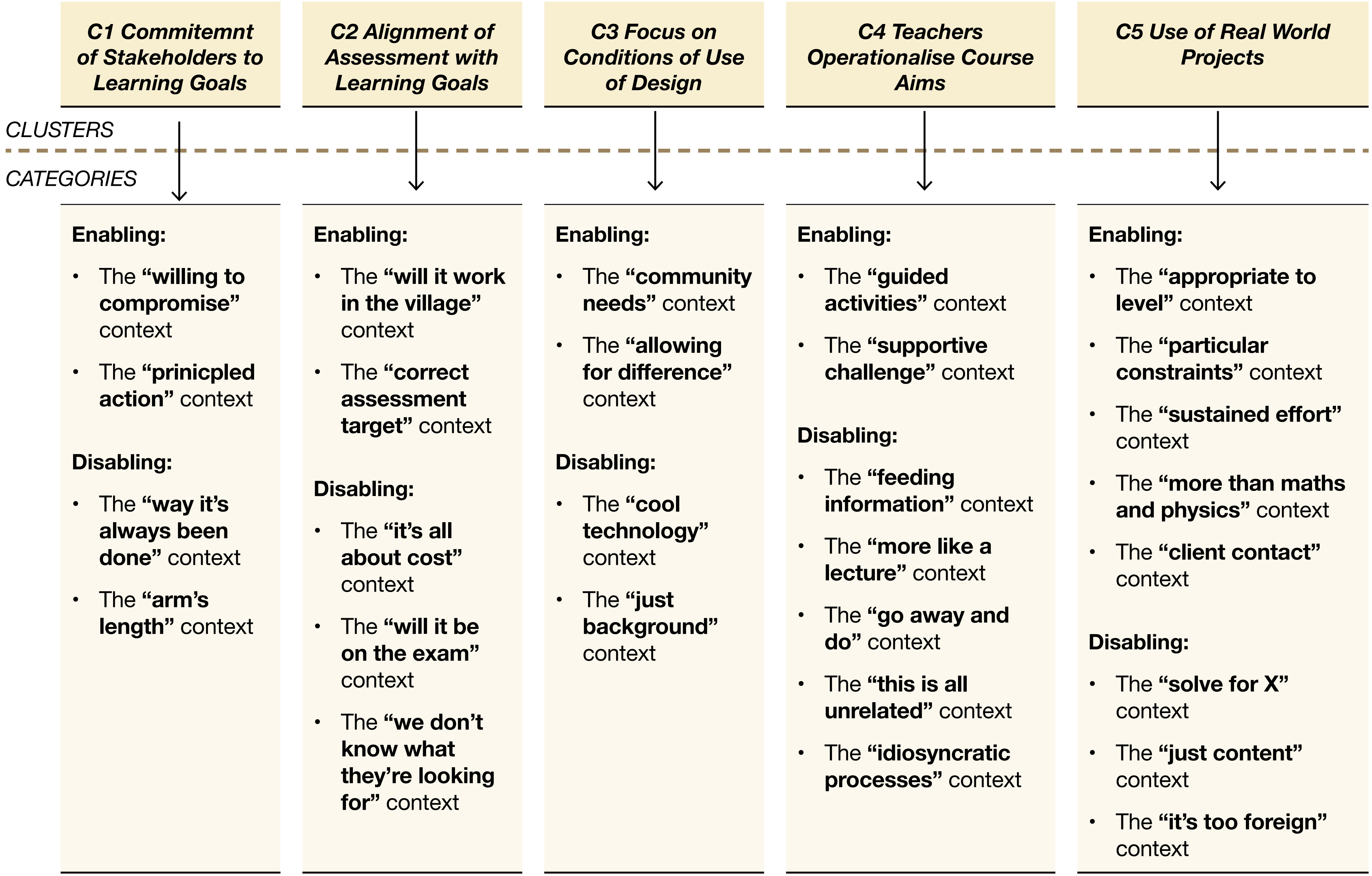
One of the first steps in the evaluation was a series of program logic interviews with the teaching staff at each of the 13 sites (University of Wisconsin, 2010; Markiewicz, 2010). As well as clarifying objectives for each site, this process allowed us to explore participants' understandings of how the EWB Challenge program was thought to work. This gave us an empirical way to start generating hypotheses in CMO form (although admittedly sketchy at this stage) which also guided our data collection activities. Although many staff identified a better understanding of sustainability as a desirable outcome, most appeared to be relying on the content of the projects alone to bring this about (Jolly, Crosthwaite, Brodie, Kavanagh & Buys, 2011). This led us to pay attention to what students learned about sustainability and how they learned it, but in the process of investigating that, our attention was drawn to a wide range of student goals and motivations – what they wanted to get out of the course – and how those things affected learning outcomes differently in different contexts. Sustainability-related mechanisms are separated from other student-goal related mechanisms in our final analysis because participating universities had requested information specifically about the sustainability goal.

Data were gathered through a range of techniques including interviews, participant observation, surveys and documentary analysis. Recurrent themes or category of event were identified through the constant/comparative method (Richards, 2005) and labelled using *in vivo* category labels. Categories of Cs and Ms were able to be grouped according to whether they were observed to have a positive or negative impact on the overall outcomes identified in the program logic interviews. Certain contexts were deemed to enable positive mechanisms, while others disabled them. Mechanisms either supported the attainment of the desired outcomes or they inhibited them (Sochacka, 2011). They were also grouped by similarity into higher level clusters which relate to both wider educational theory and observed reality and could be expected to apply across contexts. For instance, a set of contexts relating to the ways in which student actions were driven by the ways in which the course presented activities and goals were grouped together under the Context C2 *Alignment of assessment with learning goals*. This is the well-known principle of constructive alignment (Biggs, 1996). The category level factors illustrate what the principle looks like for this particular set of interventions and the labelling of 'clusters' allows comparison with other educational debates and thus transferability of the findings. The entire analysis is presented in diagrammatic form in Figures 2 and 3. From these figures a potentially very long list of CMO configurations could be generated and workshops were run with all participants to help them work through their own cases. The most common hypothesis held by educators about the sustainability outcome is represented in Table 3.

Table 3: **How the program delivered improved understanding of sustainability**

Context (C)	+	***Mechanism (M)***	=	***Outcome (O)***
• Course projects were real-world projects (C5) with real clients from developing nations such as Cambodia.		• Students reason that these communities need help and that engineers have a responsibility to come up with design solutions to identified problems which are appropriate to the culture and environment and able to be operated by local communities beyond initial implementation.		• Sustainability is automatically built into the project designs AND students acquire attitudes that take account of client need and project specificity in always striving for a sustainable solution.

Figure 2: **Clusters and Categories of Context**



CLUSTERS

CATEGORIES

C1 Commitemnt of Stakeholders to Learning Goals

Enabling:

- The **"willing to compromise"** context
- The **"prinicpled action"** context

Disabling:

- The **"way it's always been done"** context
- The **"arm's length"** context

C2 Alignment of Assessment with Learning Goals

Enabling:

- The **"will it work in the village"** context
- The **"correct assessment target"** context

Disabling:

- The **"it's all about cost"** context
- The **"will it be on the exam"** context
- The **"we don't know what they're looking for"** context

C3 Focus on Conditions of Use of Design

Enabling:

- The **"community needs"** context
- The **"allowing for difference"** context

Disabling:

- The **"cool technology"** context
- The **"just background"** context

C4 Teachers Operationalise Course Aims

Enabling:

- The **"guided activities"** context
- The **"supportive challenge"** context

Disabling:

- The **"feeding information"** context
- The **"more like a lecture"** context
- The **"go away and do"** context
- The **"this is all unrelated"** context
- The **"idiosyncratic processes"** context

C5 Use of Real World Projects

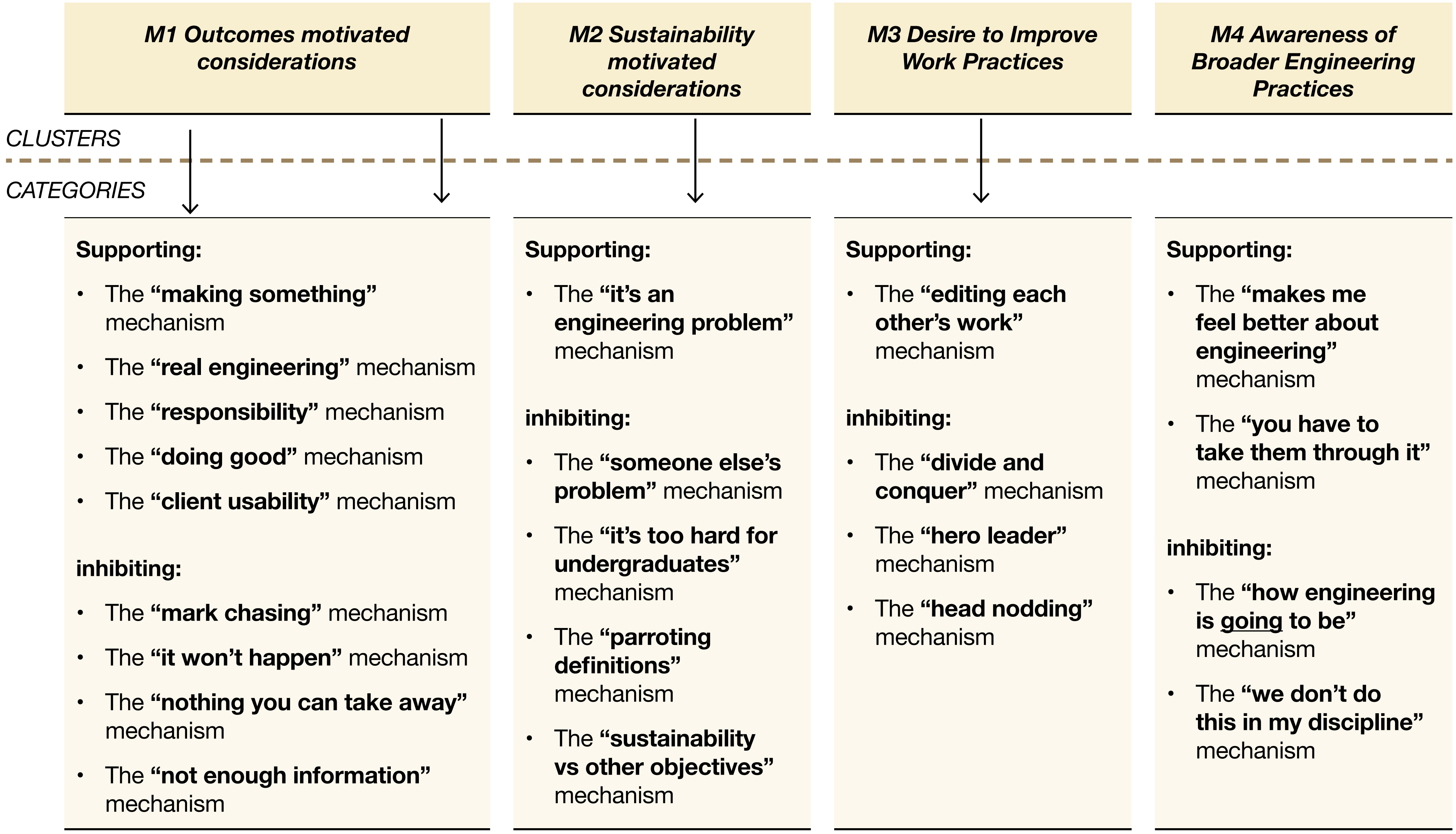
Enabling:

- The **"appropriate to level"** context
- The **"particular constraints"** context
- The **"sustained effort"** context
- The **"more than maths and physics"** context
- The **"client contact"** context

Disabling:

- The **"solve for X"** context
- The **"just content"** context
- The **"it's too foreign"** context

Figure 3: **Clusters and Categories of Mechanisms**



M1 Outcomes motivated considerations

Supporting:

- The **"making something"** mechanism
- The **"real engineering"** mechanism
- The **"responsibility"** mechanism
- The **"doing good"** mechanism
- The **"client usability"** mechanism

inhibiting:

- The **"mark chasing"** mechanism
- The **"it won't happen"** mechanism
- The **"nothing you can take away"** mechanism
- The **"not enough information"** mechanism

M2 Sustainability motivated considerations

Supporting:

- The **"it's an engineering problem"** mechanism

inhibiting:

- The **"someone else's problem"** mechanism
- The **"it's too hard for undergraduates"** mechanism
- The **"parroting definitions"** mechanism
- The **"sustainability vs other objectives"** mechanism

M3 Desire to Improve Work Practices

Supporting:

- The **"editing each other's work"** mechanism

inhibiting:

- The **"divide and conquer"** mechanism
- The **"hero leader"** mechanism
- The **"head nodding"** mechanism

M4 Awareness of Broader Engineering Practices

Supporting:

- The **"makes me feel better about engineering"** mechanism
- The **"you have to take them through it"** mechanism

inhibiting:

- The **"how engineering is going to be"** mechanism
- The **"we don't do this in my discipline"** mechanism

The nature of the data

We examined a wide range of data in considering the merits of the hypothesis represented in Table 3 about how this intervention improved students' understandings of sustainability, but we have room for only a very short selection of that data.

Firstly, we note that although course designers and instructors were inclined to include sustainability in their aspirations for student work, it was not always made explicit in course learning objectives and classroom activities. Where published course objectives did mention sustainability, it was more often than not in terms such as, "Appreciate the socio-cultural, political, environmental and economic contexts in which engineering is practised". Thus, it did not seem that students were expected to gain a very deep understanding of sustainability, and the most explicit classroom discussions of the concept we observed were reiterations of standard definitions such as that from the Brundtland Report: "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (World Commission on Environment & Development, 1987) without extended discussion of the concept. In fact one expert in the field told us that sustainability was too hard to teach to undergraduates.

Nevertheless, when asked in a survey to rate their confidence in accounting for issues of sustainability in their decision-making 81% of students professed themselves confident or very confident. At interview, however, most students we spoke to found it hard even to reiterate the Brundtland definition and offered formulations such as:

> Before I only knew about environmental sustainability and I thought that that was the only sustainability there was, but now I know the triple bottom line thing and that it has to be socially sustainable as well, and financially and environmentally, because you can't just make something environmentally sustainable and make it really, really expensive because people aren't going to use it. (Focus group, Go8A)

There was evidence of superficial approaches to sustainability in the project reports also, with many students reiterating the same stock phrases in the background section of the report and then not appearing to pay very much attention to it in the detailed design solution. We also observed in classroom observations that tutors and instructors were inclined, perhaps not surprisingly, to emphasise more technical and traditional aspects of the project, as reflected in the following exchange between an observer and a student:

> "Researcher: Why did you choose to build in material with a high embodied energy?
>
> Student: Yeah well we should think about sustainability and all that but it's all about cost really." (Observation notes, Go8B)

Students appeared to be strongest in their understanding of and commitment to sustainability where they could see it as a legitimately engineering (that is, technical) problem.

The observed outcome therefore seemed to be that at least some students learned something about sustainability but applied no critique to the concept (and therefore may not be able to transfer it to other problems) and may be inclined to consider it something to think about after all other factors (such as cost) were served.

Deciding the context/mechanism question

Empirical investigation not only articulated actual outcomes but also described some of the reasoning (on behalf of both students and instructors) that produced them, and described the contexts (socio-cultural conditions) in which they happened. However, Contexts and Mechanisms cannot just be observed, and different kinds of theory are needed to help us in analysis. Here we had a number of different theories in play for different purposes.

First, there were the participants' theories of how the intervention should work and this is important for explaining why they did the things they did and for indicating potential reasons for success and failure. With this in mind, we used in vivo labels for our categories to capture something of this level of theory, although this presented some difficulties. For instance, we saw many instances of classroom teaching, tutoring and resources which de-emphasised sustainability concerns in favour of factors such as cost. We chose to label these situations the "all about cost" context, although it could be argued that when a student tells us that, it is evidence of their reasoning about how to respond and thus should be a mechanism.

However, we also saw that almost universally students went through a process of weighing up sustainability against other objectives and could speak fluently about that process. We therefore chose to identify the "sustainability vs other objectives" as a significant mechanism producing the observed outcomes. This allows for a challenge to and modification of the participant theory represented in Table 3 that the project setting alone will result in attention being paid to sustainability. At the same time, our choice of context and mechanism was driven by higher order theories of curriculum and teaching such as Biggs' (1996) constructive alignment, which states that all of the elements of curriculum from course design and objectives to classroom activities to assessment criteria need to be coherent to provide the desired result.

Each university could now consider their use of the EWB projects according to the contexts and mechanisms in operation there. Table 4 represents just two possibilities.

Implications

While our theory of curriculum derived ultimately from educational theory, it matched well with the views of stakeholders revealed through program logic analysis, about how the intervention ought to work in their setting. However, program logic approaches (Chen, 1990; Den Heyer, 2001; Rogers, Petrosino, Heubner & Hacsi, 2000) and the closely related program theory formats (Funnel, 2000; Shadish, Cook & Leviton, 1991; Weiss, 1996) tend to concentrate on inputs and outputs and on identifying moments when a project might be monitored, rather than making explicit the factors that people involved understand to be influential in making the project work or not. It is this understanding that our model of curriculum embodies; it includes every factor that was mentioned in literature and by participants as part of the operation of the curriculum, and it thus provided a set of focusses which helped to identify contexts and mechanisms, and where and how they applied. The implication is then that such a model of operation of the intervention can be helpful to any CMO analysis. As we have shown, it does not however necessarily help in the allocation of phenomena to either context or mechanism.

Table 4: **Actual CMO configurations (CMOcs) with respect to sustainability**

Context (C)	+	***Mechanism (M)***	=	***Outcome (O)***
• Assessment and classes are presented in such a way as to reward results that have little to do with sustainability (the **“all about cost”** context C2) and which represent the engineer’s job as devising new widgets (the **“cool technology”** context C3). • Assessment weights sustainability against other criteria appropriately (the **“correct assessment target”** context C2) and engineering is represented as **“more than maths and physics”** context C5.		• Students allocate time and effort to acquiring technical skills (the **“real engineering”** mechanism M1) before those associated with sustainability (the **“sustainability Vs other objectives”** mechanism M2. • Students allocate time and effort to understanding and incorporating sustainability considerations (the **“its an engineering problem”** mechanism M2).		• Superficial treatment of sustainability considerations. • Realistic understanding of role and requirements of sustainability in design.

Conclusion

What stands out most to us from Pawson and Manzano-Santaella's (2012) discussion of the difficulties of applying the realist approach, in light of our own experience of this method, are the following issues. First, formal theory has a role to play in refining the program theories held by participants and exemplified as a CMO configuration (CMOc) in Table 3 above. Formal theory can help articulate *what* is to be explained (Tilley, 2009). In our case a theory of curriculum helped us focus on aspects of process as well as content. However, using this theory in concert with participants' own theories allowed a methodological flexibility for understanding the data (and how it functions in a causal CMOc) according to different moments of analysis. Thus, the understanding of instructors about how improved sustainability considerations could come about were influential in our decision to call "all about cost" a context rather than a mechanism. It allowed us to draw attention to contextual factors they could change fairly readily. In short, we agree with Pawson and Manzana-Santaella (2012, p.189) that Realist Evaluation needs program theory but we have argued that formal theory can help us make decisions about where in the CMO sequence any particular phenomenon fits. However, they are also right to point out that "programmes do not come in pre-ordained chunks called contexts, mechanisms and outcomes" (Pawson & Manzana-Santaella, 2012, p. 189) and it is not helpful to use formal theory to make it seem as though they do.

Acknowledgements

The evaluation discussed in this paper was a project of the Australian Learning and Teaching Council (PP2010-1647) called *Curriculum Renewal in Engineering Through Theory Driven Evaluation,* headed by Professor Caroline Crosthwaite of the University of Queensland. We also wish to thank two anonymous reviewers for thoughtful comments which helped refine our thinking.

References

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32,* 347-364.

Chen, H. (1990). *Theory Driven Evaluation.* Newbery Park: Sage.

Den Heyer, M. (2001). *The Temporal Logic Model.* Ottawa, Canada: International Development Research Centre.

Dillon, J. T. (2009). The questions of curriculum. *Journal of Curriculum Studies, 41*(3), 343-359.

Funnell, S.C. (2000, Autumn). Developing and using a program theory matrix for program evaluation and performance monitoring. *New Dimensions for Evaluation, 2000*(87), 91-102.

Jolly, L. (2014). *Curriculum renewal in engineering through theory-driven evaluation,* Sydney: Office for Learning and Teaching. Available at: http://www.olt.gov.au/resource-curriculum-renewal-engineering-through-theory-driven-evaluation

Jolly, L., Crosthwaite, C., Brodie, L., Kavanagh, L., & Buys, L. (2011). *The impact of curriculum content in fostering inclusive engineering: data from a national evaluation of the use of EWB projects in first year engineering.* Paper presented at the National Conference of the Australasian Association for Engineering Education, Fremantle, Western Australia.

Markiewicz, A. (2010). *Monitoring and Evaluation Core Concepts.* Professional Development materials used in training workshops for the Australasian Evaluation Society.

Pawson, R. (2006). *Evidence-based policy:* A realist perspective. London: Sage.

Pawson, R., & Manzano-Santaella, A. (2012). A realist diagnostic workshop. Evaluation, 18(2), 176-191.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage.

Richards, L. (2005). Handling Qualitative Data. London: Sage.

Rogers, P.J., Petrosino, A., Heubner, T.A., & Hacsi T. A. (2000, Autumn). Program theory evaluation: practice, promise and problems. *New Dimensions for Evaluation. 2000*(87), 5-13.

Shadish, W., Cook, T.D., & Leviton, L.C. (1991). *Foundations of Program Evaluation.* Newbery Park: Sage.

Sochacka, N. W. (2011). *Realistic Analysis of Socio-Technical Interventions in the Context of Urban Water Management.* PhD dissertation at the University of Queensland, Australia.

Tilley, N. (2009). What's the "What" in "What Works? Health, Policing and Crime Prevention. *Crime Prevention Studies, 24,* 119-143.

University of Wisconsin. (2010). *University of Wisconsin-Extension, Program Development and Evaluation Model.* Retrieved from http://www.uwex.edu/ces/pdande/

Weiss, C.H. (1996). Theory based evaluation, past, present and future, Paper presented at the Annual Meeting of the American Evaluation Society, Atlanta, CA.

World Commission on Environment & Development, (1987). *Our Common Future.* Oxford, England: Oxford University Press.

The case for experimental design in realist evaluation

Andrew Hawkins

ARTD Consultants

andrew.hawkins@artd.com.au

Keywords: realist evaluation, experimental evaluation, realist experimental design, transfactual mechanisms, demi-regularities, effect size, testing CMO (Context-Mechanism-Outcome) configurations, realist cost benefit analysis, youth mentoring

Abstract

This article argues the realist critique of experimental design to evaluate interventions in complex social systems is valid but incomplete. It argues for experimental approaches to testing realist theory and for estimating effect sizes.

The paper aims to provide a means for scientific understanding of the relative value of interventions in different contexts, for whom and to what effect. The paper is grounded in a realist philosophy of science and a realist approach to evaluation. It argues for the use of experimental design to test and estimate the magnitude of an outcome in a hypothesised realist Context-Mechanism-Outcome (CMO) configuration. The approach requires that program theory (rather than the program) is the unit of analysis. It also requires that context – crucial for a mechanism firing – is brought into the effect size equation, while at the same time attempts are made to control for the effects of other mechanisms.

The focus of this paper is on the general approach rather than a particular method. The approach was applied in an evaluation of a youth mentoring program. The method used was a matched-pair, pre and post-test, control group quasi-experimental design. The results of our application of the approach were limited but provided insight about the extent to which a particular mentoring mechanism, when properly targeted, could generate outcomes for certain students.

This approach to evaluation is consistent with underlying principles of scientific realism and theory testing and provides a means for generating evidence about the value of interventions in complex social systems, for whom and to what extent.

Introduction

Evaluators and policy makers are both concerned with understanding how to design and target interventions for maximum effect and understanding the relative worth of interventions. In many fields of research, “the controlled experiment is king” (Jeffery-Evans, 2012, p. 26) and the randomised control trial (RCT) is considered the gold standard, the top of the evidence hierarchy or the “the best way of determining whether a policy is working” (Haynes, 2012, p. 4). The concern here is with the relative worth of whole programs, about answering the question ‘does this program work?’ and ‘to what extent?’ While there is growing interest in the concept of mechanisms across many areas of social science (Astubry & Leeuw, 2010, p. 363), experimental evaluators tend to better meet the demands of policy makers for summative evaluations and cost benefit analyses of particular programs.

The realist does not ask ‘what works?’, but ‘what works for whom, under what circumstances and how?’ or more fully “what works, how, why, for whom, to what extent [emphasis added] and in what circumstances, in what respect and over what duration?” (Wong, Greenhalgh, Westhorp, Buckingham & Pawson, 2013). Realist evaluators are concerned with program theory and Context-Mechanism-Outcome configurations rather than entire programs. Realists investigate the contexts in which mechanisms – lying within people and society or introduced in interventions – fire to generate observable outcomes. To realists, programs work due to their effects on mechanisms or context. When considering the ‘what’, the realist position is that programs are not stable, single entities emitting some steady force for change (Pawson 2013, p. 48). “Mechanisms are the agents of change. They describe how the resources embedded in a programme influence the reasoning and ultimately behaviour of programme subjects” (Pawson, 2013, p. 115). Programs may also work because they address the context, or social structures that affect mechanisms (Astbury & Leeuw, 2010, p. 370).

This article aims to demonstrate how a policy maker, who might by sympathetic to a realist approach for understanding programs, but who will ordinarily look to an RCT for evidence of outcomes, can instead use an experimental design with a focus on program theory, to provide evidence about the value of an intervention, for whom, under what circumstances and to what extent.

While taking a realist approach, the article recognises the need for science to include the construction of falsifiable theory put to the test by experimentation (Popper, 2005) and by extension, for a realist scientific evaluation to test CMOs. This is not simply about sub-group analysis, identifying who seemed to benefit most and least in a pattern of results – either in an experimental data or a realist intra-program analysis. It is the view of this paper that a scientific, useful or portable realist CMO should be ‘transfactual’ that is, it should say something about the way the ‘real’ world operates outside a particular program or dataset.

This paper argues that experimental methods can be used to test theory, estimate the magnitude of an outcome in a hypothesised CMO configuration and assess the relative merit of an intervention for different target groups. The approach means shifting the focus of experimental analysis away from the program or intervention towards program theory. Specifically it means making realist CMO configurations, rather than a program or intervention, the unit of analysis for experimentation.

While experimentation in science is much broader than the use of control groups, this is the most commonly used experimental method in the social sciences, and while it has limitations is the one used in this article. The particular method described used was a matched-pair, pre and post-test, control group design. This is by no means the only or best means of applying the general approach, but it was feasible in the evaluation where we sought to combine realist and experimental approaches to evaluation.

The key point of the article is that if a CMO is an important and useful description of the world, then – despite any shortcomings of experimental design using control groups – we should be able to test it and observe a regular outcome pattern in most instances when we observe a C (Context) and M (Mechanism) together – ideally with, but even if we don't have evidence of the M firing. If we do not observe a regular outcome pattern on a sufficient number of occasions, we may need to refine or abandon our hypothesised CMO.

The paper also addresses a key question often asked by those commissioning evaluations of public policy, one that we expect will still be asked even if realists are successful in shifting policy makers from focusing on 'what programs work?', to asking 'what works for whom, and in what circumstances?'. This is the question of 'how big was the effect?', or 'how big an effect can we expect if we leverage this CMO in the future?'

The realist case against excluding context from experimental evaluation of interventions into complex social systems

The question for experimental design as applied to the evaluation of public policy and programs is often of the nature 'what works and to what extent?' The unit of analysis is most often the program. Typical uses of experimental design for evaluation of interventions into social systems involve a treatment group, which receives an intervention, and a control group, which is supposed to be equal to the treatment group in all factors except exposure to the intervention. The means for ensuring the equivalence of treatment and control groups is either the random allocation of a sufficient sample of participants or the matching of participants in both treatment and control groups through a quasi-experimental method, such as propensity score matching. The outcome is then measured as the difference between treatment and control groups on some key variable of interest after the intervention. Since other factors have been 'controlled for', this outcome is attributed to the impact of the intervention.

Context is often treated as a confounding variable, and attempts are made to 'control for' the impact of context on outcomes through design or statistical analysis. More recently, proponents of experimental design have identified context as something to consider when judging whether the results of a trial are applicable to populations other than those participating in the trial. However, the focus of analysis is still on the context-free impact of a mechanism or intervention. For example, in a collection of works for translating health research to public policy, reference is made to "understanding context-based factors that will have an impact on the success of interventions" (Wethington & Pillemer, 2012, p. 4) as an opportunity to contribute to translational research. The implication is clear; context is a factor that affects all aspects of an intervention rather being relevant for specific mechanisms. In the same volume, Evans conceives context as something to be addressed outside the experiment, rather than a critical part of what is being put to the test. He claims "social and behavioural science can be used to provide descriptive information on the community including family, social or political context in which interventions or policies are taking place, shedding light on the contexts in which desired changes are more likely to occur and on instances in which change is more difficult" (2012, p. 28).

Realist evaluators however, view mechanism and context as inexorably intertwined: controlling for the impact of context – as experimental designs often attempt to do – a is neither useful nor possible. The firing of a mechanism is completely dependent on context. To use a famous realist example, gunpowder does not fire when it is wet. Valid experiments are never easy to conduct, and experiments have received strong but sound criticism from realists as being ill-equipped to measure changes in complex social systems. Even if random allocation could achieve equivalent groups (at least on factors deemed important to achieving outcomes) prior to an intervention, the reality of ever-changing conditions, both within and between people, and the fact that context is part of what causes an outcome, mean an RCT that seeks to control context will often miss exactly what should be understood. The external validity of experiments will be limited as long as the unit of analysis remains the program and will be problematic whenever researchers attempt to control for the effects of context, rather than embrace context as determining whether mechanisms are activated and generate outcomes.

The common role for observation in realist and empirical social science

As with science generally, both experimental and realist approaches to evaluation rely on empirical observation. In the positivism influencing much experimental design in social science, knowledge is limited to observations of events. Realism posits that a deeper reality is knowable even if there is no such thing as final truth or knowledge.

In his 1975 landmark publication, *A Realist Theory of Science,* the realist philosopher Bhaskar (2008) argued against acceptance of this limited positivist conception of the world; "because it must be assumed, if experimental activity is to be rendered intelligible, that natural mechanisms endure and act outside the conditions that enable us to identify them" (p. 2). That is, we would not do experiments if we didn't think they told us something about the world outside the experiment.

Bhaskar argued the world is stratified into the domains of the real, the actual and the observable. The real is what exists, the structures and mechanisms that interact regardless of whether they manifest into actual entities or events, and regardless of whether we observe these or not. For Bhaskar, (2008) “the real basis of causal laws are provided by the generative mechanisms of nature… [and these are] “nothing other than the ways of acting of things”… “tendances”… [or] “powers and liabilities of a thing which may be exercised without being manifest in any particular outcome” (p. 3). This means that in the complexity of the everyday world countless mechanisms are interacting in countless contexts, with the potential to lead to actual events that we sometimes observe. But real mechanisms exist even if they aren’t obvious, or act with consistent outcomes in actual (or factual) events that we observe – the real “exist independently of and are often out of phase with the actual pattern of events” (p. 2). In other words mechanisms are transfactual; they exist at a deeper level but give rise to everyday experience because “their activities are continuous and invariant, stemming from their relatively enduring properties and powers, despite their outcomes displaying variability in open systems” (Archer, 1998, p. 195). For example, positivists may seek to understand racism by measuring the regularity of ethnic minorities being passed over for jobs. For realists, the goal is to understand the mechanism of racism, which, although invisible, really exists even if it can only be observed during actual events.

As the domain of the real is not directly accessible to observation, social scientists are required to develop their understanding of reality though the observation of actual events, even though they are concerned with the underlying generative mechanisms of events, or abstractions such as the reasoning of program participants, that are not directly observable. Post-positivism may have somewhat bridged the gap between realism and positivism by accepting realist ontology, including Bhaskar’s argument about the intelligibility of experimentation, but methods of experimentation in program evaluation have not followed suit. Positivists tend to prefer to focus on the intervention as the unit of analysis and maintain fealty towards the ideal of invariances in experimental data – summed up in Hume’s famous phrase “the constant conjunction of events” (Bhaskar, 2008, p. 3). While there is acceptance of variation in data on outcomes in experimental evaluation due to the impossibility of controlled experiments and a reliance on randomised controlled trials, this is generally considered ‘noise’ that hides the true impact of an intervention rather than integral to the theory of how something works.

Making use of experimental designs in realist evaluation

Scientific enquiry often involves developing and testing theories using experiments. A true experimental design should test a hypothesis, not an intervention. While most program evaluations using experimental design seek to test whole programs or interventions, rather than program theory, this is a problem with the application of experimentation rather than experimentation per se. These types of evaluation are referred to disparagingly as ‘black box’ evaluation (Funnell & Rogers, 2011, p. 4). An experiment, used properly, provides a means of testing whether a theory can say something useful about the way an intervention works. Problems of external validity (i.e., how likely the result of an experiment is to apply in the real world) will occur whenever the program rather than theory is the unit of analysis.

Realist evaluators argue that it is impossible to isolate and measure the impact of an intervention, or of individual mechanisms not only because of the complexity of their interactions, but because it is the interaction of context and mechanism that generates outcomes. The unit of analysis is not the mechanism, or the context but the Context-Mechanism-Outcome configuration (Pawson & Tilley 1997 p. 217)[1] . In many cases, the complexity of social systems requires us to engage in developing middle-range theories as per Merton (1949):

> Middle-range theory is principally used in sociology to guide empirical inquiry. It is intermediate to general theories of social systems which are too remote from particular classes of social behaviour, organization, and change to account for what is observed and to those detailed orderly descriptions of particulars that are not generalized at all. Middle-range theory involves abstractions, of course, but they are close enough to observed data to be incorporated in propositions that permit empirical testing (p. 39).

Realists are concerned with middle-range theories and outcome regularities, demi-regularities or simply "demi-regs" (Pawson, 2010, p.185). But it is not enough for realists to simply hypothesise or develop more sophisticated theory "it must be possible for an empirical scientific system to be refuted by experience" (Popper, 2005, p. 18). Realists require a means by which Campbell's "mutually monitoring disputatious community of truth seekers" (Pawson, 2013, p. 192) can adjudicate disputes about the value of different interventions by testing hypothesised CMO configurations. This paper argues that realists of the Pawson and Tilley (1997) school, i.e., excluding critical realists, should aim to observe outcome patterns in contexts where mechanisms are hypothesised to fire by making predictions and testing them. If a theory cannot be tested in the observable world, it will struggle to be accepted as scientific. Interventions and their mechanisms may be difficult to define precisely; theories may be 'middle range'; and the observations may be patterns rather than constant conjunctions. However, conducting this work and gradually accumulating knowledge is all part of the slow, painstaking "informed guess work" of Popper's approach to science (Pawson, 2013, p.192).

Limits to intra-program comparison for testing theory

Testing hypotheses is not the same thing as looking at the pattern of results of an evaluation and making intra-program comparisons to develop theories about what works for whom, in what circumstances, and how. A dataset from a particular evaluation may be used to construct a theory of what 'worked' for whom under what circumstances, but not for testing a theory of what 'works' for whom in what circumstances – of transfactual CMO configurations. The danger of relying on

[1] This article like much of realist evaluation sidesteps the issue of the 'emergence' of outcomes from the interaction of psychological agency and/or sociological structure (Archer, 1998, p. 356) by seeking to engage our understanding at the point of interaction using CMO configurations as the unit of analysis.

data about what happened in a program is ‘over-fitting’ the data – “the most important scientific problem you’ve never heard of” (Silver, 2012 p 166). This happens when we seek to explain something by looking for patterns in a particular set of events that does not in fact explain anything about the underlying reality that caused them. Realist analysis that relies on fitting CMOs to data as well as experimental design that has insufficient attention to theory will fall short of a scientific means of measuring the impact of interventions into complex systems. As Nate Silver (2012), comments:

> What happens in systems with noisy data and underdeveloped theory – like earthquake prediction and parts of economics and political science – is a two-step process. First people start to mistake the noise for a signal. Second this noise pollutes journals, blogs and news accounts with false alarms, undermining good science and setting back our ability to understand how the system really works. (p. 162)

There is currently no perfect solution to the question of how to measure outcomes of interventions into complex social systems, but the imperative for realists to put their theories to the test is so strong that they should manage the deficits in experimental design instead of abandoning them. In any social experiment, there will be many things affecting outcomes in addition to a hypothesized CMO. However, if the CMO does explain something about the world, and it is of sufficient importance to be worthy of scientific study, then we should be able to observe patterns in outcome data as a result of an intervention which changes context, or introduces new reasoning or resources.

Realists could employ experimental designs as a means of testing theories and providing evidence of the demi-regular outcomes of a hypothesized CMO. Just as a large effect size can be identified in a small sample, a sufficiently useful and transfactual CMO will overcome the problems of dynamic systems and non-linear outcomes by being associated with demi-regular outcomes. If we do not observe a regular pattern of outcomes on a sufficient number of occasions, we may need to refine or abandon our hypothesised CMO as a useful concept for understanding the world and designing future interventions.

Experimental design for estimating effect sizes

Realist decision makers may be willing to accept the realist logic that interventions comprise mechanisms that work in different contexts. They may find a CMO built with intra-program comparisons compelling and useful for program design and targeting; and may be willing to accept that different interventions are required for different people. However, in order to make decisions about the relative merit and cost-effectiveness of the programs they administer, they will require an answer to the question: ‘how substantial was an outcome when the mechanism fired?’ An experimental design with a control group provides some ability to measure the independent impact of a mechanism firing in context apart from the impact of the countless other mechanisms affecting observed outcomes in open systems.

In the case discussed in this paper we attempted to control for the effect of mechanisms on outcomes by the use of a treatment and control group design. The difference from many forms of experimental design was avoiding randomisation to control for context. Randomisation does not allow the theoretically important contextual factors to be brought into the equation. Instead we used a matched-pair approach to bring context into the experiment. Both the target and the control groups were constructed to have similar initial conditions, to the extent possible, in terms of the contextual factors deemed important for firing the theorised mechanism. In this case, psychological wellbeing and resources were measured using psychometric scales. Crucially, we expected that change over the period of measurement (i.e., the school year) would be affected by many things outside the intervention, such as students' natural maturation and many other mechanisms at home and school. Our method allowed us to estimate how much of the change in wellbeing was due to the mentoring mechanism by focussing on the difference in the amount of change for treatment and control group students. Estimates of the size of an O in a CMO that do not seek to isolate the effects of different mechanisms on the O may mistake changes that result from many mechanisms as evidence about the particular mechanism within a CMO.

A realist quasi-experimental design for an evaluation of youth mentoring

This section demonstrates how we used a realist approach to experimental design for evaluating a youth mentoring program and estimating the size of an outcome that could be attributed to a mechanism within a CMO.

The mentoring program allocated community volunteers as mentors to students at risk of disengaging from primary or secondary school. The physical context in which the program was delivered was quite consistent: one hour once a week on school grounds. Generally speaking, the mentors did 'fun' things with their mentees – such as cooking and playing games – as a way of developing a trusting adult – student bond. The program aimed to increase students' psychological wellbeing and levels of important psychological resources, resilience, optimism, social skills, love of learning, as well as increase school attendance and reduce problematic behaviour.

The objectives of the evaluation were to measure the outcomes and identify how they could be maximised. One important question for policy makers was, 'when the type of mentoring we have on offer is provided to the type of students we think stand to benefit most, how much do they benefit?'

The evaluation used a realist synthesis of mentoring conducted by Ray Pawson (2004) and case studies to identify potential CMOs. These focused heavily on understanding what aspects (or mechanisms) of the mentoring program worked for which students and how. We only found one relevant CMO from this theory-building stage (see Figure 1), in part because the type of mentoring the program offered was limited to the mechanism of 'affective contacts' or developing emotional resources. Other forms of mentoring, such as advocacy, coaching or directing setting, were not included (Pawson, 2004).

Figure 1: **The key Context Mechanism Outcome (CMO) configuration identified in the evaluation**

Context: Students with low self-esteem/self-confidence or poor social skills or low resilience, often manifested as shyness or acting out in class i.e. low engagement in school.

Mechanism: Regular and consistent one-on-one time with a trusted and respected non-judgemental adult who engages in activities directed by the mentee. This activates empowerment and increased self-esteem, confidence and social skills.

Outcome: Increased feelings of self-worth, i.e., self-esteem (and sometimes increased resilience and optimism) leading to observations of greater self-confidence, (and sometimes improved peer relations) and greater engagement in the classroom.

We then used a quasi-experimental method of observing changes in pre and post-intervention measures of psychological wellbeing for students allocated to a mentor, compared to a control group of similar students without a mentor. To provide a control group for the evaluation, schools were asked to nominate about double the number of students who they thought could get a mentor. A mentor was allocated when one became available and was matched to a student in the nominated group with whom they seemed to be a good fit (not necessarily the student in greatest need). It was a relatively ad hoc process. In this kind of matched-pair design students who received a mentor were in the treatment group; those who did not, were in the control group.

Measurement of outcomes of youth mentoring

We sought to measure the outcomes of mechanisms in the contexts (i.e., students with a particular psychosocial profile) from which the theory-building stage of the project led us to expect to generate outcomes i.e., our CMO[2]. The approach was in essence very simple. We used statistically matched pairs to test whether those students with lower levels of self-esteem and poor social skills who were allocated a mentor achieved greater gains in psychological resources compared to the control group of students (with similar initial levels of these psychological resources) who did not get a mentor.

What we achieved was a measurement of the effect size when we expected a mechanism had fired. If we had looked at the effect size when a mechanism actually 'fired', it is likely the estimated effect would have been larger[3]. This approach is analogous to analysing outcome data by whether someone was allocated to get a treatment (effectiveness) or actually got the treatment (efficacy). Our approach was closer to this latter intention-to-treat analysis.

We used ANOVA and *t*-tests and Cohen's *d* (Cumming, 2012) to identify and measure the effect of the mentoring mechanism in context. The results were statistically significant and large. The effect size of mentoring for the students in the treatment group whom we expected to benefit relative to similar students in the control group was large. Cohen's *d* on the net differences between treatment and control students on psychological outcome measures ranged from 0.71 to 1.38 which was significant using a one tailed *t*-test[4]. The biggest limitation for the evaluation was the small sample size. Despite planning to have matched data from over 200 students, by the end we had 143 pre-treatment surveys, and were only able to match these with 93 student post-treatment surveys of which only 13 could be matched to the control group condition. Data from other sources[5] and the indicative statistical findings supported the hypothesized CMO, but a larger sample size would be required to provide sufficient occasions to observe demi-regularities[6].

We presented data using point estimates with probability estimates (i.e., *p*-values). In hindsight, it would have been better to provide point estimates with confidence intervals. Not only are confidence intervals recognised as the best approach to reporting statistics (Cumming, 2012); the approach fits with the realist task of identifying demi-regularities. Probabilities rather than precise measures, which are judged to be either significant or not significant, are suggestive of constant conjunctions and expectations for invariant outcomes.

The findings of the evaluation were that the mentoring intervention did not work for everyone, but it worked very well for a small subset of students, those with low self-esteem and poor social skills. This was because the form of mentoring available focused on providing students with emotional resources. So while mentoring was 'fun' for nearly all students (in part because mentoring involved activities directed by the student as an alternative to being in the classroom) the particular type of mentoring provided did not meet the needs of most students.

[2] We were also required to measure the impact of the intervention as a whole. We found that while almost all mentors and students enjoyed participating and generally developed trusting and respectful relationships, and that all students improved on all measures used in the evaluation over the school year, as a cohort those who had a mentor did not improve more than control group students without a mentor.

[3] To increase the validity of the findings, we planned to observe cases where both mentor and mentee felt that mentoring had occurred as planned and actually helped the mentee (i.e., the mechanism fired), but the data was insufficient to allow us to match data collected from mentors with their mentee.

[4] Tests were significant with alpha set at 0.05, but the small sample size meant not all outcomes were statistically significant. However, it is uncertain whether the convention of setting alpha at 0.05 has as much relevance outside the laboratory for evaluations of complex social systems using mixed methods where only demi-regularities rather than constant conjunctions are expected to hold.

[5] Many teachers and school principals interviewed were 'lukewarm' in their support for the program and most had a number of examples where they believed mentoring did not lead to any change or benefit for a student but frequently reported that those that students with poor self-confidence or social skills benefited from the program.

[6] This paper is about an approach and a particular method we used to implement that approach – it does not make any claims about the effectiveness of mentoring as a result of the statistical data obtained.

Limitations in the statistically matched-pairs method

This paper is about an approach rather than a specific method, yet there were two main limitations in our evaluation relative to the ideal of matched-pair treatment and control group design. First, the resilience outcomes we measured were not exactly the mechanisms that were hypothesised to be at work (self-esteem and self-efficacy), although they were closely related[7]. With a separate theory development and experimental design phase, we may have taken separate measures of the slightly different hypothesised psychological mechanisms and outcomes. However, the biggest challenge was doing both theory building and theory testing within a single project. We had some idea of the mechanisms of mentoring from the realist synthesis by Ray Pawson (2004). Interviews with frontline program stakeholders (mainly teachers and principals at 15 schools) addressed process issues and identified what aspects of mentoring they thought worked for which students and why. We used both sources to draft CMOs. Luckily, the time needed to measure outcomes (using a standard pre and post-intervention measurement for treatment and control group students) allowed us to continue to work on theory development. It was fortunate that some of the outcomes we sought to measure were also mechanisms. This meant data on these had already been collected in pre-tests. For example, self-esteem could be a mechanism (generating other wellbeing outcomes), a context (low or high self-esteem as a starting point), and an outcome (more or less).

Second, the experimental design would have been better had we – put the CMO to the test by only providing the intervention to those students whom the theory predicted stood to benefit, but we did not have a good theory about this until halfway through the project. It would also likely have been contested by teachers in the absence of ‘evidence’. This limit to hypothesis testing raised a potential criticism that we were over-fitting the data, as might happen when developing CMOs using intra-program comparison data. What we achieved in this specific evaluation, as in many messy real world evaluations, fell short of the ideal. By using our statistically matched-pairs we made theory, rather than the program, the unit of analysis and we attempted to bring the context necessary for firing a mechanism into equation, while controlling for the effect of other mechanisms, to estimate the effect size of an outcome in a CMO configuration. The point of this paper is not that we achieved this flawlessly, but that we demonstrated how it may be achieved, in a way that is consistent both with realist theory and the principles of experimental design.

Implications for the mentoring program

The results of the evaluation suggested two main options for decision makers: develop the program so that different types of mentoring could be made available based on student needs, or target the program to students with low self-esteem and poor social skills. The first option would have been difficult as the program relied on local community volunteers for its supply of mentors, so the second was emphasised. In practice, this meant the program was only worth running in schools with a substantial number of students that fit the target group identified by the evaluation. Decentralisation of funding for student welfare programs meant school principals could decide whether or not to use their resources to fund the mentoring program or some other intervention that might better meet the needs of their students.

Potentially, with a measure of the effect size of an outcome from a CMO compared to the effect size of outcomes from other CMOs, realist cost benefit or cost effectiveness analysis may be possible. This might estimate which intervention is likely to generate the greatest benefits (or effect sizes) given the cost of the intervention and the context in which it is to be deployed.

Conclusion

Realist and experimental approaches to evaluation both involve theories about the world developed and tested through observation. When testing theories, empiricists often seek to use an experimental design that takes the impact of context out of the equation. Realists consider context as crucial to their theories and seek to observe contexts when mechanisms fire to generate outcomes. The problems of complexity and the threats to external validity of experiments in open systems are real. The practical difficulties of implementing a realist experimental design are significant. However, it is the position of this paper that if a realist CMO is worthy of scientific study, and is to inform decision making about the relative value of interventions, it should be tested experimentally and the magnitude or effect size of an outcome estimated.

This paper argues for and demonstrates a method of experimental design for testing CMO configurations. The approach makes the program theory rather than the program the unit of analysis. It brings the contextual factors hypothesised to be important for firing a mechanism into the effect size equation. A control group is used to isolate the effects of extraneous mechanisms and contexts on an outcome. It is argued that this approach is consistent with the underlying principles of realist and scientific evaluation and may facilitate more wide-spread recognition of the benefits to public policy of a realist approach to addressing questions about resource allocation.

[7] See the theory of core-self evaluations in Elliot, Kaliski, Burrus, & Roberts, (2012), p. 202.

References

Archer, M. (1998). Realism and Morphogenesis. In M. Archer, R. Bhaskar, A. Collier, T. Lawson & A. Norrie (1998). *Critical Realism Essential Readings.* London & New York: Routledge.

Astbury, B., & Leeuw, F. L. (2010). Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. *American Journal of Evaluation, 31*(3), 363-381.

Bhaskar, R. A. (2008). *A Realist Theory of Science.* London: Verso.

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge.

Elliott, D.C., Kaliski, P., Burrus, J., & Roberts, R. D. (2012). Adolescent Core Self-Evaluations. In S. Prince-Embury & D. H. Saklofske (Eds.), *Resilience in Children, Adolescents, and Adults: Translating Research into Practice.* New York: Springer Science & Business Media.

Evans, V.J. (2012). Translation in the social and Behavioural Sciences: Looking Back and Looking Forward. In Wethington E & Duniforn RE (Eds.) *Research for the Public Good: Applying the methods of translational research to improve human health and well-being.* Washington: American Psychological Association.

Funnel, S. & Rogers, P. (2011). *Purposeful Program Theory.* San Francisco: Jossey-Bass.

Haynes, L., Service, O., Goldacre B., & Torgerson, D. (2012). *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials.* Kew, London: UK Cabinet Office National Archives.

Merton, R,K. (1949). On Sociological Theories of the Middle Range, *Social Theory and Social Structure.* Simon& Schuster, NewYork: The FreePress.

Pawson, R., & Tilley, N. (1997). *Realistic Evaluation.* London: Sage.

Pawson, R. (2004). *Mentoring relationships: an explanatory review.* ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 21.

Pawson, R. (2010). Middle Range Theory and Program Theory Evaluation: From Provenance to Practice. In J. Vaessen & F.L. Leeuw (Eds.), *Mind the Gap Perspectives on policy evaluation and the social sciences. Comparative Policy Evaluation,* Vol. 16, pp.171-202). New Brunswick & London: Transaction Publishers.

Pawson, R. (2013). *The Science of Evaluation:* A Realist Manifesto. London: Sage.

Popper, K. (2005). *The Logic of Scientific Discovery.* London and New York: Routledge.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* London: Penguin.

Wethington, H. & Pillemer, K. (2012). Introduction: Translational research in the social and behavioural sciences. In Wethington E. & Duniforn RE (Eds.) *Research for the Public Good: Applying the methods of translational research to improve human health and well-being.* Washington: American Psychological Association.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J. & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC Med, 2013;* 11- 20 doi:10.1186/1741-7015-11-21.

Realist methodology in practice: translational findings from two realist syntheses

Shane Boris Pointing

The Cairns Institute,
James Cook University

boris.pointing@jcu.edu.au

Keywords: rapid realist synthesis, CCTV evaluation, homelessness evaluation framework, retrospective realist review, prospective realist synthesis

"The work will teach you how to do it."

Estonian Proverb.

Abstract

The author is currently conducting two rapid Realist Syntheses, one to identify the theoretical bases of closed-circuit television (CCTV) to reduce alcohol-related assault in the night time economy, and the other to identify dimensions of evaluation to improve the effectiveness and efficiency of a number of services in northern Australia which address homelessness and alcohol-harm reduction. The CCTV project grew out of a "completed" Realist Evaluation; the homelessness and alcohol-harm project is the foundation for a future Realist Evaluation. This paper will examine how the Realist Synthesis protocols have been applied both retrospectively, and to inform the future Realist Evaluation. Each evaluation aims to understand how specific interventions work, or don't work, using the explanatory structure of generative causation. Key findings are: that precise definitions of the programs' outcomes are crucial to retrospectively applying the Realist Synthesis methodology; that the realist methodology can embed a continuous quality improvement process in the funding organisation once these outcomes are defined, making research engagement more effective; that the outcomes (and causal mechanisms) lie at different systemic levels, both internal and external to the organisation; and that this last point is something people within the funding organisation intuitively grasp, but have had difficulty articulating.

Introduction

This paper summarises why the Realist approach to science is an important alternative to more conventional approaches in studying complex social interventions. It contrasts two current rapid Realist Syntheses, and reports on practical issues discovered in conducting each of these reviews. Examples are provided of the results to highlight how the methodology worked in practice.

The first realist synthesis (RS) is a retrospective review which attempts to refine the theoretical bases of open-space, urban CCTV systems. The second RS is a review which has been designed to inform a future evaluation of provision of social services in Cairns.

The Realist philosophy of science suggests there is a complex reality in the social world, and that complex interventions such as CCTV change the nature of this social reality at a range of different levels. It also suggests some aspects of this complexity cannot be directly measured through research methods which were designed to test the effectiveness of interventions which are applied in the same way to identical participants (Pawson & Tilley, 1997; Wong, Greenhalgh, Westhorp, Buckingham, & Pawson, 2013). Realist inquiry posits that specific ‘outcomes’ are caused by relevant ‘mechanisms’ being triggered in defined ‘contexts’. In social science these contexts are different; this will be explored below. Basically, Realism aims to understand what works for whom in what real-world context, and why. Two key methodologies in this approach are ‘Realist Synthesis’ (RS) and ‘Realist Evaluation’ (RE). Each aims to address issues of complexity in the real world.

Interventions are theories in practice (Pawson, 2003). These theories are rarely explicitly stated (Pawson & Tilley, 1997; Wong, Greenhalgh, et al. 2013). Applying these theory-based interventions relies on the actions of the people who are delivering the intervention at each stage in a chain of steps. The processes at play in each link in the chain are often not linear. They involve human beings embedded in social systems which are localised and global at the same time. Finally, interventions are open systems and change through learning. Social interventions are themselves complex systems which have been inserted into complex systems (Pawson, Greenhalgh, Harvey, & Walshe, 2005). No wonder systematic literature reviews using the Campbell protocols of review often find that the evidence of program effectiveness is mixed or conflicting, and provide few insights as to why the intervention worked or did not work when applied in different circumstances, or was implemented by different stakeholders (Pawson, Greenhalgh, Harvey, & Walshe, 2005). No wonder quasi-experimental research and evaluation designs face similar outcomes; each of these specifically aims to homogenise and flatten out differences in context (Pawson & Tilley, 1997). There is however wonder in the realist approach. It is just that it is complex — and for complex, read difficult. But the core concept of realism is simple; researchers should explicitly explain how the program theories in an intervention are supposed to work. Realist Synthesis (RS), or Realist Review, is a strategy for synthesising research which aims to disaggregate and identify the mechanisms through which complex programs work in particular social settings, or why they fail. RS aims to identify the relevant program theories.

Aiming to identify these theories is what makes RS a useful approach to summarising the evidence from across multiple studies. The methodology examines and synthesises published primary research studies on a single topic to find the main idea or ideas which inform a certain type of intervention. This is known as the 'program theory', which explains how and why deterrence works in CCTV for example. Realist Evaluation (RE) aims to understand and refine how these theory-based interventions cause outcomes in the real-world. It uses a range of quantitative and qualitative research methods to define the context in which an intervention is expected to have an impact, the mechanisms through which the intervention achieves its impact, and the outcome of introducing the intervention in that particular context.

RS and RE are both iterative methodologies, ideally identifying program theories, linking these with the chains of steps in interventions and producing outcomes which improve the life of people in the real world. This paper contrasts two current rapid RSs, and reports on practical issues which have been discovered in conducting each of these reviews.

The first is a retrospective review which attempts to refine the theoretical bases through which open-space, urban, closed circuit television systems (CCTV), which are monitored in real-time by human operators, may reduce the rate or severity of alcohol-related assaults in the night time economy[1]. This review grew out of a RE which has thus far progressed through three iterations, each focused on a different operational aspect of the system. This night time economy is in Cairns, a tropical Australian city.

The second RS is a review which has been designed to inform a future Realist Evaluation of service provision aiming to reduce homelessness and related alcohol-harm in Cairns. This review aims to broadly identify what has been found to work with clients in various age cohorts in street-based outreach, volatile substance misuse, and crisis accommodation. This improved understanding around how and why interventions work, if and where they do, will inform qualitative and quantitative research involving staff and management in a range of services which are part of one organisation.

Each synthesis has adhered as closely as possible to the RAMESES guidelines (Wong, Greenhalgh, et al. 2013; Wong, Westhorp, Pawson, & Greenhalgh, 2013) through each step of the process. In conducting the two reviews, the RAMESES protocols effectively answered process questions which arose from the research. A key mechanism in this for the author was the continued emphasis of the protocols and reasons for using a realist philosophy: "Why am I using realism to explore this question?"

As I am being so personal, some history of how and why I became a Realist may be required.

The personal process

I am a beginner in the Realist methodology, coming to research from a background in community capacity building within the public service. My research training began in the discipline of public health. Our team had conducted nearly 18 months of qualitative and quantitative field research into the problem of alcohol related assault linked to licensed premises in the Cairns NTE. Unknowingly I had been working backward from an attempt to more clearly define relevant outcomes toward an understanding of the range of contexts in which this type of alcohol related assault occur.

The original project sought to understand, through qualitative research, the ecology in which the alcohol related assaults occurred (Clough, Hayes-Jonkers, & Pointing, 2013; Pointing, Hayes-Jonkers, & Clough, 2013). The project collected and linked quantitative data from the police, the hospital emergency department, the CCTV system and other stakeholders in close-to-real-time (Pointing, Hayes-Jonkers, Stone, Brinn, & Clough, 2011). Findings were fed back to the Liquor Accord and smaller focus groups in an attempt to better inform future interventions to reduce assaults (Clough et al. 2013).

To achieve each of these aims we had to talk with people. They told us how they could link the data, and why they thought it was a good idea. They also told us what they thought the problems were and how they thought these could be addressed. These are the implicit mechanisms, which it is important to draw out, and for which Realism is such a powerful tool. For me the other mechanisms, those theories within the literature (Cherpitel, 2007; Graham & Homel, 2008; Havard, Shakeshaft, & Sanson-Fisher, 2008; Hawkins, Sanson-Fisher, Shakeshaft, & Webb, 2009; Palk, Davey, & Freeman, 2010), came later.

This original mixed methods project led to a research partnership to audit and evaluate Cairns Regional Council's CCTV system (Pointing, Hayes-Jonkers, Bohanna, & Clough, 2012). Negotiations to establish this project centred on the needs of Council to benchmark the operations of their system against international good practice. To do this we synthesised 45 peer-reviewed articles and government reports from Australia and the United Kingdom into a table of good practice, and compared the operations of the Cairns system against these benchmarks (Pointing, Hayes-Jonkers, & Clough, 2012). A reviewer of the original manuscript suggested we look at Pawson and Tilley's (1997) work on Realist Evaluation.

When I first read their summary of the Realist approach, I experienced an almost physical sensation of "clicking". The ontology of a layered structure of reality, and the epistemology of outcomes being dependent on the context in which they were sought, made intuitive and methodological sense. The article was rewritten from a realist perspective, although I still had an extremely limited understanding of what defined Realist 'context'. Defining and understanding context still presents the most challenging part of the process for me, an experience which seems to be shared by others working in my team, as well as more experienced researchers than myself (Davis, 2005; Wong, Greenhalgh, et al. 2013).

[1] For a definition of 'night time economy' and further discussion of it, see Brown (2014), also in this journal issue.

The contexts identified in that article were mapped into a logic diagram which was shown to Cairns Regional Council staff as a way to begin evaluating the effectiveness of the CCTV system in addressing alcohol-related assault in the night time economy. This is shown in Figure 1, 'Diagram of Realist Evaluation of CCTV Interventions'. Over a three year RE of Cairns Regional Council's CCTV system (Pointing, Hayes-Jonkers, & Clough, 2010, 2011; Pointing & Clough, 2013), we realised that the underlying theories of how open-space CCTV in an urban environment may reduce alcohol related assault in a night time economy had not yet been summarised into a set of Context/Mechanism/Outcome configurations. This RS aimed to specify how mechanisms associated with various dimensions of CCTV operation have generated identifiable patterns of outcomes (reduction in alcohol related assaults) in similar contextual conditions (a night time economy in a western economy).

In an initial attempt to frame a RS, these "clumps" of mechanisms and contexts were then mapped against a commentary of CCTV research (Wilson, 2008), to produce some putative theoretical domains of how CCTV might work. This is shown in Figure 2, 'Putative Diagram of CCTV Theory Domains'.

The process of undertaking two realist syntheses is described below, and the experiences gained in attempting to faithfully follow and apply the RAMESES *Realist Reviews Quality Standards and Realist Training Manual* in order to produce high quality reviews are described. This includes a commentary on how the *Quality Standards* table was applied in practice for both RS. Firstly, the retrospective CCTV review process is outlined. This project used the RAMESES documentation to guide an analysis and synthesis of 45 primary studies examining the effect of CCTV on crime reduction (Welsh & Farrington, 2009).

Figure 1: **Diagram of Realist Evaluation of CCTV Interventions**

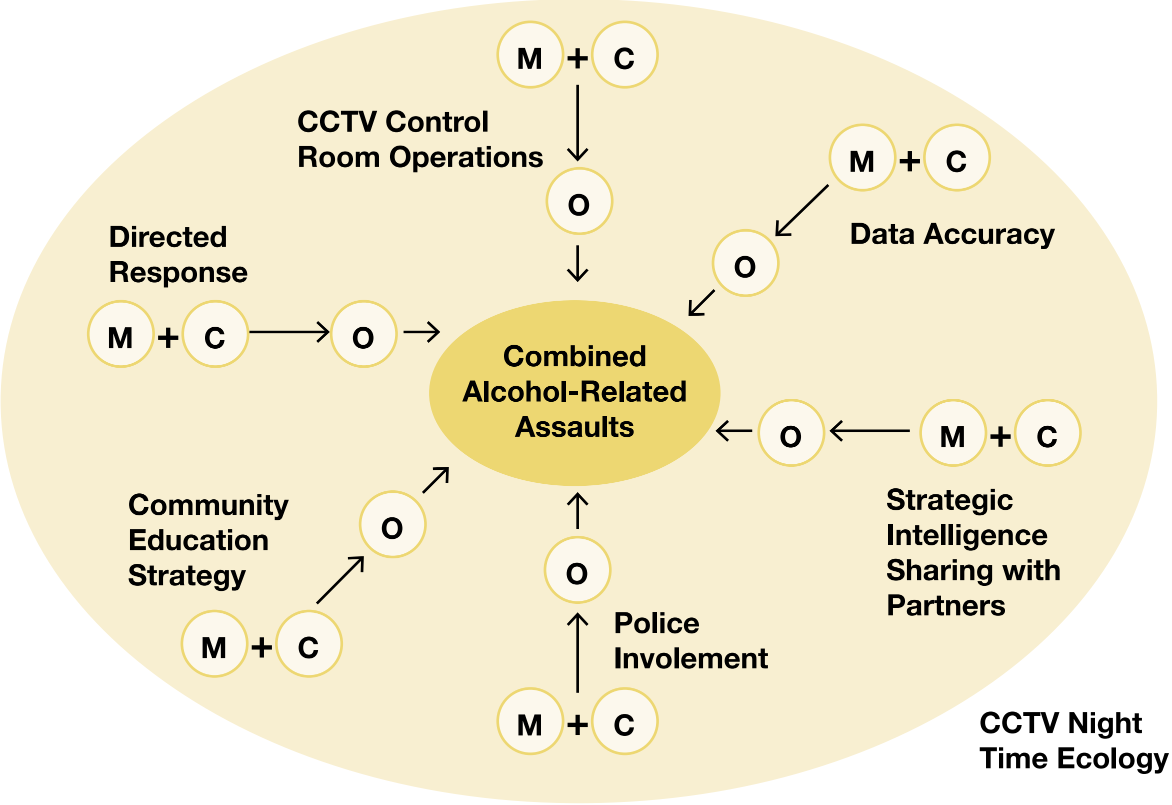
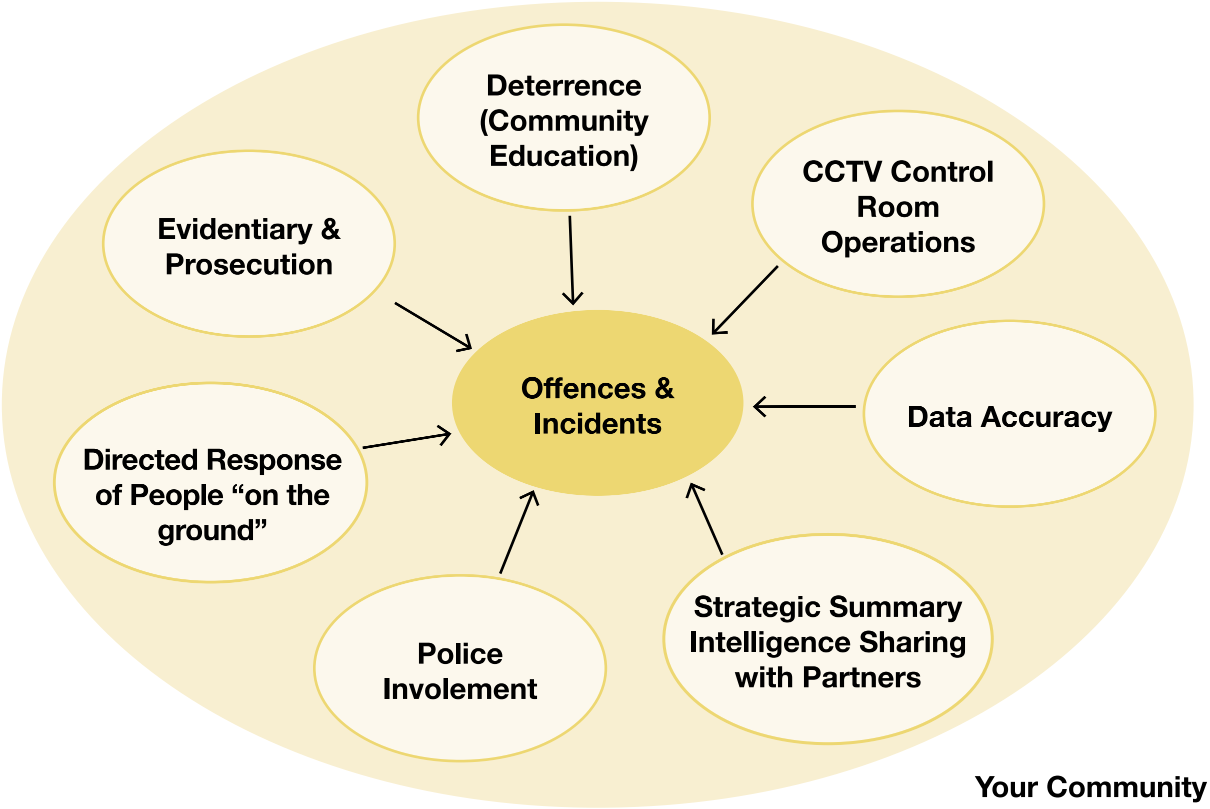

M + C
CCTV Control Room Operations
O
M + C
Data Accuracy
O
Directed Response
M + C
O
Combined Alcohol-Related Assaults
O
M + C
Strategic Intelligence Sharing with Partners
Community Education Strategy
O
M + C
O
Police Involement
M + C
CCTV Night Time Ecology

Figure 2: **Putative Diagram of CCTV Theory Domains**


Deterrence (Community Education)
CCTV Control Room Operations
Evidentiary & Prosecution
Offences & Incidents
Data Accuracy
Directed Response of People "on the ground"
Police Involement
Strategic Summary Intelligence Sharing with Partners
Your Community

Retrospective Review into CCTV and Alcohol Related Assault in Night Time Economies

Between 2006 and 2013, the Australian Government funded over $15.9 million for CCTV cameras in 29 local government areas, two States and two chambers of commerce (Commonwealth Attorney General, 2014). This does not include separate projects funded by state governments. The stated political aim of installing this open-space urban and suburban CCTV infrastructure is to reduce and prevent crime. Despite this capital spend on technology in Australia, the relevant peer reviewed literature and government reports contain little articulation of how to make these systems more effective in their stated aims, or indeed, what managers may do to improve the performance of their systems once they are installed. Further, as noted above, a widely cited systematic review and meta-analysis of whether CCTV achieves this aim, is inconclusive.

Welsh and Farrington have conducted a number of systematic reviews and meta-analyses on the crime reduction effects of CCTV (2002, 2009). They found that decreases in crime associated with the introduction of CCTV in experimental areas when compared with control areas were modest but significant. They concluded this was evidence of effectiveness was largely due to CCTVs effectiveness in reducing property crime in car parks, but that it was not effective in reducing personal crime in town centres. They also noted the importance of future research to identify the causal mechanisms linking CCTV to reductions in crime. The current RS based on the studies included in their review aimed to pull apart and understand the mechanisms which led to these different outcomes in these different contexts. Limiting the primary studies analysed through the current RS to the 44 original studies which were examined (Welsh & Farrington, 2009), kept the number of papers to be analysed to a feasible number.

Realist Synthesis protocols require the development of a Realist Matrix. In *Realistic Evaluation*, Pawson andTilley (1997) provide a range of possible CMOs for the use of CCTV. As a starting point, a CMO table was developed for the contexts and mechanisms identified in Welsh and Farrington's systematic review (see Table 1: CCTV Crime Prevention Theories identified by Welsh and Farrington (2009) as a CMO table). This first, coarse, identification of theory was drawn from Welsh and Farrington's Introduction, Discussion and Conclusions sections of their systematic review. The attempt at importing their research into a Realist framework highlights a number of issues which will be discussed below. Importantly, Welsh and Farrington (2009) note that a significant problem in interpreting the results of their review was that crucial information on the evaluations was not always included in the articles which they reviewed. This is also the case for attempts at Realist Synthesis.

Of relevance to this current paper, Welsh and Farrington (2007) found that CCTV schemes in the United Kingdom showed a significant desirable effect on crime, while those in other countries generally showed no significant effect. They partly attributed this finding to methodological issues, partly to the holistic approach adopted in the United Kingdom, and strongly suggested this may be due to a lack of public support for surveillance in the United States or Scandinavian countries compared with Britain. They attributed this lack of public support to the cultural contexts of different countries, and attributed a lower police priority or lack of deterrence to this lack of public support. The difficulty the current author had in categorising "lack of public support" as either a context or a mechanism is an example is further emphasised in the discussion of deterrence below.

Table 1: **CCTV Crime Prevention Theories identified by Welsh and Farrington as a CMO table**

Outcome	***Context***	***Mechanism***
Effect on Crime	***Theoretical Domain***	***Identified Mechanisms***
Positive	Technological determinism	Number of cameras installed (visual coverage of cameras)
Positive	Situational crime prevention (Routine Activities Theory & Rational Choice Theory)	Formal surveillance
Positive	Psycho-social processes in the monitored population	Deterrence through increased subjective probability of detection (especially if CCTV well publicised)
Positive	Natural surveillance further increases the true and subjective probabilities of detection	Increase pedestrian usage of places
Positive	Psycho-social processes in the monitored population	Encourage potential victims to take security precautions
Positive	Policing and enforcement	Increase the true probability of detection
Positive	Policing and enforcement	Direct police and security personnel to intervene to prevent crime
Positive	CCTV could signal improvements in the area	Increased social capital
Positive	Community Crime Prevention	Implemented with a range of other prevention measures
Positive	Community Crime Prevention	Increase community cohesion
Positive	Community Crime Prevention	Increase informal social control
Negative	Give potential victims a false sense of security	Relax their vigilance
Negative	Give potential victims a false sense of security	Stop taking precautions
Negative	Statistical Rates	Increased reporting and recording of crimes to/by Police
Negative	Cultural context	Lack of public support
Negative	Lack of public support	Police assign lower priority to CCTV schemes Low media coverage reduces deterrence
Neutral	Displacement	Temporal and Locational tactical (change in method), target (change in victim), and functional (change in type of crime).
Neutral	Diffusion	Offenders know CCTV in a location may result in their apprehension and move activities

After the systematic review was analysed, each original article it cited was reviewed to understand what the original research looked for and how the researchers examined CCTV. The initial examination of original articles categorised each study by country in which it took place. All were "western" cities. Each paper was examined to identify whether fear of crime and/or perceptions of safety were investigated (outcome), whether reported crime figures only were used or if the study also examined calls for service and other indicators of crime (outcome), and any other outcomes which were listed in the original studies. Each paper gave a brief overview of aspects of the ecology (e.g. night time economy, residential estate, car parks, railway stations, hospitals and nursing homes) in which the cameras were deployed, as well as the operations and management of the monitoring, or control, rooms. These were categorised by the current author as "overarching contexts".

Research which specifically focused on town centres/night time economies in 21 locations were included in the RS, with research focusing on other location types excluded from the review. Concurrently each study relating to the night time economy was categorised as to whether a theoretical basis for CCTV was detailed, and if so, what that theoretical basis was. Situational crime prevention (Clarke, 1997) was the theoretical basis most often cited (in 38 out of the 40 original studies found), with deterrence mentioned in almost all the papers but no further links in the chain of causality. That is, there was little exploration of how CCTV is thought to deter criminal activity. Real-time deployment and dispatch (Sivarajasingam, Shepherd, & Matthews, 2003), and real-time intelligence (Brown, 1995) were also mentioned in a small number of papers.

RS is an iterative process. It became clear early that studies excluded in the original project scope should be included, particularly references from one article which articulated the theoretical base of the intervention in terms of perceptions of safety (Mazerolle, Hurley, & Chamlin, 2002), and studies on camera operator performance (Donald, 2008; Gill et al. 2005). Other studies examining camera operator performance were not analysed due to time constraints (Keval, 2006; Keval & Sasse, 2006; 2008). Because deterrence was a key theoretical concept in the papers examining CCTV, and the experience of personal consequences has been found to influence deterrence of offenders, a set of studies linking camera operator performance, detection and deployment, police activity and deterrence were included (Piza, Caplan, Kennedy, & Gilchrist, 2014; Piza, Caplan, & Kennedy, 2012).

As noted above, in the original studies used by Welsh and Farrington (2009):

1. There was a paucity of papers which were theory-driven, and these programme theories were barely articulated in the studies. A number of interventions were described in detail however, and these have provided sufficient detail to identify literature which links these interventions with programme theories;

2. The description of the ecologies in which the systems operated, as well as details regarding the management and operational parameters of the CCTV systems were usually minimal. They did however provide sufficient detail to explore literature from a range of disciplines (public health, criminology, organisational psychology, drug

and alcohol interventions) in order to better define some contexts of the night time economy, and the operational contexts through which CCTV is expected to have an effect (for example, camera operator attention);

3. A range of outcomes were analysed in each paper, each of which also pointed to domains of literature in which a limited range of program theories were identified.

These findings are similar to those well documented in the Realist literature. In many papers it was possible to work backwards from the outcomes identified in each relevant paper in order to begin to disaggregate contexts and mechanisms.

The next step was to reread each original article and place the ideas and results from them into a table with a column for Outcomes, a column for Context and a column for Mechanism to begin to understand the CMO configurations. These CMO tables were adapted from a previous Realist Synthesis (Ogrinc, Batalden, & Moore, 2009). This table model was chosen for simplicity, as each original paper dealt with one overarching program; CCTV. There were however, disparate study types, and many papers required the development of a number of tables, each categorised where possible by theoretical bases. These were then synthesised according to outcomes, and concurrently synthesised by programme theory. A total of 38 initial CMO tables were constructed from the 21 original studies used. In these 38 tables, a total of 186 rows of possible CMO configurations were assembled. Of these 186 rows a total of four completed CMO configurations were possible using data contained from the original studies. Put another way, the information contained in the original studies only provided enough relevant data for the construction of four sentences in the form of, “This works for these people because...”.

The example of the theory of deterrence as it relates to CCTV is shown in Table 2: ‘Initial CMO Hypothesis Grid for understanding CCTV as a deterrent’. These are mainly drawn from two studies which most explicitly linked deterrence with the existence of CCTV systems (Piza et al. 2014; Piza et al. 2012). CMO configuration sentences remain implicit in the CCTV literature; however drawing from Piza et al. (2012), the logic for deterrence would read something like: ‘Crime is reduced when CCTV cameras are present because offenders perceive an increased level of risk in offending.’ This was the initial rough theory around how CCTV deters criminals. Following the placement of concepts into the CMO table, a range of refined configuration sentences were developed, for example: ‘Offenders are deterred when they have personal experience in being apprehended at least once before through the use of CCTV.’

The initial rough theory for failure to deter would read something like: ‘Crime is increased when offenders commit crimes when CCTV cameras are present and they avoid punishment because they adjust their perception of risk downward.’ The refined sentence could read: ‘Offenders are not deterred when there is no guarantee offences will incur punishment because they have no personal experience of CCTV footage being used to investigate or prosecute them’.

Additionally, it became clear that my use of the categorisation, ’overarching contexts’, was an inaccurate understanding of the Realist use of the term ’contexts’. The context may refer more to the psychosocial ecology contained within these spaces which were defined through the use of these spaces. The difficulty in categorising something as a context or a mechanism, or even as an outcome is shown in Table 2: 'Initial CMO Hypothesis Grid for understanding CCTV as a deterrent.'

Table 2: **Initial CMO Hypothesis Grid for understanding CCTV as a deterrent**

Theory: Deterrence		
Some Possible Outcomes	***Some potential Contexts***	***Some Plausible Mechanisms***
Lower crime	Offender deterred (rational choice theory)	Presence of cameras is salient to offender
CCTV poses an increased level of risk to offenders	CCTV detections lead to more police enforcement actions than previously	Offender deterred
Offender deterred	Previously apprehended or convicted through CCTV	Offenders attribute apprehension or conviction to CCTV
Offenders attribute apprehension or conviction to CCTV	Directed response through Camera Operator deploying on-the-ground resources to interrupt the incident. Footage successfully used to investigate and prosecute.	Camera Operator awareness and detection of incident Effective Real-time communication processes Sufficient on-ground resources available (police, private security presence)
CCTV incidents demonstrate higher arrest rates for the same offence type	CCTV generates increased law enforcement activity in target locations.	Offenders have increased perceptions of certainty of punishment.
Offenders hold increased perceptions of certainty of punishment.	CCTV footage evidence is used for more effective and efficient investigation after the incident	Offenders who have been punished adjust their perception of the certainty of punishment upward
CCTV incidents demonstrate higher arrest rates for the same offence type	Real-time intelligence by Camera Operators to police on the ground assists in arrests on-scene CCTV footage evidence is used for more effective and efficient investigation after the incident	Trust between Camera Operators and operational Police Clear communication processes

That RS is an early attempt at building theory through a realist review and is certainly under-developed, but any early attempt will appear underdeveloped in terms of middle-range theory and CMO configuration (Jagosh et al. 2012). The exploration of theoretical bases from literature continues, and the contexts are still being refined. Synthesis is continuing. I have had more than three years of sustained and concentrated exposure to this class of intervention (CCTV to reduce crime). Without this understanding of this type of program I would have floundered more than I did. I still have a headache.

Review into homelessness service delivery in order to design a Realist Evaluation in a regional (north Queensland) context

Background

Homelessness is a complex issue which affects over 100,000 people in Australia on any given night. It is the result of a range of inter-related complex risk factors, including domestic violence and abuse, family conflict and neglect, poverty, alcoholism, drug addiction, mental illness and higher housing costs[2] (Department of Families, Community Services & Indigenous Affairs, (FaCSIA), 2008; Minnery & Greenhalgh, 2007). Cairns is a community which has been found to suffer a significantly higher homelessness rate than the rest of Queensland and Australia (Australian Bureau of Statistics, 2013; Department of Communities & Queensland Council of Social Service, n.d.) and higher rates of housing related stressors (Department of Housing, 2007). Housing related stressors have been found to substantially contribute to tenancy disruption (Atkinson, Habibis, Easthope, & Goss, 2007), subsequent offending behaviour by impacted individuals (Stewart, Livingston, & Dennison, 2008), and negative impacts on individual life trajectories, including adverse contact with the criminal justice system (Homel, 1999), poor health and educational outcomes (Ancona, 2008; Bridge, Flatau, Whelan, Wood, & Yates, 2007).

This RS aims to investigate the program theories which impact on the effectiveness and efficiency of individual services in Cairns which address alcohol-harm reduction and homelessness. It will define the dimensions for a future realist evaluation aiming to efficiently enhance the quality of client service of each agency, and to improve the contribution these services make to a place-based community safety approach. Three interconnected methodologies will be used:

1. Literature review and desktop analysis (using Realist Synthesis);
2. Realist Evaluation (RE); and
3. Network analysis.

This RS aimed to build dimensions for future evaluation using the RE methodology. Many outcomes of interest to the funder are clearly defined in their service agreements. These outcomes however may be better termed as “outputs”, in that they are exclusively quantitative and use measurable descriptors such as proportion, percentage and incidence of client contacts and presentations. The realist inquiry began with the questions: ’Why are these outputs important to the outcomes?’ Again, the importance

[2] The Department of Families, Housing, Community Services & Indigenous Affairs became the Department of Social Services in early 2014.

of working backward from the outcomes required by the funding organisation was highlighted through this project. A desktop analysis of relevant policies, procedures, reports to funders, strategic plans and annual reports to define the required outcomes, contexts and interventions for each service was conducted. In the case of this RS the term “interventions” was used as a tool to categorise the broad bases of relevant program theories. These outcomes, contexts and interventions were disaggregated by client group demographics, and limited to and homelessness. The research question for analysis was: ‘How can the effectiveness and efficiency of individual services in Cairns which address youth homelessness that is the targeted services in Cairns, be improved?’

Limiting the primary studies to be analysed was in this case done through identifying the specific target client groups, and the interventions provided for these individuals by the relevant service. The management documentation provided by the services usually contained a theoretical basis as to why this form of intervention is used. This has so far been the basis for further exploration of the relevant theoretical literature. To date the synthesis has focused on formulating an initial rough theory and CMO hypothesis grid for protective CMOs and risk CMOs relating to young clients becoming homeless.

Therefore, the initial theoretical base/client base was homeless young people, or those at risk of homelessness. The first document reviewed was a recent literature review commissioned by the Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) summarising relevant research on effective interventions for working with young people who are homeless or at risk of homelessness (Barker, Humphries, McArthur, & Thomson, 2012). That review conducted a comprehensive search strategy of 22 academic journal databases, government websites, and research clearing houses as well as Google Scholar. The aim of that review was to present the available evidence regarding effective interventions and approaches to reduce youth homelessness. The report authors noted a dearth of evidence specifically applicable to interventions to alleviate youth homelessness, and widened their search to include relevant risk and protective factors, as well as a broader category of youth interventions generally. Even this did not produce many Australian studies.

Importantly, Barker et al. (2012) were aware of the realist methodology, and this report contained elements of Realist structure and analysis. Equally importantly, the report identified a lack of rigorous evaluations regarding interventions to prevent, or respond to, youth homelessness. The reasons cited for this included the transient nature of the client group and the difficulty in defining precise outcomes against which effectiveness could be measured (Barker et al. 2012). The FaHCSIA literature review was used to identify the initial rough theory which framed pathways into and out of youth crisis accommodation, a coarse summary of evidence-based outcomes related to young people entering and exiting crisis accommodation, and the contexts and mechanisms which could plausibly be associated with these outcomes. Again, the CMOs were placed into a set of tables to facilitate the initial rough conceptualisation, and assist in the disaggregation into various theoretical bases. At this stage of the project, these theoretical domains focus on how family conflict leads to pathways into and out of crisis accommodation, and on client based theories regarding the centrality of the relationship between caseworker and client.

The reference list of the FaHCSIA paper was examined and so far a small number of relevant original papers pertaining to client based theory have been further analysed. This RS is in an early stage. Again though, the number of CMO tables which have been required to be constructed is almost double the number of original articles so far analysed. There is a range of intervention program approaches documented in the youth homelessness literature, for example, 'case management', 'wraparound, community reinforcement'. The summary of the literature on these programs found that strength-based, client-focused approaches are most effective, and the quality of the relationship between case worker and client is central (Barker et al. 2012). An example CMO is shown in Table 3, focussing on the importance of the relationship between client and case-worker.

Table 3: **Initial theory and CMO hypothesis grid for agency service effectiveness**

Theory: Centrality of Relationship		
Some Evidence based Outcomes	***Some potential Contexts***	***Some Plausible Negative Mechanisms***
Engaging and maintaining clients in interventions	Establishing rapport	Reservations about program by client Distrust by client
Engaging and maintaining clients in interventions	Instability (what sort of emotional, psychological or life circumstances is the client experiencing)	Client unable to implement interventions despite commitment
Some Evidence based Outcomes	***Some potential Contexts***	***Some Plausible Positive Mechanisms***
Client is housed and maintains tenancy	Gains or regains life skills and sense of self-efficacy Practical assistance	Client engaged in interventions
Client is housed and maintains tenancy	Client engaged in interventions	Rapport between client and case worker established
Client engaged in interventions	Establishing rapport	Service and worker conveying respect of perceptions and experiences and acceptance Start where the client is: Incorporate clients perceptions and experiences into issues pertinent to the lives – prioritise issues important to the young person
Rapport established	Client trusts the service provider Client feels cared for Clients do not feel judged	Client has personal experience and perceives engagement with service will lead to a positive experience

Based on the documentation provided by the funding organisation, and their approach to client engagement, this initial rough placement of the centrality of the relationship between the client and caseworker has identified a number of dimensions around which to construct the staff interview and focus group tools for the qualitative research, and provides a causal logic chain linking the way staff engage clients with housing outcomes.

Overall experience on the use of Realist Synthesis

The *Quality Standards for Realist Synthesis [for researchers and peer reviewers]* (Wong et al. 2013a), provide ratings of the proposed RS design, ranging from 'Inadequate' to 'Excellent'. Each domain of the project is spelled out to be benchmarked. The Standards include a Table which contains criteria to assist researchers in conducting their projects. The findings below detail the authors experience in referring to criteria and standards (ranging from adequate through good to excellent), contained in this table and attempting to apply the standards to the two projects outlined above. There are a number of issues I encountered in conducting these two RS. Two of these confirm what has been extensively documented in the RAMESES protocols. These are that:

1. almost all papers which will be reviewed as part of a RS project have not been conducted within a realist ontology or epistemology and will have major data gaps, particularly in regard to underlying programme theory; and

2. the process is iterative and requires moving up and down levels of analysis, as well as excursions into disciplines and domains other than the "putative" question under study. This work is almost always valuable provided the reviewer maintains a realist mindset – how and why does this cause something relevant to happen?

The practical conclusions to addressing these issues are:

- Each Realist research project was initiated as a result of client needs for evaluation of a small-scale intervention applied by the commissioning agency. The commissioners intuitively grasp that things work for different reasons in different contexts for the same problem. The context-mechanism-outcome framework allows practitioners and commissioners to focus on the aspect of a program which impacts on their core business, and allows them to explore and communicate their implicit knowledge.

- The questions have been about service delivery in a single geographic location, addressing an issue with a defined participant base. In my experience, the research question starts singly and simply; then it is disaggregated into a series of questions related to each of the theoretical domains in the cell above to obtain an understanding of the original research question.

- Analysis is driven by a number of theories, which aim to explain the problem, or the social context in which the problem occurs and is addressed. This has required acquaintance client-engagement theory, management theory, intervention theory, and theories about the causation of the problem. It is simplified by the question of 'what works for whom and why?'

- In practice, a range of CMO tables had to be constructed based on the program theory identified in the original articles. Many of these were outcome evaluations, with underlying theory poorly articulated. Working either forward or backward from reported 'outcomes' and poorly defined 'contexts' to the relevant mechanisms has been required in both realist syntheses. This is where literature seemingly unrelated to intervention programs is required.

- My understanding of 'contexts' was originally shaped by communicating the concept to CCTV system managers and camera operators. As such, I confused and conflated operational, empirical 'contexts' with a "purist" definition of 'contexts'. I confused 'contexts' with 'factors'.

- The first iteration of the research question has always had to be broken down into a series of more refined questions, which have to be linked to the original research question.

- The process is an iterative one. Moving up and down levels of analysis from the 'empirical' to the 'real' to the 'deep' has occasionally been daunting. The analogy in the RAMESES training manual regarding 'detectives following clues' has enabled trust in following the process no matter where it has seemed to lead in the short term.

- The retrospective RS built on existing qualitative research with a range of practitioners, not all of whom are directly involved in implementing the intervention. The core business of all of them, however, is affected by the intervention (CCTV). The prospective RS aims to inform qualitative research with a range of practitioners. RS should be part of a multi-stage project (RS, RE and network analysis), which aims to work with practitioners to understand, "why do we do this and how do we do it better?", and then implement findings.

- Ideally, it would involve also working with clients to do the same, but as yet, the author has no experience in this.

- The selection and appraisal ran parallel with the analysis stage for both the retrospective and prospective reviews. The analysis of papers originally selected for review required identification of other research from wide ranging disciplines. Again, the 'detective following the clues' analogy was useful in maintaining faith in the process.

- As noted in the RAMESES quality standards, the use of any data within a paper should be based on the relevance to theory development and testing, and the rigour of the methods used. In practice, no paper which was analysed had both sufficient relevance and rigour to be adopted in sum.

Summary, conclusions and implications

This paper has summarised practical implications in conducting two realist syntheses, based on the author's personal experience. Both related to evaluations of small scale, targeted interventions in a single location. One was a retrospective Realist Synthesis (RS) of the theoretical foundations of open-space, urban CCTV; an area in which the author had extensive content expertise. The author had previously conducted a Realist Evaluation (RE) of the CCTV system, which worked back from specific outcomes required by the commissioner of the Realist Evaluation. This RE methodology for CCTV has been further used by the author with eight councils across three Australian states. The second is a prospective RS of interventions to address alcohol-related harm, street-based outreach and volatile substance misuse among indigenous people. The purpose of this Realist Synthesis is to inform and assist in the design of a theory-based RE into these issues in a specific cultural and geographic context to improve service delivery.

In both projects it has been crucial that precise, disaggregated outcomes were defined. These include proximal and distal intermediate outcomes which contributed to the outcomes required by the project commissioners. To make research engagement more effective, framing these outcomes in a way which corresponded to the funders' internal continuous quality improvement processes was necessary. The commissioners of the research had conducted their own examination of the literature, and had operated these programs for a number of years. They therefore know that the outcomes (and causal mechanisms) lie at different systemic levels, both internal and external to the organisation.

In both projects, commissioners of the projects were shown a logic model diagram for how the Realist methodology applies to their project. This led to detailed discussions regarding how mechanisms link with outcomes, and the importance of context in organisations achieving their goals. In both projects also, discussions around the definition of a range of contexts, and the Realist reasoning of breaking the programs the funders delivered into an indicative set of mechanisms specific to each of those contexts was led by the commissioners of the funding, rather than the researcher. In the case of the CCTV research, the Realist model provides a framework through which practitioners can organise their implicit knowledge. In the case of the homelessness project, the Realist model provides an organising framework through which the researcher can inquire in a deeper way to link agency practices with outcomes.

Acknowledgements

Thanks and appreciation go to Cairns Regional Council for funding the original Realist Evaluation of the CCTV system, to Anglicare North Queensland for funding the Realist Synthesis of homelessness and alcohol-related harms, and to the Australian Institute of Criminology.

References

Australian Bureau of Statistics. (2013). National Regional Profile: Cairns Local Government Area. Retrieved from http://www.abs.gov.au/AUSSTATS/abs@nrp.nsf/Latestproducts/LGA32070Population/People12007-2011?opendocument&tabname=Summary&prodno=LGA32070&issue=2007-2011

Ancona, C. (2008). What works in tackling health inequalities? Pathways, policies and practice through the lifecourse. *Journal of Epidemiology and Community Health, 62*(1), 86. doi: 10.1136/jech.2007.059634

Atkinson, R., Habibis, D., Easthope, H., & Goss, D. (2007, January). Sustaining tenants with demanding behaviour: a review of the research evidence *Australian Housing and Urban Research Insitute* [AHURI]. (pp.1-47). Melbourne: AHURI..

Barker, J., Humphries, P., McArthur, M., & Thomson, L. (2012). Literature Review: Effective interventions for working with young people who are homeless or at risk of homelessness. Department of Families, Community Services & Indigenous Affairs. Canberra: Institute of Child Protection Studies, Australian Catholic University.

Bridge, C., Flatau, P., Whelan, S., Wood, G., & Yates, J. (2007). How does housing assistance affect employment, health and social cohesion? *AHURI Research and Policy Bulletin* (Vol. 87, pp. 1-6).

Brown, B. (1995). *CCTV in town centres: Three case studies.* London: Home Office.

Cherpitel, C. J. (2007). Alcohol and injuries: a review of international emergency room studies since 1995. *Drug and alcohol review, 26*(2), 201-214.

Clarke, R. V. (Ed.). (1997). *Situational Crime Prevention: successful case studies* (2nd ed.). Guilderland, N.Y.: Harrow and Helsen.

Clough, A. R., Hayes-Jonkers, C. S., & Pointing, E. S. B. (2013). Alcohol, assault and licensed premises in inner-city areas: Scoping studies and baseline data collection for an evaluation of best-practice policing interventions augmented by collaboration with emergency medicine and local community agencies to reduce alcohol-related assault. In National Drug Law Enforcement Research Fund [NDLERF]. (Ed.). [Monograph] Series No. 45. Canberra: NDLERF.

Davis, P. (2005). The Limits of Realist Evaluation. *Evaluation, 11*(3), 275.

Donald, F. M. (2008). The classification of vigilance tasks in the real world. *Ergonomics, 51*(11), 1643-1655.

Department of Communities, Queensland Council of Social Service (n.d.). *New Ways Home: Cairns Homelessness Community Action Plan 2010-2013*. (In partnership). Department of Communities and Queensland Council of Social Service. (Funded by the Australian and Queensland Governments). Department of Communities & Quessnalnd Council of Social Service.

Department of Housing. (2007). Cairns West Community Renewal Zone Plan 2007 – 2009. Queensland Government, Department of Housing.

Department of Families, Housing, Community Services & Indigenous Affairs. (2008). *The Road Home: A National Approach to Reducing Homelessness. Canberra:* Commonwealth of Australia.

Gill, M., Spriggs, A., Allen, J., Hemming, M., Jessiman, P., Kara, D., Swain, D., Kilworth J., Little, R. (2005). *Control room operation: findings from control room observations.* London: Home Office.

Commonwealth Attorney General. (2014). Australian Government's Safer Suburbs Program. Retrieved from http://www.ag.gov.au/CrimeAndCorruption/CrimePrevention/Pages/Safersuburbs.aspx

Graham, K., & Homel, R. (2008). *Raising the bar: preventing aggression in and around bars, pubs and clubs.* Cullompton, United Kingdom: Willan.

Havard, A., Shakeshaft, A., & Sanson-Fisher, R. (2008). Systematic review and meta-analyses of strategies targeting alcohol problems in emergency departments: interventions reduce alcohol-related injuries. *Addiction, 103*(3), 368-376.

Hawkins, N., Sanson-Fisher, R., Shakeshaft, A., & Webb, G. (2009, April). Differences in licensee, police and public opinions regarding interventions to reduce alcohol-related harm associated with licensed premises. *Australian and New Zealand Journal of Public Health, 33*(2), 160-166. doi: 10.1111/j.1753-6405.2009.00364.x

Homel, R. (1999). Preventing violence: A review of the literature on violence and violence prevention. A report prepared for the Crime Prevention Division of the NSW Attorney General's Department.

Jagosh, J., Macaulay, A., Pluye, P., Salsberg, J., Bush, P. L., Henderson, J., … Greenhalgh, T. (2012). Uncovering the Benefits of Participatory Research: Implications of a Realist Review for Health Research and Practice. *The Milbank Quarterly, 90*(2), 311–346.

Keval, H. (2006, September). CCTV Control Room Collaboration and Communication: Does it Work? In University of Sussex, Human Centred Technology Workshop 2006: *Designing for Collaborative as well as Individualised Environments. Proceedings, 9th* Human Centred Technology Group Postgraduate Workshop, Department of Information, University of Sussex, Brighton. In association with Future Platforms. London, England.

Keval, H., & Sasse, M. A. (2006, July). *Man or a Gorilla? Performance Issues with CCTV Technology in Security Control Rooms.* Paper presented at the 16th World Congress on Ergonomics, Maastrict, the Netherlands.

Keval, H., & Sasse, M. A. (2008). "Not the Usual Suspects": A study of factors reducing the effectiveness of CCTV. *Security Journal, 23*(2), 134-154.

Mazerolle, L., Hurley, D., & Chamlin, M. (2002). Social Behavior in Public Space: An Analysis of Behavioral Adaptations to CCTV. *Security Journal, 15*(3), 59-75. doi: http://dx.doi.org/10.1057/palgrave.sj.8340118

Minnery, J., & Greenhalgh, E. (2007). Approaches to homelessness policy in Europe, the United States, and Australia. Journal of *Social Issues, 63*(3), 641-655.

Ogrinc, G., Batalden, P., & Moore, S. (2009). Realist Evaluation as a Framework for the Assessment of Teaching About the Improvement of Care. *Journal of Nursing Education, 48*(12), 661-667.

Palk, G. R. M., Davey, J. D., & Freeman, J. E. (2010). The impact of a lockout policy on levels of alcohol-related incidents in and around licensed premises. *Police Practice and Research: An International Journal, 11*(1), 5 - 15.

Pawson, R. (2003). Nothing as practical as a good theory. *EVALUATION-LONDON-, 9*(4), 471-490.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist Review - a new method of systematic review designed for complex policy interventions. *Journal of Health Services, Research and Policy, 10*(Supplement 1, July), 21-34.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage Publications Ltd.

Piza, E., Caplan, J., Kennedy, L. W., & Gilchrist, A. M. (2014). The effects of merging proactive CCTV monitoring with directed police patrol: a randomized controlled trial. *Journal of Experimental Criminology,* (Published online: 15 July 2014), 1-27. DOI: 10.1007/s11292-014-9211-x

Piza, E. L., Caplan, J. M., & Kennedy, L. W. (2012). Is the Punishment More Certain? An Analysis of CCTV Detections and Enforcement. *Justice Quarterly, online access first 11 Sept. 2012.*

Pointing, S. B., & Clough, A. R. (2013). Report to Cairns Regional Council; Audit and Evaluation of the open-space, urban CCTV system. Stage 2, Inner City Safety Partnership.

Pointing, E. S., Hayes-Jonkers, C.S., Bohanna, I., & Clough, A. (2012). The role of an open-space CCTV system in limiting alcohol-related assault injuries in a late-night entertainment precinct in a tropical Queensland city, Australia. *Injury Prevention, 18,* 58-61. doi: doi:10.1136/injuryprev-2011-040080

Pointing, S. B., Hayes-Jonkers, C. S., & Clough, A. R. (2010). Report to Cairns Regional Council; Pilot Audit and Evaluation of the open-space, urban CCTV system Cairns: James Cook University.

Pointing, S. B., Hayes-Jonkers, C.S., & Clough, A. R. (2011). Report to Cairns Regional Council; Audit and Evaluation of the open-space, urban CCTV system. Stage 1, Inner City Safety Partnership. Cairns: James Cook University.

Pointing, S. B., Hayes-Jonkers, C. S., & Clough, A. R. (2012). Evaluating the context of an open-space CCTV system to address assaults in an urban centre in northern tropical Australia. *Crime Prevention and Community Safety, 42,* 140-152.

Pointing, S. B., Hayes-Jonkers, C.S., & Clough, A. R. (2013). Possible Strategies for Reducing Alcohol-related Assault: Community-Based Methodology in Cairns, Tropical North Queensland (Australia). In K. Bletzer (Ed.), *Assaults: Interventions, Preventative Strategies and Societal Implications* (pp. 169-185). New York: Nova.

Pointing, S. B., Hayes-Jonkers, C.S., Stone, R., Brinn, D., & Clough, A. (2011). Is it worth emergency departments recording information about alcohol-related assault occurring in inner-city, late-night entertainment precincts? *Emergency Medicine Australasia, 23*(1), 106-107. doi: 10.1111/j.1742-6723.2010.01382.x

Sivarajasingam, V., Shepherd, J. P., & Matthews, K. (2003). Effect of urban closed circuit television on assault injury and violence detection. *Injury Prevention, 9*(4), 312 - 316.

Stewart, A., Livingston, M., & Dennison, S. (2008). Transitions and turning points: Examining the links between child maltreatment and juvenile offending. *Child Abuse & Neglect, 32*(1), 51-66.

Welsh, B. C., & Farrington, D. P. (2002).*Crime prevention effects of closed circuit television: A systematic review* (Home Office Research Study, No. 252). London: Home Office.

Welsh, B. C., & Farrington, D. P. (2007). *Improved Street Lighting and Crime Prevention: A Systematic Review.* Report prepared for the Swedish National Council for Crime Prevention.

Welsh, B. C., & Farrington, D. P. (2009). Public Area CCTV and Crime Prevention: An Updated Systematic Review and Meta-Analysis. *Justice Quarterly, 26*(4), 716-745.

Wilson, D. (2008). *Researching CCTV: Security Networks and the Transformation of Public Space.* Paper presented at the Proceedings of the 2nd Australian & New Zealand Critical Criminology Conference, Sydney.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC medicine, 11*(1), 21-35.

Wong, G., Westhorp, G., Pawson, R., & Greenhalgh, T. (2013). Realist Synthesis: RAMESES Quality Standards. Retrieved from http://www.ramesesproject.org/media/RS_qual_standards_researchers.pdf

Seeing is believing? Experiences of using systematic social observation as an evaluation method

Rick Brown

Australian Institute of Criminology

Rick.Brown@aic.gov.au

Keywords: Systematic Social Observation, observational methods in evaluation, environmental campaign evaluation, alcohol licensing evaluation, street based observation, night time economy observation, environmental visual audits.

Abstract

In conducting project evaluations there are often times when data are not available and where standard methodologies are not always appropriate or sufficient. In some cases, there can be benefit in applying Systematic Social Observation (SSO) to gain an understanding of how a project is working. This paper explores the use of simple forms of SSO as an evaluation method in such circumstances. Based on case studies from the UK involving evaluating the impact of changes in alcohol licensing laws and evaluating the impact of environmental clean-up campaigns, this paper explores design issues, problems experienced by fieldworkers and the benefits to be gained from employing SSO.

Introduction

There are times in undertaking evaluation studies when data collected for routine administrative purposes prove to be unavailable or unreliable, or when other forms of data collection provide too partial a picture, or are simply too impractical to apply in a given situation. In such circumstances, systematic social observation (SSO) may provide a useful, additional means for understanding how a particular, program, project, initiative or intervention is being implemented. Indeed, as Mastrofski, Parks and McClusky (2010) noted, “SSO may be especially desirable when the question demands detailed knowledge of situations, conditions or processes that are not otherwise well-illuminated or where there is reason to question the validity of knowledge based on other forms of observation.” (p. 228)

SSO is defined here as the routine recording of an observed event or situation, allowing for the collection of predetermined variable categories or metrics in a standardised way. The key attributes that SSO seeks to attain are reliability (the same event should be recorded in the same way every time) and replicability (the observation study should produce similar results if it were to be repeated). As such, SSO can be contrasted with other, more ethnographic or qualitative forms of observation which may focus on *interpreting* observed phenomena (Fetterman, 2010).

SSO has previously been used to collect information on a range of issues. For example, Reiss (1971) described the characteristics of SSO studies, based on his policing research and, indeed, the approach has continued to be used in policing studies (Mastrofski et al. 1998, 2010; McCluskey & Terrill, 2005; Schelnberg, 2014; Terrill & Reisig, 2003). Its use has also been extended to other fields of enquiry, including observations of the night-economy economy (Hough, Hirschfield & Newton, 2008; Homel, Carvolth, Hauritz, McIlwain & Teague, 2004; Sim, Morgan & Batchelor, 2005; Smith, Morgan & McAtamney, 2011), neighbourhood characteristics (Odgers Caspi, Bates, Sampson & Moffitt, 2012; Sampson & Raudenbush,1999; Sampson, Morenoff & Gannon-Rowley, 2002), behaviour assessment (Briesch, Chafouleas & Riley-Tillman, 2010), school environments (Wilcox, Augustine & Clayton, 2006) and physical activity (Zarrett, Sorensen & Skiles, 2010) to name just a few.

SSO as described in this paper focuses on gaining a quantitative representation of a particular event or situation. This can be particularly useful for examining how a situation changes over time, following the introduction of a particular intervention. This paper draws on the experiences of the author in undertaking SSO of different kinds of events and situations and aims to draw out both the benefits of the approach and the problems experienced when applying the approach in the field.

Background to the studies

This paper draws on two sets of studies that applied simple forms of SSO techniques in different ways, all of which were undertaken in the UK between 2005 and 2009. The first set of studies examined the impact of changes in one particular night-time economy[1] (Hartlepool, in the north-east of England) following the introduction of the Licensing Act (2003). The Act introduced a range of changes to the way in which alcohol licenses were regulated, generally liberalising the drinking environment. In particular, it increased the flexibility over opening times of licensed premises, paving the way for the potential for 24 hour opening. Two studies were undertaken in Hartlepool to gain an understanding of the workings of the night-time economy (see Brown & Evans, 2011). The first occurred in 2005, prior to the Act being occurred in 2005, prior to the Act being implemented in November of that year.

[1] The 'night-time economy' refers to leisure activity that takes place in and around entertainment districts / town centres during the evening and early hours of the morning. This typically involves activities such as dining out, consuming alcohol in bars and clubs, visiting the cinema / theatre etc. See Hadfield (2006) for an account of the rise of the night-time economy in Britain.

Among the research methods employed were street-based observations of the night-time economy over four weekends (Friday and Saturday nights). A further study in November 2009 again employed street-based observations, although over one weekend only[2].

The street-based observations were designed to gain a quantitative picture of an evening in Hartlepool. These were undertaken as part of a suite of data collection methods that included interviews with stakeholders, a survey of local residents, a survey of visitors to the night-time economy and an analysis of crime and disorder data provided by the police and the fire and rescue service. The purpose of the street-based observations was to document a typical evening in Hartlepool and to create a kind of 'natural history' of changes over the course of an evening. This kind of information was not available from existing documentary sources and the second-hand accounts of others frequenting the night time economy were considered insufficient for drawing conclusions about the workings of the night-time economy. In addition, an important aspect of the work was to examine how a night-time economy changed following the introduction of new legislation, especially in terms of the numbers of patrons frequenting the area at different times of the evening. SSO techniques were considered to be the only reliable means of assessing these changes over time.

The observation study used a data collection tool that collected information on a number of undesirable features associated with the night-time economy, including the number of discarded bottles and glasses on the street, the number of items of rubbish, the number of visibly drunk people (later dropped as a data item), the number of fights, evidence of people vomiting and evidence of public urination. Contextual information associated with the number of men and women on the street and the presence of the police were also recorded. The fieldwork involved two teams of two researchers, with one team observing each of the two main streets that formed the night-time economy area. The observation involved walking down one side of the street and then returning along the other side of the street. Each data collection cycle would take about 15 minutes. This was repeated every half an hour during the course of the evening. The researchers carried a mechanical counter in each hand (which were stowed away in their coat pockets to avoid drawing attention). These were used to count the most voluminous items – one researcher would count the number of men and the number of discarded bottles / glasses, while the other would count the number of women and the number of items of rubbish. The remaining data items were mentally recorded by both researchers and agreed upon at the end of the data collection cycle.

Ethical considerations were raised about the use of SSO in this study. However, the risks to those being observed were considered to be minimal and outweighed by the gains (in terms of knowledge acquisition) for a number of reasons. Firstly, the unit of analysis focused on the night-time economy at a particular point in time, rather than on individuals and as a result, no individual could be identified from any of the data collected. Secondly, the data collection focused on the public spaces (the streets) and did not include privately owned spaces, such as inside pubs / clubs, which could have raised more serious concerns regarding informed consent from venue managers. Thirdly, the approach of not engaging with those frequenting the night-time economy area meant that the process of data collection would have no discernible impact on those being observed.

A greater concern related to the safety of researchers operating in an area where there was a higher than average risk of random violence. This was addressed by researchers working in pairs, with a senior staff member supervising the work on the street at all times, who could make a judgement about whether conditions were suitable for the researchers to continue to operate over the course of the evening.

The second set of studies involved evaluating the effectiveness of clean-up operations undertaken by Community Safety Partnerships[3]. These were designed to improve the physical environment of local communities by removing rubbish and graffiti. Evaluations associated with ten separate clean-up operations between 2006 and 2008 involved an environmental visual audit both pre and post clean-up (Brown & Evans, 2012). The visual audits were undertaken within one week of the clean-up commencing and within a week following the clean-up. This approach to data collection was employed as it was considered to be the only reliable and cost-effective way of measuring the outcome from a clean-up operation. Alternatives, such as surveying the local community pre and post intervention were considered too imprecise (due to the difficulties of assessing perceptions of environmental change over short periods of time) and too resource intensive.

The environmental visual audits involved two researchers walking through an area designated for the clean-up and recording details of rubbish and graffiti observed in the area. Information collected on each ‘incident’ included the time, date and description of the item, a description of the location and the location on a map. A photo would also be taken of the ‘incident’. The team later developed an application that ran on a Personal Digital Assistant (the forerunner of smartphones), which automatically photographed and plotted the GPS location of the ‘incident’. There were no ethical issues considered to be problematic in these studies as the unit of analysis focused on the consequences of human endeavour (discarding rubbish), rather than any direct observation of activities – whether that be the clean-up operation workers collecting rubbish or local residents discarding rubbish.

[2] This change in methodology was at the request of the client. Indeed, the 2009 study had a slightly different emphasis with a street-based survey of patrons being deployed as the main form of data collection.

[3] In 2007, the Home Office promoted the concept of Weeks of Actions. These were multi-agency, multiple intervention initiatives, that typically incorporated a significant environmental clean-up component. An example of one such initiative in Nottinghamshire, England, which reported the removal of 343 tonnes of rubbish, was submitted for a Tilley Award. Further details can be found at www.popcenter.org/library/awards/tilley/2007/07-41.pdf (accessed 12th April 2014).

Experiences of employing SSO

The experiences of using SSO are described here in terms of the design issues associated with the methodology and the experiences of employing SSO in the field.

Design issues associated with SSO

The key design issues of concern with SSO were associated with sampling, reliability and generalisability. Other design issues noted in the literature were not relevant in these studies. For example, reactivity, in which observed subjects act differently because of the presence of a researcher has been noted as a significant risk in SSO studies, although generally considered to decline as fieldwork progresses (Mastrofski et al. 1998, 2010; Reiss, 1971; Shulenberg, 2013). The need to build rapport with observed subjects has also been noted as important (Reiss, 1971). In the studies described here, the observation focused on the environment, rather than on individuals, thereby avoiding the issues of reactivity and rapport.

The current studies were also at the *complete observation* end of the participant observation scale (Hagan, 1982), with no interaction required between the researcher and the subject of the observation (the environment), thereby limiting to a minimum the potential for bias to be introduced by the presence of the researcher. However, as noted later, in some circumstances this proved harder than expected. McCall (1978) has noted that this approach to participant observation can involve a trade-off between unobtrusiveness and the ability to ask questions about what is being observed. In the studies examined here, this was considered to be of minimal concern, given that the focus was more on recording *what* was observed, rather than understanding *why*.

Sampling

Sampling in both sets of studies was based on geography and time. In the case of the night-time economy studies, two streets in Hartlepool town centre were sampled as the principal location for pubs and clubs in the town. Fieldwork focused on Friday and Saturday nights, which reflected the time when restaurants, bars and clubs in the area were at their busiest. In the first study, fieldwork was undertaken between 8pm and 3am on each evening. Taking account of the changes in opening hours following the implementation of the *Licensing Act* (2003), the second study involved fieldwork between 10pm and 5am. The unit of analysis in these studies was a thirty minute time interval, with the aim of developing a picture of how the night-time economy area changed over the course of an evening.

In the environmental clean-up studies the sampling frame consisted of particular geographical locations (usually areas of high social housing) at which clean-up operations had been undertaken. Two points in time were sampled – a day within one week of the commencement of the clean-up operation and the same day of the week following completion of the operation. The same day of the week was sampled to account for the impact that regular refuse collection might have on the build-up of rubbish in an area. The unit of analysis in these studies was an item of rubbish or graffiti.

Reliability

Reliability in SSO studies relies (at least in part) on the observation and interpretation skills of the researcher to record data accurately against a pre-defined set of variables. As Mastrofski et al. (1998) noted, this can become problematic when observers are required to interpret and classify behaviour. For example, in studies of police practice, it proved easier to code the responses to questions such as 'Did the police handcuff the citizen?', than to code questions such as 'Were the police justified in handcuffing the citizen?' (Mastrofski et al.1998, p. 9).

In the night-time economy studies, most of the variables collected were dichotomous, based on a simple presence or absence of an issue. This avoided the need to make judgements that would have been necessary had scaled variables been used. However, one problem arose in relation to counting the number of drunk people observed on the street. It became apparent early in the fieldwork that it was difficult to discern from a distance whether someone was drunk unless they exhibited clear behaviour as such. There was also a degree of relativity to the measure. At 8pm it was easier to judge someone as drunk if their behaviour stood out as different from others. At 2am the judgement was more difficult when standards of behaviour generally seemed to have declined. A decision was therefore made to drop the requirement to count drunk people from the data collection form.

Concerns with inter-rater reliability have previously been noted in observation studies of the night-time economy, especially where different fieldworkers have been used to collect information on different nights. For example Miller et al. (2012) described how checks were made to detect significant discrepancies, and also conducted simultaneous quality assurance observations on 10% of observations. In the Hartlepool studies, inter-rater reliability was addressed by a mix of simplifying the measures to be counted, agreeing findings between researchers and using the same fieldworkers consistently over different nights.

In the environmental clean-up studies, measurement was based on the presence or absence or rubbish/graffiti. Rubbish was coded as:

- General litter – small items (cans, bottles, packaging) such as one might find in a public litter bin.
- Small fly-tips – accumulations of two or more small items such as rubbish sacks, tyres, old televisions etc.
- Large fly-tips – larger accumulations of waste in a location, such as might constitute a car load of rubbish for example.

- Large items – one or more bulky household goods such as sofas, chairs, fridges etc.
- Other – items that could not be categorised under the above, such as wood from broken down fences.

Graffiti was coded as:

- Small tag – small individual identifier (text or symbol), typically with marker pen
- Large tag – larger individual identifier, (text or symbol) typically with paint
- Unofficial mural – picture, typically with paint
- Abusive language – sexist, racist, homophobic, expletives etc.
- Other – slogans or other writing

Photos were taken of each incidence of rubbish/graffiti, which allowed for coding to be checked by a supervisor. Note that this approach focused on the presence or absence of an incidence of rubbish/graffiti and recording this allowed the researchers to identify incidents of rubbish / graffiti that had been removed following the operation, rubbish/ graffiti that had not been removed and rubbish / graffiti that was new following the pre-operation fieldwork. However, recording in this way did not account for changes in the size of piles of rubbish which could grow or diminish between the pre and post observations.

Issues over inter-rater reliability were avoided in the environmental clean-up studies by ensuring that the same researchers were used in the pre and post stages. This helped to ensure that each time, the same geographical area was covered and items were measured in the same way. items were being measured in the same way each time and ensured that the same geographical area was covered each time . However, Brown and Evans (2011) noted the potential for bias caused by researchers knowing whether fieldwork was being undertaken pre or post intervention, with the potential for them to rate differently on the basis of this knowledge. There also remained the possibility of poor inter-rater reliability between studies (despite similar training to all researchers). This potential problem proved difficult to address.

Generalisability

Generalisability has been recognised as a particular problem in qualitative, observational studies (Maxfield & Babbie, 2005), due to the often very specific context within which studies are undertaken. This can call into question the extent to which similar results would be found if the study were to be repeated elsewhere. In the studies examined in this paper, generalisability was hard to judge. Where the night-time economy studies were concerned, although undertaken in a very specific context (a town centre in the north-east of England), the universality of the changes to licensing practice suggest that the findings would have been replicated elsewhere. Indeed, other studies of the impact of the *Licensing Act* (2003) found similar results (Hewitt & Kirby, 2011; Hough & Hunter, 2008; Morleo, Harkins, Hughes, Hughes & Lightowlers, 2007; Pike, O'Shea & Lovbakke, 2008). There was, however, a question over the extent to which SSO conducted over four weekends in the first study and over just one weekend in the second were generalisable to other weekends in the year.

The environmental clean-up studies were similar in that they each involved activity by public servants to clear away rubbish and clean off graffiti in deprived areas made up largely of social housing. It would therefore follow that the findings would only be generalizable to similar types of areas subject to similar intervention of this kind. For example, similar results might not have been obtained had more affluent areas been targeted, or (as Brown and Evans (2011) suggested) had recruited local community members to undertake the clean-up work.

Experiences of employing SSO in the field

A number of practical lessons were learned during the course of undertaking SSO. These were the kind of lessons not normally found in text books, but which nevertheless had to be negotiated to complete the fieldwork successfully.

Maintaining the role of complete observer

In undertaking the night-time economy studies, the researchers intended to take on the role of complete observers (Hagan, 1982), involving minimal interaction with those under observation. However, it became clear that, despite attempts to remain unobtrusive, the researchers were noticed, especially by pub/club door-staff. This was, perhaps, inevitable, given that the fieldwork involved walking past such venues every 30 minutes. This initially raised suspicions about the researchers' behaviour, especially among the door-staff of some of the less salubrious clubs. Thus, while patrons appeared oblivious to the presence of researchers (who might only see the researchers once) the repetitive nature of the fieldwork drew attention from the more observant of the town centre's capable guardians. This was addressed by engaging with door-staff of each pub/club early in the evening to explain that a study was being undertaken for the local council. An added benefit of this engagement was that it also allowed the researchers to negotiate access to pub toilets at a time when all other public conveniences were closed. Indeed, this was a practical issue that had not been foreseen in advance.

Between tours of the night-time economy area, the researchers sat in a car just outside the fieldwork area and wrote up their notes from the previous round. On one occasion, the researchers parked on the opposite side of the road from a quiet pub that was not included in the study area. Towards the end of the evening, the researchers were confronted by an irate publican, who assumed that they were (clearly not very good) undercover officers from Her Majesty's Revenue and Customs, checking up on his business. On another occasion, the researchers returned to their car to find two women in stiletto healed shoes dancing on the bonnet of their car.

The importance of covering the same ground

The environmental clean-up studies typically involved walking the streets and alleyways of a wide area, sometimes covering as many as 2,000 households. There were times when, on the post implementation sweep, new alleys would be found that were not included in the first sweep. In such cases, it was important to retrace the original trail as faithfully as possible, even if this resulted in incomplete coverage of the area both pre and post intervention.

Fatigue

The SSO studies described here involved a considerable amount of walking – whether it be recording environmental problems on a housing estate or repeated circuits of a town centre. In both cases, researchers could often expect to walk in excess of ten kilometres during the course of a fieldwork session. In the case of the night-time economy studies, there was the added fatigue of working into the early hours of the morning.

Weather

Researchers needed to be equipped for all weather conditions, including sunshine, rain and, extreme cold, each of which could affect health and safety working conditions. From the SSO perspective, weather conditions could have a significant impact on observed subjects. For example, it rained heavily on the first evening of the night-time economy fieldwork and this influenced the number of people that were observed on the street.

Benefits of SSO

The key benefit derived from the use of SSO in the studies described here was the ability to derive measures that were not available from other sources. In the case of the night-time economy studies, the SSO approach yielded a range of findings that would otherwise have not been possible. This included showing how the numbers of people on the street increased during the evening, peaking at midnight, before declining steadily. Following the introduction of later closing of pubs/clubs (by an average of two hours) the decline in numbers occurred over a longer period, so that when the last clubs closed, there was a less pronounced increase in people on the streets. Interestingly the results showed that women tended to leave the area earlier than men, creating a change in the gender mix towards the end of the evening.

The findings from the SSO were also used to counter claims from licensees that people were simply arriving in the town centre two hours later than they were before and then staying two hours later, effectively shifting the entire night-time economy window by two hours. However, the results of the observation showed that people arrived in the town centre at a similar rate after the licensing changes as they had before.

The findings also demonstrated the detrimental impact on the immediate environment caused by the increase in alcohol and take away food litter over the course of the evening, as well as the more unpleasant prevalence of vomiting and public urination. Indeed, it was estimated that, along the 800 metres of road that formed the night-time economy area, there were 1,040 instances of vomiting per year and 1,092 instances of public urination per year. Other methods were unlikely to yield results of this kind.

Where the evaluations of environmental clean-up operations were concerned, Community Safety Partnerships, typically responsible for undertaking such activities, tended to measure success through an *output* measure of how much rubbish was removed. It was simply assumed that removing rubbish would necessarily improve the environment. The SSO approach used in these studies provided an objective measure of *outcome* – the net impact on the amount of rubbish in public spaces. Brown and Evans (2012) reported how, across the ten clean-up operations evaluated, the amount of rubbish observed actually increased by 3%. Indeed, while the cleanliness of the environment improved for five areas, it deteriorated for the remaining five. This was a completely counter-intuitive result, implying that cleaning up an area could actually make it dirtier. Brown and Evans explained this increase in rubbish on the opportunity taking behaviour of the local residents who lived in the areas subject to the clean-ups. Observing that the local authorities were removing rubbish from public spaces, some local residents decided to throw out into the street and on to parkland unwanted items that would otherwise be difficult to take to a refuse tip or would incur a charge from the council for disposal. This was evidenced by the fact that the number of bulky items observed during the post clean-up environmental audits increased.

Similar results could be observed for graffiti. In one case, the amount of graffiti tagging was found to increase from 43 instances before the clean-up to 79 afterwards, an increase of 84% in the space of just two weeks. On further investigation, it transpired that a professional graffiti artist had been hired as a diversion activity for young people during the school holiday week in which the clean-up operation was undertaken. The artist had worked with the young people on removable boards in the local community centre. This activity seemed to inadvertently encourage participants to practice their newly acquired skills in other parts of the area.

The development of bespoke software for conducting the visual audits allowed for the automatic creation of a report on each incident that included a photo, map and description. This documentation proved extremely powerful when reporting back to a local government client about the cleanliness of the area they had just cleaned. The photographic evidence proved difficult to dispute, despite claims over the amount of rubbish that had been removed.

By-products of SSO

In addition to the benefits of SSO listed above, there were a number of additional by-products derived from observing the environment in which the studies were based, which, while in no way systematic, provided useful context. One such observation related to the use of take-away establishments in Hartlepool town centre. At the end of the evening, once bars and nightclubs had closed, it was common for people to buy take-away food for the journey home. One such take-away establishment seemed to be more popular than others because they turned their music up when the clubs closed to encourage customers to continue their evening in and around the establishment. When this was later raised with the council Licensing Officer he was unaware of the practice and explained that this would clearly be in contravention of the take-away establishment's license to operate.

The use of SSO to count the amount of alcohol related litter in the night-time economy area showed how such litter could lead to people injuring themselves. During the evening, customers would stand outside the pubs in the area and, on finishing their drinks, would sometimes place the empty bottles and glasses close to the pub's wall. However, others would later pass that place and accidentally kick and smash those items, creating a carpet of broken glass that was not cleared away. Indeed, the count of alcohol related litter actually declined towards the end of the evening due to this. At the end of the evening, some women would leave the clubs bare-footed with shoes in hand, no doubt to ease their aching feet after a night of dancing. On several occasions this was observed to lead to cuts to the feet as a result of walking on broken glass.

It was also observed that raised brick flower beds, constructed between the pavement and the road on one side of the street, created a number of problems for the night-time economy. First, the bushes and trees planted in those flower beds cast shadows, creating darkness along the pavement in that location and obscuring natural surveillance. Unsurprisingly, it was noted that more public urinations tended to occur on that side of the street, as did more of the fights. Second, the brick flower beds were also used as seating for those drinking on the street and as unofficial litter bins. As a result of these observations, the flower beds were later removed by the council.

Conclusions

This paper has examined the experiences of undertaking simple SSO studies to evaluate the impact of two different types of initiatives – changes to licensing regulations on a local night-time economy and the implementation of environmental clean-up operations. Where the night-time economy studies were concerned, the fieldwork involved recording a range of information every 30 minutes over the course of the evening. The environmental clean-up studies involved recording levels of rubbish and graffiti both before and after the intervention.

Both sets of studies showed how information could be collected to show the impact of initiatives, even when other forms of data collection were not possible, producing results that provided valuable insights. In the case of the night-time economy studies, there were also a number of additional observations that, while not systematic, provided useful contextual information.

However, it is noted that there are particular challenges to be faced in the field, including how researchers interact with research subjects, fatigue and external conditions (such as imposed by the weather), which mean that SSO is not for the faint hearted. It is also important to note that SSO will not be suitable in all conditions. The studies described here benefited from the fact that the context for the research focused on open street areas and did not involve the direct recording of the behaviour of individuals. Such work has been undertaken in policing research (Mastrofski et al. 1998, 2010; McCluskey & Terrill, 2005; Schelnberg, 2014; Terrill & Reisig, 2003), but brings with it a greater set of research problems to be addressed, including negotiating organisational access, informed consent of participants and avoiding Hawthorne effects (Landsberger, 1958) that can derive from the presence of a researcher. These problems may be insurmountable in some circumstances, meaning that there are likely to be conditions in which SSO techniques may not be appropriate. For example, this could be the case when observation involves vulnerable people who are unable to give informed consent, or where the presence of the researcher may alter the dynamics of the situation to the extent that they impede the processes being examined. These problems clearly point to the need to be mindful of the circumstances in which SSO may be employed. Indeed, as with any research method, the choice of SSO will depend on an assessment of the benefits derived from the knowledge gained against the costs in terms of both resource expenditure and impact on research subjects. Nevertheless, there will be circumstances in which SSO provides a valuable contribution to an evaluation, which would not have been possible from employing other methods.

References

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behaviour assessment methods to estimate academic engagement: A comparison of systematic direct observation and direct behaviour rating. School *Psychology Review, 39*(3), pp. 408-421.

Brown, R., & Evans, E. (2011). Four years after the Licensing Act 2003: A case study of Hartlepool town centre. *Safer Communities Journal, 10*(1), pp. 39-46.

Brown, R., & Evans, E. (2012).When intervention is a load of rubbish: Evaluating the impact of 'clean-up' operations. *Crime Prevention and Community Safety Journal, 14*(1), pp. 33-47.

Fetterman, D.M. (2010). Ethnography: step-by-step, Applied Social Research Method Series, Vol. 17. Thousand Oaks, California, USA: Sage Publications.

Hagan, F.E. (1982). *Research methods in criminal justice and criminology,* New York, NY: Macmillan Publishing.

Hadfield, P. (2006). Bar wars: *Contesting the night in contemporary British cities,* Oxford, UK: Oxford University Press.

Hewitt, L., & Kirby, S. (2011). The impact of the Licensing Act 2003 on drinking habits, offences of crime and disorder and policing in England's newest city. *Safer Communities Journal, 10*(1), pp. 31-38.

Homel, R., Carvolth, R., Hauritz, M., McIlwain, G., & Teague, R. (2004, March). Making licensed venues safer for patrons: What environment factors should be the focus of interventions? [Special Section: Prevention]. *Drug and Alcohol Review, 23*, pp. 19–29.

Hough, M., Hirschfield, A., & Newton, A, (2008). *The impact of the Licensing Act 2003 on levels of crime and disorder: evaluation methods,* Research Report 04, Appendix B. London: Home Office. Retrieved from http://modgov.sefton.gov.uk/moderngov/documents/s539/Government%20evaluation%20of%20LA03%20Report%20Annex%202.pdf

Hough, M., & Hunter, G, (2008). The 2003 Licensing Act's impact on crime and disorder: an evaluation. *Criminology and Criminal Justice, 8*(3), pp. 239-260.

Landsberger, H. A. (1958). *Hawthorne Revisited*, New York: Ithaca.

McCluskey, J. D., & Terrill, W. (2005). Departmental and citizen complaints as predictors of police coercion. Policing: *An International Journal of Police Strategies & Management, 28*(3), pp. 513-529.

Mastrofski, S. D., Parks, R. B., Reiss, A. J. Jnr, Worden, R. E., DeJong, C., Snipes, J. B., & Terrill, W. (1998). *Systematic observation of public police: Applying field research methods to policy issues*, Washington, DC: National Institute of Justice.

Mastrofski, S. D., Parks, R. B., & McClusky, J. D. (2010). Systematic social observation in criminology. In A. R. Piquero & D. Weisburd (Eds), *Handbook of Quantitative Criminology* (pp. 225-226). New York: Springer.

Maxfield, M., Babbie, E. (2005). *Research methods for criminal justice and criminology*, (4th Ed.). Belmont, CA: Wadsworth Thomson.

McCall, G. (1978). *Observing the law: Field methods in the study of crime and the criminal justice system*. New York, NY: Free Press.

Miller, P., Tindall, J., Sønderlund, A., Groombridge, D., Lecathelinais, C., Gillham, K., & Wiggers, J. (2012). *Dealing with alcohol-related harm and the night-time economy*, (DANTE) [Final report Monograph Series No. 43], Canberra, ACT: National Drug Law Enforcement Research Fund (NDLERF). Retrieved from http://www.google.com/url?url=http://www.ndlerf.gov.au/sites/default/files/publication-documents/monographs/monograph43.

Morleo, M., Harkins, C., Hughes, K., Hughes, S., & Lightowlers, C. (2007). The implementation and impact of the Licensing Act 2003 in Lancashire. Retrieved from http://www.cph.org.uk/wp-content/uploads/2012/08/the-implementation-and-impact-of-the-licensing-act-2003-in-lancashire.pdf

Odgers, C. L., Caspi, A., Bates, C. J., Sampson, R. J., & Moffitt, T. E. (2012). Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method, *Journal of Child Psychology and Psychiatry, 53*(10), pp. 1009-17.

Pike, S., O'Shea, J., & Lovbakke, J. (2008). *Early experiences of the Licensing Act 2003 in the East of England and Yorkshire and the Humber regions*, (Research Report 05). London, UK: Home Office.

Reiss, A.J. Jnr. (1971). Systematic observation of natural phenomena. *Sociological Methodology, 3*(1), pp. 3-33.

Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighbourhoods. *American Journal of Sociology, 105*(3), pp. 603-651.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing "Neighbourhood Effects": Social Processes and New Directions in Research. *Annual Review of Sociology, 28*, pp. 443-478.

Schulenberg, J. L. (2014). Systematic social observation of police decision-making: The process, logistics, and challenges in a Canadian context. *Journal of Quality and Quantity, 48*(1), pp. 297-315.

Sim, M., Morgan, E., & Batchelor, J. (2005), *The impact of enforcement on intoxication and alcohol related harm*, Wellington, NZ: Accident Compensation Corporation.

Smith, L., Morgan, A., & McAtamney, A. (2011), *Policing licensed premises in the Australian Capital Territory*, (Technical and Background Paper No. 48). Canberra, ACT: Australian Institute of Criminology.

Terrill, W., & Reisig, M. D. (2003).Neighbourhood context and police use of force. *Journal of Research in Crime and Delinquency, 40*(3), pp. 291-321.

Wilcox, P., Augustine, M. C., & Clayton, R. R. (2006). Physical environment and crime and misconduct in Kentucky schools. *Journal of Primary Prevention, 27*(3), pp. 293-313.

Zarrett, N., Sorensen, C., & Skiles, B. (2010). Environmental and social-motivational contextual factors related to youth physical activity: systematic observations of summer day camps. *International Journal of Behavioural Nutrition and Physical Activity, 10*(63), pp. 1-13.

When the best laid plans go astray: a case study in pragmatic approaches to evaluation

Hayley Boxall

Australian Institute of Criminology

Hayley.Boxall@aic.gov.au

Keywords: Family group conferencing evaluation, pragmatic evaluation, reflexive evaluation, adaptive evaluation, administrative data use in evaluation

Abstract

Many researchers recognise the importance of conducting evaluations in accordance with clearly articulated plans and frameworks which are developed at the outset of the project and underpin all of the subsequently undertaken research activities. However, for reasons often beyond the control of the evaluators, research methods and data collection instruments identified in these plans and frameworks may become unfeasible or inappropriate at different stages of the evaluation period. Mitigating the potential impact of these events and circumstances can be incredibly challenging for even the most seasoned researcher, who at the end of the day is still required to answer key evaluation questions, one of which may be 'Did it work?' Using the evaluation of the Family Group Conferencing pilot project (NSW) for illustrative purposes, this paper highlights the benefits of reflexive, adaptive and pragmatic approaches to evaluation and of involving project stakeholders in the development of evaluation designs and research methods.

Introduction

It is well established that evaluation research should be undertaken in accordance with clearly articulated and agreed upon plans and frameworks (Bamberger, Rugh & Mabry, 2012; McDavid, Huse & Hawthorn, 2013; Mertens & Wilson, 2012; Morgan & Homel, 2013; White & Phillips, 2012). One of the benefits of developing an evaluation plan and framework during the early stages of the project period is that it encourages the researcher to develop a series of clear research questions and to identify which data are necessary to answer these questions. The process for developing evaluation plans and frameworks is ideally undertaken in consultation with relevant stakeholders and any established evaluation committees and working groups. This consultative process helps researchers to identify issues and challenges that may have implications for the evaluation and to develop an understanding of the social and political context within which the project is being undertaken Compton & Baizerman, 2011; Mertens & Wilson, 2012).

However, for many researchers, particularly those who are brought onto projects on a consultancy/contractual basis, evaluation plans and frameworks may be developed during the very early stages of the research period. As a consequence, these evaluation plans and frameworks may be informed by partial information and often erroneous assumptions about how the program works 'on the ground', the quality and availability of evaluation data being collected by the contracting agency and program partners and the capacity and willingness of program staff to be involved in data collection processes. Evaluators also frequently encounter barriers and challenges throughout the research period that have implications for their ability to undertake the evaluation within the specified timeframes. This includes issues associated with evaluation data collection processes and the quality of the data itself (for example, the absence of baseline data, lack of appropriate and natural comparison group, small and non-representative samples and missing data; Bamberger, Rugh, Church & Fort, 2004; Eck, 2006).

Because of these factors, evaluators may be required to amend their plans and frameworks at various stages throughout the life of the project. In such situations, it is important that researchers adopt a pragmatic approach to evaluation – that they are able to adapt and amend their evaluation approach and research design and methodology so they are still able to collect the data necessary to answer key evaluation questions. A useful framework through which this flexible approach to evaluation may be understood and grounded is the aptly named 'Pragmatism' (for an overview see Morgan, 2007). Pragmatism is an attempt to reconcile what proponents view as the false dichotomy between positivist and interpretivist/constructionist research paradigms and their related emphasis on 'pure' qualitative or quantitative research methods (Feilzer, 2010; Johnson & Onwuegbuzie, 2004; Onwuegbuzie & Leech, 2005). Instead of single-method 'pure' research methodologies, pragmatists extol the benefits associated with mixed-methods research designs. For example, Onwuegbuzie and Leech (2005) suggest that mixed-methods research facilitates collaboration between researchers and the consideration of both micro and macro factors and analyses. Further, they suggest that the 'inclusion of quantitative data can help compensate for the fact that qualitative data typically cannot be generalized. Similarly, the inclusion of qualitative data can help explain relationships discovered by quantitative data' (Onwuegbuzie & Leech, 2005, p. 383).

Importantly, under the pragmatism framework the selection of research methods is 'needs-based' and underpinned by a 'contingency approach' (Johnson & Onwuegbuzie, 2004, p. 18). As Feilzer (2010, p. 14) notes, 'pragmatists do not "care" which methods they use as long as the methods chosen have the potential of answering what it is one wants to know'. Consequently, pragmatic frameworks give evaluators the freedom to select research methods that are feasible and will provide them with the data they need to answer the research questions, rather than to satisfy an epistemological need. This in turn facilitates flexible approaches to evaluation and the evolution of research designs and methods throughout evaluation periods.

Aim of this paper

This paper aims to provide evaluators, particularly early-career researchers, with practical advice on how to undertake pragmatic evaluations in situations where there are concerns about the collection and quality of data. More specifically, using a real world case study for illustrative purposes, this paper:

- provides an overview of some of the challenges that researchers frequently encounter when undertaking an evaluation in such situations;
- describes a range of strategies identified by researchers and practitioners for minimising the impact of and addressing these concerns;
- describes how these strategies have been applied in practice; and
- outlines the subsequent impact of these strategies.

Case study—the evaluation of the Family Group Conferencing pilot project

The Family Group Conferencing (FGC) pilot project was a small-scale project implemented in New South Wales in response to recommendations made as part of the *Special Commission of Inquiry into Child Protection Services in NSW* (Wood, 2008). The project commenced operation in March 2011 and was piloted in 11 community services centres (CSCs) located across metropolitan and regional areas of NSW. The overarching aim of the FGC pilot project was to empower families to develop, implement and manage Family Plans to address the care and protection issues raised by the statutory child protection agency operating in NSW – the Department of Families and Community Services (FACS). Conferences held as part of the project were attended by parents, the children and young people (where appropriate), extended family members, service providers, Community Services Caseworkers and Managers Casework; and chaired by trained Facilitators who were independent of Community Services. Conferences were conducted in neutral community-based venues and were focused on developing strategies that could be implemented by the family to address the identified care and protection issues.

In June 2011, FACS commissioned the Australian Institute of Criminology (AIC) to undertake a process and outcome evaluation of the FGC pilot project. To assess the implementation and short-term impact of the project, the allocated research team developed a program logic model and evaluation framework that aligned with the implementation plan for the evaluation of Keep Them Safe – the NSW Government's response to the Wood Commission (Urbis, 2011). This logic model and framework formed the basis of the AIC's evaluation and informed the development of a comprehensive quasi-experimental methodology combining quantitative and qualitative research methods.

In addition to the evaluation plan for Keep Them Safe, the FGC pilot project evaluation framework was informed by two other key pieces of information – the original tender documents provided by FACS, and the AIC's evaluation of another project that was implemented as part of Keep Them Safe (see Morgan, Boxall, Terer & Harris, 2012).

Unfortunately, due to time constraints the research team were not able to consult with the relevant project stakeholders or Evaluation Working Group (EWG) established by FACS, prior to drafting the evaluation plan. The EWG was comprised of individuals who were involved in the management and delivery of the FGC project in the ten sites, including the Project Coordinator.

As demonstrated in Table 1, over the course of the research period the original evaluation plan and framework was amended and adapted in response to a series of challenges (and opportunities) that were identified by the research team. Some of these challenges and opportunities – specifically, small sample sizes, unfeasible data collection methods and concerns regarding administrative data – will be discussed throughout the rest of this paper.

Table 1: **Overview of the evaluation methodology for the FGC pilot project at the beginning and end of the evaluation period**

Original evaluation methodology	***Final evaluation methodology***
Literature review to identify good practice	Literature review to identify good practice
Observation of a small number of conferences	Observation of a small number of conferences
Semi-structured interviews with parents and family members who participated in conferences held as part of the project.	Semi-structured interviews with parents and family members who participated in conferences held as part of the project.
Interviews, focus groups and a qualitative survey to seek feedback from stakeholders involved in the project (including FACS Case Managers).	Interviews, focus groups and a qualitative survey to seek feedback from stakeholders involved in the project (including FACS Case Managers).
Analysis of administrative data collected by FACS in relation to families who: • participated in a conference held as part of the project; and • were eligible for participation in the conference but were not asked.	Analysis of administrative data collected by FACS in relation to families who: • participated in a conference held as part of the project; • consented to participate in the project but did not participate in the conference; and • were eligible for participation in the conference but were not asked.
Survey of parents and family members: • at the end of conferences held as part of the project; and • at the end of a case meeting held between the family and the FACS Case Manager.	Case studies

Challenge 1: Small sample sizes

Many projects and programs struggle to identify and maintain a consistent and sufficient flow of referrals and participating clients throughout their lifespan (for FGC specific examples see Berzin, Cohen, Thomas, & Dawson, 2008; Brady & Miller, 2009; O'Brien, 2002; Shore, Wirth, Cahn, Yancey & Gunderson, 2002). This is particularly the case for projects:

- that are new – referring agencies may be unaware of the project and the referral processes involved, or sceptical of its merits and benefits;
- that are non-mandated processes – in other words, participation is voluntary; or
- constitute a significant departure from pre-existing processes – projects such as these may require referring agencies to change the way they approach certain issues and matters and there may be some resistance to this, particularly among long-term and senior staff (Boxall, Morgan & Terer, 2012).

Lower than anticipated project referrals numbers – and subsequently small sample sizes – can have significant implications for evaluation. In particular, small sample sizes have implications for the types of analyses that may be undertaken and the subsequent rigour and external validity of the findings[1] (Bamberger, Rugh, Church & Fort, 2004). Even so, 'small-n' evaluations are frequently undertaken and have benefit insofar as they provide useful information about the project or program being examined that may not be available otherwise (Eck, 2006; White & Phillips, 2012). Further, a number of existing research methods (e.g. success case methods, general elimination methodologies) and theoretical frameworks (e.g. realism) are compatible with small-n evaluations (Brinkerhoff, 2005; White & Phillips, 2012). The point of commonality among these methods and frameworks is that they aim to identify and understand the mechanisms and processes underlying projects and programs that bring about observable change.

Other strategies for minimising the impact of small sample sizes include the use of case studies and/or multiple sources of data to examine issues from various viewpoints and angles. This strategy is often referred to as triangulation and is commonly used in mixed-methods and qualitative studies (Bamberger, Rugh & Mabry, 2012; Mertens & Wilson, 2012; White & Phillips, 2012).

The evaluation of the FGC pilot project and small sample sizes

The original tender documentation provided to the research team indicated that 60 families would be invited to participate in the FGC pilot project and proceed to a conference. Consequently it was always anticipated that the evaluation would involve the analysis of data collected and extracted in relation to only a small number of families. However, during the early stages of the evaluation it became evident that referrals to the project were lower than expected[2]. To maximise the number of families that could participate in the FGC pilot (and therefore the potential sample size), FACS decided to extend the project period for an additional few months. As a consequence, by the end of the evaluation period 29 families had been referred to the project and participated in a conference. Another 31 families were referred to the project but had not proceeded to a conference for a range of reasons described in the final evaluation report (Boxall et al. 2012).

To partially mitigate the impact of the small sample size on the evaluation, the research team developed a series of case studies that were selected using success case sampling methods. This resulted in the identification of families who did and did not appear to benefit from the project, and an analysis of the factors that may have led to their success or lack thereof (for example, the number and identity of family members who participated in the conference). These case studies were developed using a range of information collected through the evaluation, including the analysis of administrative data, and were included in the final report to highlight key findings (Boxall, Morgan & Terer, 2012).

Further, because we had collected a range of information using different data collection methods (see Table 1), we were able to problematize and explore evaluation findings from various angles and viewpoints. This increased the validity and reliability of our findings. Finally, the research team provided members of the EWG and other relevant project stakeholders with opportunities to provide their interpretation of the data and preliminary findings at different stages during the evaluation. These 'member checks' again increased the reliability of the findings (Mertens & Wilson, 2012).

Challenge 2: Unfeasible/Inappropriate data collection methods

As mentioned earlier in this paper, during the early stages of an evaluation, a researcher's understanding of the project may be limited and in some cases based on a series of erroneous assumptions. In situations such as this, research methods and data collection tools that are initially identified by evaluators may be unfeasible or even inappropriate. For this reason, it is beneficial for researchers to involve project staff and established evaluation working groups in the initial planning stages of the evaluation (Mertens & Wilson, 2012; O'Sullivan, 2012). There are numerous frameworks that explicitly encourage this process, including collaborative, empowerment, participatory and utilization evaluation. One of the functions of this engagement process is to 'reality-test' proposed research methods and advise the researchers on appropriate research approaches (American Evaluation Association, 2004). This is particularly important if the collection of data are reliant on the participation and support of individuals involved in the delivery and management of the project itself.

[1] It is important to note that there are some qualitative research methods that involve the in-depth examination of one or a small number of observations. This includes case studies, which were developed as part of the evaluation of the FGC pilot project.

[2] The process evaluation identified a number of factors that contributed to this phenomenon, including a lack of awareness of the project among FACS staff (who were responsible for referring matters in the first instance), the perceived cumbersome and time-consuming referral processes themselves and that participation in the project was voluntary (Boxall, Morgan & Terer, 2012).

Unfeasible/Inappropriate data collection methods and the FGC pilot project

Due to time constraints, the research team were unable to consult with the EWG prior to drafting the initial evaluation framework. However, shortly after submitting the evaluation framework we had an opportunity to meet with the EWG who were tasked with providing feedback on the document. Throughout this consultation process, it became apparent that there were potential issues administering surveys to; (1) parents and family members at the end of conferences held as part of the project (the intervention group); and (2) at the end of case planning meetings held with families not involved in the program (the comparison group) Because conferences held as part of the FGC pilot project could and occasionally did run for the entire day and were emotionally draining for all involved, the EWG suggested that asking parents and family members to complete a survey at the end of this process was inappropriate. Also, in order to attend conferences, family members often had to take time off work or make alternative care arrangements for their children. Consequently, family members frequently had to leave during or immediately after conferences and so did not have time to complete a survey. Similar issues were raised in relation to the administration of surveys to parents and family members at the end of case planning meetings held with FACS Case Managers[3].

In light of the issues raised by members of the EWG, the research team decided not to administer surveys to family members and parents after conferences or case planning meetings. This decision was informed by the consideration of both the potential impact of removing these research methods from the evaluation plan and of not removing them. Although their non-inclusion limited our ability to explore evaluation findings from various angles (triangulation), we determined that it was feasible to answer the identified research questions without the survey data. Further, because of issues identified by the EWG, the usefulness and reliability of this data was potentially limited anyway (i.e., non-representative sample of family members). However, more importantly, if the research team forged ahead with their plans of administering the surveys we ran the risk of jeopardising our working relationship with the EWG who may have perceived that their opinions had not been acknowledged and valued and disengage from the evaluation as a result. This was of particular concern considering that many of the research methods identified in Table 1 were dependent on the support and involvement of FGC staff and FACS Case Managers. In light of these considerations, the research team were comfortable that the potential benefits associated with removing the surveys from the evaluation methodology outweighed the drawbacks.

Challenge 3: Concerns regarding administrative data

Many program evaluations involve the analysis of administrative data collected, stored and managed by staff and agencies involved in the delivery and implementation of the project. This data may include information relating to the processes involved in the project (for example, project participation numbers), outputs (e.g. treatment and intervention plans) and also outcomes. While administrative data can be an important and integral element of any evaluation, there may be issues associated with this information. For example:

- the data may not be in a format suitable for analysis – information may only be available in hardcopy;
- there may be 'gaps' in the data – data collection protocols may have changed during the project period meaning the same information was not collected for all participants; and
- they may be some concerns regarding the quality of the data – due to inconsistent data collection procedures, multiple people inputting data into databases and spreadsheets etc (Bamberger, Rugh, Church & Fort, 2004).

All of the above identified issues have implications for the use of administrative data in evaluation and need to be managed carefully. Consequently, it is advisable that researchers request to see a sample extract from the relevant databases and speak to the contracting agency's data custodians and managers during the early stages of the evaluation. Alternatively, researchers may seek to conduct an 'audit' of existing databases and data collection processes early in the evaluation period. This process may be facilitated through the use of a quality assessment 'checklist' such as that developed by Statistics Sweden (Daas, Ossen & Arends-Toh, 2009). However, undertaking such auditing processes may be unfeasible or inappropriate, depending on the contracting agency, its data collection processes and the political context.

Concerns regarding administrative data and the FGC pilot project

The evaluation of the FGC pilot project involved the analysis of a range of administrative data collected by staff directly involved in the management and delivery of the FGC project and FACS. In particular, as demonstrated in Table 1, from the outset we planned to extract and analyse data stored in the Key information and Directory System (KiDS) for families included in the intervention and comparison groups to determine the impact of the project against a number of short-term and intermediate outcomes (e.g. re-contact with FACS).

During the early stages of the evaluation period, we were notified by the EWG and FACS data custodians that there were some issues with the KiDS data that may impact the evaluation. Specifically:

- KiDS had only been operating since the early 2000s and so did not include information in relation to families, parents and children prior to this point in time;

[3] It is important to note that all research undertaken by the AIC involving human subjects requires clearance from the Institute's Human Research Ethics Committee (HREC) which is comprised of experienced researchers in various fields who are not AIC employees. The evaluation of the FGC project involved the development of a comprehensive and detailed ethics application that acknowledged and addressed the perceived ethical issues associated with involving families not involved in the project in the evaluation.

- there was a three month time delay associated with data being entered into KiDS for children being managed by the OOHC teams; and
- information about Family and Children's Courts legal orders made in relation to families managed by FACS was not included in KiDS – rather, this information was managed by FACS legal department.

All of these factors had potential implications for the evaluation. First the absence of historical information about families meant that contextual understandings of families involved in the project (and the factors that may have influenced their success or lack thereof) were limited. Second, the time lag associated with data being uploaded for Out-of-Home-Care (OOHC) matters meant that our 'follow-up' period for these families could have been three months shorter than for non-OOHC families. Finally, the lack of information about legal orders meant that our ability to identify the impact of the project on legal outcomes (e.g. restoration orders) was potentially limited.

Because the research team had consulted with the EWG and FACS data custodians during the early stages of the evaluation, we were able to anticipate and where possible, implement strategies to manage these issues. While there was nothing we could do about the loss of historical data, we discovered that this information was, to a limited extent, being included in the referral reports completed by families and FACS Case Workers during the early stages of their engagement on the project. Consequently, the research team extracted this information from the referral reports and included it in the analysis.

Similarly, there was not much that could be done to minimise the impact of the three month delay in data being uploaded into the KiDS for OOHC matters. This said, because we identified this as an issue from the outset, we were able to stagger the data extraction dates. This meant that data was extracted for all non-OOHC matters at the end of 2013 while the same data was extracted for OOHC matters three months later. This ensured that the data extraction period for all matters was approximately the period of time. Finally, although the legal data we needed was not available through KiDS, we consulted with FACS' legal department and were able to obtain the information we needed. This data extract was subsequently 'linked' with our other data extracts using the unique case plan and child identifiers generated by FACS.

Implications for evaluation quality and rigour

Any changes that researchers make to their evaluation frameworks and plans have potential implications for the collection and analysis of data and, in turn, their ability to answer identified evaluation questions. The FGC pilot program was no exception. For example, extending the project delivery period had both positive and negative ramifications for the evaluation. On the one hand, extending the project period resulted in a larger number of families being provided with the opportunity to participate in the project, which in turn meant there was a larger population to draw from for the purpose of the evaluation. However, by extending the project period, the length of time that families could be 'followed' post-intervention was in turn limited. Consequently, the research team were only able to determine the short-term impact of the FGC pilot project.

Further, the removal of the post conference/case planning survey meant that our ability to assess the impact of the program against short-term outcomes (e.g. increased satisfaction of parents and family members with their role in decision-making processes) was limited; as was our ability to compare the outcomes of the FGC pilot against a 'standard practice' approach. Because of these identified limitations, the research team recommended that 'A future evaluation should be conducted to measure the longer term impact of FGC on care matters once the program has been fully established and data on a larger number of participants is available. Processes for monitoring outcomes from FGC therefore need to be established and/or maintained' (Boxall, Morgan & Terer, 2012, p. xvi).

However, it is important to note that some of the changes that were made to the evaluation framework and plan actually *increased* the rigour of the research and the subsequent validity of the findings. For example, as previously discussed the research team removed the post-conference/case planning survey from the evaluation methodology in part because the concerns raised by the EWG suggested that the data collected through this process may be of limited use and reliability (e.g. low response rate, meaning the external validity of the data would be low). By removing a potentially flawed and problematic research method, the research team improved the quality and reliability of the evaluation design and the subsequent findings, and also freed up additional resources for the collection and analysis of other data.

Conclusion

The FGC pilot case study described throughout this paper demonstrates that initial plans and frameworks developed by evaluators will rarely be 'perfect'. Consequently, while it is beneficial and advisable to conduct evaluation research in accordance with clearly articulated plans and frameworks, it is also necessary for evaluators to be flexible and adaptive so they can respond appropriately to challenges they encounter. This may involve removing specific research methods, including others, and amending data collection tools or processes. Consequently, these evaluation plans should be viewed as flexible working documents that will be changed, tweaked and amended as new challenges and issues emerge and the research team's understanding of the program improves. This said, decisions to amend these plans and frameworks should not be made lightly and should take a range of factors into consideration – one of the most important being, are we still going to be able to answer our identified evaluation questions if we make these proposed amendments/changes?

The case study also highlights the importance of engaging with relevant project stakeholders throughout the evaluation period, and in particular, seeking their input into the evaluation design and specific research methods. Project stakeholders may be asked to undertake a range of tasks throughout the evaluation process, such as collecting and brokering access to data and reality testing proposed research methods. Ensuring that project stakeholder's views are acknowledged and, where appropriate, integrated into the evaluation plan goes some way to ensuring that strong working relationships are maintained and potential issues identified in relation to data collection processes; early identification of data and strategies to minimise any identified issues implemented.

References

American Evaluation Association. (2004). *American Evaluation Association guiding principles for evaluators.* Retrieved from http://www.eval.org/p/cm/ld/fid=51

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation, 25*(1), 5–37.

Bamberger, M., Rugh, J., & Mabry, L. (2012). *Real World evaluation: Working under budget, time, data, and political constraints.* London: Sage Publications.

Berzin, S.C., Cohen, E., Thomas, K., & Dawson, W.C. (2008). *Does family group decision making affect child welfare outcomes? Findings from a randomized control study.* Child Welfare 87(4), pp. 35–54.

Boxall, H., Morgan, A., & Terer, K. (2012). *Evaluation of the Family Group Conferencing pilot program.* Research & Public Policy Series no. 121, Canberra: AIC.

Brady, B., & Miller, M. (2009). *Barnardos Family Welfare Conference Project, South Tipperary: Evaluation Report.* Retrieved from http://childandfamilyresearch.ie/sites/www.childandfamilyresearch.ie/files/barnardos_family_welfare_conference_service_south_tipperary_evaluation_report.pd

Brinkerhoff, R.O. (2005). The success case method: A strategic evaluation approach to increasing the value and effect of training. *Advances in Developing Human Resources, 7*(1), pp. 86–101.

Compton, D.W.& Baizerman, M. (2011). Managing evaluation: responding to common problems with a 10-step process. *Canadian Journal of Program Evaluation, 25*(2), pp. 103-123.

Daas, P.J., Ossen, S.J. & Arends-Tóth, J. (2009). *Framework of Quality Assurance for Administrative Data Sources.* Statistics Netherlands: The Hague, retrieved from http://pietdaas.nl/beta/pubs/pubs/ISI2009_paper.pdf

Eck, J.E., 2006. When is a bologna sandwich better than sex? A defense of small-n case study evaluations. *Journal of Experimental Criminology, 2*(3), pp. 345–362.

Feilzer, M.Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research, 4*(1), pp. 6–16.

Johnson, R.B.& Onwuegbuzie, A.J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), pp. 14–26.

McDavid, J., Huse, I. & Hawthorn, L. (2013). Key concepts and issues in program evaluation and performance measurement. In McDavid, J., Huse, I. & Hawthorn, L. (Eds), *Program Evaluation and Performance Measurement: An Introduction to Practice*. London: Sage Publications.

Mertens, D.M. & Wilson, A.T. (2012). *Program evaluation theory and practice: A comprehensive guide.* New York: The Guilford Press.

Morgan, A., Boxall, H., Terer, K. & Harris, N. (2012). *Evaluation of alternative dispute resolution initiatives in the care and protection jurisdiction of the NSW Children's Court.* Research & Public Policy Series no. 118, Canberra: AIC.

Morgan, A. & Homel, P. (2013). *Evaluating crime prevention: lessons from large scale community crime prevention programs.* Trends & Issues in Crime & Criminal Justice no. 458.

Morgan, D.L. (2007). Paradigms lost and pragmatism regained methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), pp. 48–76.

O'Brien, V. (2002). *Family group conference pilot project: evaluation report.* Mid-Western Health Board. Retrieved from http://irserver.ucd.ie/bitstream/handle/10197/3085/MID%20Western%20Health%20Board%20FGC%20Report%202002.pdf?sequence=1

O'Sullivan, R.G. (2012). Collaborative evaluation within a framework of stakeholder-oriented evaluation approaches. *Evaluation and Program Planning, 35*(4), pp. 518–522.

Onwuegbuzie, A.J. & Leech, N.L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), pp. 375–387.

Shore, N., Wirth, J., Cahn, K., Yancey, B. & Gunderson, K. (2002). *Long term and immediate outcomes of family group conferencing in Washington State.* Washington State: International Institute for Restorative Practices. Retrieved from http://www.iirp.edu/iirpWebsites/web/uploads/article_pdfs/fgcwash.pdf

Urbis. (2011). *Implementation plan for evaluation of Keep Them Safe. Sydney:* Urbis Pty Ltd. Retrieved from http://www.dpc.nsw.gov.au/_data/assets/pdf_file/0019/125146/Urbis_Final_KTS_Implementation_Plan_-_publication_version.pdf

White, H. & Phillips, D. (2012). *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework.* New Delhi: International Initiative for Impact Evaluation.

Wood, J. (2008). *Report of the special commission of inquiry into child protection services in NSW: Volume 2.* Sydney: NSW Government. Retrieved from http://www.dpc.nsw.gov.au/publications/news/stories/?a=33797

Improving the evaluation of crime prevention and reduction programs through research-practitioner partnerships

Anthony Morgan

Australian Institute of Criminology

Anthony.morgan@aic.gov.au

Keywords: Researcher-practitioner partnerships, Evaluation stakeholders, Stakeholder involvement, Crime prevention, Crime reduction, Criminal justice, Quasi-experimental design, Scientific methods.

Abstract

While there is widespread recognition of the importance of stakeholder involvement in evaluation, less attention has been given the role of researcher-practitioner partnerships in evaluations using rigorous scientific methods, such as quasi-experimental designs. This is particularly true for the evaluation of crime prevention and reduction programs, including those conducted by evaluators independent of program design and delivery. Reflecting on several evaluations of programs designed to prevent and reduce crime and respond to the needs of vulnerable populations in court settings, this paper highlights the benefits, challenges and lessons from working in partnership with practitioners to conduct rigorous outcome evaluations. While evaluations are often conducted by someone independent of program management and delivery; it is still important for researchers to work in partnership with policy makers, program managers and project staff to ensure evaluations are methodologically rigorous, successfully implemented and focus on delivering practical recommendations for action.

Introduction

The importance of working with stakeholders is now firmly established in evaluation theory and practice (Brandon & Fukunaga, 2014). Involving stakeholders in an evaluation is believed to offer a number of benefits, such as increasing support for evaluation and the use of evaluation findings, and for this reason is a common feature across a range of different approaches to evaluation. Despite this, few areas of evaluation are as divisive as determining the optimal nature and extent of stakeholder involvement in an evaluation, and the level of involvement encouraged therefore varies depending on the approach that is adopted (Hall, 2008).

The aim of this paper is to demonstrate the value of partnerships between researchers and practitioners in those situations in which an external evaluator, independent of program design and delivery, has been commissioned to undertake the evaluation. In particular, this paper argues that the relationship between these external evaluators and agencies commissioning evaluation should be viewed as a partnership, and that working in partnership can increase the likelihood that rigorous scientific methods will be used to measure program effects and that, where they are used, will be more likely to be successfully implemented. The paper begins by reflecting on the standard of crime prevention and criminal justice evaluation in Australia and describing some of the barriers to high quality evaluation. Using several recent evaluations as case studies, the benefits and challenges associated with researcher-practitioner partnerships in evaluations involving quasi-experimental methods are then described. The paper ends by outlining some key lessons for working collaboratively with evaluation stakeholders.

Evaluating strategies to prevent and reduce crime

There has been rapid growth in the evaluation of strategies designed to prevent and reduce crime (National Research Council, 2005; Tilley & Clarke, 2006; Weatherburn, 2005). While there is certainly room for improvement, government agencies are now more likely than ever to allocate funds within program budgets for evaluation research, with a view to commissioning independent evaluators to provide an objective and impartial assessment of the effectiveness, efficiency and appropriateness of policies and programs (New South Wales, Department of Premier & Cabinet, 2013; Western Australia, Program Evaluation Unit, Department of Treasury, 2014). This is reflected in the amount of evaluation activity that is undertaken by government researchers, academics and consultants (as well as internal evaluations conducted by practitioners), the number of researchers working in evaluation, and the level of procurement activity managed by criminal justice authorities.

There are several different approaches to measuring the impact of programs designed to prevent and reduce crime. Selecting an appropriate evaluation design and research method requires consideration of the characteristics of a program, the purpose of the evaluation, the available options, and the views of key stakeholders (English, Cummings & Stratton, 2002; National Research Council, 2005). However, experimental (especially quasi-experimental) and observational methods are the most common approaches used in crime prevention and criminal justice research (Idriss, Jendly, Karn & Mulone, 2010). Given the influence of experimental methods, the Scientific Methods Scale (SMS) was developed to assess the quality of outcome evaluations in crime prevention and criminal justice research. The SMS forms the basis of systematic reviews of crime prevention methods undertaken by the Campbell Collaboration (Farrington, Gottfredson, Sherman & Welsh, 2006; Sherman, Farrington, Welsh & MacKenzie, 2006), while similar criteria have been used by the Washington State Institute of Public Policy (Lee et al. 2012). The SMS is primarily focused on ensuring the highest possible level of internal validity and drawing valid conclusions regarding the causal relationship between interventions and the outcomes observed. The scale ranges from a correlation between a program and a measure of the outcome (level one) through to randomised control studies (level five), which are widely (but not universally) regarded as the gold standard for evaluation research (Farrington et al. 2006; Tilley & Clarke, 2006). A research design that achieves level three on the SMS, with measures of the outcome (usually a reduction in crime) pre and post intervention and an appropriate comparison group against which to compare results (a quasi-experimental design) is considered the minimum design for drawing valid conclusions about the effectiveness of a strategy (Farrington, et al. 2006; Sherman, Gottsfredson, MacKenzie, Eck, Reuter & Bushway, 1998).

Applying this standard, large-scale systematic reviews have shown that there is an accumulated body of high quality research demonstrating the effectiveness of crime prevention and criminal justice strategies (e.g. Lee et al. 2012; Sherman et al. 1998; Sherman, Farrington, Welsh & MacKenzie, 2006). Most of these reviews have acknowledged the need to improve the standard of evaluation practice, with many studies failing to meet the criteria for inclusion.

However, the number of Australian initiatives included in the systematic reviews, meta-analyses and databases describing effective (and ineffective) interventions are relatively small when compared to other countries (Morgan & Homel, 2013). This is particularly apparent when considered alongside the level of crime prevention activity (as illustrated by the Australian Crime and Violence Prevention Awards and the number of commonwealth, state and territory funding programs), the financial resources invested in the criminal justice system (Steering Committee for the Review of Government Service Provision [SCGRSP], 2014), and the volume of evaluation research that has been conducted. There has only been a handful of randomised control studies conducted in Australia (Lind et al. 2002; Jones, 2011; Mazerolle, Antrobus, Bennett & Tyler, 2013; Sherman, Strang & Woods, 2000). A review of community-based crime prevention strategies suitable for local government to address a number of common crime types in NSW found that, despite the emphasis on the important role of local government in crime prevention in this country for more than two decades, fewer than 20 Australian studies met the criteria for inclusion (Morgan, Boxall, Lindeman & Anderson, 2012). Similarly, a recent research project by the Australian Institute of Criminology exploring the evidence in support of police crime prevention in Australia has also revealed gaps in the amount and quality of evaluation research (Morgan & Mann, forthcoming). There is a large body of Australian evidence for certain responses, including random breath testing (Hendrie, 2003; Shults et al. 2001), community-based and regulatory responses to alcohol-related violence (Morgan, Boxall, Lindeman & Anderson, 2012; National Drug Research Institute, Curtin University of Technology, 2007), cautions and conferencing (Morgan & Mann, forthcoming) and a significant body of research (much of which has been undertaken by the NSW Bureau of Crime Statistics & Research) into the impact of different court initiatives and sentence options.

There are several possible explanations for the variable quantity and quality of Australian studies evaluating the impact of strategies to reduce and prevent crime. It is likely to be a combination of factors such as insufficient funding for robust experimental designs (which can be costly), a reluctance among program managers to subject their programs to rigorous evaluation for fear of negative results, disagreement between evaluators and program managers on questions of attribution, problems accessing meaningful data on program effects, challenges identifying suitable comparison groups or areas (particularly outside of institutional settings or when evaluating place-based strategies), or the limited skills, knowledge or experience among those entrusted with evaluation (Morgan & Homel, 2013). It may also be due to ideological differences about the best way to assess the effectiveness of strategies to prevent and reduce crime (Tilley & Clarke, 2006).

Improving evaluation standards through partnership approaches

This paper argues that some of these common barriers to rigorous outcome evaluations, particularly evaluation based on quasi-experimental research designs, can be overcome by establishing and maintaining effective partnerships between evaluators and practitioners (in this paper practitioner collectively refers to policy makers, program managers and project staff). This is not a new concept – the importance of working with stakeholders in evaluation is well established (Brandon & Fukunaga, 2014; Plottu & Plottu, 2009). Stakeholder involvement is a common principle underpinning many different approaches to evaluation (Brandon & Fukunaga, 2014), For example, utilisation-focused approaches are underpinned by a commitment to meeting the needs and requirements of the intended user/s (which informs the decision about the evaluation approach and methods) and the participation of these users in decision-making at each stage of the evaluation (Patton 2008). In participatory approaches to evaluation, stakeholders are involved in all aspects of the evaluation and empowered to own the evaluation process, fulfilling roles normally assigned to the evaluator and building evaluation skills and knowledge in the process (Greene, 2006). Reflecting the widespread acceptance of the important role of stakeholders in evaluation, there has even been some acknowledgement of the need to involve stakeholders as part of quantitative approaches to measuring the impact of programs, including areas such as education research (Datta, 2006).

While these approaches differ in terms of the philosophical position they adopt and their rationale for involving stakeholders, there are some common themes in terms of the range of benefits that are believed to result from involving stakeholders in the process of planning an evaluation, collecting and analysing data and reporting findings and recommendations. These include increasing the likelihood that evaluation findings will be used, facilitating access to better quality data and contextual information and enhancing the perceived credibility and validity of the evaluation design, methods and results among key stakeholders (Braga, 2013; Brandon & Fukunaga, 2014; Bryson, Patton & Bowman, 2011; Cullen, Coryn & Rugh, 2011; Patton, 2008; Plottu & Plottu, 2009). Conversely, failing to engage with stakeholders can lead to missed opportunities or result in evaluations that are inaccurate or insensitive to the needs of stakeholders, which may actually deter future investment in evaluation (Bryson, Patton & Bowman, 2011).

Despite the important role of stakeholders in evaluation, and the potential value of researcher-practitioner partnerships, little attention has been given to these arrangements within criminal justice and crime prevention evaluations. For example, a report by the National Research Council (2005) put forward a number of recommendations to improve the evaluation of anticrime programs, with a particular focus on increasing the number of experimental and quasi-experimental designs. While the report argued that both evaluators and policy makers needed to do more to increase the amount and methodology quality of impact evaluations, little attention was paid on how the two parties might better work together. Similarly, commentators arguing for criminal justice evaluations that are methodologically rigorous have tended to be preoccupied with debates about the relative strength of different research designs and methodological approaches, particularly in terms of maximising internal and external validity (Braga, Welsh & Bruinsma, 2013; Idriss et al. 2010; Tilley & Clarke, 2006). While there have been a number of studies exploring research utilisation by criminal justice agencies (Lum, Telep, Koper & Grieco, 2012), and others exploring the impact of researcher-practitioner partnerships on evaluation findings (Petrosino & Soydan, 2005; Welsh, Braga & Hollis-Peel, 2012), there has been limited research into researcher-practitioner partnerships in criminal justice research and, as a result, little is known about the factors that contribute to successful partnerships for evaluation (Alpert, Rojek & Hansen, 2013). This paper attempts to address this gap.

Benefits of working in partnership with policy makers, program managers and project staff

There are different arrangements in terms of who is responsible for an evaluation and their relationship to the program being evaluated. Petrosino and Soydan (2005) reviewed more than 300 individually-focused crime reduction programs and identified three categories of evaluation teams, each with a number of sub-categories: internal (program developer/creator, program/agency staff, government evaluator), external (academic researchers, private research firm, foundation/not for profit) and collaborative approaches (academic/practitioner, academic/government). The Australian Institute of Criminology (AIC), which is frequently engaged by both Commonwealth and state and territory government agencies to evaluate programs designed to prevent and reduce crime, has typically been engaged as an external evaluator, independent of program design and delivery. As an external evaluator, the AIC is tasked with making an objective and impartial assessment of the impact of the program being evaluated, drawing upon its specialist expertise to design and implement rigorous evaluation methodologies to produce valid, transparent and defensible findings.

In performing this role, researchers involved in these evaluations have frequently approached the relationship with the agency commissioning the evaluation as a partnership. This meant working closely with that agency (and other partners) during the planning stages of the evaluation, data collection and analysis and during the reporting of evaluation findings. Several examples of both published and unpublished evaluations conducted by the AIC are used to describe these partnership arrangements, as well as the benefits that resulted:

- Evaluation of the Queensland Murri Court (Morgan & Louis 2010): An Indigenous sentencing court for both adult and youth defendants that allowed greater input from the Aboriginal and Torres Strait Islander (ATSI) community into the sentencing process.

- Evaluation of the Queensland Special Circumstances Court Diversion Program (SCCDP) (unpublished): A court diversion program for defendants who were homeless or who had a mental illness or intellectual disability.

- Evaluation of alternative dispute resolution initiatives in the care and protection jurisdiction of the NSW Children's Court (Morgan, Boxall, Terer & Harris, 2012): A new model of dispute resolution involving the use of conciliation and mediation to resolve child protection disputes before the Children's Court.

- Evaluation of Indigenous drug and alcohol treatment programs (unpublished): Six community-based residential rehabilitation programs in four jurisdictions designed to reduce substance misuse, particularly among ATSI people.

- Policing licensed premises in the ACT (Smith, Morgan & McAtamney, 2011): A strategy to reduce alcohol-related crime in the ACT's main entertainment precinct, involving intelligence gathering, education for licensees, proactive enforcement of liquor licensing and high visibility policing.

- Partnership between the AIC and CrimTrac (recently commenced): A new program of research focused on the evaluation of information systems and services for law enforcement agencies.

In the majority of these evaluations, the AIC was commissioned through a procurement process to conduct an evaluation of the program, which had important implications for how the evaluations were managed and the nature of partnership arrangements. For this reason, most of the following discussion centres on partnerships between practitioners and evaluators independent of program design and delivery.

Planning the evaluation

Identifying and involving evaluation stakeholders early in the process of planning an evaluation provides a range of benefits, irrespective of the approach to evaluation being used (Brandon & Fukunaga, 2014). However, there are certain benefits that are particularly relevant to the types of evaluations that are the focus of this paper. First, stakeholders can assist with determining what outcomes can and should be attributed to the program being evaluated and, therefore, which outcomes should be the focus of an evaluation. This helps guide the evaluation, determine appropriate evaluation questions, and ensures that there is agreement regarding the scope of the evaluation well in advance of data collection and reporting (avoiding future disagreements). Involving stakeholders in discussions about how a particular program and its component parts contribute to desired outcomes can also empower and encourage the meaningful participation of stakeholders in evaluation, which can 'dramatically increase the chances that the evaluation will meet utility, feasibility, propriety and accuracy evaluation standards' (Donaldson & Lipsey, 2006, pp. 65-66).

One way of approaching this is to involve stakeholders in the development of a program logic model and evaluation framework. A program logic model describes the main activities that will be delivered as part of a program, and the relationship between these activities and the hierarchy of short, intermediate and long-term outcomes. This helps to determine what outcomes can be reasonably attributed to the program and makes explicit the underlying theory about how a program contributes to these outcomes (Funnell & Rogers, 2011) – hypotheses that can be tested using rigorous scientific methods. In addition to being a powerful communication tool for criminal justice evaluation (Willis & Tomison, 2014, this issue), the logic model also forms the basis of an evaluation framework, which details the specific evaluation questions and performance indicators that will be measured (and how) as part of the evaluation. This enables stakeholders to have input into the design of the evaluation, including data collection processes, which can encourage buy in, facilitate access to data and identify and address potential barriers to conducting the evaluation.

This approach has been used extensively in AIC evaluations. Undertaken in consultation with key stakeholders, the development of program logic models has helped to resolve important conceptual questions (and concerns) about attribution. In an evaluation of six Indigenous drug and alcohol treatment programs, the logic model that underpinned the evaluation (developed with input from program operators) was used to demonstrate the hypothesised link between improved health and wellbeing – the direct consequence of effective treatment – and reduced recidivism. This was particularly helpful in convincing key stakeholders of the need to measure reoffending for programs that did identify this as an explicit objective. Using a quasi-experimental research design (comparing program completers and non-completers), the evaluation concluded that there was evidence that treatment had a positive impact on reoffending, helping to address a significant gap in the available evidence base. More recently, the use of a program logic approach has been used to determine what outcomes could be attributed to a number of law enforcement data services provided by CrimTrac, and persuade policing agencies of the need to measure these important outcomes. These logic models were developed with considerable input from both CrimTrac and their police agency partners.

Involving stakeholders in the evaluation planning process can also help to overcome resistance to the use of quasi-experimental research designs and the challenges associated with identifying suitable comparison areas. In the evaluation of alternative dispute resolution for care and protection matters in the NSW Children's Court, the evaluators worked closely with the evaluation working group to develop an innovative solution to challenges associated with determining an appropriate comparison group comprising matters that did not go through the alternative model of decision making. This was necessary because the program was rolled out state wide (i.e., there was no natural comparison area). Given the significance of the reforms, which were introduced in response to a Royal Commission, the working group was reluctant to restrict access to the program group for the purpose of evaluation. The research team therefore negotiated a retrospective comparison group, including appropriate parameters and a process for collecting additional data for this purpose. The final report on the evaluation compared the outcomes for matters referred to ADR and this comparison group, including a cost savings comparison, and the findings from this analysis were instrumental in informing a recommendation to continue to support the use of ADR – a recommendation that was supported by the then Attorney General (Morgan, Boxall, Terer & Harris, 2012).

Similarly, in the evaluations of the Queensland Murri Court (Morgan & Louis, 2010) and the Queensland Special Circumstances Court Diversion Program, the evaluators worked closely with the Department that commissioned the evaluation and the program Steering Committee (comprising representatives from each agency involved in the program) to identify a suitable comparison group of offenders who did not participate in the program. There was resistance to this approach among some stakeholders – due largely to a limited understanding of the rationale for the comparison group – and various practical barriers to accessing data on offenders in the comparison group, but these were overcome through careful negotiation, by explaining both the strengths and limitations of the different options, and working collaboratively with court partners to develop an appropriate solution. Working with these stakeholders during these early stages to determine the comparison group was important because these same stakeholders were responsible for reviewing the findings from the evaluation. Had the evaluators simply decided on the eligibility criteria for the comparison group in isolation and without consultation, the process of collecting the data and the results from any analysis would likely have been met with significant resistance and scepticism. As it was, the collaborative and consultative approach was particularly helpful given the eventual findings from the evaluation of the Queensland Murri Court (which is discussed later).

Conducting the evaluation

A collaborative approach to evaluation recognises the potential of involving program managers and staff in the data collection process, rather than only seeing them as sources of data for the purpose of the evaluation (Cullen, Coryn & Rugh, 2011). Involving program staff in this way can help to increase the perceived credibly and validity of the findings. It is common for there to be significant gaps in the data required to measure the outcomes from crime prevention and criminal justice programs. Administrative databases – where they exist – are used extensively to measure the impact of criminal justice programs. However, these databases are often not established for the purpose of evaluation, and there are limitations to their use. For example, the evaluations of the Queensland Murri Court and the Queensland Special Circumstances Court Diversion Program (SCCDP) were undertaken as part of long-term partnership with the Department of Justice and Attorney General, and followed earlier evaluations of the Queensland Drug Court. Recognising the need to utilise court data to measure the impact of the Queensland Murri Court, and to integrate court data with other information on important risk factors for future offending, the Department developed a new system for recording information on all court participants and the comparison group. The AIC was instrumental in helping to design this database, working closely with the Department to build a system that would eventually be used for all court innovation programs across the state. This was a time consuming process, but was critical in ensuring that adequate data were available to measure the impact of the Murri Court and Queensland SCCDP (and other programs) on recidivism.

In the evaluation of alternative dispute resolution (ADR) in the NSW Children's Court, there was no pre-existing administrative database for care and protection matters. This meant that initially data were not available to measure the impact of ADR in the time required to finalise care matters or the proportion of matters that resulted in a court hearing. The AIC worked very closely with the evaluation working group to establish a process for extracting information from the hardcopy case files in a systematic way, and to negotiate a process by which the information was extracted by program staff. This could only have been achieved by working in partnership with the evaluation working group, and ensured access to data for the analysis of key outcomes and cost-savings.

Further, as part of this same evaluation, a number of other stakeholders were involved in data collection. Conference chairs (registrars and mediators) completed a report on the outcomes of each conference for the purpose of the evaluation, and also disseminated a brief satisfaction survey to conference participants – whom the evaluators would have struggled to survey and achieve a similar response rate. In the evaluation of policing responses to alcohol related violence in and around licensed premises in the Australian Capital Territory (ACT), the AIC worked with ACT Policing to develop and implement a 'place of last drink' form for all alcohol-related incidents attended by police. In the evaluation of the Queensland SCCDP and evaluation of Indigenous drug and alcohol treatment programs, case managers completed a client questionnaire designed by the AIC at program entry and exit to measure change in the health and wellbeing of participants. In each of these examples, program staff had better access to the information source (i.e. program participants) than the evaluators and were able to capture information that would have far more difficult to collect otherwise. Where possible, these evaluation processes were embedded as part of the design of the program, thereby minimising the additional impost on program staff. This also requires that appropriate data collection protocols be developed and agreed to ensure the validity and reliability of the data being collected by third parties.

Reporting on findings

Working in partnership with relevant stakeholders as part of the process of reporting findings from the evaluation can help to overcome a major limitation of experimental and quasi-experimental research designs – namely, that they provide evidence of whether an outcome has been achieved, but not the reasons why the outcome has or has not been observed. By providing stakeholders with an opportunity to consider and respond to preliminary findings, such as by holding a workshop with these stakeholders prior to submitting the final report, it is possible to identify alternative explanations or additional information that may be crucial to an understanding of the results. This is particularly useful where the results are unexpected or potentially negative. In the example of the Queensland Murri Court, initial results showed a negative finding – offenders sentenced in Murri Court were more likely to be sent to prison. Following extensive consultation with the evaluation working group, and working with the Department to re-examine the data, a valid explanation was identified. Offenders appearing before the Murri Court were more likely to be serving a term of imprisonment for earlier offences at the time of being sentenced, which meant they were also more likely to receive a further custodial penalty (i.e., an increase to the original imprisonment term). Once the data were re-analysed using modified parameters to take this issue into account, the true impact of the Murri Court on sentencing outcomes could be determined (although the final result was still not entirely positive as offenders sentenced in the Murri Court were still just as likely as the comparison group to receive a custodial sentence). This approach minimised any delay and ensured a resolution could be identified quickly, before the results were disseminated to a broader audience. Importantly, experience shows that evaluation stakeholders are less likely to question or criticise the technical quality of an evaluation where they have been involved in the process, even if the findings are unexpected or negative (Brandon & Fukunaga, 2014). This means any discussion around results can focus on explaining and understanding the findings – as was the case with Murri Court. In the end, the results of the evaluation were mixed, and input from a range of stakeholders, particularly representatives from the Indigenous community, was used to help explain the findings, as well as to measure a number of important outcomes that could not easily be quantified, such as the impact on the partnership between the court and Indigenous community.

Another opportunity for collaboration at the time of reporting and communicating findings is in developing evaluation recommendations. While evaluators must control the process, ensure that any recommendations are substantiated by evidence from the evaluation and give adequate consideration to the likely reaction to the recommendations from different stakeholders (Bryson, Patton & Bowman, 2011), working with program staff to develop recommendations can increase the likelihood that the findings from an outcome evaluation will be used to inform decision making. In the evaluation of ADR in the NSW Children's Court, a workshop was held with members of the evaluation working group to discuss and agree on the proposed recommendations (which were provided in advance of the workshop). Much of this discussion was focused on the reasons that one of the two models evaluated was not as effective as the other (based on results from the quasi-experimental evaluation) and potential strategies for how that model might be improved. Rather than attempt to influence these recommendations in a negative way, this approach resulted in recommendations that were more practical, useful and offered a clearer course of action.

Limitations associated with research-practitioner partnerships for evaluation

The discussion above described a range of benefits associated with adopting a partnership approach to evaluating strategies to prevent and reduce crime using rigorous scientific methods. However, there are limitations and challenges associated with this approach that need to be acknowledged. First, collaborative approaches to evaluation are time consuming and resource intensive, requiring additional resources to manage the partnership and a commitment from program staff to dedicate additional resources to support evaluation processes (Brandon & Fukunaga, 2014). In the examples described throughout this paper, significant time was devoted to establishing effective partnership arrangements with the representatives of the agency commissioning the evaluation (and other partners, where necessary) – although there was a substantial return on this investment in terms of ensuring the successful implementation of the evaluation.

The most obvious and perhaps important criticism of researcher-practitioner partnerships is the actual or perceived impact on the objectivity and validity of the research findings. Views regarding the appropriateness of stakeholder involvement in evaluation employing experimental and quasi-experimental research designs are mixed (Brandon & Fukunaga, 2014; Pollitt, 2009). Some authors argue that the two are not mutually exclusive and that evaluators should involve stakeholders in the evaluation process (Patton, 2008; Plottu & Plottu, 2009), while others have suggested that there are occasions in which evaluations based on rigorous scientific methods should be conducted independently of program staff (Pollitt, 2009). There is an inherent tension that comes with being commissioned by the people responsible for the design, management and/or implementation of a program to conduct the evaluation. Providing an objective and independent assessment of the success (or not) of that program can be a source of conflict between program managers and evaluators, and pressure to adjust or soften unfavourable findings is not uncommon (Tilley & Clarke, 2006). While previous research into participatory approaches to evaluation has suggested they enhance the validity and credibility of the findings (at least among the stakeholders involved in the evaluation), they also acknowledge the potential for bias (Brandon & Fukunaga, 2014; Cullen, Coryn & Rugh, 2011).

The findings from empirical studies on the impact of researcher-practitioner partnerships on the technical quality and potential for bias in evaluations of strategies designed to prevent and reduce crime have been mixed (Petrosino & Soydan, 2005; Welsh, Braga & Hollis Peel, 2012). Petrosino & Soydan (2005) reviewed 12 meta-analyses on offender treatment programs and found that all but one of these previous studies had observed higher effect sizes when evaluators were involved in the design or delivery of the program. They also conducted their own meta-analysis of almost 300 randomised control trials of individually focused crime reduction and observed a similar result. Importantly, effect sizes were also higher for evaluations that involved collaboration between researchers and practitioners (i.e., where evaluators were not directly involved in program delivery). However, Welsh, Braga and Hollis-Peel (2012) conducted a meta-analysis of more than 40 experimental and quasi-experimental studies into the effects of police crime prevention strategies and did not find a relationship between evaluator involvement and program effects. Two explanations have been provided for the possible relationship between evaluator involvement and program effects – the high fidelity theory, which argues that involving researchers in program design and delivery may increase program fidelity, and the cynical view, which suggests there is pressure on the evaluator to deliver positive results and therefore some level of bias (intentional or unintentional) (Petrosino & Soydan 2005; Welsh, Braga & Hollis-Peel, 2012). While the cynical view cannot be discounted (Eisner, 2009), and past research has been less concerned with partnership arrangements than with evaluator involvement in program design and delivery, it is still important that steps be taken to minimise the risk of bias. This might involve strategies such as offering the opportunity for external oversight of an evaluation by an independent researcher (Braga, 2013), and ensuring the evaluator remains in control of the evaluation process (Cullen, Coryn & Rugh, 2011).

A balanced approach to evaluation

The focus of this paper has been on improving the standard of outcome evaluations for strategies designed to prevent and reduce crime through researcher-practitioner partnerships. The emphasis has been on increasing the use of rigorous scientific methods to measure program effects. However, adopting a quasi-experimental design to evaluate programs does not preclude the use of qualitative data collection methods, which are another mechanism for engaging stakeholders as part of the evaluation. In all of the evaluations described in this paper, a mixed methods approach was adopted, combining the quantitative measurement of key outcomes (e.g. reoffending) with qualitative methods. Combining quasi-experimental research designs with field study (e.g. interviews) is a common approach to understanding social change, particularly among pragmatists (see Boxall, this issue), and offers a number of important benefits (Hall, 2008). First, these interviews can confirm (through triangulation) the findings from the quantitative analysis, which can add weight and credibility to the findings. Second, they can help to provide explanation and context for the findings from any quantitative analysis of key outcomes. Third, these interviews can address evaluation questions that cannot be answered using quantitative methods. Fourth, these interviews will usually be a primary source of information required for a process evaluation and developing a comprehensive understanding of the program being evaluated, which will directly inform recommendations. Fifth, incorporating these qualitative approaches can help to achieve buy in from stakeholders who are less supportive of quantitative methods. And finally, qualitative interviews can help to address the common criticisms of quasi-experimental methods for lacking external validity by developing a more detailed understanding of the mechanisms that underpin interventions and the context in which these mechanisms are applied (Tilley & Clarke, 2006).

Factors contributing to successful partnership arrangements for evaluating strategies to prevent and reduce crime

Partnership approaches to evaluation are not that dissimilar to partnerships in other contexts, including those that exist as part of program delivery. A recent review of empirical studies exploring stakeholder involvement in evaluation identified many of the criteria for effective partnership working as being relevant to evaluation (Brandon & Fukunaga, 2014). Therefore in thinking about how best to manage the relationship between evaluators and the agencies that commission an evaluation, it is helpful to reflect on the qualities of effective partnerships and governance arrangements required (to varying degrees, depending on the nature of the partnership) for partnerships to function effectively.

Table 1: **Criteria for effective partnerships and their relevance to evaluation**

Criteria for effective partnership working	***Relevance to evaluation***
A clear mission and agreement on the objectives of the partnership	Evaluations were less likely to encounter resistance or practical barriers to accessing data when evaluators and program staff shared similar goals for the evaluation, particularly in terms of making a valid assessment of the effectiveness of the program.
Good knowledge and understanding of one another's roles and responsibilities	Agreement on roles and responsibilities for all parties involved in the evaluation at the commencement of the evaluation, and documenting this in either a project implementation plan or contract, helped to ensure that there was no confusion about who was responsible for undertaking evaluation activities, particularly with regards to data collection.
A high level of trust between partner agencies, including members that work well together, respect one another and are committed to ensuring the partnership succeeds	The examples in this paper involved evaluators and program staff that had established a positive relationship built on trust. This helped to minimise any suspicion among program staff about the motives of evaluators pressing for more rigorous designs and overcome reluctance to share data and support the evaluation.
Strong leadership, including local 'champions'	In the evaluation of the Murri Court and SCCDP and the evaluation of ADR in the NSW Children’s Court, there was a champion in the agency that had commissioned the evaluation who advocated for the evaluation and encouraged other stakeholders to participate. This was a key factor in their success.
Adequate resourcing, including staff having enough time away from agency core business to provide input to the partnership	In each of the examples presented in this paper, evaluators were adequately resourced to undertake a high quality evaluation. However, for collaborative approaches to evaluation to function effectively, program staff also needed to be able to invest the necessary time to undertake tasks to support the evaluation, including data collection.

Table 1: **Criteria for effective partnerships and their relevance to evaluation *Continued***

Criteria for effective partnership working	***Relevance to evaluation***
Partnership structures that are relatively small, businesslike, with a clear process for making decisions and a focus on problem solving	As described throughout this paper, a working group comprising representatives from the various parties was established early in each evaluation to oversee the development, implementation and ongoing review of the evaluation. This group was responsible for facilitating access to data and personnel, providing input on key outputs as they were produced, providing input into the final report recommendations and developing solutions to any problems (such as access to data), as they arose.
Data sharing policies and protocols	The evaluations described in this paper required access to data from a range of sources, and it was important to ensure the privacy and confidentiality of that data. There needed to be clear policies and protocols for the sharing of information between the evaluators and program staff (typically managed as part of a contract). Where program staff were collecting data on behalf of the evaluators, there were processes in place to ensure the data collected were both valid and reliable.
Continuity in partner representation and participation and documentation of processes and decision-making	Program turnover can affect the progress of evaluations as well as program delivery. While not ideal, this was encountered in a number of evaluations. Where there was appropriate documentation and effective transition arrangements, the impact on the evaluation was minimal.

Source: Gilling, 2005; Homel, 2006; Morgan & Homel, 2011; Rosenbaum, 2002.

Table 1 describes a number of criteria for effective partnership working in crime prevention and criminal justice settings, based on an established body of literature. The relevance of these criteria to evaluations employing rigorous scientific methods is also highlighted, based on the examples discussed in this paper. This shows that for researcher-practitioner partnerships to work effectively as part of an evaluation, there needs to be clear agreement regarding the use of quasi-experimental research designs (or rigorous scientific methods more broadly), clear roles for the various parties in supporting the evaluation, a high level of trust between the evaluator and agency commissioning the evaluation and adequate resourcing to support the approach to evaluation.

Conclusion

This paper began by arguing that researcher-practitioner partnerships can help improve the design and implementation of evaluations conducted by evaluators independent of program design and delivery using rigorous scientific methods. Then, reflecting on several evaluations of programs designed to prevent crime and respond to the needs of vulnerable populations in court settings, the paper identified a number of important benefits that result from for working in partnership with policy makers, program managers and project staff. Involving stakeholders can assist with:

- determining what outcomes can and should be attributed to the program being evaluated and should therefore be the focus of an evaluation;
- designing the evaluation, including data collection processes, which can encourage buy in, facilitate access to data and identify and address potential barriers to conducting the evaluation;
- overcoming resistance to the use of quasi-experimental research designs and the challenges associated with identifying suitable comparison areas;
- providing assistance with the collection of data, particularly from hard to reach populations, to enable key outcomes to be measured;
- identifying alternative explanations or additional information that may be crucial to an understanding of the results; and
- developing recommendations that are more practical, more useful and offer a clear course of action to make improvements to the program.

There are certain challenges and limitations associated with this approach, but experience has shown that these are far outweighed by the benefits described above and that problems such as bias may not be as common as sometimes believed.

In an era of growing emphasis on accountability, government agencies will continue to engage evaluators independent of program design and delivery to make an objective and impartial assessment of the impact of crime prevention and criminal justice programs. Experimental research designs (particularly quasi-experimental designs) are also likely to remain the sought after standard for many evaluations of strategies to prevent and reduce crime. Evaluators involved in evaluations that employ quasi-experimental designs would benefit greatly from embracing the importance of working cooperatively with commissioning agencies, program staff and other key stakeholders, and viewing the relationship as a partnership. There is, after all, very good reason that stakeholder involvement and participation has become such a dominant theme in evaluation theory and practice.

References

Alpert, G. P., Rojek, J., & Hansen, J.A. (2013). *Building bridges between police researchers and practitioners: Agents of change in a complex world.* Washington DC: US Department of Justice.

Brandon, P. R. & Fukunaga, L. L. (2014). The state of the empirical research literature on stakeholder involvement in program evaluation. *American Journal of Evaluation, 35*(1), 26-44.

Braga, A.A. (2013). Embedded criminologists in police departments. *Ideas in American Policing,* No. 17, pp. 1-20.

Braga, A. A., Welsh, B. C. & Bruinsma, G. J. N. (2013). Integrating experimental and observational methods to improve criminology and criminal justice policy. In B. C. Welsh, A. A. Braga & G. J. N. Bruinsma (Eds.). *Experimental criminology: Prospects for advancing science and public policy* (pp. 277-298).

Bryson, J. M., Patton, M. Q., & Bowman, R. A. (2011). Working with evaluation stakeholders: A rationale, step-wise approach and toolkit. *Evaluation and Program Planning, 34,* 1-12.

Cullen, A. E., Coryn, C. L. S., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluation, *American Journal of Evaluation, 32*(3), pp. 345-461.

Datta, L. (2006). The practice of evaluation: Challenges and new directions. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.). *The Sage Handbook of Evaluation.* (pp.419-438). London: Sage Publications.

Donaldson, S. L., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.). *The Sage Handbook of Evaluation,* (pp. 56-75). London: Sage Publications.

Eisner, Manuel. (2009).. No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology, 5* (2), pp. 163–184.

English, B., Cummings, R., & Stratton, R. (2002). Choosing an evaluation model for community crime prevention programs. In N. Tilley (Ed.). *Evaluation for crime prevention* (pp. 119-169). Monsey, NY: Criminal Justice Press.

Farrington, D. P., Gottfredson, D. C., Sherman, L.W., & Welsh, B. C. (2006). The Maryland Scientific Methods Scale. In L. W. Sherman, D. P. Farrington, B. C. Welsh & D. L. MacKenzie (Eds.). *Evidence-based crime prevention* (pp. 13-21). London: Routledge.

Funnell, S., & Rogers, P. (2011). *Purposeful program theory: Effective use of theories of change and logic models.* San Francisco: Jossey Bass.

Gilling, D. (2005). Partnerships and crime prevention. In N. Tilley (Ed.). *Handbook of crime prevention and community safety* (pp. 734-756). Cullompton, UK: Willan Publishing.

Greene, J. C. (2006). Evaluation, democracy and social change. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.). *The Sage Handbook of Evaluation* (pp. 118-140). London: Sage Publications.

Hall, R. (2008). Applied social research: Planning, designing and conducting real-world research. South Yarra: Palgrave MacMillan

Hendrie, D. (2003). Random breath testing: *Its effectiveness and possible characteristics of a 'best practice' approach.* Crawley: The University of Western Australia.

Homel P. (2006). Joining up the pieces: what central agencies need to do to support effective local crime prevention. In J. Knutsson & R. Clarke (Eds.), *Putting theory to work: Implementing situational prevention and problem-oriented policing* (pp. 111-139). New Jersey: Prentice Hall.

Idriss, M., Jendly, M., Karn, J., & Mulone, M. (2010). *International report on crime prevention and community safety: Trends and perspectives,* 2010. Montreal: International Centre for the Prevention of Crime.

Jones, C. (2011). Intensive judicial supervision and drug court outcomes: Interim findings from a randomised control trial. *Crime and Justice Bulletin,* No. 152, pp. 1-16.

Lee, S., Aos, S., Drake, E., Pennucci, A., Klima, T., Miller, M., Anderson, L., Mayfield, J., & Burley, M. (2012). *Return on investment: Evidence-based options to improve statewide outcomes,* April 2012, Olympia, Washington: Washington State Institute for Public Policy.

Lind, B., Weatherburn, D., Chen, S., Shanahan, M., Lancasar, E., Haas, M., & De Abreu Lourenco, R. (2002). *New South Wales Drug Court Evaluation: Cost-effectiveness.* Sydney: NSW Bureau of Crime Statistics and Research, Attorney General's Department. Retrieved from NSW Bureau of Crime Statistics and Research, Attorney General's Department website: http://www.bocsar.nsw.gov.au/agdbasev7wr/bocsar/documents/pdf/l15.pdf

Lum, C., Telep, C. W., Koper, C. S., & Grieco, J. (2012). Receptivity to research in policing. J*ustice Research and Policy, 14*(1), pp. 61-95.

Mazerolle, L., Antrobus, E., Bennett, S., & Tyler, T. R. (2013). Shaping citizen perceptions of police legitimacy: A randomised field trial of procedural justice, *Criminology, 51*(5), pp. 33-63.

Morgan, A., Boxall, H., Lindeman, K., & Anderson, J. (2012). *Effective crime prevention strategies for implementation by local government.* Canberra: Australian Institute of Criminology (AIC).

Morgan, A., Boxall, H., Terer, K., & Harris, N. (2012). *Evaluation of alternative dispute resolution initiatives in the care and protection jurisdiction of the NSW Children's Court.* Canberra: Australian Institute of Criminology (AIC).

Morgan, A., & Homel, P. (2013, July). Evaluating crime prevention: Lessons from large-scale crime prevention programs. *Trends and Issues in Crime and Criminal Justice* Series No. 458, pp. 1-12. Canberra: Australian Institute of Criminology, Australian Government.

Morgan, A., & Homel, P. (2011). *Model performance framework for local crime prevention,* Canberra: Australian Institute of Criminology (AIC).

Morgan, A., & Louis, E. (2010). *Evaluation of the Queensland Murri Court: Final report,* Canberra: Australian Institute of Criminology (AIC).

Morgan, A., & Mann, M. (in press). Police crime prevention in Australia: Findings from a review. *Trends and Issues in Crime and Criminal Justice.*

National Drug Research Institute, Curtin University of Technology. (2007). *Restrictions on the Sale and Supply of Alcohol: Evidence and Outcomes,* Perth, WA: Curtin University of Technology.

National Research Council. (2005). Improving Evaluation of Anticrime Programs. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id=11337

New South Wales (NSW), Department of Premier & Cabinet. (2013). *NSW Government evaluation framework. Sydney:* NSW Department of Premier and Cabinet.

Patton, M. Q. (2008). *Utilisation-focused evaluation* (4th ed.).Thousand Oaks, CA: Sage.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology,* 1(4), pp. 435-450.

Plottu, B., & Plottu, E. (2009). Approaches to participation in evaluation: Some conditions for implementation. *Evaluation, 15*(3), pp. 343-359.

Pollitt, C. (2009). Stunted by stakeholders? Limits to collaborative evaluation. *Public Policy and Administration, 14*(2), pp. 77-90.

Western Australia, Program Evaluation Unit, Department of Treasury. (2014). *Evaluation Guide,* Department of Treasury, Government of Western Australia, Perth: Author.

Rosenbaum, D. (2002). Evaluating multi-agency anti-crime partnerships: Theory, design and measurement issues. In N. Tilley & N. J. Monsey (Eds.). *Evaluation for Crime Prevention.* Criminal Justice Press/Willow Tree Press. Retrieved from National Criminal Justice Reference Service Library website https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=195631

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (2006). *Evidence-based crime prevention* (2nd ed.), London: Routledge.

Sherman, L.W., Gottsfredson, D., MacKenzie, D.L., Eck, J., Reuter, P., & Bushway, S. (1998). *Preventing Crime: What Works, What Doesn't, What's Promising.* Washington, DC: US Department of Justice.

Sherman, L. W., Strang, H., & Woods, D. J. (2000). *Recidivism patterns in the Canberra Reintegrative Shaming Experiments (RISE).* Canberra: Australian Institute of Criminology.

Shults, R. A., Elder, R. W., Sleet, D. A., Nichols, J. L., Alao, M. O., Carande-Kulis, V. G., … Task Force on Community Preventive Services. (2001). Reviews of evidence regarding interventions to reduce alcohol-impaired driving. *American Journal of Preventive Medicine, 21*(4), 66-88.

Smith, L., Morgan, A., & McAtamney, A. (2011). *Policing licensed premises in the Australian Capital Territory.* Canberra: Australian Institute of Criminology (AIC).

Steering Committee for the Review of Government Service Provision [SCRGSP], (2014). *Report on Government Services 2014,* Canberra: Productivity Commission, Commonwealth of Australia.

Tilley, N., & Clarke, A. (2006). Evaluation in criminal justice. In I. F. Shaw, J. C. Green & M. M. Mark (Eds.), *The Sage Handbook of Evaluation.* (pp.512-535). London: Sage Publications.

United Nations Evaluation Group [UNEG], (2005). *Norms for evaluation in the UN System,* Vienna: UNEG, Author.

Weatherburn, D. (2005). Critical criminology and its discontents: A response to Travers, critique of criminal justice evaluation. *Australian and New Zealand Journal of Criminology, 38*(3), pp. 416-420.

Welsh, B. C., Braga, A. A., & Peel, M. E. (2012). Can 'disciplined passion' overcome the cynical view? An empirical inquiry of evaluator influence on police crime prevention program outcomes. *Journal of Experimental Criminology, 8* (4), pp. 415–431.

Program evaluation in a cross-cultural context: Action research, program logic and youth justice in Thailand

Matthew J Willis Australian Institute of Criminology Honorary University Fellow, Northern Institute, Charles Darwin University matthew.willis@aic.gov.au	Adam M Tomison Australian Institute of Criminology Honorary Professor, Australian Catholic University

Keywords: cross cultural program logic, participatory action research, cross cultural evaluation, multilingual evaluation, juvenile justice evaluation, Juvenile Justice Reform Project, Justice for Our Youth project.

Abstract

Using the evaluation of a multi-faceted juvenile justice project as a case study, we demonstrate how applying an action research approach to program logic development provided a way of arriving at shared understandings of evaluation in a cross-cultural, cross-language context. The paper explores work undertaken by the Australian Institute of Criminology for the Thailand Department of Juvenile Justice and Observation to support the evaluation of the Justice for Our Youth (JOY) project, a complex project aimed at improving outcomes for young offenders by improving the quality of service offered by the Department. We describe a workshop conducted in Thailand where the authors provided capacity-building for Thai officials in program monitoring and evaluation and then worked with the officials to apply the learning from this part of the workshop to developing program logic models and identifying data and information needs for the JOY program evaluation. The utility of Participatory Action Research (PAR) and program logic approaches to working in a cross-cultural, cross-language context are discussed and their application to other cross-cultural situations is considered. The authors conclude that PAR can provide a valuable and appropriate model for establishing mutual understanding and trust in such contexts, but also recognise that the realities of difference and distance can reduce the ability of evaluators to apply PAR in a way that represents best practice.

Introduction

There are many publications and manuals that attempt to provide ideal or best practice approaches to conducting evaluations in community and cross-cultural settings. The reality of evaluation practice is that often the 'ideal' is not possible or practicable, with evaluators constrained in terms of project parameters, resources (time, funding and access to participants) and the approaches able to be employed. Typically, evaluators must 'cut the cloth' of the evaluation to take into account imposed constraints and other conditions present in the evaluation environment, while maintaining an ethical approach, doing their best to meet the needs of stakeholders and participants and enabling sufficient data to be collected to demonstrate project success.

In this paper the authors demonstrate the application of an action research approach based around the development and use of program logic models, as part of an evaluation of a multi-faceted juvenile justice project being developed in Thailand. By describing the process used, the paper aims to provide insights about the benefits and limitations of Participatory Action Research approaches to generating shared understandings of evaluation in a cross-cultural, multi-language context. The aim is to give a real world (in situ) example of conducting evaluation in such contexts to inform evaluation practitioners and others involved with cross-cultural engagement. Some insights are provided into how the practice of evaluation can occur successfully in a real world situation where situational constraints affect both the evaluators and participants.

Background

The focus of this paper is an evaluation project being conducted by the Australian Institute of Criminology (AIC) for the Thailand Department of Juvenile Observation and Protection (DJOP). The DJOP provides services for children and young people entering the justice system, including juvenile detention and post-release services. The Department is empowered under Thai legislation to undertake investigation of juveniles' circumstances and characteristics and assessment to determine individual service and support needs.

Between 2009 and 2013 DJOP, together with a number of other Thai agencies, implemented the Juvenile Justice Reform Project (JJRP), which aimed to address service improvements in screening and classification, rehabilitation programs, pre-release preparation and post-release support. This Project resulted in the development of a range of assessment and classification tools, intervention programs for juveniles at each stage of the juvenile justice process and a network of government and non-government agencies providing services and programs for young offenders.

Consultation and analysis following the Juvenile Justice Reform Project identified areas where further work was needed to realise the outcomes established through the Project. A needs analysis showed that a majority of Probation Officers lacked the knowledge and skills to properly use the risks and needs assessment structured interview form developed through the JJRP and some were reluctant to use the form. The analysis also showed DJOP lacked standardised objective psychological assessment tools to assess mental health problems and traits associated with the criminal behaviours of young people. The Department also lacked a fully developed throughcare – focused intervention program provided for juveniles in detention training schools. 'Throughcare' refers to the treatment and supports provided to prisoners and detainees from their reception into custody that continue after release into the community (Borzycki & Baldry, 2003). Finally, the intervention programs being used did not fully address criminogenic needs and risks for the young people, and did not utilise community resources, leading to extended periods in residential placements and have inadequate continuing care after release.

The Justice for Our Youth (JOY) project was created to fill gaps remaining after JJRP. JOY was established explicitly around principles of enhancing the participation of stakeholders and greater involvement of DJOP staff, which issues were identified through the JJRP process. There are three sub-projects within the overall JOY project:

1. Building staff capacity in risk and needs assessment through a program of training for experienced Probation Officers, and the solidification and transfer of skills from these Officers to less experienced Probation Officers.

2. Development of mental health and behavioural assessment tools for juveniles with mental health issues and complex problems (through standardized adaptation of existing standard psychometric tests developed outside Thailand).

3. Development of seamless intervention and rehabilitation programs for juveniles (using an established approach known as to provide throughcare outcomes.

The JOY project commenced in mid-September 2013 and will continue for approximately three years, until mid-October 2016.

The Evaluation Process

Achieving quality evaluation outcomes requires the use of evaluation processes that are well-grounded and well-designed. Effective practice principles suggest that each evaluation should be tailored to fulfil the specific purpose for which it is required and to meet the needs of the various stakeholders involved (Stufflebeam, 2004, Tomison, 2000). Often there are multiple purposes behind the evaluation, multiple dimensions to the intervention or program being evaluated, and multiple stakeholders whose needs must be considered. Different stakeholders will often have disparate interests and may well require different evaluation 'products' and styles of communication (Williams & Tomison, 2013).

The research question(s) and the level of explanation required determine the methodologies and research tools used and the degree of experimental rigour that is desired and/or possible (Frechtling, 2007; Funnell & Rogers, 2011; Tomison, 2000). The type of data and information required to answer the research questions, and the practicalities of gathering the necessary data and information, are also essential in shaping the methodology to be applied to each evaluation. Not infrequently, interventions may be hampered by limited access to statistical data or records. Sometimes, what is available may not capture the reality of participants' experiences or the issue under consideration. Nevertheless, funding bodies and other external stakeholders often require quantitative measures, 'hard data', to ensure the methodological rigour of any findings, and to enable comparisons over time and between programs and/or communities. The challenge when undertaking work in applied community settings, particularly when working with local communities of different cultural and language backgrounds, is creating evaluation evidence that is meaningful to funding bodies and other stakeholders. The use of multiple methods and data sources to 'build a picture' over time typically improves the ability of the evaluation to demonstrate program success (Pawson & Tilley, 1997; Tomison, 2004).

In order to establish the key questions, measures and sources of data and information for programs and evaluations, many evaluators use a variant of a 'program logic' or 'theory of change' framework (e.g. Clark & Anderson, 2004; Funnell & Rogers, 2011). Program logic is a process of developing a conceptual model of how, and why, a program is expected to work (Funnell & Rogers, 2011; McLaughlin & Jordan, 1999). While models can vary, program logic modelling typically results in a diagram that depicts the inputs and activities required for the program to operate, the assumptions and external factors that influence the program's development and implementation, and the outcomes expected from the program in the short-term, medium-term and long-term (McLaughlin & Jordan, 1999). Program logic models can be valuable tools for communicating and disseminating information about a program, identifying the key elements of a program, informing measure of success, and for identifying data and information needs (McLaughlin & Jordan, 2004).

Program logic frameworks are necessarily influenced by those who are involved in their development. Frameworks developed in isolation by the evaluator and the staff of an intervention or program may lack the depth, rigour and application to specific contexts and circumstances that can be achieved if a wider group of stakeholders, including community participants, are involved in the framework development process. This is particularly important where there are cultural or language differences between the evaluator/researcher, community members and other stakeholders.

Ideally then, a shared understanding of key concepts and intentions should be developed between the evaluator(s) and the range of stakeholders for the evaluation - this will be critical to the evaluation's success. Taking a culturally sensitive approach to building this shared understanding will also help to overcome some of the potentially negative consequences of conducting cross-cultural evaluation and research. These can include imposing Western perspectives and conceptual frameworks on the evaluation process and its findings, reducing the non-Western participants to being passive objects of examination, and developing findings based on shallow information and flawed or limited understandings (see Liamputtong, 2010).

A key element for the AIC in conducting an effective evaluation of JOY was to ensure that a culturally respectful process was developed that was flexible enough to be used with the range of diverse professional groups involved with the project (and the evaluation), which considered their varied experiences as participants in evaluations. The process respected the expertise of the DJOP staff and their knowledge of the context and circumstances in which the JOY project was developed and implement. Key considerations in framing the project including ensuring the aims, outcomes and success measures of the project were accurately defined and that all participants:

- shared an understanding of the project, its aims and objectives;
- agreed (through a collaborative process) on the nature and goals of the evaluation;
- contributed to developing an agreed understanding about how and why data would be collected; and
- had roles in the evaluation process that aligned with their knowledge and experience and allowed them to contribute to the overall success of the JOY project.

Frameworks

The evaluators began the initial steps of building a framework for the evaluation using mixed methods approaches within a realist evaluation framework (Pawson & Tilley, 1997). A combination of quantitative and qualitative methodologies appeared to present the best option for capturing the data and information needed for the evaluation. Quantitative methodologies would support the need for highly tangible data-driven measures, such as reductions in offending behaviour that are typically sought by evaluation stakeholders. Qualitative methodologies would allow for the capture of deeper levels of information about outcomes, such as changes in perceptions or the quality of engagement between DJOP and young people. The evaluators considered that qualitative techniques would be particularly valuable given the cross-cultural context of the evaluation. Qualitative approaches support the gathering of information linked to subjective experiences, situational meanings and allow for levels of interpretation and flexibility that can help to bridge gaps between parties in cross-cultural situations (Liamputtong, 2010).

In circumstances where researchers and evaluators come from positions of cultural dominance, traditional positivist methodologies can override the perspectives of participants and deny them agency (Liamputtong, 2010). This typically occurs when the evaluand is a program or intervention involving Indigenous (first nations) peoples, who come to the research or evaluation process from an inherently disempowered position. The AIC evaluators considered that qualitative methods, used within a realist evaluation framework, would support the agency of the Thai staff participating in the evaluation. At the same time realist approaches, with their emphasis on understanding context and refining theories of how an intervention produces change, offered the best prospect for collaboratively generating knowledge of the project and the environment within which it is implemented (Pawson & Tilley, 2007).

Participatory Action Research

The evaluators' initial considerations for the evaluation framework were based primarily on reading of documents describing the basis and establishment of the JOY project. The AIC and DJOP agreed that further development of the framework and implementation of the evaluation process should be realised through collaboration between the parties. While mixed methods approach and realist perspectives appeared to the evaluators to provide an appropriate basis for the evaluation, progressing without forming a strong collaborative approach would undermine the inclusionary objectives that these methods and perspectives were intended to support.

A Participatory Action Research (PAR) approach was chosen as the foundation methodology for working with program participants to develop an agreed program logic model that would inform and support evaluation of the JOY project and its sub-projects. PAR is an approach to research that brings together those directly involved in an intervention or project that is being evaluated - generally staff of an organisation or members of a community - and those tasked with conducting the evaluation (Whyte, Greenwood & Lazes, 1991). Using PAR approaches, direct stakeholders are involved throughout the evaluation, from initial planning and design to interpretation and dissemination of the findings. Based on a fluid process founded on iterative cycles of Planning, Acting, Observing and Reflecting; PAR has been found to be a very useful approach for such human service evaluations, particularly where the focus is on program and service improvement rather than accountability. PAR supports the active engagement of those directly involved with delivering services and allows their knowledge and experience to contribute directly to the foundations of the evaluation (Brydon-Miller, Kral, Maguire, Noffke & Sabhlok, 2011; Burns, 2006; Crane & O'Regan, 2010; Hecker, 2008; McTaggart, 1989; Stringer, 2014).

Such approaches allow project teams to develop their own evaluation frameworks and ways of working. Having enacted planned actions, PAR explicitly encourages teams and their members to build in processes for collecting their observations and experiences of practice, to reflect on them with their team (i.e., to assess what is working or not working, or how can the project or outcomes be improved) and using the outcomes of these processes to guide a continuous improvement approach to program delivery and evaluation (Kemmis & McTaggart, 2000; McTaggart, 1989).

PAR is based on the development of a partnership between the evaluator, funding body and participants, with participants taking an active role in developing and informing the evaluation. That is, the intent is that project teams 'own' their projects and the evaluation. PAR is expressly designed to be participatory and collaborative, with evaluators in the role of 'walking alongside' and 'doing with' rather than 'doing to' a project. The notion of 'doing with' embraces the fundamental idea that project teams are not merely objects of examination for the evaluation, or passive actors whose outputs are scrutinised, but active participants in the process of evaluating.

While PAR is most often used with qualitative research, it fits with a multiple methods approach and is flexible enough to underpin many evaluation models and approaches, and the inclusion of quantitative data where it is available, as was the case for the JOY project. The action research approach is also well-suited to operating in a range of cultural contexts, and facilitates culturally secure evaluation practice by its very nature of being designed around inclusionary processes (Brydon-Miller et al. 2011; Hecker, 2008; McTaggart, 1989).

Evaluating the JOY Project, Thailand

The Australian Institute of Criminology (AIC) conducts criminological research across a range of topic areas, including evaluations of crime-related interventions and programs. During 2013, the Thailand DJOP contacted the AIC to commence a process of engaging the AIC to provide expertise to support the evaluation of the JOY project. Through negotiations it was agreed that the AIC would conduct two workshops in Thailand to build the capacity of DJOP staff in program monitoring and evaluation, to improve their understanding and ability to collect data and information for the evaluation of JOY, to undertake the data analysis for the evaluation; and to prepare an evaluation report of JOY and its three sub-projects.

In February 2014, the authors travelled to Bangkok to deliver a workshop to approximately twenty DJOP staff, representing most of the staff working on the JOY project. The primary aim of the workshop was to build the capacity of DJOP staff in program monitoring and evaluation and to develop program logic models for each of the JOY sub-projects with the active involvement of program participants. The first part of the workshop focused on general principles and best practice issues, while later sessions adopted an action research approach to apply these principles and practices to developing evaluation frameworks for each of the three sub-projects.

The workshop helped to refine both AIC and DJOP understandings about the aims of the JOY sub projects, their intended outcomes, how the success of these outcomes could be measured and what data and information could be collected to inform these measures. The workshop also represented a critical milestone in the application of an action research approach for the development, implementation and evaluation of the JOY project.

Explicit intended outcomes of the process were for staff participants to develop:

- a shared understanding of the role each person would play;
- the importance of that role for achieving both program and evaluation objectives;
- a clear understanding of which information would be collected, how it would be collected and who would be responsible for its collection; and
- how that information would be used to populate the evaluation framework and program logic model.

This initial ‘skilling up’ and capacity-building for staff participants formed the basis for establishing a positive engagement with AIC staff allowing for further relationship-building to occur over time. It also enabled the building of relationships between AIC and DJOP staff and assisted in establishing the trust and mutual understanding needed to support the active exchange of information and sharing of insights fundamental to a PAR approach.

Conducting the workshop required a quite measured and structured form of delivery from the AIC facilitators, as most of the DJOP staff had little or no understanding of English. Throughout the three days, the workshop was continually translated into Thai by the head of the DJOP research section, who was also leading the JOY project. The AIC facilitators used strategies such as re-stating and questioning to regularly check for clear understanding of the information being provided. Throughout the workshop there were periods where the DJOP staff discussed issues among themselves in Thai, providing only summaries of the discussions in English. Welcoming these discussions, rather than trying to impose a requirement that all discussions be translated into English was both respectful to the DJOP staff and supportive of richer levels of engagement by the Thai speakers than would otherwise have been possible.

Program logic and PAR

Following initial training and capacity-building sessions, the workshop was then focused on working with DJOP participants to describe and refine how the evaluation would operate in practice, using PAR and the development of program logic models for each of the three JOY sub-projects. Before each session, the facilitators prepared a program logic diagram that was partly completed on the basis of information known to the facilitators through background documentation on JOY (translated from Thai into English), together with knowledge of long-term criminal justice outcomes drawn from experience on other evaluations.

This proved to be very useful in a number of ways. In a purely practical sense, having the diagrams part-completed helped move the discussion along and overcame the time constraints that came with working across languages. Information that had been prepared in a succinct way suited to the model could be directly interpreted and allowed discussions around the elements of the model to proceed from a clearer foundation. Part completion also allowed those unfamiliar with logic model approaches before the workshop to engage more readily with concept of program logic by presenting those concepts in a more tangible way than a blank template would have allowed. Working from partial models also indicated in an explicit manner the extent of the facilitators’ familiarity with the resources, influences and external environmental factors affecting the project and the types of data and information that could be useful for the evaluation. While this approach helped to establish the facilitators’ credibility and knowledge of youth justice systems; providing incomplete program logic models that required the stakeholder input and evaluator/participant discussion for completion proved to participants how crucial their input and expert knowledge was to developing the evaluation.

Discussions held during the workshop allowed the evaluators to gain insights into the forms of information and data held on young offenders to support case management, internal reporting between DJOP head office and the provinces, and DJOP's relationships with other criminal justice system agencies. The discussions also shed light on DJOP's expectations of its staff and the ways in which DJOP staff engaged with the families of young offenders. Understanding the expectations of staff was important for the evaluation as the anticipated outcomes of two of the sub-projects include improvements in the knowledge and skills of DJOP staff.

Further, in evaluating any staff-delivered intervention, it can be valuable to understand the organisation's expectations of its staff as this will help evaluators establish performance measures and assess whether staff have met organisational targets in implementing the intervention. This may be less of an issue when evaluations are conducted within the same cultural group. In the absence of specific information assumptions will generally be made about organisational expectations based on societal norms. These expectations can then be tested at later stages of the evaluation. When working across cultures, particularly with different systems of government, it becomes more critical for evaluators to understand the contributions of organisational perceptions and societal norms. Being able to gain some insights into how the DJOP saw its relationships with young people and their families, and how the organisation viewed the responsibilities of young offenders, was important for identifying sources of information for the evaluation. The overall objectives of the JOY project include reduced juvenile offending, as well as better life outcomes for young offenders. These outcomes would include reintegration with families and communities and engagement in education or employment. While these outcomes would be familiar to evaluators working with the Australian criminal justice system, discussions allowed the evaluators to identify subtle differences in the Thai context related to differing expectations about when and how young people would engage with education and employment. These differences were able to be captured and will be explored further, using PAR approaches, as the evaluation progresses.

JOY aims to achieve its objectives through improved juvenile justice services, such as individualised case management services delivered from the young person's first contact with the agency through to their return to the community, including support in the community. Young people are clearly integral to the work of DJOP and being able to include their voices, and those of the families and communities they return to, is an important element of ensuring a well-rounded evaluation. It is also important that DJOP is able to provide the evaluators with qualitative and quantitative evidence of outcomes in the form of risk assessments and case management plans that can be analysed to identify the impacts of the JOY sub-projects. The insights gained through the workshop assisted the evaluators to understand how they, and DJOP, could access information from and about young Thai people. These insights have informed questions about how indicators of successful reintegration to communities should be conceptualised and measured in the context of Thai family structures and familial and societal expectations of young people and their roles.

Overall, the workshop provided an ideal starting point for applying PAR approaches to this evaluation project. The workshop became a tangible manifestation of the PAR model of reflection, data/information collection and action (see Baum, MacDougall & Smith, 2006). Participants were able to reflect on aspects of the JOY project, its outcomes and information needs and a number noted how much they had learned about their own project through this reflective process. The process provided a platform for sharing information about the project and the work of the DJOP; some of this information then required action to adjust the program logic models and data collection matrix. For example, information concerning the program's objectives and expected outcomes that had been necessarily included in the project establishment documentation created for the funding provider, were not necessarily able to be operationalised in a way that could be measured. Within the workshop context, the facilitators and DJOP participants were able to brainstorm data collection options that would be achievable (and better placed) ways of meeting the needs of key stakeholders. Using a PAR approach, these options were incorporated into the logic models and led to a new cycle of reflection, collection and action.

The PAR cycles occurring within the workshop served as a microcosm of the PAR processes that will be applied throughout the evaluation, both by DJOP and the AIC. The authors anticipate that a PAR approach will help both organisations manage the substantial challenges of implementing a long-term multi faceted project with wide-ranging outcomes and diverse stakeholders, and evaluating this project in a cross-cultural, cross-language and geographically distant context.

Participation and Power

One of the distinctive features of PAR is that it allows for open and respectful communication that bridges power relationships by deliberately sharing power between the parties and imparting equal value on the contributions of researchers and those being researched (Baum, MacDougall & Smith, 2006). This makes the approach highly suitable for working in a cross-cultural situation where the researchers or evaluators are not necessarily fully versed with cultural protocols or expectations. PAR approaches help to make explicit the notion that an external evaluator is bringing a particular knowledge or expertise to the relationship, but is not doing so from a position of superiority.

In the case of working with stakeholders in an Asian country, this helps to overcome any notion that Western perspectives may be inherently superior, or that the Asian partner does not have knowledge or expertise to contribute to the working relationship. In the case of evaluations involving programs operating in Australian Aboriginal and Torres Strait Islander communities, a PAR model has been shown to assist in overcoming potential barriers created through mistrust, gives appropriate recognition to Indigenous knowledge and perspectives, and fosters information collection that accords with Indigenous communication styles and cultural constraints on knowledge sharing (Williams & Tomison, 2013). Applying these notions to yet another cultural situation, Letiecq and Bailey, (2004) noted the importance of building and reaffirming relationships when conducting research with Native American (American Indian) groups. In common with Indigenous Australians, Native Americans have experienced historical injustices and misuse of traditional knowledge. Conducting research through PAR approaches has been found to be an effective way of working through these barriers and fostering trust and mutual exchange of knowledge (Letiecq & Bailey, 2004), as we have found with the JOY evaluation.

Constraints to Practice

As noted at the outset, this paper aimed to provide an operational example of attempts to apply a PAR process to the evaluation of a multi-faceted government project in a cross-cultural, cross-language context. The evaluators adopted PAR approaches within a realist evaluation framework and these approaches helped to progress the evaluation in an inclusive, collaborative way that assisted with overcoming some of the cultural and language differences between evaluators and program stakeholders and participants. At the same time, the process highlighted some of the limitations of trying to apply genuinely collaborative and interactive approaches across these differences and across distance.

In an ideal application of PAR, evaluators and participants would meet regularly to reflect on their experiences and plan next steps. There would then be a process of implementation and observation, followed by further reflection and planning. While the JOY evaluation is able to incorporate these cycles, there are some very real limitations on the capacity of evaluators and participants to meet and discuss. While this can be partly overcome through regular video-conferencing between the AIC evaluators and key DJOP staff, it will not be practical to include the larger group of participants in those discussions. Further, while these video-conferences can be conducted regularly, this will have to be held at fairly fixed intervals that may not align with times and stages that would be best suited to observation and reflection. The video-conferences may also be more structured and less conducive to free-ranging discussion than might be ideal for a PAR approach.

Differences in language will also tend to limit the scope of discussion and prevent sharing of some of the more subtly nuanced elements of reflection that might otherwise benefit the evaluation. Language differences will also tend to require the discussions to be structured and involve both participants and evaluators preparing some of the information material in advance of the discussions. This will also tend to constrain discussion and raises the potential for the evaluation process to be more fully guided and led by the evaluators than would be ideal. This carries an inherently related risk of the process being overly influenced by the evaluators' Western perspectives and modes of working.

Some of the basic tenets of PAR require participants to keep and build records of reactions, judgements, impressions and to establish evidence of changes and improvements (McTaggart, 1989). The capacity of participants to do this, in a way that can be shared with evaluators, is limited in a cross-language context as records cannot be directly shared and there can be substantial costs involved in translating records. The need for translation may limit the scope to which records can be used to capture reactions and impressions. Apart from these issues, participants will be working in roles where they will face competing priorities and time pressures will tend to reduce the extent of their record-keeping. The problems of distance and different political and social environments within which the evaluators and participants are operating will reduce the ability of the evaluators to positively influence these aspects of the PAR process.

One further consideration that may arise in other cross-cultural evaluations, much more than it has for the JOY project, is the issue of power imbalances between evaluators and participants. This will typically arise when researchers and evaluators from mainstream organisations, particularly government agencies, work with Indigenous peoples. It is important to recognise and respect the impacts of past colonisation experiences and detrimental government policies as research practices that engendered distrust and current experiences of disadvantage is incumbent on evaluators. PAR approaches can help with empowering Indigenous peoples in the context of these impacts. Work on the JOY project raised the potential for evaluators to perpetuate a ‘methodology of imperialism’ (Said 1995, cited in Liamputtong 2010: 22) by imposing Western perspectives and paradigms on Eastern participants. However, the JOY participants are educated and qualified government staff members who are relatively empowered compared to participants in other cross-cultural contexts, such as members of Indigenous communities. As has been described above, the evaluators deliberately adopted approaches to reduce imbalances as much as possible within the constraints imposed by the contract and evaluation environment in order to set up a process of real partnership between the AIC and DJOP who each brought valuable knowledge and expertise to the evaluation process.

Conclusions and Implications

Participatory ways of working are particularly important in cross-cultural contexts as they afford deeper levels of knowledge to emerge and become incorporated into the evaluation process, knowledge that may not otherwise transfer readily across cultural and linguistic divides. External evaluators, while bringing a different perspective and expertise, cannot have the same nuanced understanding of an intervention or its targets as those engaged continually in developing and implementing an intervention and in resolving the myriad of issues that typically arise in conducting this work. By adopting a true partnership approach, including the knowledge, expertise and insights of intervention participants in the evaluation, a more richly informed and valid evaluation outcome can be produced. Importantly, the PAR approach supports respectful ways of working together and can bridge distrust and apprehension between evaluators and stakeholders.

The approach taken in the JOY evaluation, and the initial workshop described here, would have utility in a wide range of cross-cultural situations. It highlights the value of capacity building and PAR approaches combined with the use of structured tools, such as program logic models, when conducting evaluations in cross-cultural contexts. PAR and program logic approaches allowed program stakeholders, who held rich and deep knowledge about the program and its target audience, to share their knowledge with the evaluators in a mutually beneficial and trusting way that placed explicit value on the knowledge shared by all participants. The cycles of reflection, information collection and action implemented through PAR allowed this knowledge to be gathered, checked and evaluation processes adjusted in a way that effectively disabled some of the conceptual barriers that can arise when working across different languages and cultures.

The structured and succinct nature of program logic models proved to be an effective tool for communication during the JOY workshop that could be similarly applied in other cross-cultural and cross-language situations. Program logic models help to bring together large amounts of information in an accessible format by capturing the key elements and activities required for a program to operate effectively. The presentation of expected outcomes allows stakeholders to readily engage with and to help define a project's outcomes, and that critical reflection can build greater understanding of the program and the true nature of its objectives among both stakeholders and evaluators.

Yet it is recognised that there are limitations to the application of PAR approaches within the practical constraints of this evaluation, and with other evaluations operating in cross-language contexts and over distance. Evaluators and participants can work together to overcome those limitations, but the extent to which difference and distance limit the ability to freely share information, observations, reactions and perceptions may depend on how best practice PAR approaches can be utilised. Despite this, PAR provides an appropriate model for making genuine attempts to ensure collaborative, inclusive and respectful evaluation practice.

While there is still considerable work to be done on the evaluation of the JOY project, the foundations put in place during the workshop in Bangkok and the continuing application of participatory ways of working together are expected to yield positive benefits throughout the life of the evaluation. Feedback from workshop participants indicated clearly that they found the process worthwhile and considered themselves as active participants involved in the development and operation of the evaluation. That is, a sense of ownership was achieved. While the evaluators recognise that Thai culture emphasises politeness and respect, particularly for foreign visitors invited to share their knowledge and expertise, this feedback was encouraging and has been reinforced in subsequent communications.

As the evaluation progresses, the authors will gain clearer indications of whether the approach described in this paper was effective. As data and information becomes available, it will yield insights into whether the goals of shared understanding and engagement are being realised. A key benefit of PAR is that it provides a model for collaborative engagement to find ways of improving the process. Also, it is expected that by continuing to work through cycles of planning, implementation, review and improvement with our Thai colleagues, there remains considerable scope for achieving effective outcomes in this cross-cultural evaluation. In the longer term, we believe the approaches adopted will increase the likelihood that the evaluation findings will adequately and reliably reflect the outcomes that have been achieved through the project.

References

Baum, F., MacDougall, C. & Smith, D. (2006). Participatory Action Research. *Journal of Epidemiology and Community Health, 60*(10): 854-867.

Borzycki, M., & Baldry, E. (2003). Promoting integration: The provision of prisoner post-release services. *Trends & issues in crime and criminal justice, No. 262*. Canberra: Australian Institute of Criminology.

Brydon-Miller, M., Kral, M., Maguire, P., Noffke, S., & Sabhlok, A. (2011). Jazz and the Banyan Tree: Roots and Riffs on Participatory Action Research. In N. Denzin & Y. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (4th ed., pp. 387-400). California: Sage Publications.

Burns, D. (2006). Evaluation in Complex Governance Arenas: the Potential of Large System Action Research. In B. Williams & I. Imam (Eds.), *System Concepts in Evaluation: An Expert Anthology* (pp. 181-196). California: Edge Press.

Clark, H. & Anderson, A.A. (2004). Theories of Change and Logic *Models: Telling Them Apart.* Paper presented at the American Evaluation Association, Atlanta, Georgia, 3-6 November 2004. Retrieved from http://www.theoryofchange.org/library/presentations

Crane, P., & O'Regan, M. (2010). *On PAR: Using Participatory Action Research to Improve Early Intervention.* Canberra: Department of Families, Housing, Community Services and Indigenous Affairs, Australian Government. Retrieved from http://www.dss.gov.au/our-responsibilities/housing-support/publications-articles/homelessness-youth/on-par-using-participatory-action-research-to-improve-early-intervention?HTML=

Frechtling, J. (2007). *Logic Modeling Methods in Program Evaluation.* California: John Wiley and Sons.

Funnell, S., & Rogers, P. (2011). *Purposeful Program Theory.* San Francisco: Jossey-Bass.

Hecker, R. (2008). Participatory action research as a strategy for empowering Aboriginal health workers. *Australian and New Zealand Journal of Public Health, 21*(7), 784-788.

Kemmis, S., & McTaggart, R. (2000). Participatory Action Research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (2nd Ed., pp. 567-605). California: Sage Publications.

Letiecq, B.L., & Bailey, S.J. (2004). Evaluating From the Outside: Conducting Cross-cultural Evaluation Research on an American Indian Reservation. *Evaluation Review,* 28(4), 342-357.

Liamputtong, P. (2010). *Performing Qualitative Cross-Cultural Research.* Cambridge: Cambridge University Press.

McLaughlin, J.A., & Jordan, G.B. (1999). Logic Models: A Tool for Telling Your Program's Performance Story. *Evaluation and Program Planning, 22*(1): 65-72.

McLaughlin, J.A., & Jordan, G.B. (2004). Using Logic Models. In J.S. Wholey, H.P. Hatry, & K.E. Newcome. (Eds.), *Handbook of Practical Program Evaluation*. (pp.7-33). San Francisco, CA: Jossey-Banks.

McTaggart, R. (1989). *16 Tenets of Participatory Action Research. Paper presented to the 3er Encuentro Mundial Investigacion Participatva* [The Third World Encounter on Participatory Research], Managua, Nicaragua, 3-9 September 1989. Retrieved from http://www.caledonia.org.uk/par.htm

Pawson, R. & Tilley, N. (1997). *Realistic Evaluation.* London: Sage Publications.

Stringer, E. (2014). *Action Research* (4th Ed.). California: Sage Publications.

Stufflebeam, D. (2004). *Evaluation Design Checklist.* Michigan USA: The Evaluation Center, Western Michigan University. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/evaldesign.pdf

Tomison, A.M. (2000). Evaluating Child Abuse Prevention Programs. *National Child Protection Clearinghouse Issues in Child Protection no.12.* Melbourne: Australian Institute of Family Studies. Retrieved from http://www.aifs.gov.au/nch/pubs/issues/issues12/issues12.html

Tomison, A.M. (2004). *Current Issues in Child Protection Policy and Practice: Informing the Northern Territory's child protection review.* Darwin: NT Department of Health and Community Services. Retrieved from http://digitallibrary.health.nt.gov.au/prodjspui/handle/10137/121

Whyte, W.F., Greenwood, D.J. & Lazes P. (1991). Participatory Action Research: Through Practice to Science in Social Research. In W.F. Whyte (Ed.), *Participatory Action Research.* California: Sage Publications.

Williams, E. & Tomison, A.M. (2013). Monitoring and Evaluating Community based Interventions for Children and Families in the Asia-Pacific Region. In R.N. Srivastava, R. Seth & J. van Niekerk (Eds.), *Child Abuse and Neglect: Challenges and Opportunities* (pp.159-172). New Delhi: JayPee.

Looking back, moving forward: the place of evaluation at the Tangentyere Council Research Hub

Matthew Campbell

Tangentyere Research Hub

matthew.campbell@tangentyere.org.au

Denise Foster

Tangentyere Research Hub

denise.foster@tangentyere.org.au

Vanessa Davis

Tangentyere Research Hub

vanessa.davis@tangentyere.org.au

Note: This paper, at the request of the authors, did not go through an academic peer review process. However, the authors are happy to present their views where others can see them, including other Indigenous researchers and evaluators.

Keywords: Tangentyere Research Hub, role of Elders in Aboriginal evaluation, Alice Springs Town Camps, Aboriginal evaluation, Aboriginal control of research, Aboriginal control of evaluation, Tangentyere Council.

Abstract

Tangentyere Council is a very important organisation for Aboriginal people living in Alice Springs. The Tangentyere Council Research Hub has been going now for more than ten years, based on two core philosophies: ‘researching ourselves back to life’ and ‘no survey without service’. We rely on Aboriginal research practices and Aboriginal understandings of evaluation as part of this practice. In this paper we will talk about our organisation and its history as well as the research work we do. What we hope to do is to set out what evaluation means for us in terms of the work that we do, and how being aware of this helps us to keep our knowledge and organisation strong. We also hope that it might help others working with Aboriginal people to think differently about evaluation and maybe even approach it differently in the future.

Tangentyere Council

Tangentyere Council has a long history in Alice Springs. It began operating in the early 1970s and was first incorporated in 1979. It is the major service delivery agency for 17 of the 18 Housing Associations known as 'Town Camps' in Alice Springs (Tangentyere Council, 2014).

Tangentyere Council was established by Aboriginal people from what were called the 'fringe camps', areas of Crown land around Alice Springs where Aboriginal people lived but to which they had no formal title. These fringe camps were as old as Alice Springs itself. Tangentyere was formed so that we could get legal tenure for the land we were living on in order to obtain essential services and housing. Since its incorporation in 1979, it has grown into an organisation with a workforce of 243, 61% of whom are Aboriginal.

It is difficult to estimate the population of the Town Camps; the estimated population in Tangentyere's 2005 Mobility Report was between 1,765 and 2,065 people (Foster, Mitchell, Ulrik & Williams, 2005). Given the expansion of Town Camp housing infrastructure, particularly since 85 new houses were built under the Strategic Indigenous Housing and Infrastructure Program (SIHIP), it is hard not to imagine there has been an increase. The service population is likely to be larger again, with people from remote communities all over central Australia coming into Alice Springs, which is the only large service centre in the region.

Each Town Camp comprises an Indigenous community, based on language and kinship groups, which often had its origins prior to the formation of Tangentyere. The majority of Town Camps have Arrernte residents, who are the Traditional Owners of Alice Springs and its immediate surrounds. Many Town Camps also have residents who belong to language groups whose traditional lands are found across central Australia, but who have moved to Alice Springs over a period of time for various reasons. Town Camp residents often have strong links with remote communities and there is substantial mobility between bush and town.

While Town Camps are located in Alice Springs, residents are often culturally and linguistically isolated from the services available. Provision of services by Tangentyere Council, often in partnership with government and other non-government organisations, means that Town Camp residents have access to services which they would otherwise miss out on.

Tangentyere Council was set up to gain access to land and to provide housing and other vital infrastructure, and since its inception the management of this housing was a core function. However since the takeover of Aboriginal housing in 2007 by the Commonwealth government as part of the Northern Territory Intervention (Commonwealth Ombudsman, 2012), Tangentyere Council no longer has any role in the day to day management of housing. Today, Tangentyere Council runs a range of family and youth services, a night patrol, day patrol and youth patrol, a research hub, an art centre, an aged and community care program, a community banking facility and five not for profit enterprises.

This background has been provided to show that Town Camp residents are, and have always been, determined to be in control of their own affairs, assert their rights to a maintain their distinctive Aboriginal identities, and to build and maintain institutions to help them to do this. It is within this milieu that the Tangentyere Council Research Hub came to life in the early years of the 21st century, with its initial motto of 'researching ourselves back to life'. As a research space dedicated to making a difference in the lives of Town Camp residents, evaluation has come to take on critical importance.

In the following section we identify a number of things about our research and evaluation that we think are important. These points are summarised from internal discussions we held to reflect on the place of evaluation in our work.

The importance of control over research

When we set up the Research Hub, we took control of our own research. One of the things we saw so much of was other people controlling research in our communities. When they control the research then they don't have to do things the right way. We already knew, from our own way of understanding research, that doing things the right way is very important. It is wrong for people from other places to come in, ask questions, learn things and then go away and get qualifications from the knowledge that really belongs to Aboriginal people. We have talked to many people who are unhappy about this, especially when these people didn't come back and report on what work they'd done and what they'd found. For us, this is a breach of trust in the knowledge making process. Taking control of research means that we decide what is important, we decide what questions need to be asked, we decide the process. When we do this, we know how to do it in the right way, and this means that not only do we learn new things, but we strengthen ourselves as professional researchers. Also, when we are in control we strengthen our own community because using our own knowledge is strength.

Research is not about finding things out, it's about making a difference

Our original motto for the Research Hub was 'researching ourselves back to life'- our work is not just about 'finding things out' but about 'making a difference'. For us, doing research is only worthwhile if it makes a difference to the people in the camps. Making a difference means making our people, families and camps stronger.

What we mean by evaluation

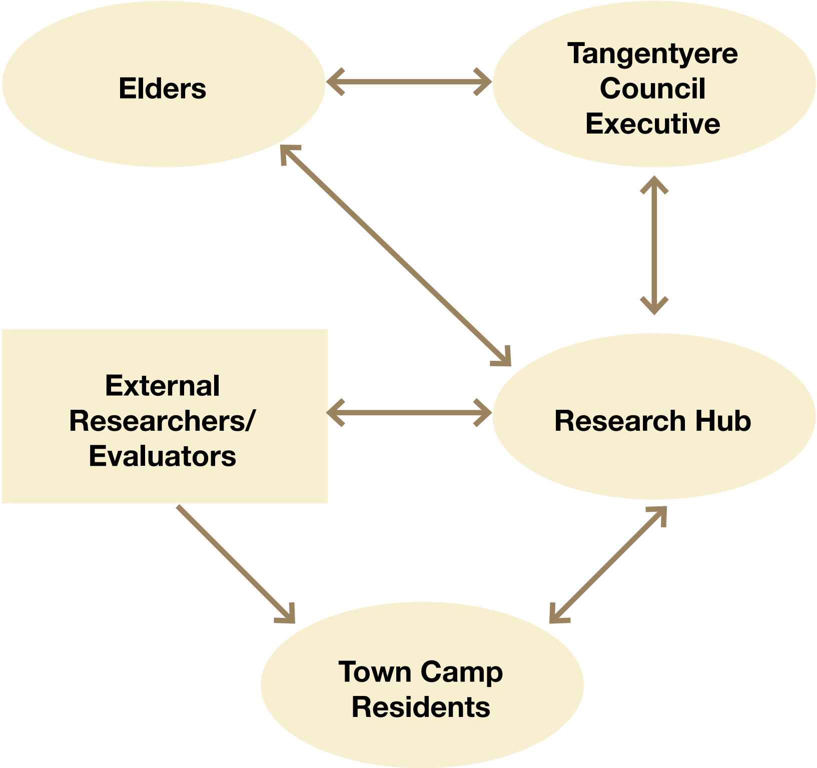
Sometimes Government wants to do evaluations to see how things are working. Often they send people in to do research or evaluation and then these people go away and make decisions. This is not our way. What we mean by evaluation is the way we find better ways to make a difference in our lives and in our community. We are still learning about the different ways ‘making a difference’ can be understood (and that is part of our ongoing evaluation). One of our key values is doing work that makes a difference. It is our processes of evaluation that have helped us to learn this, and to keep trying to make it part of our work. One of the keys for this is how we work with the Elders. They are an important part of teaching us our evaluation.

We have made two drawings for this article. Figure 1 shows the right way to do evaluation, making connections before starting the research, and getting direction from the Elders. Figure 2 shows the way that an evaluation can be done the wrong way, with external researchers/evaluators just trying to go straight to Town Camp residents.

The role of Elders in knowledge making

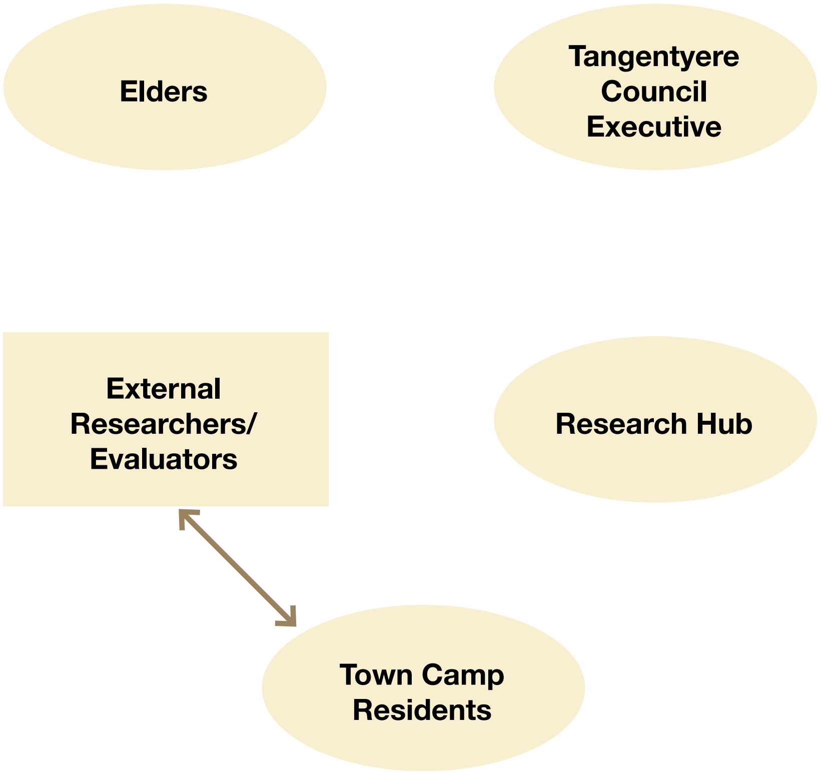
The most important thing to understand is that the Elders are the most important people in our community. Showing respect for them is how we keep our families and communities strong, and we rely on their experience and wisdom to guide us, especially when we are trying new things or are unsure about what we should do. Part of our job in research is to go to the Elders for guidance. When we do this we show them that their knowledge and experience is important, and at the same time they give us their blessing to do the work we do. This makes sure that other people also see that we are doing our research the ‘right way’, and because we are doing it the right way they can support us and get involved too. One of the important things we have learned from the Elders is that we need to be connecting up our work with the past.

Figure 1: **Doing evaluation the right way**


Elders
Tangentyere Council Executive
External Researchers/ Evaluators
Research Hub
Town Camp Residents

1.	External evaluator wishes to evaluate a program for Town Camp residents
2.	External evaluator contacts Tangentyere Research Hub
3.	Research Hub seeks approval from Tangentyere Executive to proceed
Decision making point	
4.	External evaluator works with the Research Hub to devise questions and methodology
Decision making point	
5.	Researchers liaise with the right elders to discuss the project
6.	Researchers (with or without external evaluators) do on ground work with Town Camp residents
7.	Researchers and external evaluators analyse data and come up with tentative findings
8.	Tentative findings are shared with elders and Executive - changes made where necessary
Decision making point	
9.	Findings are reported to the elders, the Executive and the Town Camp residents
10.	Findings are reported externally

Figure 2: **Doing evaluation the wrong way**


Elders
Tangentyere Council Executive
External Researchers/ Evaluators
Research Hub
Town Camp Residents

1.	External evaluator wishes to evaluate a program
2.	External evaluator devises questions
	Decision making point
3.	External evaluator approaches people in town camps
4.	External evaluator analyse data and come up with findings
	Decision making point
5.	External evaluator presents findings to external audience (and perhaps to residents)

Connecting up with the past

Evaluation for us is less about ‘looking back to learn for next time’; instead, it is focused on ‘making sure we are connecting up with the past’. When we say ‘connecting up with the past’ it might make it sound as if we are only interested in what went before – something like doing history work, but this is not what we mean. What we mean is, whatever we do, we have an important job to *show* how what we are doing builds on the work that was done before, not just by us as researchers, but everyone in the Town Camps. People in the Town Camps want to see how the work we do builds us up and keeps us strong as Aboriginal people, and so we need to be able to show people this.

People might worry that this might mean that we do not consider how evaluation assists us to do things differently in the future. This is not the case; we are always looking both forwards and backwards. We need to look forward, but we also have to think about what we are doing now. What we do creates what comes next. If we did not think about this, we would just keep on making mistakes. However, because our job is also to build up our communities, we need to build on what was done in the past. The old people fought those fights so we can be here doing what we are doing today. We respect them by connecting our work with theirs. We are only standing here today because of those who came before us. Without them, we are nothing. Evaluation helps us to remember we are continually building the foundation upon which those who come after us will stand.

How do we do this connecting work?

It's a bit hard to talk easily about how we do this connecting work as it is really just a part of our day to day lives. But it is important that we try to show how we do it, because we think it is something that other researchers should be doing too – other researchers can help us to do this work, or they can be helping others in other places. Most of our evaluation takes place as we do our work. We are going backwards and forwards to our Elders all the time, talking about the projects we do; we are learning from them about how to do research properly. They are always talking about how important it is to talk to the right people, and they teach us how to build or keep connections between people and families. This is important in Aboriginal knowledge making. They also tell us if our research is valued, and what we might think about doing next time. Most of our evaluation occurs in our day to day conversations. Some people might think that proper evaluation should be more formal; however our authority to do this work comes from these conversations. These Elders are helping us to do the important work of connecting what we do with what we've done before, and with what others have done before. As we move on to become Elders in our community, we will take up this role.

Evaluation helping us to focus on what is important

Evaluation is also part of helping us to keep our minds on our relationships with the Town Camps. We are not really separate from the Town Camps and we rely on them, and they rely on us. Evaluation has helped us to be very clear about what is important to us and the people we work with, and the main thing is good trusting, respectful relationships. If we don't have this then we can't do our work. And the thing for us is, if we do the wrong thing that doesn't just affect 'this' project; it affects every project we might want to do in the future. It's because of this that we are very careful about how we do things, the questions we ask, and how we give information back. We are not like outside researchers, who can come in, do the wrong thing and go away. We will be living here forever, so we cannot do anything that makes a problem.

Evaluation to keep our organisation strong

Another important part of evaluation is thinking about how does it help to make our organisation strong? Tangentyere is interesting because it is made up of all the Housing Associations that set up each Town Camp in the first place. This means Tangentyere can do things that the each little Town Camp cannot. It can be really hard when people from different Town Camps have different ideas and opinions. The Research Hub is one place where we talk to people from all the camps, and can put their stories together so we can tell others about them. Part of what we do is bring the different Town Camps together, showing that we can be different, but at the same time there are things that keep us connected. We know we are doing our work well when it plays a part in Tangentyere talking up strong in discussions about issues in Alice Springs. At the moment, the value of our work is being recognised through projects looking at alcohol management and chronic disease management. In both of these projects, our work contributes to Tangentyere being able to play a leading role in getting the stories of the people of the Town Camps heard.

Final reflections

Evaluation is something that we do all the time in our work; we can’t separate it from our research work. Above all, we value three main things: our relationships with our Elders, and how our work can strengthen our old people and their position as the knowledge authorities in our communities; our relationships with people in the Town Camps; and making sure we connect up our work with the past.

The Elders are the people who help us to understand our place in the world, and guide us to do work that makes a difference for the people in our communities – the Town Camps of Alice Springs. At the moment, most of our evaluation takes place behind the scenes, in all the conversations we have around our work. As we go on, we want to get better at capturing this process, but as we do so, we will never lose sight of the fact that we are here today, doing our work in our own way. For us, this is evidence that our work is valuable, not only to us, but to our families and communities as well.

For others, we want you to know that you can help us to do this work. You can do this by listening to us about how research should be done, and giving us the time to talk properly and respectfully with our Elders. If you can do this, then it means that research is being done properly in our way. If you can do this, then you will help us as individuals and you will help our communities; but you might also find that you learn something new about how our knowledge is made too.

For us, the past is so important. Whatever is happening now is happening because of what happened before. If we forget this we are in danger of not learning, but also in danger of breaking our connections. This connection with the past is an important part of our strength and who we are as Aboriginal people. As Aboriginal people we have had to fight many fights to have our rights recognised. Those old people who fought those fights did them for us, and for the people who come after us. If we forget them, we undermine our own strength. We also know that we are not in this alone. We recognise the people who fought, and who continue to fight alongside us. Understanding the past and learning from it, connecting it with what we do today, is something we need to do in every project. For many of us, we are still fighting to make our lives better.

References

Commonwealth Ombudsman. (2012). Remote housing reforms in the Northern Territory. Canberra, Australia. Retrieved from Commonwealth Ombudsman's website: http://www.tangentyere.org.au/publications/research_reports/DKCRC-Report-9-Population-and-Mobility-in-the-town-camps-of-Alice-Springs.pdf

Foster, D., Mitchell, J., Ulrik, J., & Williams, R. (2005). Population and Mobility in *the Town Camps of Alice Springs,* A report prepared by Tangentyere Council Research Unit, Desert Knowledge Cooperative Research Centre, Alice Springs. Retrieved from: http://www.tangentyere.org.au/publications/research_reports/DKCRC-Report-9-Population-and-Mobility-in-the-town-camps-of-Alice-Springs.pdf

About Us. (n.d.) Tangentyere Council. Retrieved from http://www.tangentyere.org.au/

Aboriginal contributions to the evaluation of housing (and to postcolonial theory)

Michael Christie	Matthew Campbell
Northern Institute Charles Darwin University michael.christie@cdu.edu.au	Tangentyere Council matthew.campbell@tangentyere.org.au

Keywords: Aboriginal evaluation, Indigenous evaluation, housing reference groups, ground up evaluation, evaluation policy impact, double participation, cross-cultural evaluation

Abstract

Housing Reference Groups (HRGs) began to be established in remote Northern Territory (NT) Aboriginal communities in 2009 when the Northern Territory Government compulsorily acquired remote Aboriginal housing and closed down 75 Aboriginal Housing associations. In this highly contested context, we were invited to undertake an evaluation of the HRGs. Through both open discussion and semi-structured interviews, we learnt from the Aboriginal people we worked with to see a much wider, more structural understanding of housing and its governance. This in turn led us to reflect upon Aboriginal contributions to the theory and practice of evaluation, and their various relations to received theories such as social justice, pragmatist philosophy and ethics, and Frierian conscientization. This collaborative ground-up evaluation process contributed to our own ongoing practices of evaluation, and possibly to some slight but reverberating changes in government policy and practice.

Introduction

Remote Housing NT is a division of the NT Department of Housing which was established to manage the delivery and improvement of Aboriginal public housing in the Northern Territory. It is a system through which the NT and Australian governments are implementing their National Partnership Agreement on Remote Indigenous Housing (NPARIH; 2009). The introduction of Remote Housing NT saw all Aboriginal housing previously managed under Aboriginal Housing Associations converted into public housing. This entailed the government acquisition through compulsory leasing of Aboriginal land for public housing. Two construction company consortia known as Alliances were engaged to deliver the Strategic Indigenous Housing and Infrastructure Program (SIHIP), an element of NPARIH (2009) in which houses will be built, rebuilt or refurbished. Housing Reference Groups (HRGs) were set up in 73 communities, a number of Town Camps and other 'living areas' 'to work with government and ensure local communities had input into decisions about housing in their community'. HRGs were to advise on, but not to allocate, housing. Allocation and repair and maintenance decisions are ultimately to be made by the Department.

The 'Consultation for Better Housing' project was an element of a much larger Australian Research Council project called 'More Than a Roof Overhead' (MTRO). MTRO sought to take a broad, whole-of-system and interdisciplinary approach in research on the delivery and management of remote Indigenous housing where the key challenge is to design, build and manage housing as an integrated and sustainable system. Our sub-project was to look at the consultation and engagement mechanisms that existed in the NT and the extent to which they were facilitating dialogue between housing users and administrators.

We write this paper as two academics who have been involved in grounded research and evaluation work in Aboriginal contexts for 40 and 20 years respectively. In this project we were determined from the beginning that this would not be a conventional evaluation, where we would go in with the questions already prepared, and to which we would write down the responses. We had worked together before on a range of projects around Indigenous community engagement which had led us to understand the research process (and the evaluation that emerges from it) as a work practice in which participants collectively and iteratively define the issue being researched while undertaking action to produce change (Christie, 2013a). As such we had moved from understanding research as a process of 'finding things out' to a social process in which the world changes through our acts. In this sense we understand knowledge as action and performative, rather than representational. This orientation led us to fundamentally rethink what it means to be a researcher, making a decisive move from being a 'judging observer' to being a (albeit privileged) 'participant in collective action' (Addelson, 1993). Taking the position of a 'participant in collective action' also entails thinking about methodology, as we must work to ensure our research is generative – producing tangible change in the world. We will return to these points later in the paper.

This paper first outlines the methods we used, and then summarises the findings, recommendations and conclusions. This is followed by a section where we reflect quite specifically on the contributions of Aboriginal people to our thinking about practices of evaluation. We also make some comments on how these principles of evaluation are echoed by and extend theoretical positions coming out of the academic (enlightenment) tradition. We conclude with a short discussion of some surprising insights and outcomes of the project.

Method

There was no escaping the fact that ours was not an arms-length evaluation. We had been invited to undertake the evaluation because our research showed a history of theoretical and practical engagement with Aboriginal people that was seen by government as being both just and productive. Everyone, right up to senior bureaucrats, accepted that Aboriginal people in remote places had been disempowered since the NT Emergency Response. Houses they used to own and manage collectively, were not, in their eyes, public housing. Yet that was exactly what they had become, according to the government. So often we heard of the shock of Aboriginal residents being told by government workers: 'This is not your house any more'. Everyone therefore had an interest in the HRGs being effective, even if only as a way of addressing the range of difficult conversations in remote communities that had emerged around housing as a result of the takeover. We were welcomed by the Aboriginal residents and HRG members (and the public servants) because they could see we were there to listen and to help.

In the first phase of the research, a group of Aboriginal elders were brought together to a 'workshop' to discuss the housing in their communities and homeland centres, past present and future, and their views and recommendations for the community-based HRGs. They were paid for their work. In the second phase, we conducted semi-structured interviews with senior bureaucrats in the Department of Housing, with 'street-level bureaucrats' (Lipsky, 1980) working to organise housing and HRGs in remote communities, and Aboriginal members of HRGs and senior local authorities. The senior managers of the Department of Housing (which had contributed funding to the research) were keen to see the interviews in the second phase focus upon specific areas of interest. We negotiated with them around these areas, and together we came up with what we came to know as the 'focus area' list. This list included: evaluate the size and representatively of the HRGs, their selection and review processes; their Terms of Reference, roles and governance arrangements, payment and time commitment, and frequency of meetings; the future of the HRGs after SIHIP; the relationship of the HRG to other bodies like the shire 'Local Boards' etc; and the government's communication and feedback structures and processes. We were told not to open the question of sitting fees for HRG members, as remuneration was 'out of the question'. Although the focus area list appeared like a set of questions, in practice we used it as a starting point for discussions and each interview unfolded differently.

We began by talking with senior NT Government bureaucrats within the Department of Housing. The interviews were semi-structured, using the focus area list as a basis. The interviews were taped and transcribed; the transcriptions and a summary we made from them were returned to the interviewee for comment and changes if required. At the conclusion of each interview we asked if there were other people we should seek to speak to about the same set of questions. In this way we were led to others within the Department dealing with remote housing at all levels, from those responsible for management of housing or for policy development and implementation, to those engaged in day to day, face to face interactions with Aboriginal residents.

In a parallel process we talked to Aboriginal members of HRGs in both remote communities and Alice Springs Town Camps. Again the same process applied, of recorded semi-structured interviews conducted using the focus areas list. We paid HRG participants for their time if they were not otherwise in paid employment while doing the interview. We consider that people in such interviews are providing us with access to their knowledge, which is valuable and which needs to be recognised in the form of payment. Our experience shows us that Aboriginal people view their knowledge differently from non-Aboriginal people, and that acting in good faith requires recognising these perspectives. Our experience also shows us that Aboriginal people recognise and appreciate this.

Our interviews were conducted either as one-on-one interviews or interviews with small groups. The particular configuration depended on our relationship with the person being interviewed and what made them feel comfortable. We conducted interviews with bureaucrats in their offices, and with Aboriginal people in our offices, over the phone and in community centres. In some situations we knew the Aboriginal people we were interviewing through previous work; they were comfortable talking to us in ways that they might not use with people they did not know.

Once comfortably ensconced wherever we were, we were told stories. These stories ranged from the old days and the ancestral metaphors for housing, to the way housing manages people and the difficulties experienced by bureaucrats in finding a quorum for an HRG meeting, from dissatisfaction about communication practices to stories about how the change in housing management has created myriad problems that are beyond the scope of tenancy agreements and repairs and maintenance regimes. The diversity of these stories indicated to us that housing, and the HRGs as a consultation tool, could not be talked about simply or instrumentally, and recommendations to improve their functioning would address only a small part of the problem. Indeed each person we interviewed talked about quite different things. Yet it was apparent that their perspectives were all connected, and we hoped that through our generative analysis approach we might be able to make some practical contributions that would assist people wherever they found themselves within the system.

Our approach was founded on the assumption that our research process had the potential to make connections between the stories of the participants regardless of where they sat within the housing system. We theorised that our process would assist participants to see things that previously they could not see, and hopefully begin to make changes in their practices based upon seeing things in new ways. We made a transcription of everyone’s interview which we sent back to them accompanied by a summary in which we identified what we thought were the major points to emerge. We invited each participant to review this material and get back to us with any changes, adding things, removing things or clarifying whether our summaries made sense. We asked Aboriginal participants if they would like us to revisit them with the view to making any required changes to their interviews.

We also looked at these summaries collectively and tentatively put together another story (what we called the mega-summary) that sought to draw together the disparate range of issues, practices and concerns that emanated from the field. We then distributed the overall summary. Our hope was that this would elicit feedback and interest in particular issues or areas that we could then focus on with the participants. Our rationale at this point was not to try to develop a coherent story ourselves, but to draw the other participants in to the knowledge making process.

However, interestingly and frustratingly only one person got back to us with feedback. In hindsight, it is possible that the approach we took was not the most appropriate way to elicit feedback, particularly from the Aboriginal participants whose grasp of English may not have been great or who were uncomfortable coming back to us with concerns. However, the lack of response was consistent from all (bar one) of the people interviewed. Undeterred, we continued interviewing more people using the same process. In the end we interviewed 15 people (bureaucrats and Aboriginal HRG members), and analysed and summarised over 100,000 words of transcription, from which we developed a draft report. This again was distributed with invitations for feedback and comment, with a particular focus on the findings and our proposed recommendations. Again we did not receive any feedback. We also distributed the report to other researchers within the MTRO project, who thought that our approach, findings and recommendations were interesting and useful for the purposes of the project overall.

Findings

The final report for government, community, and the public was in the form of a '1-3-25': one page of 'main messages', three pages of executive summary, and a full report. This was to make the report accessible to a range of people, some of whom would be interested in the detail, others only in the 'take home' messages. The bulk of the report was built on direct quotes, organised thematically, from people interviewed in the second stage of the research (Campbell & Christie, 2013), but the report encompassed findings and analysis from both phases of the research. Overall, nine Department of Housing bureaucrats and 14 Aboriginal people participated in the project.

The seven main messages of the report were summarised as follows:

1. In Aboriginal communities good housing and its negotiated deployment are seen as crucial determinants of health, wellbeing, and local governance. Much community distress is attributed to poor housing and bad allocations.

2. Cultural authorities in remote communities and town camps play an often hidden but ongoing role in decision making about housing, which should be recognised and integrated into HRG processes.

3. The current policy of 'advice only' creates many problems, both at the community level and at the interface between government and community. Careful on-the-ground negotiations and decision making within HRGs would very seldom end in disagreement, so the advice-only policy is potentially harmful and unwarranted.

4. From the Aboriginal perspective, decisions about housing are not separate from decisions about health, education, employment, community development or economic development. The work of HRGs should therefore be integrated into wider collaboration between senior (and other) community members, and all levels of government.

5. Miscommunication and lack of communication around new works, allocations and repairs, and lack of timely responses results in much distress and acrimony. The establishment of more immediate and effective communication and accountability processes (for example a 1800 phone number), would improve the engagement of community members and the effectiveness of HRGs.

6. Issues around housing cannot be solved by the Department of Housing alone. A 'whole-of-government' approach is necessary to match the 'whole-of-community' approach.

7. Senior community members who work with governments on decision making around housing and other issues should be remunerated for their knowledge, authority and insights.

Reflections on Aboriginal evaluation and its relation to 'theory'.

As academics and researchers we are constantly confronted with 'theory', which can be both interesting and exciting. However, if we are not careful, it can distract us from our engagement with the real world. We may be tempted to use 'northern theory' (Connell, 2007) as a lens through which to make sense of what Aboriginal (and other marginalised) groups are trying to tell us. Such an approach commonly sees researchers using the received theory to pre-empt the possibility of Aboriginal people sharing with us anything new, or different, or challenging to our cherished assumptions. In this paper we are attempting to articulate quite specific Aboriginal contributions to our thinking about housing (and evaluation). In this section we draw attention to five insights provided by the Aboriginal participants in this research and try to link them to academic theory and practice which has helped us to clarify their messages, critique the received theory, and create something new. One of the interesting things to note here is that the Aboriginal people we interviewed were from Arnhem land and Alice Springs Town Camps, people who have very different histories, languages and housing contexts, yet the stories they tell have strong connections.

1. First, we note something holistic about the Aboriginal understanding of housing and governmentality. People manage houses, houses manage people. HRGs are there to help people manage houses, but evaluating the HRGs simply as technical mechanisms, employs a top-down model which pathologises (or condemns) the impoverished and overcrowded occupants of (what is now) public housing. Understanding HRGs solely or primarily as technical mechanisms assists in obscuring injustices that may be perpetuated and, importantly, may foreclose possible action towards a more systemic solution. Thus our Aboriginal participants remind us of the 'social justice' approaches,

aimed at unearthing underlying causes for social problems and suggesting ways of going on together (see for example Weinberg, 2008). Even further from what the structural theorists avow, our Aboriginal participants remind us that we must work towards social justice by paying attention to our land, our environment and our place in the world.

> … Children just grow up, because of the land, and the old people… they are born gifted, with a talent which only the wise people, and which only the land can provide. And when they are born on the land they are chosen to be certain leaders in the tribes. Straight after the Wet Season when we sit down by the beach and look at the sea around the small islands of the hunting grounds of the reefs where we hunt turtles and the certain signs in the skies tell the stories, of clouds sitting in the air after people have eaten …. Actually, it tells the story that we are the right people of that country. (Guyula, 2010, pp. 18-19)

2. We were struck by the contrast between the rather disciplinary approach of some of the bureaucrats …

> You have a responsibility to put back into your community. So I would hope that people would put their hands up, and want to be involved, and when we set up an HRG that they'd come along and have a presence, and be actively involved... (Territory Housing employee cited in Campbell & Christie, 2013, p. 23)

…with the patient storytelling on the part of the Aboriginal elders, gently involving us in what years ago, we called 'conscientisation' (Friere, 1972). Back then of course, our focus was on helping Aboriginal students and their communities achieve what we took to be an in-depth understanding of the world, and the perception and exposure of social and political contradictions. Now the elders use traditional agreement making methods to 'conscientise' us and the people who fund and act upon our research. We see their subtle good-humoured storytelling as requiring us to back off from our easy assumptions about the functionality of HRGs and see them more clearly in their historical socio-political contexts.

3. We had our 'focus area' list (which looked at the workings of the HRGs- their constitution, selection, induction etc.) which we used to frame our interviews and discussions to focus on the work at hand, however we maintained a non-prescriptive attitude toward how each discussion unfolded. In our interviews with Aboriginal HRG members each question seemed to lead to a story about people, place and housing, which only indirectly addressed the problem of HRG efficacy. More correctly, the Aboriginal participants could see a quite different problem, of which the HRGs were a manifestation, and seemed to do with finding ways to go on together in good faith working through these housing and related dilemmas such as health, education, and environment. So much of the discussion was not about housing or HRGs, but on ways in which people come together and address the problems and remain accountable to the solutions. There is no top-down solution.

Reminded once again of the America pragmatist philosopher John Dewey's book *The Public and its Problems,* we were being guided to see the poor state of Aboriginal housing, the fractious HRG meetings, the frazzled street-level bureaucrats, the disappointed and disempowered elders, all as effects rather than causes of the problems of not making agreement on ways to go on together. We were directed away from making judgements about the effectiveness of specific mechanisms, rules or provisions. Rules and regulations work differently in different places depending upon the good will and the problem of the moment. Overall we came to see the dislocation between residents of Aboriginal housing, and government bureaucrats delivering the housing as an effect of not working together in good faith rather than a problem with the HRGs per se (although improving the HRGs would certainly help). It wasn't a fundamental dislocation, there was good and bad on both sides of the divide, but our research findings led us towards reading the process from the (Aboriginal) view of conflict resolution, rather than from the (government) view of improving the workings of an ordered governmental structure. As a result, our report made no differentiation between Aboriginal and non-Aboriginal voices in the quotations which made up most of our evaluation report since everyone saw Aboriginal housing equally – although differently – as a difficult issue.

4. Everyone had good words to say about the hapless bureaucrats whose job it is to drive or fly to very remote Aboriginal communities, dealing with dirt roads, extreme heat, tiny planes, and violent storms as they set up meetings, find a quorum, keep focus, report back to government, bring often bad news of unpopular decisions, and explain government policy. These, we were told from people at all places within the system, are committed people doing the best of a very difficult job – the sort of working-in-good-faith which is so valued by the Aboriginal participants. What became clear to us, after triangulating and revisiting, was the extent to which the success of the Street Level Bureaucrat depended upon their moment by moment discretion in often difficult settings. They sometimes bent rules or ignored them, they had creative ways of becoming quorate, they did their best to keep the meetings focussed, and when they went off on a tangent they went to talk to people who might help (they knew, and had ongoing relations with many people in the community). This was also true of the HRG members, who, for example, worked hard to have the ideas of the traditional landowners included in the decision making even when they had been excluded from membership. Talking about the way that negotiations around housing in their communities can and do work, the Aboriginal people prosecuted a vision of negotiation reminiscent of the work of Lipsky (1980), who wrote of street level bureaucrats as the policy makers, because in a real sense the only "policy" that the public experiences is that which is mediated through their contact with street level bureaucrats. In this sense, "policy" is the cumulative effect of the individual decisions made by street level bureaucrats. It was clear that in the reflections of the Aboriginal participants, these street level workers on both sides of the divide were the true policy people, but there was more to it than that. They were not just clever bureaucrats, but decent people of good faith who, no matter how much they were positioned so by the system for which they worked, were not unconcerned, judging observers (Addelson, 1993). They were in the field, confronted by and having to deal with a range of issues, not all of which were related to housing.

5. Finally, we set out, or were sent out, to conduct an evaluation of the system of HRGs. Within the wider research brief, our research questions were phrased, quite naturally, as technical questions. How big? How often? What training? Those technical questions were reworked, by the Aboriginal people we talked to, into ultimately moral questions. The stories – some funny, some sad, some outrageous – all seem to move the ostensibly technical question of housing and its administration towards an ethical question of working honourably together. This was done through a very natural but subtle narrative technique locating the actors in the broad socio-economic and political situation of remote Aboriginal people living on their own land with their own histories (Christie, 2013b).

A surprising end

So much of this project was disappointing. We had begun by taking a generative approach to our research; we didn't want simply to come up with a report, but to actually change practice. We have unashamedly come to see ourselves as activist researchers – we are participants in the world, not detached observers, and we want to make a difference. Our big question, and the one we continue to ask ourselves, is 'how do we do this responsibly?' How do we create the space and processes for ourselves and for others to do "work that is respectful of the creativity of others as they enact truth and take part in making the meaning of the world"? (Addelson, 1994, p. 8). In terms of enacting our process, we talked to as many people as we could. We asked them how we might be able to help. We transcribed the interviews and sent them back to be extended, shared or changed. We asked if there are people or groups we could talk to or work with to actually implement some of the proposed changes. We heard nothing. We sent out reminders, but received nothing back. We made the report and sent it out expecting to get some feedback – maybe from senior bureaucrats telling us we hadn't done what we were paid to do. Nothing.

Some months later, we were approached by government to undertake another, much larger project which involved (among other things) looking carefully at which different community-level advisory organisations could better work together or be amalgamated (what we had called the 'whole-of-government – whole-of-community' approach). We were asked to investigate the possibilities of making payments to senior community members who were spending increasing amounts of time doing indispensable facilitation work for government workers. We expressed surprise at what seemed to be a complete change around in policy around remuneration, and were told in passing, in a meeting with government, that this new thinking had came out of our housing report. Someone, somewhere, had been listening, thinking and reading, invisibly to us but clearly within the corridors of power.

We have spent much time reflecting on how the Aboriginal evaluation responses had somehow brought about this surprising change. Perhaps the ongoing effort in keeping the process public, or having Aboriginal voices prominently telling an alternative story about housing, generated an awareness that previously was absent. Either way, a critical revelation is that there are other ways to change policy than through the more traditional method involving membership within the 'policy development' system.

Understanding how this particular change might have come about is an important question, and given the circumstances of this change our attention is drawn to what Kathryn Pyne Addelson calls "double participation" (1994). As academic researchers with an interest in generative methodologies, we worked hard to be inclusive, providing the other participants in the project with numerous opportunities to contribute to the process, the findings, the report and further action. As it turned out, this did not seem to work in the way we imagined it would. We are prompted then, to think more deeply about the notion of, *'providing others with opportunities'* to be involved. How are we positioned? How might this affect the work we do? And what might emerge as a result? As academic researchers whether we like it or not, we are agents of governance. As such, rightly or wrongly, we assume the right to know in our own terms. As professionals we have a lot of power to frame how things are understood, and what is to be included and excluded. We assumed at the outset that other participants would want to be involved in this knowledge making process; after all, they knew much more about what went on within HRGs than we did. However it turned out that it was our double participation that appeared to make the decisive difference, our ability to function both in the day to day world of bureaucrats and housing residents, where we learnt about HRGs and how they worked, but also our ability to turn this into another kind of story, one that operated in other contexts. It was this double participation that allowed the voices and perspectives of Aboriginal people to be heard and to make a difference.

It is hard not to feel that there is something a little bit wrong about this, because it was our professional power that ultimately appeared to make the difference. It is important to note that if this is the case, it was our professional power used responsibly. That is, we did not seek to create an authoritative account on our own terms. Our attempt at using a generative process privileged the views and methods of the people and the contexts we encountered in an iterative process, and it was this that allowed us to produce the report we did.

In the end, it appears that all those unanswered emails, (apparently) unchecked texts and the simple public 1-3-25 may have finally contributed to enough momentum, visibility and implicit agreement to precipitate a slight change in policy and a call for a new consultancy which entailed 'development' rather than just 'research'. As noted, this understanding relied on the validity and rigour of our process, bringing diverse stories together in a way that made sense and illuminated something about the HRGs and the processes that surround them that had hitherto been invisible. We still find it puzzling, upon reflecting on Aboriginal contributions to our evaluation, how hard it is to distinguish between what we had been told about good housing, and about good evaluation. This seems to lead towards new, more holistic thinking about evaluation. Aboriginal participants refused to be blamed for poor participation in HRGs and required everyone to view the technical problem of HRGs as the moral problem of working together across cultures productively and in good faith around the inseparable problems of community housing, health, education and environmental management in the here and now.

References

Addelson, K.P. (1993) Knowers/Doers and their moral problems. In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies.* (pp. 265-301). New York, NY: Routledge.

Addelson, K.P. (1994) *Moral Passages: Toward a Collectivist Moral Theory.* New York, NY: Routledge.

Addelson, K. P. (2002) The Emergence of the Fetus. In C. Mui, & J. Murphy (Eds.) *Gender Struggles: Political Approaches to Contemporary Feminism.* (pp.118-136). Lanham: Rowman & Littlefield.

Campbell, M., & Christie, M. (2013). *More than a Roof Overhead: consultations for better housing outcomes.* Retrieved from http://www.cdu.edu.au/centres/yaci/docs/MTROreport2.pdf

Christie, Michael. (2013a). Generative and 'Ground-Up' Research in Aboriginal Australia., *Learning communities: International Journal of Learning in Social Contexts,* (13), 3-13.

Christie, Michael. (2013b). Talking home and housing: The ethnographer brought back down to earth. *Learning Communities: International Journal of Learning in Social Contexts,* 12, 29-34.

Connell, R. (2007). *Southern Theory: the Global Dynamics of knowledge in social science.* Cambridge, UK: Polity Press.

Council of Australian Governments. (2008). *National Partnership Agreement on Remote Indigenous Housing,* Retrieved from http://www.federalfinancialrelations.gov.au/content/npa/housing/remote_indigenous_housing/national_partnership.pdf

Dewey, J. (1927/1954). *The Public and its Problems.* Athens, Ohio: Swallow Press.

Freire, P. (1972). *Pedagogy of the oppressed.* Harmondsworth, Middlesex: Penguin.

Guyula, Yingiya. (2010). The Story comes along and the children are taught *Learning Communities: International Journal of Learning in Social Contexts Australia* (2), 18-22.

Lipsky, M. (1980). *Street level bureaucracy: dilemmas of the individual in public services.* New York: Russell Sage Foundation.

National Partnership Agreement on Remote Indigenous Housing (2008). National Indigenous Reform Agreement. Retrieved from http://www.federalfinancialrelations.gov.au/content/npa/housing/remote_indigenous_housing/national_partnership.pdf

Weinberg, M. (2008). Structural Social Work: A Moral Compass for Ethics in Practice. *Critical Social Work,* (9), 1.

Evaluation Methods for Vulnerable Populations: The Case for the Deaf and Hard of Hearing

Kathryn Cairns

Centre for Program Evaluation
University of Melbourne

kathryn.cairns@unimelb.edu.au

Patrick McLaren	**Janet Clinton**	**Ruth Aston**
Centre for Program Evaluation University of Melbourne	Centre for Program Evaluation University of Melbourne	Centre for Program Evaluation University of Melbourne

Keywords: evaluation ethics, participatory evaluation, collaborative evaluation, evaluation with Deaf stakeholders, evaluation with hard of hearing stakeholders, real-time captioning

Abstract

Conducting research and evaluation with vulnerable populations requires deliberate and mindful adherence to ethical standards and principles such as those outlined in the Belmont report, the Nuremberg Code and Declaration of Helsinki, the Australian Code for the Responsible Conduct of Research, the National Statement on Ethical Conduct in Human Research, as well as other professional standards of practice. When working with these populations, the principles of beneficence and non-maleficence, in particular, must be embedded from the project outset through to the dissemination of the findings. However, operationalising these principles can pose a challenge in practice. In this paper, we will present a case study of an evaluation of a real-time captioning program for Deaf/hard of hearing students, to illustrate the implementation of an inclusive and participatory evaluation methodological framework, in which adherence to ethical standards and principles was first and foremost.

Introduction

Research exploring literacy levels within the Deaf/hard of hearing population has revealed a gap in literacy attainment among this population, with Deaf/hard of hearing students typically demonstrating weaker literacy skills when compared to their hearing peers (King & Quigley, 1985). For example, a Victorian study found many Deaf/hard of hearing school leavers achieve a literacy level roughly equal to a Year 6 student (Walker & Rickards, 1992). Further, Brett (2010) has reflected on the challenges faced by Deaf/hard of hearing students due to the dominance of spoken language as the mode of instruction in academic settings, and noted that functional difficulty with spoken language excludes many students from participating in classroom learning activities. Research such as this highlights the need for assistive support services and technologies to help reduce the inequality in access to education and ensure that Deaf/hard of hearing students have an equal opportunity to reach their learning potential.

Literature Overview of Real-Time Captioning in the Classroom

Captioning is a form of assistive technology which has received attention in recent decades due to its potential to improve literacy and language skills amongst Deaf/hard of hearing populations. Research has demonstrated that classroom captioning can increase the comprehension of spoken material for Deaf/hard of hearing students (Boyd & Vader, 1972; Markham, 1989; Murphy-Berman & Jorgensen, 1980; Stinson, Stinson, Henderson & Miller, 1988). To illustrate this, in a study conducted by Boyd and Vader (1972), a group of Deaf/hard of hearing students were shown an educational television program with and without captions which were generated by a stenographer. Comprehension of the program's content was significantly higher after exposure to captioning (Boyd & Vader, 1972). Stinson et al. (1988), reported similar results among Deaf/hard of hearing students who were using captioning when compared with students using manual interpreting alone. The effects of captioning have also been observed with students for whom English is a second language. A study conducted by Markham (1989), examined the comprehension of video material by 76 university-level English as a Second Language (ESL) students and found greater comprehension with the captioned segment. Markham (1989) concluded that these findings support the proposition that learning capability is improved by simultaneous processing of different sensory modes.

In addition to comprehension, it has been suggested that captioning may generate an increased sense of inclusion and participation in the classroom for Deaf/hard of hearing students. Youdelman and Messerly (1996) explored the perceptions of students, teachers and note takers on the effectiveness of caption-like technology (computer-assisted note taking), and found that Deaf/hard of hearing students understood the lesson content better, could keep up with the pace of the lesson, and were better able to summarise the content of the lesson. The impact of this could lead to enhanced academic performance, but also to a more inclusive classroom environment with students able to participate more fully in the classroom as they can understand what is happening.

Real-time captioning (RTC) is a rapidly developing technology that differs somewhat from other forms of captioning. Whereas typical captioning technology is retroactively added, often in a simplified form, in transcripts or as subtitles (see Daelemans, Hothker & Sang, 2004), RTC is generated in real time with a delay of only seconds, thus providing viewers with access to captioned content almost instantaneously. As research indicates that Deaf/hard of hearing students experience difficulties understanding content delivered in English, it is conceivable that RTC within a classroom context could potentially contribute to improved literacy by providing verbatim text to supplement spoken language. As a consequence RTC provides students with an opportunity to connect written and spoken language in a concrete way, and within an interactive context. While clearly an important area of research, the technology is in its infancy, and thus there has been no research to date examining how RTC affects the participation and performance of Deaf/hard of hearing students in educational settings.

Evaluation Case Study: RTC in Secondary Schools

To explore the value of RTC for participating students and educators, an evaluation was commissioned by a governmental body, to evaluate a pilot program implemented in 2011-2012. The pilot program used a captioning infrastructure developed by a private captioning provider, to deliver captions of teacher talk in real time for Deaf/hard of hearing students in the classroom. The captioning infrastructure produces RTC by transmitting a teacher's speech to a remote respeaker who translates the spoken content into verbatim text using speech recognition software. This text can then be returned to the student in the classroom via iPad or laptop within 7 seconds. The evaluation was focussed on investigating the impact of the pilot program using this technology for Deaf/hard of hearing students in years 10 - 12 across 8 metropolitan and regional facilities. The aim of the evaluation of this pilot program was to explore the impacts of RTC on:

a. Comprehension and academic performance outcomes, such as language and literacy levels; and

b. Participatory outcomes such as inclusion, engagement in the classroom environment and school attendance.

The challenge for the evaluator

As alluded to earlier, working with populations like the Deaf/hard of hearing requires that particular attention is paid to ensuring their participation is meaningful, and that the evaluation results in some benefit for the community. Disengaged and vulnerable populations tend to be one of the hardest to engage in evaluation, but paradoxically they are often the groups that stand to benefit the most from evaluation findings and recommendations. Thus the imperative for developing inclusive evaluation methods is clear for these groups. Additionally, evaluators and researchers are ethically obligated to ensure participants do not experience harm by virtue of their engagement with the evaluation. Beattie (2001) speaks directly to this, noting the complexity of the interplay between ethics, deafness and education.

Several ethical challenges were present from the outset of this evaluation. These ranged from engaging and communicating with the multiple stakeholder groups within the pilot, developing appropriate evaluation procedures and protocols, and minimising participant burden and fatigue. To meet these challenges, we embedded ethical and professional standards of practice, which guided evaluation procedures, within the overarching evaluation framework and methodology. These standards are briefly described below, followed by an overview of the evaluation framework.

Ethical Standards and Professional Codes of Practice

The Belmont Report, Nuremberg Code and Declaration of Helsinki are considered to be the cornerstone of medical human research ethics. These principles are arguably equally applicable to evaluation practice within the education sector, particularly the principles of beneficence and non-maleficence. Therefore, these became a touchstone for ethical practice within the context of the evaluation. In addition, professional standards for ethical conduct of research and evaluation practice guided the development and implementation of the evaluation. These included: the *Australian Code for the Responsible Conduct of Research;* the *National Statement on Ethical Conduct in Human Research;* the *Guidelines for the Ethical Conduct of Evaluations,* developed by the Australasian Evaluation Society ([AES], 2013), which are designed to suit the cultural, social and institutional contexts of evaluation in Australia and New Zealand; the Program Evaluation Standards (2nd Edition, Sage 1994); and the *American Evaluation Association's Guiding Principles for Evaluators* (2004). Further, as the Deaf community, typically, consider themselves Culturally and Linguistically Diverse (CLD), it was necessary to adhere to guidelines for working with CLD populations as outlined in the *Standards for Educational Assessment and Psychological Testing* (American Evaluation Research Association [AERA], 1999) within the context of conducting psychological or educational assessments. The evaluation was also subject to review by the University of Melbourne Human Research Ethics Committee. The ways in which these standards and guidelines have been used to inform practice will be outlined in further detail below.

Finally, the evaluation approach was conceptualised from a social justice perspective. Donna Mertens (2009), a researcher and program evaluator who has pioneered the philosophy of *transformative evaluation,* posits that evaluation should transform outcomes, and it should, where possible, result in improvements in the evaluand and the participants. In the case study presented in this article, the conduct of the evaluation was focussed on enhancing the RTC program for the benefit of the participating Deaf/hard of hearing students, with an emphasis on allowing the participants to have a voice through the evaluation.

An Inclusive Evaluation Framework

To meet the ethical challenges posed by the context for this evaluation, we employed an evaluation framework which was participatory, collaborative, innovative, ethical, rigorous, and based on continuous feedback to stakeholders with a view to supporting program enhancement. The framework was underpinned by *The Framework for Program Evaluation of Public Health Initiatives,* developed by the Centers for Disease Control and Prevention (CDC&P, 1999). The model provides an overarching framework for the evaluation through the application of six steps ('stakeholder engagement'; 'program description'; 'focussing the evaluation design'; 'data collection'; 'justifying conclusions' and 'using and sharing lessons learned'). The framework is also clearly aligned to the aforementioned ethical and professional standards, to ensure the accuracy, reliability and validity of the evaluation process and findings. A mixed-methods research design was utilised to guide data collection and analysis. Such an approach combines qualitative and quantitative methods of inquiry, incorporating the strengths of both methods to better understand the research questions and strengthen the research design. In essence, a mixed-methods approach is; "… generative and open, seeking richer, deeper, better understanding of important facets of our infinitely complex social world". (Green, 2007, p. 20).

Finally, the evaluation framework was designed to be inclusive in nature. Briefly, an inclusive evaluation involves the systematic investigation of the merits or worth of a program to promote social change, with a focus on the involvement of all potential stakeholders of the project, with a particular focus on engaging those who have typically been under-represented (Mertens, 2009); in this instance, the Deaf/hard of hearing students in the pilot program. To support inclusive evaluation practice, evaluators must learn about the group that is under-represented in two ways: (1) by familiarising themselves with the literature, in this case research conducted with the Deaf/hard of hearing and the CLD within the context of education settings; and (2) by interacting with members of the community and wider stakeholders in a meaningful way (Mertens, 2009).

Methods and Principles

In the section that follows, we will outline the evaluation aims and methods and describe how we utilised the above evaluation framework to adhere to each of the principles of *respect for persons, justice, beneficence and non-maleficence* as outlined in the Belmont report (1979). In addition, reflections on the lessons learned through working with this population will be shared.

Summary of Evaluation Aims and Methods

The key aims of the evaluation were to explore the impacts of RTC on Deaf/hard of hearing students' comprehension, academic performance outcomes, inclusion, engagement and behaviour in the classroom environment. Further, we sought to examine unintended outcomes for key stakeholder groups, including teachers and parents. A brief overview of each method that was adopted within the evaluation is given below, as a comprehensive discussion of evaluation methodology or findings is beyond the scope of this article. The evaluation reference group was involved in shaping and providing feedback on the development of each of the data collection instruments used within the evaluation. This section will then be followed with a commentary on how the evaluation methods and process adopted relate to each of these evaluative principles, with a particular focus on how this population was given the opportunity to have a voice in the evaluation. Further, we will discuss the ways in which the evaluation team promoted the use of data and subsequent conclusions and recommendations to inform the pilot program and future programs in the field of captioning in the classroom to benefit Deaf/hard of hearing students.

Literature Review: To ensure the rigor of an inclusive evaluation it is necessary to develop familiarity with the Deaf/hard of hearing population. As part of this process, the evaluation team conducted a review of the literature, focused on the impact of captioning on learning and participatory outcomes, particularly for the Deaf/hard of hearing. This literature review served as a foundation that guided the development of the broader evaluation methodology.

Online Survey with Students and Teachers: An online survey was delivered to participating teachers and students to gather background information, and experiences with and perceptions of the pilot program. The surveys also included a number of psychological and psychosocial constructs (for students) and questions targeted towards gathering the perceptions of RTC, transcripts and their implementation in the classroom.

Computer Assisted Telephone Interviews (CATI): CATI were used to survey parents of students involved with the RTC pilot. The interview protocol was designed to gather information about the parent's linguistic background, their understanding and perceptions of RTC and its impact on their child, and their overall satisfaction with the pilot.

Interviews with Program Staff: Semi-structured interviews with program staff were conducted to elicit information about the effectiveness and efficiency of the development and implementation process for the RTC pilot from the perspective of various program stakeholders.

School Data Audit: An electronic data collection template, designed to elicit information from each participating student on their achievement, behaviour, and attitude over the course of their involvement with RTC was sent to each facility coordinator to complete for each student. The key evaluation contact at the school then populated this template with relevant documentary information for each of the students participating in the pilot.

Language and Literacy Assessments: These assessments were conducted in two waves. Initially, the Compass test (Australian Council for Educational Research [ACER], n.d.) was chosen to assess the language and literacy skills of the students. Compass is an online literacy and numeracy assessment (the numeracy component was not administered in this instance) specifically designed for adolescents and young adults who have had limited or disrupted exposure to formal education. The second wave of testing, in response to feedback from the reference group and key stakeholders, utilised the Woodcock Reading Mastery Test, a standardised assessment of language literacy, to gather normed data on participating students.

Analysis of Transcripts: The evaluation team was provided with a selection of classroom transcripts (that is, the printed version of the captions that appear onscreen during the lesson), which were then analysed for evidence of change in teacher practice or interactions within the classroom over the course of the term since the introduction of captioning. A rubric-based coding scheme was developed as informed by the literature and program stakeholders and was used to analyse the transcripts. The rubric had four main components: structure of the lesson; clarity of communication; clarity of communication for captioning; and technology and presence of captioning in the classroom.

Given the richness and scope of the multiple data sources, the triangulation of data from all these sources enriched the overall evaluation findings, allowed for multiple stakeholders to have a voice throughout the evaluation, and ultimately led to a greater level of confidence in the accuracy of the conclusions generated. The following sections provide a discussion on how the evaluation team upheld ethical principles throughout the implementation of these data collection procedures. Challenges and lessons learned in this process are also included.

Commentary on Adherence to Ethical Principles

As mentioned earlier in the article, the ethical principles that inform the majority of research as outlined by the Belmont report are discussed in this section, within the context of the evaluation case study.

Respect for persons

The principle of respect acknowledges individuals' and groups' autonomy and their right to make choices, to hold views and to take actions based on their values and beliefs (Belmont Report, 1979). Within the context of this evaluation case study, this meant ensuring that all those with a stake in the program and its evaluation had an opportunity to engage with the evaluation, and to share their views on the program. Considerable effort was made to ensure the reference group for the evaluation was inclusive. It thus comprised of stakeholders from the evaluation commissioner (Executive director, program manager, and technical advisors); participating schools (principals, coordinators, teachers of the Deaf/hard of hearing); and parents of students involved in the program. The evaluation team members who sat on this group contributed expertise and experience in the fields of evaluation, education and psychology and Deaf education. In assembling this group of stakeholders, we aimed to ensure our approach was appropriate for the population of interest and that we were cognisant of any particular concerns or issues for this population. As a consequence, we hoped to facilitate a greater level of engagement in the evaluation process from all stakeholders and evaluation participants.

At the initial stages of the evaluation, the reference group were involved in a program logic workshop where the key stakeholders came to a shared vision with regard to the desired outcomes for the program, ways in which implementation could support the achievement of these outcomes; and the appropriate measures for the evaluation with a view to inform the development and improvement for future roll-out of the program. Thereafter, regular reference group meetings were held approximately every two months, with meetings focused on the evaluation process and development of protocols and instruments, and the immediate dissemination of any emergent findings. Finally, the reference group was involved in a summative workshop in which we presented the draft evaluation findings and asked them to reflect on these within the context of their own experience. Based on this workshop, we revised the original program logic model, and also added some recommendations to the final evaluation report to ensure these were consistent with the lived experiences of those involved in the project.

The principle of informed consent poses a challenge when working with CLD populations such as Deaf/hard of hearing groups, as traditional evaluation practices and procedures rely on written or oral communication and expressions of consent. In this case, the consent process needed to include the communication of information about the evaluation in a mode that all participants could understand. Thus, the evaluators needed to consider not only the preferred language and communication style of the participants, but also their literacy level. There was significant variation in the degree of deafness and preferred mode of communication across participants, with some students requiring everything to be translated into Auslan and others being able to comprehend English relatively easily by lip-reading, or hearing with the assistance of their cochlear implant. Further, there were considerable differences in the literacy level of participants which necessitated that protocols such as plain language statements and consent forms were written at a reading age that was accessible for all participants. In addition to ensuring that written communications were commensurate with the literacy level of the participants, information about the evaluation was also provided in person with an Auslan translator present when requested by the participants. To adhere with ethical guidelines regarding research with minors, the parents/guardians of the participants also received written information about the evaluation and were required to give their consent for their child to participate.

Justice

The principle of justice focuses on the equitable treatment of research participants, as well as ensuring an equitable distribution of benefits and burdens. As evaluators, we have a mandate to consider effective ways to balance the burdens of engaging in the evaluation process with the potential benefits. Considerable effort was made to collaborate with the reference group to create protocols that collected valuable information without impacting on participants in a negative way. In addition, data collection procedures were developed with a view to ensuring that data was collected and analysed in a 'fair' way, such that the conclusions drawn from the data accurately reflected the outcomes and experiences of Deaf/hard of hearing participants involved in the program. Great care was taken to ensure that all participants, irrespective of their literacy level or communication preferences, were given equal opportunity to engage in a meaningful way with the evaluation.

Beneficence and non-maleficence

Fundamentally this evaluation was commissioned to determine the efficacy of RTC as an assistive technology to support the learning and achievement of Deaf/hard of hearing students. However, the notion of formative evaluation and continuous improvement was also central to the evaluation philosophy, which supports the principles of beneficence and non-maleficence. These principles assert that research and evaluation should seek to maximise benefits of the research project for the participants, whilst simultaneously minimising risks to the research subjects (Belmont Report, 1979). Deaf/hard of hearing populations are at greater risk of research-related harms, by virtue of their minority status, therefore it is critical that evaluators and researchers are mindful of this to ensure the evaluation design seeks to minimise harm and maximise benefit.

The main benefit of participation for the students in this evaluation was the opportunity to access a new technology which could support their engagement and inclusion in the classroom, and facilitate their learning more broadly. Further, through their participation in this evaluation, students also had the opportunity to contribute to the development and improvement of services for the broader Deaf/hard of hearing community. Evaluation participants trialled the effectiveness of a new technology designed to increase access to the spoken word, and ultimately to improve comprehension and learning, and, critically, provided their feedback on ways in which this technology and program could be improved; which information can subsequently be used to inform program development.

In relation to minimising potential harms resulting from the evaluation of the project, all data collected from students was anonymised. This data was not made available to teachers for review, despite some requests for this information, to preserve the authenticity of the consent process. Further, we triangulated several sources of achievement data (rather than just using one source of data) collected across time periods rather than relying solely on one observation point. The findings were also communicated with the necessary qualifying language to prevent misinterpretation or misuse of the results. Most importantly, the reference group was involved throughout to ensure that the evaluation was designed and implemented in a way that minimised harms to the participants.

To further illustrate adherence to these ethical principles, an example is provided below in the design and implementation of protocols to minimise participant fatigue in the language and literacy testing of participants.

Language and Literacy Testing Procedures

In the conceptualisation phase of the evaluation, it became evident that there is a dearth of appropriate assessment tools available to support accurate and meaningful assessment of the language and literacy skills of Deaf/hard of hearing students at a secondary school level. Given the reported delays in these skills for this population relative to their hearing peers, it was important to identify a tool that would fairly and accurately assess the student's literacy skills. After extensive research, and in consultation with the reference group, we chose to utilise an online assessment called Compass, which is produced by the Australian Council for Educational Research (ACER, n.d.).

Compass was selected for several reasons. First, it is delivered online, a space with which Deaf/hard of hearing students are potentially as conversant as their hearing peers. Second, the test contains engaging young adult stimulus material and lines of questioning for senior school students. Third, the test is not designed to be diagnostic in itself, and should be used in conjunction with other observation points, which was compatible with the mixed methodology employed in the overall program evaluation. Finally being a virtual, 'off-the-shelf' product, the instrument was cost-effective (ACER n.d.). The test was also confirmed as being suitable by a Deaf/hard of hearing education academic.

In a later stage of the evaluation, the evaluation commissioners requested that the students were assessed using a standardised measure of literacy. The Woodcock Reading Mastery Test (WRMT) was selected for this purpose, due to its robust psychometric properties, particularly test-retest reliability, as well as its ability to assess a variety of linguistic skills. As outlined by the *Standards for Education and Psychological Testing* (AERA, 1999), test administration should follow the standardised procedure for administration outlined by the test unless a situation arises, such as working with CLD populations, which dictates that an exception should be made, and any exceptions made must be documented (Standard 5.1 and 5.2) (American Educational Research Association [AERA] , 1999).

It is important to acknowledge the challenges of using norm-referenced tools such as the WRMT testing in CLD populations. Issues such as content bias, linguistic bias and disproportionate presentation in normative samples need to be addressed when using such tools to inform evaluative judgments (Laing & Kamhi, 2003). For instance, content bias occurs when a test assumes shared experiences, concepts or vocabulary between examinees. As the WRMT was normed using a United States sample, it was examined for possible content bias. While there were a small number of items in the *Reading Comprehension* selection that would be considered unfamiliar for Australian students, overall, the test was determined to be appropriate for use.

In addition to reviewing the test for content bias, the evaluation team consulted with Auslan interpreters to identify potential linguistic bias that may be present for native Auslan signers. As some linguistic issues were identified in the conduct of the test, such as the need to respond to questions with one word, which may be difficult for Auslan signers who may use more than one word to describe the meaning of a concept, it was necessary to create a protocol to ensure equal fairness of testing for Auslan signers, English speakers and bilingual students.

The evaluation needed to ensure that all examinees understood the test instructions, and were able to respond fairly. This language requirement presented the evaluators with an obstacle, as team members suitably qualified to administer the WRMT were not proficient in Auslan. It was essential to enlist the support of Auslan interpreters, who were provided for all students for whom oral communication was not their primary mode of communication. However, this scenario presented additional complications in relation to the testing environment which, as specified by the standards, should only include the administrator and participant. To minimise the impacts of the presence of the interpreter, all interpreters were made aware of the testing protocols and given clear instruction regarding their involvement.

Finally, standard procedures for the WRMT were amended to take the reading level of a year 6 student as a starting point. This was informed by research indicating the reading levels of Australian Deaf/hard of hearing school leavers are approximately at this level, by amending the starting point, the potential for floor effects was minimised (Walker & Richards, 1992).

Both tests, when amended and administered as described above, were found to be suitably engaging and sensitive in assessing the language and literacy skills of the students in this study. However, there remain obvious limitations in the use of such instruments in the absence of meaningful norms for the population of interest. The identification of appropriate assessment tools for this population remains an important area for the research agenda in Deaf/hard of hearing education.

Conclusions

This paper has described the ways in which an inclusive evaluation methodology, underpinned by ethical and professional standards and a philosophy of transformative evaluation, was utilised to facilitate engagement with a CLD population, namely Deaf/hard of hearing secondary students.

Engaging the Deaf/hard of hearing population effectively in evaluation and research is an important step in improving services and outcomes for this population, who are often marginalised and poorly served in the community and society as a whole (Beattie, 2001). It is through the continued development of new methodologies that are inclusive and underpinned by ethical and professional standards, that members of this population will be enabled and empowered to have a voice in the ongoing development of services, programs and policies of which they are beneficiaries. Through this process, policy makers and service providers will be better informed as to how best to meet the needs of this population (Mertens, 2009).

The evaluation was designed to better inform program developers, academics and educators about the challenges Deaf/hard of hearing students face in education, and the potential for an assistive technology, namely RTC, to help students navigate and better manage these challenges. In addition to illustrating their challenges, participants also shared what worked well for them and enabled them to engage in education. Overall, the evaluation process sought to ensure that the participants' engagement in the evaluation served to improve the RTC program, accurately capture the impact of program, and generate information about the challenges and enablers in education for Deaf/hard of hearing students in an Australian context. To do so, the evaluators adopted Merten's (2009) view of evaluation, ensuring that the evaluation itself led to better outcomes for vulnerable and marginalised populations which in this case were Deaf/hard of hearing students in selected secondary schools.

As a consequence of this evaluation, the evaluation commissioners have a strong foundation on which to base further research into the use of RTC. They also have evidence from which to develop an improved model for the provision of assistive technology services in education to maximise access to the curriculum for students who are Deaf/hard of hearing.

References

Australian Council for Educational Research (ACER). (n.d). *Compass Literacy and Numeracy Test.* Retrieved from http://www.acer.edu.au/tests/compass/compass-overview1.

American Evaluation Association. (2004). *Guiding Principles for Evaluators.* Retrieved from http://www.eval.org/p/cm/ld/fid=51.

American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational Assessment and Psychological Testing.* American Educational Research Association: Washington: AERA.

Australasian Evaluation Society (AES), (2013). *Guidelines for the Ethical Conduct of Evaluations.* Retrieved from http://www.aes.asn.au/images/stories/files/membership/AES_Guidelines_web.pdf.

Beattie, R.G. (2001). *Ethics in deaf education: The first six years.* San Diego: Academic Press.

Brett, M. (2010). Challenges in Managing Disability in Higher Education, Illustrated by Support Strategies for Deaf and Hard of Hearing Students. *The Open Rehabilitation Journal,* 3, 4-8.

Boyd, J., & Vader, E. (1972). Captioned Television for the Deaf, *American Annals of the Deaf , 117,* 34-37.

Centres for Disease Control and Prevention. (1999, September 17). Framework for Program Evaluation in Public Health. *Morbidity and Mortality Weekly Report,* *48*(RR11), 1-40.

Daelemans, W., Höthker, A. & Sang, E.T.K. (2004). Automatic sentence simplification for subtitling in Dutch and English, Retrieved from http://www.clips.ua.ac.be/~walter/papers/2004/dhs04.pdf

Green, J. (2007). *Mixed Methods in Social Inquiry.* San Francisco: John Wiley & Sons.

King, C.M., & Quigley, S.P. (1985), *Reading and Deafness.* San Diego: College-Hill Press.

Laing, S.P., & Kamhi, A. (2003), Alternative Assessment of Language and Literacy in Culturally and Linguistically Diverse Populations, *Language, Speech & Hearing Services in Schools, 34,* (1), 38-41.

Markham, P. (1989). The Effects of Captioned Videotapes on the Listening Comprehension of Beginning, Intermediate and Advanced ESL Students. *Educational Technology, 29*(10), 38-41.

Murphy-Berman, V., & Jorgensen, J. (1980). Evaluation of a Multilevel Approach to Captioning Television for Hearing-Impaired Children. *American Annals of the Deaf, 125,* 1072-1081.

Mertens, D. (2009). *Transformative research and evaluation.* New York: Guilford Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research. (1979), *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Sanders, J. (1994). *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs.* The Joint Committee on Standards for Educational Evaluation. Thousand Oaks, CA: SAGE Publications.

Stinson, M., Stinson, R., Henderson, J., & Miller, L. (1988). Perceptions of Hearing-Impaired College Students Toward Real-Time Speech to Print: RTGD and Other Educational Support Services. *The Volta Review, 90,* 336-338.

Walker, L., & Rickards, L. (1992). Reading Comprehension Levels of Profoundly, Prelingually Deaf Students in Victoria. *The Australian Teacher of the Deaf, 32,* 32-47.

Youdelman, K., & Messerly, C, (1996).Computer-Assisted Notetaking for Mainstreamed Hearing-Impaired Students. *The Volta Review, 98*(4), 191

Informed consent in evaluation: informed of what, exactly?

Emma Williams

Northern Institute
Charles Darwin University

emma.williams@cdu.edu.au

Keywords: informed consent, evaluation ethics, evaluation informed consent, front end review, back end risk, layered power

Abstract

Australia's 2007 National Statement on Ethical Conduct in Human Research includes a requirement for informed consent from research participants. This paper discusses how the dynamics of bio-medical research, which underlie the national statement and processes aligned to it, differ from the dynamics often encountered by researchers conducting program evaluations with particular reference to two areas: the multiple layers of power that interact in a typical program evaluation, and the risks experienced by evaluation stakeholders in the late stages of the process. Implications for informed consent in evaluations are outlined and future steps proposed.

Introduction

This paper examines areas in which the ethics of evaluation differ from the ethics of other types of research, focusing particularly on 'informed consent'. The paper demonstrates how the context of evaluation impacts on calculations of risk and benefit, and why 'informed consent' may need to be re-conceptualised for evaluators.

The emphasis here is not on personal ethics, although they are acknowledged as vital for ethical practice. Rather, the focus is on how the dynamics of evaluative research present challenges to the way that 'informed consent' is regarded in ethics review committees' decision-making aligned to Australia's National Health and Medical Research Council's National Statement on Ethical Conduct in Human Research (NHMRC, 2014e). The National Health and Medical Research Council (NHMRC) procedures and forms were developed originally to assess biomedical risks and later extended to social science research, with a recent document focusing on quality assurance and evaluation activities (NHMRC, 2014b).

Although the extension of medical ethics procedures to research in other fields was driven first in America (Ticheloven & Willems, 2013), it has since occurred in many countries around the world. Social scientists in a number of countries including Australia, Canada and England have raised the issue of disparities between medical research approaches and those used in other fields, noting that the risks and benefits of their research were not sufficiently well recognised by committee deliberations aligned to bio-medical research models (e.g. Israel, 2004: Social Sciences & Humanities Research Ethics Special Working Committee, 2004; Calvey, 2008; Hammersley, 2010; Academy of Social Sciences, 2013).

Although consultation continues to occur with a range of social scientists as well as other researchers (e.g. NHMRC, 2014a) and progress has been made in addressing some concerns (e.g. Zimmerman, 2013), less attention has been paid in the past to the ethics of evaluation (Berends, 2007; Chesterton, 2003).

A recent initiative to address issues associated with ethical conduct of evaluations was the NHMRC's *Ethical Considerations in Quality Assurance and Evaluation Activities* released in March 2014. It noted that, 'Human Research Ethics Committee (HREC) review processes are often not the optimal pathway for review of these [evaluation and quality assurance] activities' (NHMRC, 2014a, p. 1). Although identifying higher risk criteria such as the use of placebos and control groups, the document focuses on quality assurance and evaluations relying solely or in large part on the analysis of existing data rather than requiring research with human subjects/participants.

While some evaluations are undertaken without undertaking research activities with human participants, this paper focuses on those that do. It identifies ethical issues which distinguish research in the context of evaluation of social programs and interventions from many other forms of research involving human participants, with special emphasis on what is required for informed consent.

Background

While ethical decision-making has been discussed for centuries, many of the current guidelines and regulatory mechanisms encountered by evaluators and other researchers took shape after the Second World War. Revulsion at Nazi-auspiced medical abuse led to the Nuremberg Code in 1947, followed by the Declaration of Helsinki on ethics for human experimentation issued in 1964, and since updated a number of times. Intended to be an international guideline rather than a legally binding document, not every country has agreed to all aspects of the most recent version of the Declaration (World Medical Association, 2013). However, many countries have issued their own codes and guidelines that include mechanisms for local enforcement.

In some cases, these were driven by medical and/or social science research scandals. Making research funding contingent on the approval of independent ethics review committees, in the form of Institutional Review Boards, was initiated in America in the 1970s in part due to the Tuskegee syphilis research project (Heintzelman, 2003). The

Tuskegee study used approximately 400 African American men as a control group and did not inform participants of remedies that were increasingly identified as effective for their illness over the course of the study, resulting in unnecessary deaths.

Social science research projects in the 1960s and early 1970s also raised serious ethical issues (Ticheloven & Willems, 2013). Two American studies often cited as bringing about greater ethical oversight were the Milgram experiment (Milgram, 1963) and the Humphreys (1970) 'tearoom trade' project. The Milgram experiment deceived participants into believing they were delivering painful electric shocks to other participants. In Humphreys' 'tearoom trade' project, he copied the licence plate numbers of men frequenting public toilets for sexual encounters and then tracked some to their homes, later interviewing them as a social health surveyor on topics such as their marital and occupational status.

Subsequent calls for such research to be reviewed by independent committees to ensure ethical practice led to the US *National Research Act* of 1974 (National Commission for the Protection of Human Subjects of Biomedical & Behavioral Research, 1979) and the development of institutional review boards for non-medical research. The review boards were empowered to require researchers to submit research designs for approval before commencing. This "front end" process has since intensified in the degree of scrutiny mandated for such review, and has also been extended to a broader range of research fields. (See Hammersley, 2010 for a somewhat vituperative discussion of this phenomenon.)

Australia has followed these precedents, although the development of Australian ethical regulatory mechanisms appears less scandal driven.

> Before…1985, ethical issues in social and behavioural research were recognised by Australian associations in sociology, psychology and anthropology, who offered guidance to their members on the ethical conduct of research… NHMRC extended the jurisdiction of IECs [Institutional Ethics Committees, the precursors to Human Research Ethics Committees] to include non-medical projects in 1986… [reflecting] the shifting ground of health care delivery... As research institutions began to implement the requirements for… behavioural health research, many extended the requirements of the NHMRC to other types of social and behavioural research. In addition, funding bodies began to require approval by an IEC as a condition for considering social science grant applications in areas related to health or health care (NHMRC, 2014c).

Further development has occurred in the decades since, although there is not enough space here to address it in detail. Working together, the Australian Health Ethics Committee, the Australian Research Council and the Australian Vice-Chancellors Committee conducted consultations with institutions and researchers to produce the 2007 *National Statement on Ethical Conduct in Human Research,* with an updated version released in 2014 (NHMRC, 2014e).

Consultation and further development is continuing (NHMRC, 2014a), with a new distinction proposed between activities that are relatively 'low risk' and could benefit from a less onerous process than that required by higher risk research, similar to the 'light touch review' noted in the UK Research Ethics Framework (Economic & Social Research Council, 2012). Evaluation has been identified by NHMRC as an area of relatively low risk compared to research.

> … QA [Quality Assurance] and evaluation commonly involve minimal risk, burden or inconvenience to participants, and, while some level of oversight is necessary, Human Research Ethics Committee (HREC) review processes are often not the optimal pathway for review of these activities. (NHMRC, 2014b, p. 2)

The relationship between 'evaluation' and 'research' has been explored from multiple perspectives in the literature, (e.g. Levin-Rozalis, 2003; Fain, 2005). Rogers (2014) sets out four ways in which the relationship has been conceived:

- Research and evaluation may be seen as dichotomous, and the differences between them emphasised.

- Evaluation and research may be seen as mutually exclusive domains with an area of potential overlap where systematic data collection and analysis are the basis of evaluative judgments.

- Research may be seen as a subset of evaluation in that evaluations involve research, but also other types of activity.

- Evaluation may be seen as a subset of research as it is only one of many ways of investigating questions and reporting findings.

Rogers notes that these perspectives need not be mutually exclusive, and the choice of which perspective to employ will depend upon context.

Although *Ethical Considerations in Quality Assurance and Evaluation Activities* (NHMRC, 2014b) notes that Quality Assurance, evaluation and research can be seen as "a continuum of activity" (NHMRC, 2014b, p. 2), overall the document appears to treat evaluation and research as largely mutually exclusive domains, with a limited degree of overlap. As noted above, this paper addresses evaluations which fall into that area of overlap.

After an overview of requirements for informed consent, the paper looks at how the social science research is impacted by an ethical review process based on a biomedical research model. A section on the special dynamics of evaluation leads to one specifically on informed consent in evaluation, what is required for it and the degree to which it is addressed in current ethical guidelines.

Informed consent

Informed consent lies at the heart of research ethics. The 1947 *Nuremberg Code* begins:

> The voluntary consent of the human subject is absolutely essential… the person involved should have… sufficient knowledge and comprehension of the elements

> of the subject matter as to enable him to make an understanding and enlightened decision. (Carlson, Boyd & Webb, 2004, Appendix 1, paragraph 2)

Multiple versions of the Helsinki Declaration state that an individual's right to decide whether or not to participate takes precedence over the benefit of increased scientific knowledge. This precedence is not absolute; all ethical decision-making requires potential benefits to be balanced against potential risk, but what Sleat (2013, p. 15) calls the 'participant protection model… is at the heart of the ethical regulation of the biomedical sciences and…has often either influenced or been directly imported as the model for thinking about similar regulation of the social sciences'[1].

Informed consent is not an issue reserved for research; Skene and Smallwood (2002) note that medical practitioners are now expected to provide patients with an increasing range of information, including more details on outcomes such as possible side effects of treatment.

> Courts in Australia and England have begun applying a tougher standard to the information that doctors should give their patients… recent judgments in both English and Australian courts suggest that judges are moving away from accepting what "reasonable doctors" might do to supporting what "reasonable patients" might expect. (Skene & Smallwood, 2002, p. 39)

While medical practitioners might once have been reluctant to share information that might distress patients and therefore be regarded as causing them harm, the trend now is towards providing all available information, to enable patients to better understand their options.

In research, what is required for fully informed consent is that 'there must be full disclosure of the purpose to which the research will be put, the nature of the information sought from the participant, and the motivations of the researcher in seeking this particular information' (Sleat, 2013, p. 16). The Australian National Ethics Application Form (NHMRC, 2014d) asks: if consent will be sought from all participants; if they have capacity to give consent, and what mechanisms, assessments and tools will be used to determine capacity to give informed consent; how participants will be informed about the project and choose whether or not to participate; and whether there are consequences to non-participation or later withdrawal. Answering such points may seem simple, but Wiles, Crow, Charles and Heath (2007) note:

> While at first glance informed consent appears a relatively straightforward issue involving the provision of appropriate information to enable people to make informed decisions about participation in a research project, a closer examination of the issues involved reveals that the process is far from straightforward… (Section 1.4)

Issues identified in Wiles et al. (2005, 2007) include: a lack of consensus about what comprises 'informed consent'; whether the concept is or should be the same across different fields of research and methodological frameworks; the tension between providing adequate information and overloading potential participants with so much information they cannot absorb it; and the factors that affect participants' desire to engage with information (including format of presentation, language and literacy factors but also interest in the topic and a consequent reluctance to fully engage with discussion of risk factors). One of the most difficult issues is deciding how much information to provide on possible research outcomes, especially as these may not be known until the research has progressed substantially. Heintzelman, discussing a bio-medical project, notes:

> The first major ethical issue to be considered is informed consent, which refers to telling potential research participants about all [emphasis added] aspects of the research that might reasonably influence their decision to participate. A major unresolved concern is exactly how far researchers' obligations extend to research subjects. (Heintzelman, 2003, para. 6)

A recent book on ethics in evaluation also emphasises the value of voluntary consent. One of the most famous ethical formulations is 'first, do no harm', often popularly although erroneously attributed to Hippocrates (Sokol, 2013). However, as Morris notes in writing about evaluation ethics, the principle of non-maleficence or 'do no harm' is placed second in priority in a hierarchy of five ethical principles for helping professions (Morris, 2008). The first principle was 'respect for autonomy', which underlies informed consent[2]. With 'respect for autonomy', Morris notes, 'an evaluator's desire to collect information… does not necessarily [sic] override someone else's right not to provide some or all of that information, depending on the circumstances'.

Informed consent is a required element in the institutional review of research ethics in Australia, conducted by Human Research Ethic Committees aligned to the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2014e). The current model used at many Australian research institutions requires researchers to fill out forms, with an on-line version available (NHMRC, 2014d), which asks them to 'Describe the consent process, ie how participants or those deciding for them will be informed about, and choose whether or not to participate in, the project' (2014d, 6.6.1.1.3).

The informed consent question – including procedures for obtaining (and documenting) informed consent – is embedded in a longer form, which also requires researchers to provide information on:

- the research project title and description;
- who will undertake the research, their qualifications and training
- funding/support for the project;

[1] There is no space here to discuss how ethical theories such as principlism, consequentialism and utilitarianism, deontology and virtue ethics, and the ethics of care have impacted on such developments. However, see Carpenter, 2013; Simons, 2006; Treasury Board of Canada Secretariat, 2006, for examples.

[2] The final three, in descending order of priority were: 'beneficence'; 'justice', which included both procedural and distributive fairness; and 'fidelity', i.e., keeping commitments to people.

- any previous reviews;
- the type of research proposed, research plan, risks and benefits identified, andhow monitoring will be done;
- a description of participants, how they will be recruited and their consent requested, followed by a section on participants with specific vulnerabilities
- confidentiality and privacy issues; and
- in a final section, more detailed questions for certain types of projects — most medical, but also including workplace research, overseas research, and research involving Aboriginal and Torres Strait Islander participants.

The answers provided by the researcher(s) go to a committee where they are reviewed and comments made. The time required for this can be substantial, particularly in institutions where committees meet at relatively infrequent intervals. Typically no member of the research team is present at the review; reviewers' comments are provided in writing, and the researcher(s) have to respond to identified concerns before the project can advance.

Ethics review and social science research

Where evaluation is seen as a subset of research, it is typically viewed as a subset of social science research; evaluations of social programs and interventions commonly involve social science research methods. As well as analysing existing program/ service data, evaluators often collect new data through a range of social science research methods tailored to the research being undertaken (e.g. surveys, focus groups, interviews, projective techniques, observational methods including participant observation techniques, etc.). The appropriate technique often depends upon the topic; for example, criminological evaluations tend to use the techniques of criminological research.

The relationship between social science and bio-medical research[3] has been widely discussed, including the effect on social science research of institutional review mechanisms originating from bio-medical models. Van den Hoonard (2013) notes, for example that since 2000, an average of ten publications a year have discussed the difference between biomedical and sociological paradigms. He states that 'the biomedical paradigm offers nothing that might be even remotely helpful to sociologists in their search for ethics in research' (2013, p. 23).

Van den Hoonard's views are more extreme than many others. Some note improvements have occurred within the past decade. In 2004, a working committee of those working in the humanities and social sciences in Canada stated:

> If there is a fundamental problem we can identify, it is that the granting agencies' desire to create a regulatory structure to deal with the stereotypical clinical trial

> has resulted in a document and set of structures that assume different modes of research involving different relationships and different concerns than most social science and humanities researchers seek and encounter. (Social Sciences & Humanities Research Ethics Special Working Committee, 2004, p. 10)

Informed consent in particular was noted as an issue:

> Informed consent is a universally important component of respect for the autonomy of research participants, but the approach to consent in the present TCPS [Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans] is narrowly conceived and does not fit many modes of inquiry in the social sciences and humanities. SSHWC [the Social Sciences and Humanities Research Ethics Special Working Committee] recommends the idea of consent (and default expectations about the way it is obtained) be considered further, with a view to making the TCPS better include and reflect the diversity of ethical relationships between researchers and participants. (Social Sciences & Humanities Research Ethics Special Working Committee, 2004, p. 6)

However, by 2013, a Canadian representative was able to assure social scientists that the 2010 version of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2) incorporated changes based on social scientists' critiques and was now more responsive to research approaches in multiple disciplines (Zimmerman, 2013).

Other social scientists have pointed out the positive benefits of ethics review, with Hunter (2013) stating that 'ethical review forms are one of the few [administrative] processes that I actually find helpful, that provoke me and my students to be better researchers… (2013, p. 19). Nevertheless, Hunter finishes her sentence by noting that the review board may obstruct certain types of research, a theme echoed by other researchers (e.g. Wynn, White, Thomson & Israel 2013). Concerns raised by social scientists to ethics review by institution review boards also include:

[3] The National Health and Medical Research Council, as attested by its name, focuses primarily on bio-medical research risks and benefits, although many other forms of research are subject to its forms and procedures for approval. A recent document (NHMRC, 2014a) sets out five major categories of research:

- research infrastructure, i.e., establishing a biobank or database;
- health research;
- clinical research with four sub-categories;
- laboratory/basic science research including most genetic research; and
- one category that lumps together all other forms of research, listed in the document as 'Arts, Social sciences, Humanities, Business, Education, Law, Engineering and Computing' research (NHMRC, 2014b, p. 8).

Regulated mandatory review is not needed and may itself be unethical

Some see ethics review boards as being more about minimising institutional risk than ensuring ethical research, and want to see discussion of ethics separated from practical concerns about 'risk' (Emmerich, 2013, p. 12) and a 'more discursive encounter between researcher and reviewer/committee' (Emmerich, 2013, p. 13). Others object to the very notion of ethical regulation:

> My view is that any form of ethical regulation *in this context* [author emphasis retained, speaking of social science research as the context] is itself unethical, because it damages the quality of research and infringes the legitimate autonomy of researchers, without there being any counterbalancing need for it (the situation is different in some areas of medicine). (Hammersley, 2013, p. 6)

'Front end' reviews do not work well for much social science research

As noted above, ethical systems with bio-medical origins, such as the current NHMRC ethical review process, are 'front loaded', i.e., approval is given at an early stage of the project and typically required before the project can formally begin. Research questions, forms and procedures are expected to be developed before the research commences. As noted regularly in the literature, much social science research is dynamic and contingent, with the research model evolving over the course of the study (e.g. Calvey, 2008). Participatory and empowerment methodologies particularly struggle with 'front end' ethical review, as in these approaches the research design and even evaluation questions emerge from the research, rather than being pre-determined. (See Simons, 2006, for more on this issue.)

'Risk' and 'power' in social science contexts differ from bio-medical contexts

There are substantial differences in risk and power between medical and social science research contexts. A clinical study, for example, typically involves people who have come to a medical practitioner seeking help because the practitioner has expert professional knowledge they do not have, and medical treatments such as new drugs being tested can pose genuine physical risks. In such situations, with such vulnerable participants, asymmetry of knowledge/power, and risks of worsening health or even death, strict guidelines are essential.

However, in social science contexts, 'in the vast majority of studies the potential for causing physical harm to the participants is clearly lower than in biomedical sciences' (Sleat, 2013, p. 16). Social science risks are more commonly ones of inconvenience with potential in some cases for participant stress or offence, and in particularly severe cases harm to reputation. (These are, of course, still worthy of concern.)

Similarly, there is generally much less asymmetry in power – it is relatively common for social scientists, in fact, to seek to research those with greater power than the researchers, if not as individuals, then because of their role, such as holding public office (Boyd, 2013). Sleat similarly notes that in the social science research context ‘it is the researcher who is often in the weaker position and the participant in a position to potentially harm’ (2013, p. 16).

More discipline-specific committees are required

Some authors point out the importance of discipline-specific committees, noting that committee members may struggle with calculating the relative risks and benefits of research in areas unfamiliar to them.

Criminological research provides a good example, as it poses special challenges such as dealing with illegal activities; real expertise is required to accurately understand how risks can be managed and benefits realised. With regard to criminological research techniques which have been well established over decades, Israel (2004) notes that some Australian ethics review committees ‘seemed to overestimate both the magnitude and the probability of risks’ (p. 65), with some seeking to reject techniques well-accepted by research committees overseas. Israel notes that criminology-specific committees perform better in this area.

The special dynamics of evaluation

Evaluators conducting program evaluation requiring research contact with human participants share many of the issues of social science researchers.

The ‘front end’ ethics review process, where research questions are expected to be formulated before going into field, does not align well with the frequent use by evaluators of program logic, program theory and other processes where research questions are formulated and refined through engagement with participants and stakeholders. Many other evaluators use participatory and empowering approaches which, as noted above, are not well suited to a ‘front end’ approach. Some types of evaluation approaches such as ‘developmental evaluation’ (Patton, 2010), designed to be applied in contexts where programs are constantly evolving in response to dynamic contexts, face special challenges in complying with a ‘front end’ approach.

However, these are relatively minor issues compared to other aspects of evaluative research. Rogers (2014) notes in her “dichotomous” view of evaluation and research that research ‘is seen as… more controlled by the researchers – evaluation is seen as… more controlled by those funding or commissioning the evaluation’ (Rogers, 2014, section 1, para. 1). Similarly, Michael Quinn Patton notes that typically in research, questions ‘originate with scholars’ and the research ‘quality and importance [are] judged by peer review’, while in evaluation ‘questions originate with key stakeholders and primary intended users’ with ‘quality and importance judged by those who will use the findings to take action and make decisions’ (2014, section 5).

This means, for example, that the researcher who fills out the forms for the Human Research Ethics Committee may well be constrained in research design and timing by the dictates of the evaluation 'commissioner'[4], who may or may not have much research or evaluation experience and expertise. The ethical dilemmas this places on evaluators have been cited quite widely in the evaluation literature, including Scougall, 2006; Roorda and Peace, 2009; Williams, Guenther and Arnott, 2011.

Another issue is that often the time allowed for the project to go into field is truncated relative to other types of research (Scougall, 2006; Roorda & Peace, 2009, Guenther, Williams & Arnott, 2010), so there is often pressure to go through the ethics process quickly or avoid it altogether. This has sometimes resulted in academic institutions withdrawing from responding to opportunities to undertake evaluations, because the time given to undertake the evaluation and submit the final report is insufficient to go through an institutional board ethical review process, or because the time frame for the research appears in itself unethical (Williams et al. 2011).

Part of the dilemma is that while evaluators can be considered to have a degree of power with regard to evaluands (i.e., evaluation subjects, participants), in many cases evaluation commissioners have a degree of power with regard to evaluators. As Simons (2006) notes:

> Evaluation involves at least four levels of social-political interaction – with government and other agency policy makers who commission evaluation; with participants in the programmes, policies and institutions evaluated; with the evaluation profession; and with the wider audiences to whom evaluators in a democratic society have a responsibility to report. Evaluation has to operate in this multilayered context of different interests... it is not surprising that ethical dilemmas arise... (p. 213)

Morris (2008) cites a useful distinction between 'ethical dilemmas' and 'mixed dilemmas'. For Morris, 'ethical dilemmas' require an evaluator to balance conflicting principles, such as respecting individual confidentiality versus providing information for public benefit. It can often be difficult to decide in such circumstances what course of action is most ethical. 'Mixed dilemmas', on the other hand, require an evaluator to maintain an ethical principle in the face of external pressure to abandon it. The example given by Morris is an evaluator being pressured by a stakeholder to write up more positive results for a program that are warranted by the evidence.

> It is usually clear to the individual in a mixed dilemma what it is, from an ethical perspective, he or she should do. The problem is that the ethical course of action is often a risky course of action for that individual. The evaluator who refuses to bend in response to stakeholder pressure for positive results might find him- or herself the target of a subtle (or not so subtle) smear campaign waged by the aggrieved stakeholder… (Morris 2008, Chapter 1, third section)

Again, the evaluation literature contains many examples of such mixed dilemmas (e.g. Guenther, Williams & Arnott, 2010; Markiewicz, 2008, 2010). A survey of Australian Evaluation Society members with 132 respondents (Turner, 2003) found that when members identified ethical challenges and dilemmas, they were very likely to be 'mixed dilemmas'. Those cited by Turner included:

- Managers or funders trying to influence or control evaluation findings, sometimes including pressure on evaluators for positive results (cited repeatedly), sometimes including pressure to provide "dirt" on a program
- Conflicts between the organisation's needs and those of the client (when working as an internal evaluator)
- Political interference
- Dissemination or suppression of reports
- Requests to use information gathered for one purpose (e.g. program improvement) for a different purpose (e.g. accountability)
- Unilateral changes to terms of reference midstream or at time of reporting an evaluation and dealing with the implications for quality and relevance of data collected. (2003, para. 5)

As Markiewicz (2008) notes, pressure may be applied to make more positive reports, or alternatively to make a negative report on a program even where the evaluator's evidence indicates that the program is achieving successful outcomes. The pressure applied in these situations can be quite intense, with real implications for the evaluator's future work and income.

Moreover, the evaluator is not the only one to feel the impact of these decisions. Findings from an evaluation can lead to programs being renewed, reshaped, expanded or terminated. Particularly in tight economic environments, evaluation utilisation can result in the loss of jobs and the withdrawal of community services, with both staff and program clients impacted.

Further, evaluators can be pressured to not reveal findings to participants, e.g. Williams 2011. In 2010, a workshop held by the Australasian Evaluation Society (Markiewicz, 2010) in New Zealand elicited many cases from those attending of submitting positive reports on programs and services, but being forbidden to share them with the evaluation participants. The evaluators spoke of the participants' anger when their program or service was terminated, assuming the evaluator must have submitted a negative report. Workshop participants noted the degree to which this poisoned the relationship between evaluators and participants, damaged trust, and created a toxic environment for future research work. (This type of impact, i.e., spoiling the potential

[4] A term often used in Australia to denote the person or agency contracting/commissioning the evaluation.

for future research and lessening community trust in researchers, was one of the issues that led to the current system of institutional ethics review after projects such as Humphreys' (1970) and Milgram's (1963) research, as discussed above.)

Importantly, these risks tend to occur at the 'back end' of the evaluation research process. Forms and procedures developed from a bio-medical model of research focus on risks incurred during the data collection stage, and consider issues such as potential participant burden during that stage; even social science research echoes this pattern. However, in many evaluations, the most important risks often come for the evaluator at the reporting stage of the evaluation, and for participants and community members at the utilisation stage.

Not all ethics review committee members may be familiar with the special dynamics of much contracted program evaluation, hence Berends' (2007) proposed training and initiatives in this area. Reviewers working within a 'front end' system based on bio-medical research models could struggle particularly in two areas: recognising the risks that may be incurred after data collection and analysis has been completed, and understanding the special role of the evaluation commissioner. As Chesterton notes:

> To what extent could or should the commissioner of an evaluation be able to control the nature and focus of an evaluation? In one sense, the answer to this latter question is quite clear — the purchaser decides what he or she will purchase and spends accordingly. When this is put into a broader context of consequences, duties, obligations, rights, justice, and care, involving a range of stakeholders as well as the commissioner and the evaluator, and the use of public money, the answer is not so clear. (2003, p. 57)

Given these dynamics, what are the implications for informed consent in evaluation?

Informed consent in evaluation

Calvey notes that an issue faced by social science researchers is 'what I refer to as the 'consent to what' problem, in that social research is often contingent and all probabilities cannot be covered by the consent form' (Calvey, 2008, p. 907). This problem is particularly acute in evaluation. What exactly is the 'what' that evaluation participants need to understand before they can provide informed consent?

For example, as noted by Patton (2014) above, one of the differences between 'evaluation' and 'research' (when taking a dichotomous view of their relationship) is that evaluation informs decisions. If the decision-making stage is considered an integral part of the evaluation, should human research ethics committees require it to be discussed with participants in securing their informed consent? Should informed consent extend to information about potential participant risks related to the utilisation stage – taking into account that this stage is not under the control of the researcher/evaluator and utilisation decisions typically cannot be predicted too far in advance?

As many risks of become evident only as the evaluation progresses, Simons (2006) has proposed “rolling consent”:

> Gaining informed consent is an ethical imperative in any research study and signing an informed consent form is a familiar formal procedure in many contexts. Yet for many [in evaluation contexts] this is not consent at all, as it is never possible to know what will transpire as the programme unfolds in a precise socio-political context. Whatever forms are signed, “free and fully informed consent” needs to be realized through the additional different procedure of “rolling consent” - renegotiating consent with each person and/or site once a greater awareness of the context and structure of the study is known… (p. 26)

How does this notion align with current institutional ethics review procedures? Also, a critical aspect of voluntary consent is the capacity of the participant to withdraw from the research. Given the risk to participants may only become apparent at the end stage of evaluations, at the submission and utilisation stages, what are the realistic options for participants to withdraw at that point? How should evaluators deal with identifying and taking out perspectives which may by that time have permeated much of the report document? While it is certainly possible to take out specific sentences, it would be disingenuous to claim that all of the knowledge gained could be forgotten.

Finally, how do evaluators practise the ethical principle of ‘fidelity’ to their commitments to vulnerable stakeholders in a context where external parties wield such power, particularly in the later stages of the evaluation, such as submission (typically involving final payments to the evaluators) and reporting?

At the very least, it seems that the information required to ensure fully informed consent by evaluation participants would include reference to contextual factors, including identifying the evaluation commissioner and the evaluation questions, the potential outcome of the evaluation and particularly the likelihood of its utilisation by others for decision-making. This information would have to be updated over the course of the evaluation.

Professional evaluator guidelines

Such issues are not always – or even often – addressed in evaluators’ professional codes and guidelines. In spite of some early aspirations (Keith, 2003), no international code of evaluation ethics has yet been developed, although a burgeoning of national and regional evaluation societies, associations and networks around the world led in 2012 to the development of the International Organization for Cooperation in Evaluation (Kosheleva & Segone, 2013). Most national or regional guidelines focus on other aspects of professional practice and refer only obliquely to participant consent. Three sets of professional guidelines, however, are relevant, those of the United Kingdom Evaluation Society, the American Evaluation Association and the Australasian Evaluation Society guidelines.

The United Kingdom Evaluation Society's *Guidelines for good practice in evaluation* address many aspects of participant consent in considerable detail, although the term is not mentioned explicitly. The guidelines require that participants get an explanation of the purpose and methods of the evaluation, how data will be checked, stored and disseminated, including the right to see the evaluation agreement before the evaluators access the programme. The guidelines also require that participants have opportunities to discuss and question issues with the evaluators, with independent arbitration in cases of disputes. However, these guidelines are aspirational suggestions on the United Kingdom Evaluation Society (UKES) website. Members are not required to be bound by them, and there are no sanctions within the UKES for not following them.

The American Evaluation Association (AEA) guidelines refer specifically to informed consent.

> Evaluators should abide by current professional ethics, standards, and regulations regarding risks, harms, and burdens that might befall those participating in the evaluation; regarding informed consent for participation in evaluation; and regarding informing participants and clients about the scope and limits of confidentiality. (AEA, 2004, p. D2)

Other aspects of the American guidelines provide more guidance in aspects of informed consent relevant to evaluation, noting that evaluators should seek a 'comprehensive understanding of the important contextual elements of the evaluation [including] timing, political and social climate, economic conditions… (AEA, 2004, p. D1). The relationship to the evaluation commissioner is also noted:

> … Evaluators necessarily have a special relationship with the client who funds or requests the evaluation… that relationship can also place evaluators in difficult dilemmas when … client interests conflict with the obligation of evaluators for systematic inquiry, competence, integrity, and respect for people. In these cases, evaluators should… determine whether continued work on the evaluation is advisable... (AEA, 2004, p. E4)

As with the UK guidelines, there are no penalties proposed for those who do not follow them; the American Guiding Principles are meant to guide rather than constrain members' activities. In contrast, the Australasian Evaluation Society Inc. (AES) notes on its website that:

> In deciding to become a member of the AES you are committing to two things:
>
> - to abide by the AES Code of Ethical Conduct, and
> - to support the AES Guidelines for the Ethical Conduct of Evaluations. (downloaded July, 2014)

Both documents were recently updated, and two points from the updated *Guidelines for the Ethical Conduct of Evaluations* (2013b) relate to informed consent:

> Point 11. Evaluators should identify themselves to potential informants or respondents and advise them of the purpose and use of the evaluation[5] and the identity of the commissioners of the evaluation.

> Point 12. The informed consent of those directly providing information should be obtained, preferably in writing. They should be advised as to what information will be sought, how the information will be recorded and used, and the likely risks and benefits arising from their participation in the evaluation. In the case of minors and other dependents, informed consent must be sought from parents or guardians.

The updated documents were intended to be only the first stage in revising the ethical guidelines; interactive materials were to be added to the website, including vignettes (see Desautels & Jacob, 2012; Morris, 2008 for examples) and case studies similar to those elicited at the Wellington workshop. It was intended that the issues identified in this paper would be addressed in the vignettes and case studies. Although staffing and resource changes have stalled the development of these additional resources, the guidelines that do exist are binding on AES members, with potential for sanctions if complaints are made about them not being followed.

In Australia, there is some degree of direction for the hundreds of Australian members of the Australasian Evaluation Society on the importance of securing participant consent and what is required for it. How is this complemented by the informed consent requirements of Human Research Ethics Committees in Australia?

NHMRC guidelines on research and evaluation

As noted above, consent in the forms used by the NHMRC, such as the online National Ethics Application Form (NHMRC, 2014d), deals with issues such as ensuring participants understand the risks and obligations imposed by participation, and also understand that participation is voluntary. Potential payments, obtaining consent for minors and dependents, and other related issues are also covered. Except in rare, defined instances, researchers are expected to obtain the informed, voluntary consent of participants for all aspects of the research. The *National Statement* notes that: ‘The ethical and legal requirements of consent have two aspects: the provision of information and the capacity to make a voluntary choice’ (NHMRC, 2014e, p. 12). However, a recently released document dealing specifically with quality assurance and evaluation notes that ‘importantly, QA and evaluation commonly involve minimal risk, burden or inconvenience to participants’ (NHMRC, 2014b, p. 2). The document notes that evaluation ‘is undertaken to generate outcomes that are used to assess and/or improve service provision’ (NHMRC, 2014b, p. 2) and, citing the Australian Evaluation Society (2010) Guidelines, defines evaluation as:

> … a term that generally encompasses the systematic collection and analysis of information to make judgements, usually about the effectiveness, efficiency and/or appropriateness of an activity. The term is used in a broad sense to refer to any set of procedures, activities, resources, policies and/or strategies designed to achieve some common goals or objectives. (2014b, p. 2)

[5] This may of course change over the course of the evaluation, and continued updates may well be required.

While stressing the need to ensure that participants need to be protected from physical, spiritual, social harm and distress, and have mechanisms to voice concerns, the document offers advice on 'opt out' procedures, 'a method used in the recruitment of participants into an activity where information is provided to the potential participant regarding the activity and their involvement and where their participation is presumed unless they take action to decline to participate' (NHMRC, 2014b, p. 3). Due to the low risk nature of evaluation, it is noted that rather than going through the usual Human Research Ethics Committee procedures, ethical decision making for evaluation projects can be delegated to an HREC Chair, individual member or sub-committee, or to a special 'low risk committee' (NHMRC, 2014b, p. 4).

Conclusion

Informed consent is recognised as a cornerstone of ethical research practice. Although the NHMRC defines evaluation as generally 'low risk' (NHMRC 2014b), this paper has presented a view that the consequences of evaluation involving human research present a number of risks different to those incurred in (other) research. While most research ethics are designed to protect participants against risks incurred at the data collection stage, e.g. receiving a new drug or participating in a social science investigation, evaluative risks tend to occur at the 'back end' of the evaluation process, subsequent to the data collection stage.

The context of program evaluation is – not always, but more often than not – that it is contracted by someone who pays the evaluator and receives the evaluation findings as evidence for decision-making. This context leads to four layers of risk. Evaluators typically report particularly high risk at the report submission stage, with pressure (from a minority of commissioners[6]) to alter findings in ways not warranted by the evidence. Participants, on the other hand, may experience greater risks at the utilisation stage, as dissemination decisions are made and as changes are made to programs and services. For program staff, this stage can include changes in employment, including job loss. The next 'layer' at risk comprises program/service recipients, even if they were not evaluation participants whose informed consent was sought or required. They are affected by changes in operation, and sometimes by the termination of the service or program they had been accessing. Finally, some of the practices noted above – i.e. evaluation commissioners/finders requiring evaluators not to let participants know of positive evaluation findings and then terminating the program – can result in a long term lack of trust at the wider community level, with resistance to future research and evaluation.

In this context, the NHMRC document classifying evaluation as 'low risk' and generally not requiring a full Human Research Ethics Committee (HREC) review provides an opportunity to deal with evaluative risks more appropriately and expeditiously. One of the issues raised above, the difficulty of reconciling the time required for HREC review with the frequent need to get evaluations designs approved quickly, is resolved. The option of an alternative mechanism for ethical approval also offers opportunities to address the special risks posed by evaluation, which differ from bio-medical risks but also from the risks posed by other forms of social science research. Considering the four areas discussed above:

- Regulated mandatory review similar to that used for bio-medical research has been acknowledged by the NHMRC as less well suited to evaluation, perhaps partly in recognition that evaluation contract opportunities often cannot be approved through standard HREC processes in the time required.
- The 'front end' review process originating in bio-medical research contexts presents issues for social scientists due to the more dynamic and contingent nature of much social science research, as documented above, and evaluation shares that dynamic. However, the risks to social science participants typically arise in the data collection stage, as with bio-medical models, while risks to evaluation participants are often greater at the reporting stage and after the report has been submitted.
- 'Risk' and 'power' in social science contexts differ from bio-medical contexts, where there is often an asymmetry of power between researcher and research subject. Research participants in social science research may be possess equal or greater power than the researcher, and may pose risks to them. In evaluation, the agency commissioning/funding the project often possesses significant power.
- More discipline-specific processes may be enabled by the NHMRC determination that evaluation is 'low risk', and should be approved using different mechanisms than those used for bio-medical, social science etc. research.

With regard specifically to consent by evaluation participants, the minimum required to secure informed consent (as already noted above) would require the evaluator to provide potential participants with the identity of the evaluation commissioner, a list of the proposed evaluation questions (updated throughout the project as required), the potential outcome of the evaluation and particularly the likelihood of its utilisation by others for decision-making (again updated throughout the project as required).

However, unless the commissioning body is also willing to sign off on avoiding pressure to distort findings and/or to prevent them from being available to participants, fidelity to the requirements of informed consent will remain problematic. This is where the new NHMRC approach provides an opportunity for improved practice. Evaluation specific mechanisms could look at innovative approaches such as the more 'discursive' approach recommended by Emmerich (2013), involving both the evaluator(s) and the

[6] It is worth noting that many evaluations are conducted without harm to the evaluator or participants. As in other forms of research, risk is calculated through a consideration of likelihood and consequence, and takes the minority of evaluations posing harm to evaluators and participants into account. Also, the Wellington workshop noted above indicates that the number of evaluations posing such risks are not insignificant. Further, it is worth noting that pressure on researchers to provide certain findings, or not to report others, is found in other forms of research, including some pharmaceutical research, defence research, etc., but it is a feature of much program evaluation.

evaluation commissioner with the review committee or representative in discussing the ethics of the project. Discussions in this new model could address the importance of:

a. the commissioner not applying pressure to alter findings and produce conclusions not in accordance with the evidence; and

b. enabling evaluation participants to receive findings, at least in summarised form.

Such commitments would go a long way to ensuring truly informed consent in evaluation.

It is appreciated that this would be a departure from current practice, but it is hoped that this paper may lead to further dialogue with evaluation commissioners, users and evaluand representatives, as well as evaluation practitioners/researchers, to explore the issues of such an approach to evaluation ethics and informed consent in evaluation. In fact, it is even possible that if this approach proves successful, over time it may lead to changes in the ethical review of social science, bio-medical and other types of research.

References

Academy of Social Sciences. (2013, October). *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03. Retrieved from http://acss.wpengine.com/wp-content/uploads/2013/11/pb3_genericethicsprinciples.pdf

American Evaluation Association. (2004). Guiding Principles for Evaluators. *American Journal of Evaluation, 27*(3), [September 2006], pp. 293-294. Retrieved from http://www.samea.org.za/documents/AEA_guiding_principles_for_evaluators.pdf

Australasian Evaluation Society Inc. (2013a). *Code of Ethics.* Retrieved from http://www.aes.asn.au/images/stories/files/membership/AES_Code_of_Ethics_web.pdf

Australasian Evaluation Society Inc. (2013b). Guidelines for the ethical conduct of evaluations. Retrieved from http://www.aes.asn.au/images/stories/files/membership/AES_Guidelines_web.pdf

Berends, L. (2007). Ethical decision-making in evaluation. *Evaluation Journal of Australasia 7*(2): 40-45.

Boyd, K. (2013, October). Response (2) to 'Responsible to Whom? Obligations to Participants and Society in Social Science Research'. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03, (pp. 20-22). London, UK: Academy of Social Sciences.

Calvey, D. (2008). The Art and Politics of Covert Research: Doing `Situated Ethics' in the Field. *Sociology 42*(5): 905-918.

Carlson, R., Boyd, K., & Webb, D. (2004). The revision of the Declaration of Helsinki: past, present and future. *Journal of Clinical Pharmacology 57*(6): 695-713.

Carpenter, D. (2013, October). Generic Ethics Principles in Social Science Research Discussion 'Stimulus' Paper. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research.* [Professional Briefings], Issue 03, (pp. 3-6). London, UK: Academy of Social Sciences.

Chesterton, P. (2003). Balancing Ethical Principles in Evaluation: A Case Study. *Canadian Journal of Program Evaluation 18*(1): 49-60.

Desautels, G., & Jacob, S. (2012). The ethical sensitivity of evaluators: A qualitative study using a vignette design. *Evaluation 18:* 437-450.

Economic & Social Research Council [ESRC]. (2012). *Framework for Research Ethics 2010, revised September 2012* Swindon: ESRC. Retrieved from ESRC website http://www.esrc.ac.uk/_images/framework-for-research-ethics-09-12_tcm8-4586.pdf

Emmerich, N. (2013, October). Summary of Symposium 1. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research.* [Professional Briefings], Issue 03, (pp. 9-14). London, UK: Academy of Social Sciences.

Fain, J. (2005). Is There a Difference Between Evaluation and Research?. *The Diabetes Educator 31*(2): 150-155.

Guenther, J., Williams, E., & Arnott, A. (2010). *The politics of evaluation: evidence-based policy or policy-based evidence?* Paper presented at NARU Public Seminar Series, Darwin.

Hammersley, M. (2006). 'Are ethics committees ethical?' Qualitative Researcher 2: 4. Retrieved from http://www.cf.ac.uk/socsi/qualiti/QualitativeResearcher/QR_Issue2_06.pdf

Hammersley, M. (2010) 'Creeping Ethical Regulation and the Strangling of Research'. *Sociological Research Online 15*(4): 16.

Hammersley, M. (2013, October). Response (1) to 'Generic Ethics Principles in Social Science Research'. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03, (pp. 3-6). London, UK: Academy of Social Sciences.

Heintzelman, Carol, A. (2003). The Tuskegee Syphilis Study and Its Implications for the 21st Century. *Social Worker, 10*(4). Retrieved from http://www.socialworker.com/tuskegee.htm

Humphreys, L. (1970). *Tearoom Trade: Impersonal Sex in Public Places.* Chicago: Aldine Publishing Company.

Hunter, R. (2013) Response (2) to 'Responsible to Whom? Obligations to Participants and Society in Social Science Research'. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03, (pp. 18-20). London, UK: Academy of Social Sciences.

Israel, M. (2004). *Ethics and the governance of criminological research in Australia.* NSW Bureau of Crime Statistics and Research, Attorney General's Department.

Israel, M. (2013). Rolling back the bureaucracies of ethics review. *Journal of Medical Ethics 39*(8): 525-526.

Keith, G. (2003). *The Canadian Evaluation Society (CES) Experience in Developing Standards for Evaluation & Ethical Issues.* Paper presented in The 5th European Conference on the Evaluation of Structural Funds in Budapest, Hungary.

Kosheleva, N., & Segone. M. (2013, December). EvalPartners: Facilitating the development of a new model of voluntary organization for professional evaluation to support the development of national evaluation capacities. *American Journal of Evaluation,* July 10, 2013, 34(4), pp. 568-572. Retrieved from http://dx.doi.org/10.1177/1098214013493656

Levin-Rozalis, M. (2003). Evaluation and research: Differences and similarities. *Canadian Journal of Program Evaluation 18*(1): 1-31.

Markiewicz, A. (2008). The political context of evaluation: what does this mean for independence and objectivity? *Evaluation Journal of Australasia 8*(2): 35-41.

Markiewicz, A. (2010). *Can Evaluation be Politically Grounded, Policy Relevant, Participatory AND Objective and Independent?* Australasian Evaluation Society International Conference, Wellington, New Zealand. September 2010, Retrieved November 2011 from http://aes.asn.au/conferences/2010/Presentations/Markeiwicz,%20Anne.pdf

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology 67:* 371-378.

Morris, M. (Ed.). (2008). Evaluation Ethics for Best Practice: *Cases and Commentaries.* New York: Guilford Press.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (April 18, 1979). *The Belmont Report: Ethical Principles and Guidelines for the protection of human subjects of research.* [Regulations and Ethical Guidelines]. US Department of Health, Education & Welfare. Retrieved from the US Department of Health & Human Services website http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

National Health & Medical Research Council. (2014a). *Building a new application form for use in human research: Consultation on the structure and content of the form.* Draft document, May 2014. Retrieved from https://www.nhmrc.gov.au/_files_nhmrc/file/research/clinical_trials/discussion_paper_human_research_application_form_may_2014_140606.pdf

National Health & Medical Research Council. (2014b). *Ethical considerations in Quality Assurance and evaluation activities.* Retrieved from http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e111_ethical_considerations_in_quality_assurance_140326.pdf

National Health & Medical Research Council. (2014c). *History of ethics and ethical review of human research in Australia.* Retrieved 7 July 2014, from http://www.nhmrc.gov.au/health-ethics/human-research-ethics/history-ethics-and-ethical-review-human-research-australia.

National Health & Medical Research Council, Australian Research Council & Australian Vice-Chancellors' Committee. (2014d). *National Ethics Application Form, Version 2.2.* Retrieved from http://www.nhmrc.gov.au/health-ethics/human-research-ethics-committees-hrecs/hrec-forms/neaf-national-ethics-application-for

National Health & Medical Research Council, Australian Research Council & Australian Vice-Chancellors' Committee. (2014e). *National Statement on Ethical Conduct in Human Research 2007,* (Updated March 2014). Retrieved from https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72_national_statement_march_2014_140331.pdf

Patton, M. Q. (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use.* Guilford Press, New York. Retrieved from http://tei.gwu.edu/courses_approaches.htm#developmental_evaluation

Patton, Michael Quinn. (2014). *Evaluation Flash Cards: Embedding Evaluative Thinking in Organizational Culture.* St. Paul, MN: Otto Bremer Foundation, ottobremer.org. Retrieved from http://www.ottobremer.org/sites/default/files/factsheets/OBF_flashcards_201402.pdf

Rogers, P. (2014). *Ways of framing the difference between research and evaluation: Better evaluation.* Retrieved from Better Evaluation website (Week 19, May 9, 2014)http://betterevaluation.org/blog/framing_the_difference_between_research_and_evaluation

Roorda, M., & Peace, R. (2009). Challenges to implementing good practice guidelines for evaluation with māori: A pākehā perspective. *Social Policy Journal of New Zealand 34:* 73-89.

Scougall, J. (2006). Reconciling tension between principles and practice in Indigenous evaluation. *Evaluation Journal of Australasia 6*(2): 49 - 55.

Simons, H. (2006). Ethics in evaluation. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The SAGE Handbook of Evaluation* (pp. 213-232). London UK, SAGE. Retrieved from http://www.uk.sagepub.com/gray3e/study/chapter12/Book%20 chapters/Ethics_in_Evaluation.pdf

Skene, L., & R. Smallwood. (2002). Informed consent: lessons from Australia. *British Medical Journal 324*(7328): 39–41.

Sleat, M. (2013, October). Responsible to whom?: Obligations to participants and society in social science research. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03, (pp. 15-18). London, UK: Academy of Social Sciences.

Social Sciences & Humanities Research Ethics Special Working Committee. (2004). *Giving Voice to the Spectrum.* Ottawa: Interagency Advisory Panel on Research Ethics. Retrieved from http://www.pre.ethics.gc.ca/english/workgroups/sshwc SSHWCVoiceReportJune2004.pdf

Sokol, Daniel K. (2013). First do no harm. [Revisited]. *British Medical Journal 347:*f6426

Ticheloven, A. & Willems, M. (2013). The development of ethics in medical and social sciences in the last half of the twentieth century. Social Cosmos – URN:NBN:NL:UI:10-1-114217

Treasury Board of Canada Secretariat. (2006). *Professional ethics and standards for the evaluation community in the Government of Canada.* Retrieved from website: http://www.tbs-sct.gc.ca/cee/career-carriere/pesecgc-enpcegc/pesecgc-enpcegc-eng.pdf

Turner, D. 2003. *Evaluation Ethics and Quality: Results of a Survey of Australasian Evaluation Society Members,* Australasian Evaluation Society, Retrieved November 2011 from http://aes.asn.au/about/Documents%20-%20ongoing/ethics_survey_summary.pdf.

Van den Hoonard, E. W. C. (2013, October). Are We Asked to 'Other' Ourselves?: Social Scientists and the Research Ethics Review Process. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings], Issue 03, (pp. 23-28). London, UK: Academy of Social Sciences.

Wiles, R., Heath, S., Crow, G., & Charles, V. (2005). *Informed Consent in Social Research: A Literature Review.* NCRM Methods Paper Series 001, ESRC National Centre for Research Methods.

Wiles, R., Crow, G., Charles, V., & Heath, S. (2007). Informed Consent and the Research Process: Following Rules or Striking Balances? *Sociological Research Online 12*(2). doi:10.5153/sro.1208

Williams, E., Guenther, J., & Arnott, A. (2011). *Beyond informed consent: how is it possible to ethically evaluate Indigenous programs?* Paper presented at the NARU Public Seminar Series, Darwin.

Wynn, L., White, K., Thomson, C., & Israel, M. (2013) 'The Expanding Disciplinary Scope of Research Ethics Committees - An Inquiry into Need & Resistance'. Macquarie University ex ARC Discovery Project.

World Medical Association. (2013). *Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects.* Retrieved from http://www.wma.net/en/30publications/10policies/b

Zimmerman, S. (2013, October). Summary of Symposium 3. In Academy of Social Sciences (Ed.), *Generic Ethics Principles in Social Science Research,* [Professional Briefings] Issue 03, (pp. 44-45). London, UK: Academy of Social Sciences.

Measuring the unmeasured in educational programs: filling in the blanks through evaluation?

John Guenther

Cooperative Research Centre for Remote Economic Participation, and Flinders University

john.guenther@flinders.edu.au

Keywords: performance measures, performance indicators, measuring education, social circus, Families and Schools Together, Measurement Framework for Schooling in Australia, qualitative outcomes indicators

Abstract

The indicators of performance put forward as measures of achievement at a state and territory level in Australia reflect to some extent the priorities of those jurisdictions. These are revealed in the annual reports of departments, usually under headings of targets and corresponding outcomes. It may seem reasonable to assume that these performance measures line up with stated objectives, and with what matters on the ground. But do they? This paper argues that while the aims of education are broad, the measures of education are narrow. Philosophically, a good education is one that has social, developmental, intellectual and economic aims. However, even though some of those broad aims are reflected in national, state and territory foundational documents and reports, they are not reflected in Australian measurement or reporting frameworks, which seem to suggest that it is neither practical nor cost-effective to collect such data.

The purpose of this paper is to demonstrate that it is possible to measure educational success in ways that support the broader goals of education and schooling in Australia. To this end, the paper draws on two evaluation case studies in the field of education (run within schools), to highlight ways that the unmeasured aspects of educational activity can be measured. One case highlights the significance of social capital in a school-family partnership program, and the other demonstrates the psycho-social benefits of an alternative education program for ‘at risk’ children. These cases illustrate what can be measured and they provide useful data to fill in the blanks of what is not measured at a system level. However, the cases also raise bigger questions about what should be measured and reported as indicators of what matters to educational stakeholders.

Introduction

Over recent years there has been a lot of emphasis on measuring education. The introduction of the *My School* website in 2010 brought this into sharp focus and caused (and still does) something of a stir among those who saw the data measured being used for the production of league tables to unfairly rank schools on the basis of academic performance (Redden & Low, 2012). But what appears to emerge on the surface from the *My School* product actually comes from deeper roots that define and shape what education should be for and about. In this paper, I explore these foundations from the starting point of the 2008 *Melbourne Declaration on Educational Goals for Young Australians,* with an eye to uncovering the broader philosophical foundations of a 'good' education, and how these are played out in the Australian context. Next, I turn to the literature on issues of measurement in education and why some aspects of education are measured and some are not. I then consider how a good education might be considered from the perspective of four departments of education annual reports from 2013. What this will show is a fairly narrow interpretation of 'good education' with an even narrower set of measures. The *Measurement Framework for Schooling in Australia 2012* states that the measures used are 'cost effective, practical to collect, and take account of the burden and impact that data collection may place on students, schools and schooling systems' (ACARA, 2012, p. 5). However, do the measures actually reflect policy makers' real (and narrow) priorities, and have less to do with practicalities or burden?

In order to counter the 'practicality' argument, I present two evaluation case studies that demonstrate in the context of alternative education programs, how other indicators can be identified. The first is a social circus program conducted in Tasmanian schools and the second is a family strengthening program conducted in a Northern Territory school. I contend that while these evaluations used qualitative methodologies, the identification of indicators allows for a valid quantitative approach to be taken that could easily measure other indicators of a good education. But these alternative indicators are seldom measured as confirmed by the examination of departmental annual reports presented here. But why is this so?

Background

I begin with the broad philosophical question about what makes a 'good' education. Following this, the review turns specifically to the Australian context, showing how those philosophies are played out at a national level. Finally, attention is drawn to the specific goals and outcome measures for four state and territory departments of education.

What makes a 'good' education?

The philosophical and theoretical bases for educational strategic policy directions are diverse. The associated strategies and performance measures reflect sometimes divergent views about what education is really for. This paper does not allow for a detailed discussion of the theoretical and philosophical foundations of education. However, it may be helpful to briefly outline some of the key foundations on which departmental visions, goals and objectives, are based. What I am trying to get to is, as Biesta (2009) asks: 'what constitutes good education?'

There is a social and societal rationale for education.

There are various social theories that underpin educational systems. Education has been seen as a vehicle for social control (Dewey, 1938; Payne, 1927) and for the promotion of citizenship (Gutmann, 2009; McCowan, 2010). Others have described education as transformative and emancipatory (Freire, 1970; Oakes et al., 2013). Education too, is seen as a process that builds 'social capital' (Coleman, 1988) and is a product of 'cultural capital' (Bourdieu, 1983) which in turn maintains class divisions (Reay, 2010).

There is a developmental rationale for education.

The international discourse around education and development suggests strongly that better education leads to increased levels of development (Hanushek & Woessmann, 2007; Keeley, 2007; OECD, 2012a). The empirical evidence that education and learning is related to a range of benefits including social equity (Field et al. 2007; OECD, 2012b), health (Ross & Mirowsky, 2010), justice and reduced criminal behaviour (Lochner, 2011; Machin et al. 2011), employment, economic and developmental (Hanushek & Woessmann, 2009; OECD, 2012a), family and individual outcomes (Schuller et al. 2004) is readily available in an array of literature. The hope of education is that it leads to a better life, particularly for those living on the margins of society. Leadbeater (2012, p. 23) suggests that education 'offers them a hope that their place in society will not be fixed by the place they were born' and that through education people can 'remake their lives'.

There is a knowledge and skills rationale for education.

There is a view that knowledge is an end in itself, that one of the primary aims of education is epistemic (Robertson, 2009), and that for educators it is reasonable to expect that it is 'possible, and desirable for people to *know and do* things, and to engage in and take seriously the fruits of *rational inquiry,* where such inquiry is understood to involve the pursuit of *truth'* (Siegel, 2010, p. 283). Such pursuit of knowledge forms a foundation for students to be able to make appropriate moral choices and therefore become good citizens (Feldman, 2009; Halstead, 2010).

There is an individual and economic rationale for education.

The focus on individualism has its roots in Greek philosophy and was further developed by Enlightenment philosophers such as Kant and Rousseau, who emphasised individual autonomy and individual freedom. (for a discussion of the historical development of philosophies of education see Carr, 2010). The arguments of liberalist education philosophers suggest that 'schools should encourage competition between individual students and prepare students to live independent lives in society, respecting their uniqueness and distinct capabilities' (Portelli & Menashy, 2010, p. 421). Individualism is also reflected in the economic theories of Adam Smith (1904) which in turn is reflected in what could be described as free market capitalism. The economics of education has come to the fore in recent decades. A notable contribution to the field was Becker's (1964) work *Human Capital: a theoretical and empirical analysis, with special reference to education,* which brought together ideas of return on investment in education and distribution of income on the basis of educational attainment. Internationally, policies are built around assumptions of economic growth flowing from education and training: 'skills have become the global currency of 21st-century economies' (OECD, 2012a, p. 10).

Philosophies of education played out in Australia

While elements of the above discussion can be found in literature that spans decades in Australia, a significant marker in education occurred in 2008 with the *Melbourne Declaration on Educational Goals for Young Australians.* It articulates two main objectives:

> Goal 1: Australian schooling promotes equity and excellence
>
> Goal 2: All young Australians become successful learners, confident and creative individuals, and active and informed citizens. (Ministerial Council on Education, 2008, p. 7)

The goals articulated in the Declaration are also broadly consistent with a philosophy of education that goes beyond a focus on academic performance and transition to employment. They represent education as a vehicle for individual and social achievement, for an inclusive and respectful society that supports the development of knowledge and skills, but not to the exclusion of other personal and social imperatives. The definition of equity is not about conformance to the norm. Rather the goals express a need for diversity such that 'schooling contributes to a socially cohesive society that respects and appreciates cultural, social and religious diversity.' (p. 7)

In short, the goals represent an array of educational, epistemic, moral and political aims (Brighouse, 2009; Robertson, 2009). The 2012 National Education Agreement, which builds on the Declaration, specifies five outcomes of education that in turn determine the key performance measures of education. These are:

> (a) all children are engaged in and benefiting from schooling; (b) young people are meeting basic literacy and numeracy standards, and overall levels of literacy and numeracy achievement are improving; (c) Australian students excel by international standards; (d) schooling promotes the social inclusion and reduces the educational disadvantage of children, especially Indigenous children; and (e) young people make a successful transition from school to work and further study. (Standing Council on Federal Financial Relations, 2012, p. 6)

The Measurement Framework for Schooling in Australia (ACARA, 2012) articulates how these outcomes are to be measured. Essentially, this document distils the outcomes into three main areas: participation, achievement in the National Assessment Program and attainment. The specific indicators include enrolments, attendance, and participation in assessments, levels of literacy and numeracy, school completion and attainment and achievement of young people in other learning pathways. Equity 'measures are not separately listed in the Schedule of Key Performance Measures but are derived, for reporting purposes, by disaggregating the measures for participation, achievement and attainment where it is possible and appropriate to do so' (p. 6). This last point is important as it means the focus is not on equity and diversity but on the other outcomes listed, which may be seen to promote conformance to the norm. The Measurement framework makes no attempt to measure the broader goals of education as outlined earlier in the broad discussion about what makes a 'good' education or the goals of education as outlined in the Melbourne Declaration.

Measurement of education

Why do systems measure what they do in education? At the school level, the issues of measurement – as reflected in schools' student reports for example – are about ensuring that students learn what they are taught. Measurement is essential for student feedback, teacher professional development and informing parents about student progress. This kind of measurement underpins what Hattie (2009) describes as 'visible teaching and learning'. Even with this as a given, the issues of how assessments and tests should be used are controversial and contested. Some scholars (Anyon, 2010; Oakes, 2005; Oakes et al., 2013) argue that testing which 'tracks' – or streams – students has a deleterious effect on student learning and equity. The drive for accountability is another factor that determines what is measured. As an American proponent of educational accountability, Taft (2012) argues that:

> Aligned systems of academic standards, assessments and accountability are one essential component of an effective strategy for raising the bar for student learning … and giving all students an opportunity for success after high school. (p. 4)

The language of transparency and accountability, effectiveness and efficiency underpins assumptions about why education should be measured. The 2012 Measurement Framework is premised on the 'accountability requirements established in the National Education Agreement and *Schools Assistance Act 2008*' (ACARA, 2012, p. 1). On the surface, the argument for accountability is one of equity. Zanderigo et al. (2012) contend that:

> The transparency and accountability mechanisms are aimed at improving outcomes and equity for all students by using nationally comparable school performance data to build a substantive evidence base to support future improvements. (p. 3)

All other things being equal, there is some merit to this argument, provided that accountability measures do not lead to 'tracking' effects (Schütz et al., 2007). Making schools accountable for student performance – creating competition by providing funding incentives on the basis of differential outcomes – may well have the effect of driving consumer choice away from poor performing public schools to higher performing private schools where low income families are effectively excluded due to high fees. School behaviour may also change to take advantage of accountability systems. Figlio and Loeb (2011) suggest that schools' strategies may change in response to accountability measures, for example teaching to the test, and the use of strategies to exclude poor performing students at test times. As can be seen, all things are not necessarily equal and trying to disaggregate cause and effect in a complex system is very difficult, as Jenson (2013) revealed in his study of competition, autonomy, choice and markets in Queensland.

The above discussion suggests that accountability frameworks set the agenda for what is measured in education. Frameworks that promote measures such as attendance rates, academic performance and retention rates, result in those things being privileged. While indicators of equity are collected, in Australia at least, the incentives associated with accountability frameworks do not directly reward increases in social inclusion or ethnic diversity for example. The 2012 Review of Funding for Schooling (Gonski et al., 2012) argued that evidence 'shows that strengthening equity in education can be cost beneficial' (p. 108) and supported needs based funding models to promote equity. However, its Terms of Reference were focused on funding arrangements rather than measurement frameworks, even though it did discuss accountability. If equity was of primary concern in the measurement framework, surely equity outcomes and targets (not just equity indicators) would be incorporated into the table of measures listed under the Schedule of Key Performance Measures of the Framework (ACARA, 2012, p. 7).

In summary, Australia has a measurement framework that argues for the importance of equity but does little to promote it. What do count are measures related to student academic performance (in standardised test results) and participation (through attendance and retention measures). We would therefore expect to see those priorities reflected in the official positions of education departments. In the next section that proposition will be tested.

Education Department objectives, outcomes and performance measures in four states

Table 1 is a condensed summary of the vision, mission, goals, targets, strategies and measures of four jurisdictional education departments. The analysis is based on publicly available annual reports for the year 2012-13. The purpose of the table is to highlight the priorities of departments of education and how they line up with the priorities of a good education as discussed above. For all states, the vision/mission statements and supporting goals line up fairly closely with the measures associated with the direction presented in the Measurement Framework discussed above. There is a strong focus on skills, knowledge, innovation, economic prosperity, productivity, achievement and success.

South Australia deviates somewhat from the other jurisdictions listed in that its vision includes a democratic, equitable and cohesive society. This vision indicates a broader understanding of what a good education is in terms of its societal goals. However, when it comes to measurement, the indicators are all about academic performance. Similarly, Western Australia includes the motto ‘excellence and equity’ but makes no mention of this in its performance measures. It adds cost per full time student to the mix of performance measures, which perhaps links to the goal of a ‘capable and responsive organisation’ but the link is not entirely direct. The Tasmanian mission statement does include the aim of ‘contributing positively to the community’ but none of its measures reflect that goal. The Northern Territory vision statement, targets and measures are perhaps the most congruent of all the jurisdictions presented. However, they are narrowly focussed on a limited number of academic/school related outcomes.

One of the key drivers for the indicators chosen appears to be the Measurement Framework for Schooling in Australia, which is clearly reflected in the reported performance measures for each of the jurisdictions. There is no attempt to measure the social or equity aims that are included in the mission or goals of South Australia, Western Australia or Tasmania. Further, for these jurisdictions there is no apparent logic that connects these aims to the performance measures. In summary, the performance measures appear to be narrowly focused on the knowledge and skills rationale for education, underpinned by a Human Capital theoretical position that sees economic benefit from schooling.

Given that three of the four jurisdictions seem to put education forward more holistically than that, it could be reasonable to ask why measures for these broader purposes and outcomes of education are not included. The Measurement Framework argues that equity measures can be derived from disaggregated indicators such as Indigenous status, gender, location, language background, socio-economic background and disability. However it could be argued that these disaggregated measures simply indicate how well equity groups achieve the performance measures. They do not in themselves indicate how equitable, inclusive, respectful, democratic or cohesive the education provided intrinsically is, or how it contributes to society in those ways.

Table 1: **Departmental mission, vision and goals**

Vision or mission (in abridged form)	***Supporting goals or targets***	***How do they get there?***	***Reported performance measures***
Northern Territory: Through quality strategies, programs, people, partnerships and systems we will grow **educated, skilled and smart** Territorians.	• Compulsory schooling: 90% attendance for every child at school; every child up to 17 in school, training or employment pathways. • 20% increase in NTCET completions • 3% increase in NMS for non-Indigenous students in NAPLAN; 9% for Indigenous. • 20% completion for VET in schools certificate	• Quality strategies and programs, people and partnerships, systems and support	**All groups:** • Student enrolments and attendance rates **Primary and middle years:** • NAPLAN performance **Senior Years:** • VET Certificate completions • Enrolments in school-based apprenticeships • Qualification for NTCET
South Australia: **democratic, equitable, prosperous and cohesive** society.	1. Every child achieves their potential 2. Excellence in education and care 3. Connect with communities 4. A successful and sustainable organisation	• Children and young people are at the centre; • Quality teaching and learning; • Schools engage with families and the wider community; • Responsive to the needs of students and the workforce	• Early childhood AEDI indicators • Year 1 literacy • Aboriginal early years (literacy) • NAPLAN measures • Understanding Aboriginal culture (schools teaching Aboriginal Cultural Studies) • Science and Maths (Students with a TER including science or maths) • SACE or equivalent

Table 1: **Departmental mission, vision and goals** *Continued*

Vision or mission (in abridged form)	***Supporting goals or targets***	***How do they get there?***	***Reported performance measures***
Western Australia: **provision of quality education,** whatever their ability, wherever they live, whatever their background.	• Success for all students • Distinctive schools • High quality teaching and leadership • A capable and responsive organisation • "excellence and equity"	• Create opportunities for success; • Meeting the needs and aspirations of students to build strong communities; • High quality teaching and leadership; • Using resources wisely and making open and transparent decisions; • Build community confidence	**Effectiveness indicators** 1. Rates of participation in education; 2. Retention in public schooling; 3. Secondary graduation rates; 4. Student achievement in literacy; 5. Student achievement in numeracy **Efficiency indicators** • Costs per student FTE
Tasmania: successful, skilled and innovative Tasmanians. Mission: To provide every Tasmanian with the opportunity to continue to learn and reach their potential, to lead fulfilling and productive lives and to contribute positively to the community.	**Key drivers:** • successful learners, innovative workforce, inspired leadership, dynamic learning environments, community confidence. • Underpinning **values:** excellence, equity, respect and relationships	• Quality programs, integrated services, engaged families; Implement Australian Curriculum; • quality resources; organisation wide framework for literacy and numeracy; • networks of schools	**Primary and secondary years** • % of Kinder and prep students achieving expected development, literacy and numeracy outcomes; NAPLAN measures • % gap in Indigenous students achieving expected outcomes **Senior years:** • 120 credit points in education and training; Some vocational education and training TCE completions; Tertiary Entrance ranks; Year 10-12 retention rates

Sources: Department for Education and Child Development, 2013; Department of Education (NT), 2013; Department of Education (Tas), 2013; Department of Education (WA), 2013

It may be argued that these things are not measured because they are difficult to put a number on. This is the position that the Australian Bureau of Statistics takes in its *Measures of Australia's Progress* documents (ABS, 2013). However, it could be equally argued that perception measures could be used. I will return to this later, but for the moment I want to consider two cases of evaluations involving schools in two of the jurisdictions tabulated above, where the measure of outcomes extends beyond academic performance and attendance.

Findings from two alternative education programs

Two case studies highlight how measures other than attendance and academic performance can be measured. The activities were quite different and in different jurisdictions, but conducted by the same evaluator. The focus here is not on the outcomes of the evaluation but on the development of alternative measures.

Case study 1: Social circus in Tasmanian schools

As part of an evaluation of the *Communities for Children (CfC)* program in Burnie, Tasmania the evaluator was asked to conduct a detailed evaluation of two social circus activities. Social circus is a subset of community arts, which uses circus skills to promote artistic expression as a vehicle for a range of social and educational outcomes for participants. The activities in this case were designed to work with parents, carers and their children in an effort to engage them in an alternative educational activity that would build bridges between schools and families and offer opportunities for creative expression and learning success outside the classroom. The activities were evaluated during 2011. CfC was designed to support and build capacity in vulnerable families.

The two activities under investigation were designed to support families and students to connect with their schools and vice versa. There was also an expectation that parents would engage with their children in the activities. This was a prerequisite for the children to be able to participate. There were however, two major differences in the ways the activities were run. The first was conducted over six weeks at a hall away from school and involved a family support service. The second was conducted over four weeks at a primary school and did not include involvement of a family support service.

The outcomes of the two activities differed somewhat but there was substantial overlap. Outcomes common to both were categorised as: an experience to remember; interaction within family; physical activity; self-confidence; and skills and capacity. In the first activity, which was more intensive, additional outcomes were recorded. They were categorised as: better behaviour from children; changed attitude; mutual support; rapport, relationship, social networks and trust.

The purpose of this case study is not to detail what these outcomes mean. Rather it is to highlight that they were measured. How were they measured? The tools used for both evaluations included: evaluator observations using photographs and video recordings; a reflective process with circus trainers (using a journal); a focus group with organisational stakeholders; a focus group with participants at conclusion and a survey of participants on completion. By combining the data from the various sources it was possible to demonstrate how the outcomes were perceived through the lens of participants, trainers, family support workers and the camera (as taken by the evaluator).

Case study 2: Family strengthening program in a Northern Territory school

Families and Schools Together (FAST) is a family strengthening program that helps build positive relationships between parents, schools and the broader community. At its core is an evidence-based eight week program designed to facilitate social support, greater parental self-efficacy, better child behaviour and improved educational outcomes. In the Northern Territory, the program has been run since 2002 in a variety of locations from remote communities to urban contexts. A 2013 evaluation of the program run in Alice Springs highlighted social capital outcomes that emerged from the program. The evaluation found that FAST contributed to a growing sense of community within the school. A recurring theme emerging from the data was the significance of improved social networks and support fostered during the eight week program.

The link between parent-school engagement and learning outcomes was not direct, but what could be surmised from the data was that better engagement led to better understanding of family needs, which in turn led to more responsive teaching, better targeted resources, and subsequently better learning outcomes. Over the longer term, for the school, FAST affirmed a culture of school-family-community engagement that permeated the school environment. The school's reputation in the community was enhanced. FAST Parents reported a sense of belonging and allegiance to the school that was reflected in their willingness to support the school as *their* school, not as *the* school.

So how were these findings measured? This was a largely qualitative study. Interviews were conducted, themes were analysed and quantitised[1]. The questions were built on a pre-existing theory of change which had been built up from previous evaluations. The themes (in order of importance) were: engagement and partnership; relationships with school; empowerment; social networks; support for struggling families; identity and confidence; attendance, performance and behaviour; and communication. There were some differences in the relative importance of some themes for different groups. For example, staff commented more on the importance of attendance, performance and behaviour than the other groups. Families commented more on 'support for struggling families' and the FAST team saw empowerment as more important.

Discussion

The intention in this paper is not to provide a detailed analysis of the case studies presented above. Nor is it to consider the impact of the measures identified above for commissioned evaluations. Rather the intention is to demonstrate that, based on the literature which describes a ‘good education’ beyond the fairly narrow indicators presented in Table 1, it is possible to measure other aspects of education that contribute to a broad range of outcomes.

Developing additional measures of a good education.

Essentially, the above case studies demonstrate the potential for outcomes such as engagement, partnerships, social connectedness, parental involvement, behaviour change, identity and confidence, social support and social capital to be measured. While the methods were qualitative, it would not be too difficult to change the instruments so that they assessed the indicators quantitatively. The value of the qualitative process is that it builds a theory that can be tested – which is consistent with grounded theory approaches in social research (Charmaz, 2006, 2011). In the past I have developed tools that do this kind of testing quite successfully without the need for complex or overly long tools (see for example Guenther, 2011; Guenther & Arvier, 2010; Guenther & Boonstra, 2009; Guenther & Falk, 2000). Schools in general are quite good at eliciting perception data from parents, students and teachers. Further, there are plenty of precedents for the measurement of social characteristics such as social capital and social inclusion (Grootaert et al., 2004; Harpham, 2008; Narayan & Cassidy, 2001), social connectedness and conflict (Bond et al., 2007; Cornwell et al., 2008) to mention just a few. The point of this discussion is to confirm that qualitative measures such as those used in the evaluations cited earlier can be measured quantitatively as well. They could form part of a suite of alternative measures of a ‘good’ education that would better reflect the broader purposes of education beyond the prevalent individualistic, knowledge and skill based measures that comprise the Australian Measurement Framework.

It is reasonable to ask why the current measurement and reporting frameworks do not allow for a broader measurement of important aspects of educational goals such as social inclusion, equity, or civic participation. Is it because, as the 2012 Measurement Framework suggests, the measures of these elements of a good education are impractical? Or is it because the current measures reflect what is important for policy makers?

The importance of alternative measures

Why should any of this matter? After all, it could be argued that the vast majority of students happily conform to or embrace an education that fits neatly within a Human Capital Theory framework where returns from education are based on individuals investing in knowledge and skills for a productive economy. The challenge to this is that education should be for all, not just the majority. In my work as an evaluator

[1] A term that Tashakkori and Teddlie (1998) used to describe the process of turning qualitative data into quantitative data.

and researcher in education I have seen how marginalised students – such as those with learning disabilities, those from struggling families, those in contact with the criminal justice system, and Indigenous young people living in remote Australian communities – are further marginalised by measurement systems that fail to recognise their strengths, and which relegate them to alternative educational programs.

If, as the three departments of education shown in Table 1 profess, education is about equity, democracy, inclusion, respect and social cohesion, then these elements of a good education ought to be measured. It is one thing to have a *Closing the Gap* agenda – as Australian governments do have currently (Council of Australian Governments, 2011) – but it is quite another thing to measure performance of interventions and policies against the outcomes which it supports. (Appendix 1 in Atelier Learning Solutions, 2012 offers an extensive list of alternative indicators for National Partnership programs). This is a concern that has been identified often in evaluation reports and reviews. Sometimes it is put down to a lack of data or evidence (Wilson, 2014). Sometimes it is due to measures being associated with outputs rather than outcomes (Atelier Learning Solutions, 2012). I am arguing here that – in the case of education – the approach which disaggregates equity groups from the measures that matter (attendance, academic performance, attainment and retention) simply diverts attention from equity and has the potential to further marginalise those who are already marginalised.

Schools that promote ethnic and cultural diversity, that support students with learning or physical disabilities, that promote civic participation, and that foster social inclusion could very easily be identified by tweaking the current Index of Community Socio-educational Advantage (ACARA, 2013) so that it reflected equity or social inclusion indicators. Using this measure, schools which measure improved social inclusion and equity as outcomes (rather than needs) could be rewarded. The two case studies highlight – based on evaluation findings – how success can be reconceptualised and measured to support these goals.

Conclusion

This paper has set out to lay a foundation for what makes a ‘good’ education in Australia. It shows—despite the multiple purposes of education—that in Australia the dominant rhetoric is that successful education is largely about individual knowledge and skills for economic benefits. This is reflected in the accountability and measurement frameworks that currently exist in Australia. This in turn is reflected in annual reports where performance measures are almost all related to academic performance, attendance and attainment.

This is all well and good if, like the majority of Australians, we take these assumed purposes as a given. My concern in evaluation and research in education, however, has often been about minorities. The two evaluation case studies of alternative education programs demonstrate the array of alternative indicators that may be important beyond academic performance, such as equity and inclusion. While those evaluations drew on qualitative methods, there is ample precedent for the quantitative measure of similar outcomes.

Alternative measures do matter – and not just for equity groups. They matter because the important outcomes for many people taking part in alternative education programs are not necessarily academic ones. Even for those in standard programs, the significance of outcomes such as social cohesion, civic participation, identity formation, equity, democracy, and the development of strong and productive social networks should not be underestimated. Education without these underpinning values runs the risk of perpetuating and promoting social inequalities. In this paper what I have hopefully achieved is something like filling in the blanks of missing information in the patchwork of evidence that contributes to our understanding about what makes a good education in Australia.

Acknowledgement

The work reported in this publication was supported by funding from the Australian Government Cooperative Research Centres Program through the Cooperative Research Centre for Remote Economic Participation (CRC-REP). The views expressed herein do not necessarily represent the views of the CRC REP or Ninti One Limited or its participants. Errors and omissions remain with the author.

References

ABS. (2013). *Measures of Australia's Progress 2013:* Is life in Australia getting better? Retrieved from http://www.ausstats.abs.gov.au/ausstats/free.nsf/vwLookupubject/1370.0~2013~MAP%202013%20Summary%20Brochure~13700_2013_MAP_Brochure.pdf/$File/13700_2013_MAP_Brochure.pdf

ACARA. (2012). *Measurement Framework for Schooling in Australia 2012.* Australian Curriculum Assessment and Reporting Authority Ed. Sydney: ACARA. Retrieved from http://www.acara.edu.au/verve/_resources/Measurement_Framework_for_Schooling_in_Australia_2012.pdf.

ACARA. (2013). *About ICSEA.* Australian Curriculum Assessment and Reporting Authority. Retrieved from http://www.acara.edu.au/verve/_resources/Fact_Sheet_-_About_ICSEA.pdf.

Anyon, J. (2010). What "Counts" as educational policy? Notes toward a new paradigm. In S. Semel (Ed.), *Foundations of Education: the essential texts* (pp. 81-100). New York: Routledge.

Atelier Learning Solutions. (2012). Phase 1 National Evaluation: *Final Report on the Analysis of SSNP Activity and Evaluation Effort.* Atelier Learning Solutions Pty Ltd. Retrieved from http://docs.education.gov.au/system/files/doc/other/phase_1_evaluation_final_report.pdf

Becker, G. (1964). *Human capital: A theoretical and empirical analysis with special eeference to education.* Chicago: University of Chicago Press.

Biesta, G. (2009). Good education in an age of measurement: on the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability, 21*(1), 33-46. doi: 10.1007/s11092-008-9064-9

Bond, L., Butler, H., Thomas, L., Carlin, J., Glover, S., Bowes, G., & Patton, G. (2007). Social and School Connectedness in Early Secondary School as Predictors of Late Teenage Substance Use, Mental Health, and Academic Outcomes. *Journal of Adolescent Health, 40*(4), 357.e9-357.e18. doi: http://dx.doi.org/10.1016/j.jadohealth.2006.10.013

Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital (R Nice, Trans.). In Reinhard Kreckel (Ed.), Soziale Ungleichheiten (*Soziale Welt, Sonderheft 2*) (pp. 183–198.). Göttingen: Otto Schwartz & Co.

Brighouse, H. (2009). Moral and political aims of education. In H. Siegel (Ed.), *The Oxford handbook of philosophy of education* (pp. 35-51). Oxford: Oxford University Press.

Carr, D. (2010). The philosophy of education and educational theory. In R. Bailey, C. McCarthy, D. Carr & R. Barrow (Eds.), *The SAGE handbook of philosophy of education* (pp. 37-53). London: Sage Publications.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis* Thousand Oaks: Sage.

Charmaz, K. (2011). Grounded Theory Methods in Social Justice Research. In N. Denzin & Y. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (Vol. 4th Edition, pp. 359-380). Thousand Oaks: Sage Publications Inc.

Coleman, J. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*(Supplement S), 95–120.

Cornwell, B., Laumann, E. O., & Schumm, L. P. (2008). The Social Connectedness of Older Adults: A National Profile. *American Sociological Review, 73*(2), 185-203. doi: 10.2307/25472522

Council of Australian Governments. (2011). *National Indigenous Reform Agreement (Closing the Gap*). Retrieved from http://www.federalfinancialrelations.gov.au/content/national_agreements/indigenous_reform/National_Indigenous_Reform_Agreement_from_13_Feb_11.pdf.

Department for Education and Child Development. (2013). *Annual Report 2012.* Adelaide: Government of South Australia. Retrieved from http://www.decd.sa.gov.au/docs/documents/1/DECDAnnualReport2012.pdf.

Department of Education (NT). (2013). *Annual Report 2012-13.* Darwin: Northern Territory Government. Retrieved from http://www.education.nt.gov.au.

Department of Education (Tas). (2013). *Annual Report 2012-13.* Hobart. Retrieved from https://www.education.tas.gov.au/documentcentre/Documents/DoE-Annual-Report-2012-2013.pdf.

Department of Education (WA). (2013). Annual Report 2012-13. East Perth. Retrieved May 2014 from http://det.wa.edu.au/detcms/cms-service/download/asset/?asset_id=14340938.

Dewey, J. (1938). *Experience and education.* New York: Kappa Delta Pi.

Feldman, R. (2009). Thinking, reasoning and education. In H. Siegel (Ed.), *The Oxford handbook of philosophy of education* (pp. 67-82). Oxford: Oxford University Press.

Field, S., Kuczera, M., & Pont, B. (2007). No more failures: Ten steps to equity in education. Paris: OECD.

Figlio, D., & Loeb, S. (2011). School accountability *Handbook of the Economics of Education* (Vol. 3, pp. 383-421).

Freire, P. (1970). *Pedagogy of the oppressed.* New York: Continuum Publishing Company.

Gonski, D., Boston, K., Greiner, K., Lawrence, C., Scales, B., & Tannock, P. (2012). *Review of Funding for Schooling, Final report.* Canberra: Department of Education, Employment and Workplace Relations. Retrieved February 2012 from http://www.deewr.gov.au/Schooling/ReviewofFunding/Documents/Review-of-Funding-for-Schooling-Final-Report-Dec-2011.pdf.

Grootaert, C., Narayan, D., Jones, V. N., & Woolcock, M. (2004). *Measuring Social Capital: An Integrated Questionnaire.* World Bank

Guenther, J. (2011). *Evaluation of FAST Galiwin'ku program.* Ulverstone: Cat Conatus. Retrieved from http://www.fastnt.org.au/documents/File/Galiwinku_FAST_evaluation_report.pdf.

Guenther, J., & Arvier, M. (2010). Mobile Family Resource Service for Parents and Young Children. In E. Bell & J. Merrick (Eds.), *Rural Child Health: International Aspects* (pp. 125-134). New York: Nova Biomedical Books.

Guenther, J., & Boonstra, M. (2009). *Adapting Evaluation Materials for Remote Indigenous Communities and Low-Literacy Participants.* Paper presented at the 12th Australasian Conference on Child Abuse and Neglect, Perth, Western Australia. 15-18 November 2009. Retrieved from http://www.catconatus.com.au/docs/091117_APCCAN_FAST.pdf.

Guenther, J., & Falk, I. (2000). *Measuring trust and community capacity: social capital for the common good.* Launceston: University of Tasmania, Centre for Research and Learning in Regional Australia.

Gutmann, A. (2009). Educating for individual freedom and democratic citizenship: In unity and diversity there is strength. In H. Siegel (Ed.), *The Oxford handbook of philosophy of education* (pp. 409-427). Oxford: Oxford University Press.

Halstead, J. (2010). Moral and citizenship education. In R. Bailey, C. McCarthy, D. Carr & R. Barrow (Eds.), *The SAGE handbook of philosophy of education* (pp. 253-268). London: Sage Publications.

Hanushek, E. A., & Woessmann, L. (2007). *The role of school improvement in economic development.* National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w12832.

Hanushek, E. A., & Woessmann, L. (2009). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *National Bureau of Economic Research Working Paper Series, No. 14633.*

Harpham, T. (2008). The Measurement of Community Social Capital Through Surveys. In I. Kawachi, S. V. Subramanian & D. Kim (Eds.), *Social Capital and Health* (pp. 51-62): Springer New York.

Hattie, J. A. (2009). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement.* Abingdon: Routledge.

Jensen, B. (2013). *The myth of markets in school education.* The Grattan Institute. Retrieved from http://grattan.edu.au/static/files/assets/de60db0d/myth_of_markets_in_school_education.pdf.

Keeley, B. (2007). *Human capital: How what you know shapes your life.* Paris: OECD Publishing.

Leadbeater, C. (2012). *Innovation in education: Lessons from pioneers around the world.* Dohar: Qatar Foundation.

Lochner, L. (2011). Nonproduction benefits of education: Crime, health and good citizenship. In E. A. Hanushek, S. J. Machin & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. Volume 4, pp. 183-282). Amsterdam: North-Holland.

Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education*. *The Economic Journal, 121*(552), 463-484. doi: 10.1111/j.1468-0297.2011.02430.x

McCowan, T. (2010). Can schools make good citizens? In R. Bailey (Ed.), *The philosophy of education* (pp. 86-98). London: Continuum International Publishing Group.

Ministerial Council on Education, Employment, Training and Youth Affairs,. (2008). *Melbourne declaration on educational goals for young Australians.* Melbourne: Curriculum Corporation. Retrieved from http://www.curriculum.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf.

Narayan, D., & Cassidy, M. F. (2001). A dimensional approach to measuring social capital: development and validation of a social capital inventory. *Current sociology, 49*(2), 59-102.

Oakes, J. (2005). *Keeping Track: How schools structure inequality.* New Haven: Yale University Press.

Oakes, J., Lipton, M., Anderson, L., & Stillman, J. (2013). *Teaching to change the world* (4th Edition ed.). Boulder: Paradigm Publishers.

OECD. (2012a). *Better skills, better jobs, better lives: A strategic approach to skills policies.* OECD Publishing. Retrieved from http://skills.oecd.org/documents/OECDSkillsStrategyFINALENG.pdf.

OECD. (2012b). *Equity and quality in education: Supporting disadvantaged students and schools.* OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264130852-en.

Payne, E. G. (1927). Education and social control. *Journal of Educational Sociology, 1*(3), 137-145. doi: 10.2307/2961744

Portelli, J., & Menashy, F. (2010). Individual and community aims of education. In R. Bailey, C. McCarthy, D. Carr & R. Barrow (Eds.), *The SAGE handbook of philosophy of education* (pp. 415-433). London: Sage Publications.

Reay, D. (2010). Sociology, social class and education. In M. Apple, S. Ball & L. Gandin (Eds.), *The Routledge international handbook of the sociology of education* (pp. 396-404). Abingdon: Routledge.

Redden, G., & Low, R. (2012). My School, Education, and Cultures of Rating and Ranking. *Review of Education, Pedagogy, and Cultural Studies, 34*(1-2), 35-48. doi: 10.1080/10714413.2012.643737

Robertson, E. (2009). The epistemic aims of education. In H. Siegel (Ed.), *The Oxford handbook of philosophy of education* (pp. 11-34). Oxford: Oxford University Press.

Ross, C. E., & Mirowsky, J. (2010). Why education is the key to socioeconomic differentials in health. In C. Bird, P. Conrad, A. Fremont & S. Timmermans (Eds.), *Handbook of medical sociology* (Sixth Edition ed., pp. 33-51). Nashville: Vanderbilt University Press.

Schuller, T., Preston, J., Hammond, C., Brassett-Grundy, A., & Bynner, J. (2004). *The Benefits of Learning: The impact of education on health, family life and social capital.* Abingdon: Routledge Falmer.

Schütz, G., West, M. R., & Wößmann, L. (2007). *School Accountability, Autonomy, Choice, and the Equity of Student Achievement: International Evidence from PISA 2003.* Vol. No. 14: OECD Publishing. Retrieved from http://www.eric.ed.gov/PDFS/ED503829.pdf.

Siegel, H. (2010). Knowledge and truth. In R. Bailey, C. McCarthy, D. Carr & R. Barrow (Eds.), *The SAGE handbook of philosophy of education* (pp. 283-295). London: Sage Publications.

Smith, A. (1904). *Inquiry into the nature and causes of wealth of nations* (Fifth Edition ed.). London: Methuen and Co. Ltd., Library of Economics and Liberty.

Standing Council on Federal Financial Relations. (2012). *National Education Agreement.* Retrieved from http://www.federalfinancialrelations.gov.au/content/npa/education/national-agreement.pdf.

Taft, B. (2012). Will educational accountability enhance the ability of schools to foster student academic growth? In T. Lasley (Ed.), *Standards and Accountability in Schools* (pp. 4-11). Thousand Oaks: Sage Publications.

Tashakkori, A., & Teddlie, C. (1998). Mixed Methodology: *Combining Qualitative and Quantitative Approaches.* Thousand Oaks, CA.: Sage.

Wilson, B. (2014). *A share in the future: Review of Indigenous Education in the Northern Territory.* Retrieved from http://www.education.nt.gov.au/__data/assets/pdf_file/0007/37294/A-Share-in-the-Future-The-Review-of-Indigenous-Education-in-the-Northern-Territory.pdf.

Zanderigo, T., Dowd, E., & Turner, S. (2012). *Delivering School Transparency in Australia: National Reporting through My School, Strong Performers and Successful Reformers in Education.* OECD Publishing.

NOTES

NOTES

CHARLES DARWIN UNIVERSITY | Northern Institute

The *Learning Communities* International Journal of Learning in Social Contexts is published by the Northern Institute, Charles Darwin University.

The journal aims to publish articles which advance our knowledge and understanding of social contexts in Australia and internationally.”