Measuring the unmeasurable: Evaluations of complex programs in the Northern Territory

NARU Public Seminar, 8 October 2009, Darwin

Presented by Dr John Guenther, Dr Allan Arnott and Associate Professor Emma Williams

Introduction

The purpose of this paper is to offer a counter to the argument that changes emerging from social programs are often unmeasurable. We offer a critique of the kinds of indicators used to evaluate outcomes of different types of programs and ask the question: 'Are these indicators a true reflection of outcomes?' The critique is based on a number of evaluations carried out by members of Charles Darwin University's Social Partnerships in Learning consortium over recent years. These evaluations have included a range of social programs funded by several government and non-government organisations. Many of the programs were designed to address health and well-being, community safety, family function, education and community capacity issues.

The paper commences with a review of relevant literature related to evaluation methodologies. It then goes on to consider traditional approaches to measurement of 'performance'. Examples are drawn from four recent Northern Territory Government departmental annual reports. Some of SPiL's recent evaluation work is then described as a backdrop to a discussion about the use of appropriate tools and indicators for complex evaluations. We conclude by suggesting that, provided outcomes are ascertained through culturally informed processes and correctly attributed to program activities, there is no reason why outcomes should be unmeasurable.
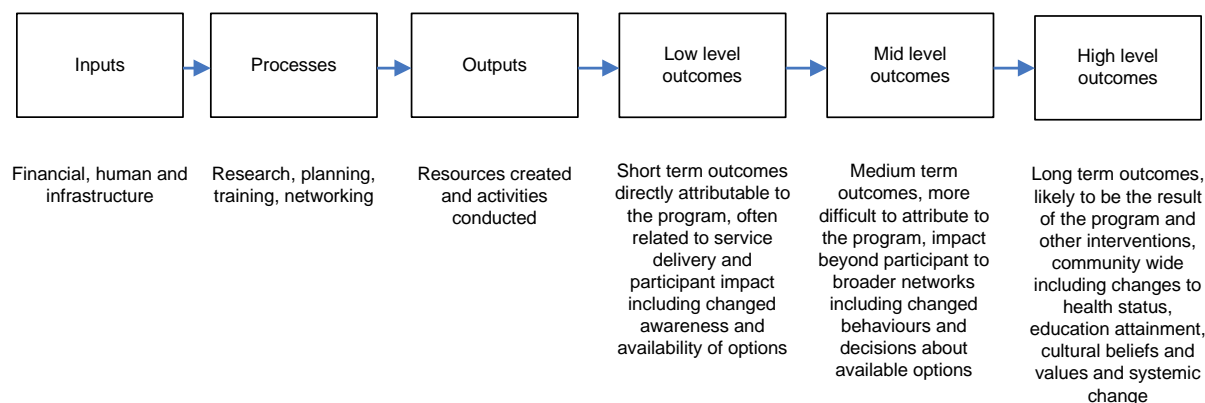
Evaluation methodologies

In this section the intent is to ground the paper in a discussion about evaluation methodologies. We first discuss the role of program logic before considering relevant literature on indicators, data and evidence. We also present a review of the literature about complicated and complex evaluations.

Program logic as a tool

Program logic may be based on 'theories of change' assumptions—emphasising the theoretical foundations of a program; or an 'outcome approach' which emphasises the causal linkages between outputs and outcomes; or an 'activities approach' which emphasises intended work plans (Patton 2002; W.K. Kellogg Foundation 2004). One of the perceived benefits of a program logic approach is that it builds an 'evidence base' (Pawson 2002). An important use of a logic model is to assist with the identification of indicators that will measure the intended outcomes. The choice of indicators is critical for determining impact. There is a tendency in some program evaluations to incorrectly ascribe process and output statements to outcomes (Mitchell 2000). Program logic as a methodology is not without its critics. There is a risk that use of program logic may prescribe an outcomes framework that ultimately is not valid. Stufflebeam and Shinkfield (2007) warn that evaluators using this approach may 'focus attention on theory developed early in the program and later discover that the program has evolved to be a quite different enterprise from what was theorised at the outset'. These are valid criticisms but as a tool for helping to identify anticipated impact, it is still a valuable construct. Hence one of the first tasks of any evaluation can be to develop a program logic model. A sample generic logic model is shown below in Figure 1. Generally, logic models would show a direct and explicit connection between outputs and low level outcomes. There should, however be reasoned assumptions underpinning the progress of impact towards high level outcomes.

Figure 1. Generic program logic model

Inputs	→	Processes	→	Outputs	→	Low level outcomes	→	Mid level outcomes	→	High level outcomes
Financial, human and infrastructure		Research, planning, training, networking		Resources created and activities conducted		Short term outcomes directly attributable to the program, often related to service delivery and participant impact including changed awareness and availability of options		Medium term outcomes, more difficult to attribute to the program, impact beyond participant to broader networks including changed behaviours and decisions about available options		Long term outcomes, likely to be the result of the program and other interventions, community wide including changes to health status, education attainment, cultural beliefs and values and systemic change

Indicators: data and evidence

Monitoring for the purpose of evaluation and reporting is frequently used as a tool for building accountability into program management. Patton (2008) suggests that while this may be a good thing, care must be taken to ensure that indicators reflect the required outcomes:

> *The potential positive contribution of performance monitoring is captured in the mantra that what gets measured gets done. Well-developed and appropriate indicators both focus attention on priority outcomes and provide accountability for achieving those outcomes. The shadow side of performance indicators is that measuring the wrong thing means the wrong thing gets done.* (p. 257)

There is sometimes a perception among program managers that data is numerical evidence. Hence, the kinds of measures frequently used for reporting purposes are largely nominal (and sometimes ordinal) in nature. A quick glance at Appendix 1 confirms this—where almost all of the measures given are represented in a count or percentage of something. Stake and Schwandt (2006) note that quality in evaluation is frequently conceptualised in terms of what is measured:

> *Among the most common measurement constructs associated with judging the quality of the provision and performance of programs and policies are values, goal attainment, effectiveness, efficiency, productivity, functions, treatments, needs, performance outcomes, units, context, input, process, product, dependent and independent variables, side-effects, program theory, program logic, and so forth... These constructs and their measurements are weighted in terms of their importance.* (p. 407)

While there is sometimes a good argument for the simple indicator as a representation of outcomes, often in complex evaluations the apparently simple can be more confusing than clarifying. Skate and Schwandt (2006)make just this point.

> *Representations oversimplify, leave out some aspects of quality in order to signify others, displace the complex with the simple, and so forth. Yet, incompleteness is less a worry than obfuscation. Some representations are just plain confusing.* (p. 414)

Indicators then, need to be carefully thought out from a variety of perspectives before any one (or a set of them) is settled on. For example, the perceptions of 'success' in an intervention can be variously interpreted depending on the point of view taken. Clients, service providers and funders may each have their own view of what success is. Hence, insufficient 'identification of the effects on different groups of program recipients will hide such differences and prevent users of the evaluation findings from considering equity issues' (Hatry and Newcomer 2004:554).

Mixed method approaches are one way of addressing these concerns. Stufflebeam and Shinkfield, in their review of evaluation approaches (Stufflebeam and Shinkfield 2007:189) suggest that it is 'almost always appropriate to consider using a mixed methods approach'.

> *Investigators look to quantitative methods for standardized, replicable findings on large datasets. They look to qualitative methods for elucidation of the program's cultural context, dynamics, meaningful patterns and themes, deviant cases, and diverse impact on individuals as well as groups.* (p. 188)

Evidence and data are not the same. Data collected for an evaluation for example, may ultimately have no meaning or utility. Glasby et al. (2007:434) suggest that 'we need to embrace a broad definition of evidence, which recognises the contribution of different sorts of knowledge to decision making'. They point out that:

> *...the challenge is not one of choosing between different sources of evidence, but of finding ways to synthesise and integrate different types of evidence in a meaningful and practical way to inform decisions...* (p. 434)

According to Glasby et al. evidence that counts for decision making should be based on: theoretical, empirical and experienced evidence. Thus, to have utility, evaluation evidence must be informed by and contribute to theory; it should say what has occurred and how outcomes are perceived. Further, the utility of the evidence must consider the cultural context in which it is both gathered and used. Arney et al (2009), commenting on utilisation of evidence in policy and practice in the Australian child and family welfare sectors acknowledge the importance of policy, practice and research cultures to this end. They omit a further important factor, which is related to the client culture. Evidence for good practice arguably should also address the culture into which interventions are implemented. Briskman (2007:149) alludes to this issue when she says that an important reason for conducting research in Indigenous contexts is to 'have voices heard that have been previously marginalised in the research literature and the public domain'. What Briskman does not say though, is that this form evidence requires some translation—not only in terms of language, but in terms of divergent worldviews. Good evidence from a policy perspective may have absolutely no utility from a local Indigenous perspective.

Evaluations: what makes them complicated or complex?
Evaluations can be divided into those that are simple, those that are complicated and those that are complex. The simplest evaluations could be said to employ a linear logic where causality follows predictably from inputs through to outcomes. Patton (2008:376) suggests that complexity occurs when there is a 'low certainty' about the outcomes that a program will achieve and 'low agreement' about how outcomes should be achieved.

Rogers (2008), following arguments presented by Glouberman and Zimmerman (2002), differentiates between simple, complicated and complex evaluations. Complicated evaluations are those where interdisciplinary and cross-jurisdictional governance structures result in more negotiation is required for agreement on evaluation parameters to occur, there are multiple and simultaneous causal strands, and different causal mechanisms occur in different contexts. Complex evaluations on the other hand are those where outcomes are achieved through non-linear feedback looks and where outcomes are emergent—and where measures cannot be determined in advance. She suggests that:

> *...it is complex interventions that present the greatest challenge for evaluation and for the utilization of evaluation, because the path to success is so variable and it cannot be articulated in advance.* (p. 31)

Rogers (2008) proposes that for complex interventions an evolving logic model may be required or alternatively 'a series of logic models can be developed alongside development of the intervention, reflecting changes in the understanding' (p. 39).

Complexity however, is not just about predicting outcomes or their causes through a single strand or simultaneous or multiple cause and effect diagrams. Evaluations are also complex because of the context. That is, depending on context, a theory of change model may work well in one context and not in another. Burton et al. (2006:307) suggest a number of context factors that contribute to complexity. These include (among others): History of previous attempts at involvement; the socio-demographic profile; the state of local voluntary and community sector; availability of resources; and timing of interventions.

Further, while at the outset, a program may be envisaged as simple, through the course of implementation, it may become complex. Mason and Barnes (2007) make the point that:

> *Programmes and projects change and develop over time (particularly where an element is participation of users in further design) and, more importantly, it may not be possible to make explicit connections between each element of the change process, no matter how detailed the initial work. It is only once data collection is under way with services, their stakeholders and users that such detail of operation and implementation becomes apparent. Thus, programme theory becomes refined while it is being explored.* (p. 159)

They go on to conclude that this refining process may cause a problem for policy makers wanting to know 'what works' for the sake of building an evidence base. They see the emergent nature of program theory to be a process of knowledge building:

> *Policy-makers should not be looking to evaluators simply to present them with evidence of 'what works', but be open to a dialogue about the way policy initiatives work in practice and to reflect on the consequences of adopting different approaches to achieving positive change.* (p. 168)

Many of the evaluations discussed in this paper are both complicated and complex. We would argue that almost all evaluations carried out in an Indigenous context (as are the SPiL evaluations described here) are necessarily complex. This is in part because of the reasons outlined in the literature above, including the contextual factors pointed out by Mason and Barnes, but more specifically because of the often disparate worldviews of the evaluands, the funding bodies, the evaluators, the auspicing bodies. This issue is raised in the context of multicultural health evaluation in California (Ngoc Nguyen et al. 2003):

> *The cultural value orientations and philosophical worldviews that evaluators bring to a project often determine the whole process of research and evaluation, including: what questions are asked, how programs are designed, what program aspects are evaluated, how effectiveness is assessed, how data are interpreted, and what results are highlighted and disseminated.* (p. 3)

We would argue that application of this understanding goes well beyond the need for 'cultural competence' in evaluation (Botcheva et al. 2009).

Programs and outcomes measures in government funded programs

We now turn our attention for a moment to the contemporary use of performance indicators and what they do and do not tell us. The discussion here is based on a review of a selection of annual reports from four Northern Territory Government departments: The Department of Health and Families (DHF), the Department of Justice (DoJ), the Department of Employment and Training (DET) and the Department of Police, Fire and Emergency Services (NTPFES). These are selected because of the SPiL team's work in programs that intersect with the four departments. The review of performance indicators shown in annual reports is undertaken because the reports are publicly available and should represent a reliable source of data. The reports also reflect the goals and objectives of the organisational units.

We pose the following questions. Do the performance indicators accurately reflect the goals and objectives of the organisational units concerned? Is there a connection between organisational outputs and their

goals/objectives? We want to examine whether the goals and objectives, as they are stated, are realistically measurable. Is a completely different set of indicators required to measure the impact of programs and their impact? The reader is referred to Appendix 1 (page 13), which tabulates a selection of performance indicators together with divisional goals and objectives.

Connections between performance indicators and goals
Table 1 is an attempt to summarise the nature of performance measures and the aims they are associated with. A more complete tabulation of indicators is given in Appendix 1. Readers should refer to this table when considering the subsequent commentary.

Table 1.   Nature of reported performance measures and corresponding divisional aims

Department	Selected divisional aims	Nature of reported performance measures
DHF	Improved health and wellbeing of those… who require acute or specialist care.	Service delivery measures related to hospital admissions
DHF	Individuals, families and young people are able to maintain social independence and overcome crises. Children are protected from harm and children's optimal development is promoted.	Service delivery measures related to children and family services as well as administration of financial support
DoJ	NT Correctional Services contributes to individual and community safety… A safe, secure and humane correctional system.	Measures related to operational capacity and offending behaviour while prisoners are in custody
NTPFES	Enhanced community safety and protection.	Measures of operational hours delivered plus measures of perceived safety and satisfaction with police service
NTPFES	Effective and efficient response services.	Measures of operational hours delivered plus measures of perceived satisfaction with police service and timeliness of response
DET	Improved educational outcomes for all students, particularly Indigenous students, in all key learning areas.	Service delivery measures related to schools and enrolments, cost of education and national benchmark achievements
DET	Maximising training opportunities for Territorians.	Service delivery measures related to hours of training provided, cost of training and conformance to QA standards

Some observations about performance indicators
We make a few brief observations about the data presented in Table 1.

- Performance indicators in annual reports are frequently summarised in terms of quality, quantity and timeliness.
- The indicators are almost always described in numerical terms (in the form of a number or percentage).
- The majority of performance indicators are related directly to service delivery measures (as they relate to quantity, cost and timeliness).
- Most of the measures used assume an uncontested connection between the indicator and the divisional aim.
- Some of the performance indicators assume a direct connection between an objective described as a strength (e.g. health, safety) and a deficit measure (hospitalisations, offending behaviours).
- The reports do not appear to suggest what the target for each performance indicator is (except where forecast measures are offered as a comparison with actual measures);
- Most indicators report from the perspective of the organisation rather than the client.

With regard to the latter point we note that NTPFES use community satisfaction measures and perceptions of safety to report on quality. While there are limitations associated with these measures they do reflect

the client perspective and therefore offer an alternative perspective that could be used to validate other measures.

In some cases the measurements appear to be almost meaningless, irrelevant or too crude to measure. For example, several measures shown in Appendix 1 are shown as either 0% or 100%. These measures would suggest that no further improvement is possible. We would suggest that a more refined indicator would be appropriate.

In other cases the objective statements used to articulate divisional aims are somewhat ambiguous and somewhat difficult to measure. For example, what does the DET objective to 'maximise training opportunities…' mean? At what point are training opportunities maximised? What is the measure of training opportunity? Is it related to training places available? Is it related to ability to meet the demands of industry?

The lack of benchmarks or standards means that the reader will never know how far the division is from achieving its stated aims. The presentation of a value such as '3 250 clients accessing crisis support services' gives little indication of what is a desirable figure. If the number of clients accessing services declines does that mean there are fewer clients or fewer available places? If the number goes up, does it indicate that the division is better able to meet demand or does it mean that demand is increasing? As a *performance* indicator, such numbers do little to indicate a) whether goals are being achieved or b) whether things are getting better or worse. Another case in point is the relationship between 'improved educational outcomes' for students and the cost of education, which DET reported was $14,499 in 2007-2008. Does this cost education performance indicator imply that as cost goes up, outcomes improve—or is it the reverse?

Improving the connection between divisional outputs and outcomes
Are the issues described here, then, a case of the objectives being too difficult to measure? Or are they more about a failure to make connections between outputs and outcomes and then to find appropriate indicators that will demonstrate those outcomes? We believe it is more the latter than the former.

The organisational goals and objectives (see Appendix 2, page 16) could well provide a starting point for a discussion about making the connections between divisional activities and their performance measures. As an example, consider the stated DHF goals:
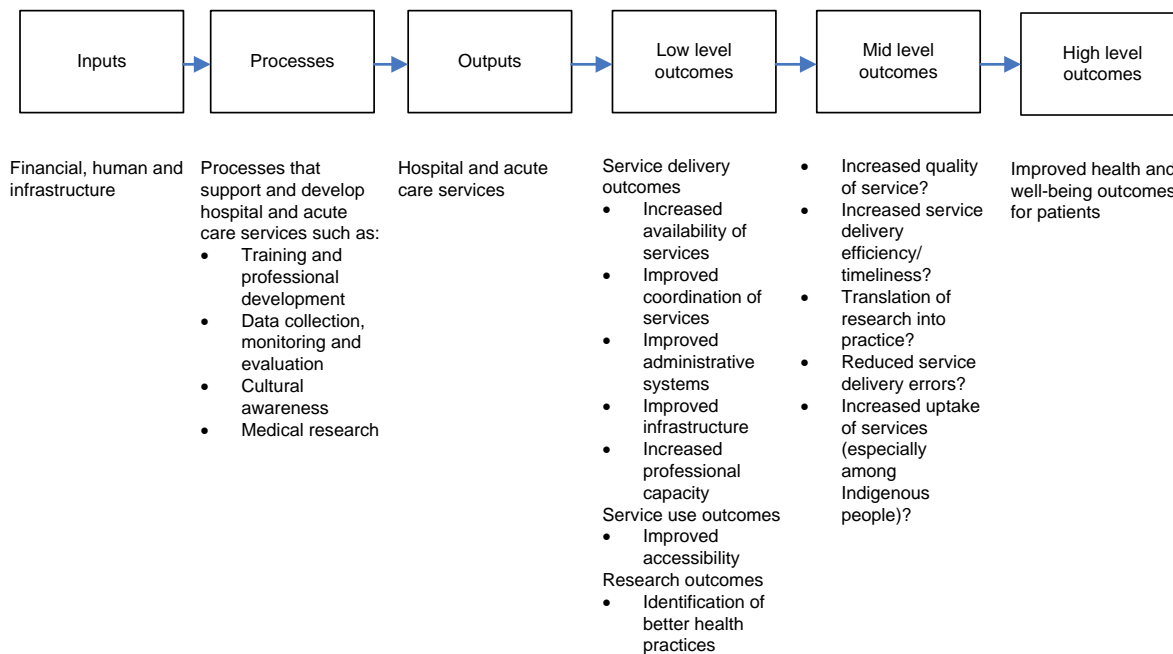
> *We promote, protect and improve the health and well-being of all Territorians in partnership with individuals and the community.*

This goal effectively represents the high level outcomes of the Department. These outcomes are clearly articulated in terms of:

- Health promotion;
- Health protection;
- Health improvement;
- Well-being improvement;
- Engagement with individuals and communities.

Taking the first of the DHF examples shown in Table 1, which are drawn from the Acute Care Division, the connection between the divisional activities and these high level outcomes can be considered. For example, the goal of achieving improved health and well-being of those in need of acute care is directly linked to the third and fourth dot points listed above. A logic modelling process may show the connection to be as follows in Figure 2:

Figure 2. Possible logic model based on DHF Acute Care activities and objectives

Inputs	Processes	Outputs	Low level outcomes	Mid level outcomes	High level outcomes

| Financial, human and infrastructure | Processes that support and develop hospital and acute care services such as:
• Training and professional development
• Data collection, monitoring and evaluation
• Cultural awareness
• Medical research | Hospital and acute care services | Service delivery outcomes
• Increased availability of services
• Improved coordination of services
• Improved administrative systems
• Improved infrastructure
• Increased professional capacity
Service use outcomes
• Improved accessibility
Research outcomes
• Identification of better health practices | • Increased quality of service?
• Increased service delivery efficiency/ timeliness?
• Translation of research into practice?
• Reduced service delivery errors?
• Increased uptake of services (especially among Indigenous people)? | Improved health and well-being outcomes for patients |

From here we could then consider indicators of the low, mid and high level outcomes. However, it is first important to identify the definitional assumptions associated with outcomes. For example, what constitutes improved health? What constitutes well-being? If patients are the object of these outcomes, how do they perceive this outcome? Having established an accepted definition of health and well-being, the question then needs to be asked: how are we going to measure it? Given too, that any number of factors could contribute to the high level outcomes, are there mid level outcomes that could act as pointers to the higher level outcomes and which can reasonably be expected to contribute to these?

Rather than simply relying on a range of service delivery output measures to measure performance we would suggest that a range of other monitoring and evaluation tools could be employed to better reflect the performance of a division as it aims to achieve its mission or objective. These would necessarily include a number of qualitative and quantitative measures that connected the division's activities to higher level outcomes. For example, we would suggest:

- Measuring the perceptions (in terms of satisfaction with level of care and awareness of available treatment options) of service users—and particularly Indigenous service users;
- Reporting service failures (not just service achievements);
- Measuring the translation of research and evaluation into practice (for example how many and what kind of research project findings are taken beyond research and applied); and
- Identifying and recording measures of quality of care, from a patient perspective (particularly from an Indigenous patient perspective).

This is by no means an exhaustive list. The point of these indicators and associated measures would be to monitor performance, not against outputs, but rather against outcomes. The issue from a reporting perspective then is not about measuring the unmeasurable, but rather about measuring outcomes using appropriate indicators. As we can see, most of the indicators used in Annual Reports do not reflect outcomes—they simply reflect activity—and are therefore a poor indication of performance.

As we have conducted complex evaluations in the Northern Territory, we have tried to come to grips with the issue of indicators that measure outcomes that are attributable to the activities of programs. The next section of this paper goes on to briefly review some of these evaluations.

Findings from evaluations of complex programs
Recent and current SPiL evaluations, including brief summary of findings and processes to date

Family Violence Partnership Program evaluations
CDU's SPiL team conducted a series of program evaluations under the banner of the Australian Government funded Family Violence Partnership Program. Nine initiatives were evaluated separately against a set of criteria developed jointly by the Northern Territory Government and SPiL in an overarching scope of work. The evaluations are briefly described in a paper presented to NARU in 2008 (Guenther 2008). The evaluators used and created a variety of data sources including:

- Customised databases for individual initiatives;
- Individual and focus group interviews;
- Existing data sources such as PROMIS (Police), IJIS (Justice), CCIS (Child Protection);
- Evaluator observations and field notes; and
- Workshops and meeting notes.

While these evaluations provided recommendations about the individual activity, a point of difference in this project was that the evaluation team were asked to prepare a report titled 'Towards an investment framework to reduce family violence in the Northern Territory' (Arnott et al. 2009). This then synthesised the combined findings of each individual evaluation.

Evaluation of Safe Places program
The introduction of 'Safe Places' in the Northern Territory arose in part out of the recommendations of the *Little Children Are Sacred Report* (Wild and Anderson 2007*)*. Construction of community-based Safe Places commenced in 2008. Most were opened in the first half of 2009. The purpose of the evaluation of the new Safe Places was primarily to inform the Northern Territory and Australian Government on the ongoing development and future potential of these facilities. Given that the facilities have so recently been opened, the primary focus of the evaluation is on processes and outputs rather than outcomes. The evaluation of the program was carried out over a short period of time under less than ideal conditions. Site visits were conducted at a number of remote locations and data was collected from operational information as well as through interviews with stakeholders.

Evaluation of Remote Aboriginal Family and Community Workers (RAFCW) program
An evaluation of the Remote Aboriginal Family and Community Workers program is currently underway. The program is designed to provide a more responsive service to families at risk in remote communities, particularly where children's welfare is a concern by employing, training and supporting local Indigenous people to work as family support workers in their own communities. The purpose of the evaluation of the RAFCW program is primarily to inform the Northern Territory Government and Australian Government on the future development of the program.

Evaluation of East Arnhem Communities for Children program
SPiL has been involved as a local evaluator for Anglicare-Northern Territory's Communities for Children program, which covers a large area of East Arnhem Land. The program's activities are delivered primarily in remote communities. We soon recognised that traditional tools to evaluate the program would be inadequate. Further, the budget allowed for the evaluation was largely inadequate for the purpose of conducting on-site field work, partly because of the cost of travel to remote communities and partly because of the time and specialist expertise required to conduct evaluations in a remote, cross-cultural context.

Evaluation of Akeyulerre Healing Centre
The Evaluation Service Plan states that:

The Evaluation will determine the effectiveness of Akeyulerre including processes, outputs and outcomes, and make recommendations for future service delivery. The evaluation will contribute to an understanding of how traditional knowledge can be used in service delivery.

In practical terms for Akeyulerre this translates into three main aims: 1) To tell a story to Government; 2) To help Akeyulerre to learn; and 3) To tell a story back to the Arrernte community for which the service is designed. SPiL is partnering with the Tangentyere Research Group (TRG) to conduct the evaluation. TRG will be collecting data from local Arrernte people using storytelling and interviewing techniques in language. SPiL will gather data from mainstream service providers.

Evaluation of the Northern Territory Differential Response Trial

The introduction of a Differential Response Framework (DRF) is one of a number of important reforms to the child protection system in the Northern Territory. The aim of the DRF is to enable a more flexible response to protective and broader child welfare concerns by engaging and facilitating greater inter-agency collaboration and supports for vulnerable children and their families through what is described as a Targeted Family Support Service. The Trial was based in Alice Springs. The evaluation developed using a 'community of practice' approach, drawing on the community's learning and input from other stakeholders, including clients. Attempts to use more traditional measurement tools largely failed for a number of reasons.

Development of evaluation tools for Families and Schools Together (FAST)

FAST (Families And Schools Together) is an eight-week, early intervention/prevention program, designed to strengthen family functioning and so build protective factors in children. The model has a built-in evaluation component based on a set of pre- and post- surveys designed to assess participant, school and coordinator perceptions. However in most remote Northern Territory communities, Indigenous people speak English as a second, third or fourth language and the cultural norms and worldview is quite different to mainstream sites where the program is often run. The program staff have recognised the weakness of the usual tools and together with a team from SPiL are developing an alternative set of tools for use in remote sites. The development process is still underway, one year after the project began.

Discussion

Managing complexity in cross-cultural domains

While not strictly about measurement, the issue of managing complexity has had a significant impact on our ability to collect data. The following discussion then is foundational for us.

One of the dilemmas faced by our team relates to the problem of differing worldviews. Translational issues arise when evaluation strategies are developed in one cultural domain and then administered in a different cultural domain. Concepts and ideas sometimes do not translate well from the mainstream to remote Indigenous contexts—or vice versa. A good example of this arose in our recent evaluation of Safe Places. On the one hand the documented purpose for which Safe Places were funded, was to reduce the impact of violence in communities. However, when we asked remote community respondents about this, some respondents talked about the importance of cultural maintenance instead.

Strong partnerships with locally-based organisations have been a prerequisite for many of the evaluations we have conducted. One of the advantages of working with an organisation in a location where we were effectively foreigners was that it gave us leverage to work in that community. It has made introductions easier: 'we are working with Organisation A on an evaluation…'. This then gave us an entree in that community. The partnerships we are talking about here are not necessarily formalised in any way. Nor are we necessarily talking about evaluation partners (though that helps too sometimes).

Where possible we encourage program staff to capture data, particularly as it relates to activities. Having local staff collect data can be a time consuming process that requires additional support not usually required when the researchers are supposedly in control. However, if locally based staff can be persuaded to see data collection and recording as an important and valuable task, the chances of capturing relevant

and accurate, meaningful data are much higher than would otherwise be the case. Local staff are in some ways the keys to measuring the unmeasurable. Particularly with remote Indigenous staff, SPiL members believe that it is an important part of our role to empower these workers by giving them the tools and skills to document and be involved in assessing their own work.

We have long recognised that a significant amount of time is required for relationships with partners to develop. One of the reasons that indicators are unmeasurable is that the holders of the data are reluctant to release that information. Trust is therefore required. This takes considerable time to build. In one evaluation we are currently working on, it has taken us the best part of a full year to get to this point. The organisations involved, having seen us stick with them, are now much more willing to share information with us than they were in the early part of the relationship.

When is an indicator not an indicator?

As noted in the earlier review of annual reports a number of indicators used to measure performance have very little to do with measuring outcomes. In order to have some utility for evaluation purposes an indicator of program impact needs to reflect the intended outcomes of the program. This is where the logic model comes in. The theory of change expressed in the model should then direct the choice of indicators. Some of these outcomes would typically include:

- Increased availability of services;
- Increased awareness of services among clients;
- Better networks among service providers;
- Improved accessibility to services among clients; and
- Increased organisational capacity.

Note then, that these outcomes are all premised on changes occurring, either among clients or organisations delivering services. The indicator used to reflect these outcomes must then show the changes that have occurred. While some of the outcomes imply a quantitative change, such as increased availability of services (which could for example be measured by reporting on the number of discrete services), most are largely qualitative in nature, such as awareness, networks and accessibility. These cannot be measured with numbers. 'Better networks' for example is defined by the users of networks, so it the only way to determine whether networks are better is to ask those involved in the networks. Similarly, the nature of awareness change can only be determined by asking service users (and potential service users) how their awareness has changed—if it has.

So when is an indicator not an indicator? We would suggest that an outcomes indicator is not an indicator if:

- It does not measure the outcome it is intended to or it measures output/activity rather than outcome;
- It cannot be measured;
- It cannot reflect ongoing change;
- Its interpretation is ambiguous; or
- At low level other factors will contribute to the change anticipated.

Therefore, we believe that the following are not good indicators, for example:

- Measure the number of participants attending a program (rather than what it does to the participants);
- Measure the opposite of an outcome (e.g. ill-health in a health promotion program);
- Measure an outcome that is the product of multiple interventions (e.g. school attendance in an education program where there are other activities designed to encourage school attendance);

- Record an outcome that could reasonably be achieved without the intervention (e.g. when 100% of something is achieved); and
- Measure something that is not likely to change or is too crude a measure to show change in the time frame provided.

Sometimes we find that indicators need to be revisited to make them more appropriate for the program.

Evaluation tools and approaches to fit the complex evaluation context
We have experimented with a number of different evaluation tools and processes to fit the various contexts of evaluations. These tools have been designed to capture outcomes as well as activities. We are well aware of the shortcomings of many tools and instruments we have tried

Community of practice approaches
In two recent evaluation projects we have employed a 'community of practice' (Wenger 1998) approach. One project was an evaluation within the East Arnhem Communities for Children program for the Aboriginal Resource and Development Service, based in Darwin, where the focus was to test a workable internal evaluation process that was consistent with their 'discovery education' process. The other was for an evaluation of the Northern Territory DHF Differential Response Framework trial, which was designed to offer alternatives to Child Protection services through a 'Targeted Family Support Service'. In both cases the evaluation findings emerged to a large extent from the learnings of the work teams.

Story gathering processes
A project we are working on at the moment is an evaluation of the Akeyulerre Healing Centre in Alice Springs. The Centre is built around traditional healing for Arrernte families. The evaluation is funded by DHF. We grappled with how to capture the meaning of healing in that context and felt that the best way of capturing the Arrernte meaning was to partner with Arrernte researchers from the Tangentyere Research Group (TRG) to gather stories about the outcomes of the activities, from an Arrernte perspective. SPiL's role in the evaluation will be to gather stories from mainstream service providers. The evaluation outputs will be produced as a partnership between TRG and SPiL.

Reflective practice tools
In the Raypirri Rom (Gove Peninsula) project we evaluated as part of the FVPP suite of evaluations, the work team engaged in a series of critical reflection activities to identify strengths and weaknesses. The exercise was repeated over time to assess the growth of individuals and their progress towards overcoming barriers and applying skills to problem solving.

Risk assessment tools
Another tool, developed by the Raypirri Rom team, was an incident risk assessment tool. Raypirri workers marked on an ordinal risk scale, the perceived level of risk before and after an intervention. Analysis of the data showed the relative change in risk level that had resulted from the team's work.

Community perception tools
One tool that has been used by our team is one we call a 'thermometer' tool. Thermometer stylised gauges are used in a community or group context to assess perceptions about the nature and extent of a particular problem. For each issue, the thermometer scale is graded from 'small' to 'big' and after the group has discussed the issue, consensus is reached about where on the scale, the issue fits. The measure can then be repeated over time to gauge changes in perception over time.

Culturally sensitive visual tools
In a project sponsored by Families and Schools Together (FAST) members of the SPiL team have been developing an instrument for use in remote Indigenous contexts. The instrument will ultimately replace a complex psychometric tool that assesses changes in participant perceptions and should be adaptable to a number of remote contexts with minor adaptations. The FAST evaluation instrument uses visual and tactile 'feeling cards' to assess emotional response to a number of relational scenarios. The tool also uses a felt-

board activity to assess the strength of social relationships. The tool will be used before and after an eight week program to determine the extent to which the program has resulted in change.

Customised databases

In two projects under the FVPP evaluations (Safe Families based in Alice Springs and Raypirri Rom), members of the SPiL team supported the local program staff by developing customised databases to collect and summarise data. We also provided training and support to ensure that local staff could use the tools. At the end of the program the collected data was returned to us for analysis and this was used to present findings about outcomes from the projects.

Summary

The purpose of the foregoing discussion about tools and approaches has been to demonstrate the wide variety of instruments available for use in Indigenous contexts. As evaluators we have found that there is any number of ways of measuring the unmeasurable. The approaches described seldom employ quantitative assessment of outcomes that show a percentage increase or decrease in a particular indicator. They do however demonstrate outcomes and they are based on sound and arguably reliable sources of evidence.

Conclusion

This paper set out to offer a counter to the argument that changes emerging from social programs are often unmeasurable. The notion of unmeasurable outcomes arose for the SPiL evaluation team out of both an examination of performance indicators required of government funded programs and a perception among service providers that the kinds of outcomes they see from their activities are largely unmeasurable. Many perceive that outcomes and change can only be measured with numbers.

This view is reinforced when performance indicators of Northern Territory Government annual reports are examined. What we find from these is that mostly these reports measure performance as indicators of activities, rather than outcomes. Further, the performance indicators are often crude and cannot possibly hope to measure change over time. There is little if any commentary in the reports about what the numbers actually mean.

SPiL's evaluations of complex programs in the Northern Territory have shown that those apparently unmeasurable outcomes can indeed be measured. However, the measures are not necessarily numerical. Similarly the tools that we have used or are trialling are not the traditional, tick the box, closed question, survey forms. They are designed to work for the context and take into account the complexity of programs delivered into cross-cultural situations. We acknowledge that the tools we have developed do not always work. However, we feel that this is no reason to stop trying to develop culturally appropriate tools that effectively measure the impact of programs.

Further information

Cat Conatus
http://www.catconatus.com.au/res.html

CDU Social Partnerships in Learning consortium
http://www.cdu.edu.au/centres/spil/

John Guenther
john@catconatus.com.au
0412 125661

Allan Arnott
allan.arnott@cdu.edu.au
0448 686 953

Emma Williams
emma.williams@cdu.edu.au
0412 283 268

Appendix 1: Selected performance indicators and measures from Northern Territory Government departmental annual reports

Table 2. Selected performance indicators and their measures, taken from Northern Territory Government department annual reports (2007-2008)

Northern Territory Government Department	Division	Aim	Selected performance indicators	Measure 2007-2008
Department of Health and Families	Acute Care	"Improved health and wellbeing of those in the Northern Territory community who require acute or specialist care" (Department of Health and Families 2008:57)	Hospital separations Average length of stay Elective surgery waiting list admissions Emergency procedures Beds accredited by the Australian Council on Healthcare Standards Elective surgery waiting times Category 1: admission within 30 days	99 884 5.6 days 6 061 7 727 100% 80%
Department of Health and Families	Community Services	"Individuals, families and young people are able to maintain social independence and overcome crises. Children are protected from harm and children's optimal development is promoted" (Department of Health and Families 2008:75).	Subsidised child care places Child care subsidy payments made on time Clients accessing crisis support services Days of crisis support Notifications of child harm Child protection assessments investigated Children in care during the year Proportion of children on a care and protection order who had one or two placements leaving care after less than 12 months (quality) Investigations of reports commenced Category 1: within 24 hours of notification (timeliness) Supported accommodation places HACC urban services reviewed against service Standards CSTDA urban services reviewed against service Standards Pensioner concession recipients Applicant able to access pensioner concessions within 14 days Individuals receiving community-based public mental health services Public mental health services accredited Community education and community development activities Utilisation rate of sobering up shelter bed hours	4 082 95% 3 250 170 260 3 678 1 841 804 91% 73% 135 0% 0% 20 206 100% 4 737 100% 249 31%
Department of Health and Families	Health Services	"The burden of ill health in the community and the need for hospitalisation are reduced" (p. 105).	Funded government managed rural community health centres Resident child population <five years participating in Growth Assessment and Action Program Proportion of Aboriginal babies born with low birth weight Proportion of screened Aboriginal children less than 5 years who are underweight	52 3170 13.1% 12.4%

Northern Territory Government Department	Division	Aim	Selected performance indicators	Measure 2007-2008
Department of Justice	Northern Territory Correctional Services Custodial Services	"NT Correctional Services contributes to individual and community safety…" "A safe, secure and humane correctional system" (Department of Justice 2008:57)	Quantity: Prisoner utilisation (operational capacity) Quantity: Daily average number of prisoners Quality: Participation in education Quality: Rate of assaults of prisoner on prisoner Quality: Rate of assaults of prisoner on officer Timeliness: Sentences completed in accordance with conditions of order	103% 875 Not Available 3% 0% 100%
	Community Corrections		Quantity: Offenders under supervision (daily average) Quantity: Order commencements Quality: Successful completion of Community Corrections orders Timeliness: Pre-sentence report and Parole Board reports completed on time	1313 1083 62% 100%
Northern Territory Department of Police, Fire and Emergency Services	Community Safety and Crime Prevention	"Enhanced community safety and protection (NT Police Fire and Emergency Services 2008:21)"	Police hours Community safety and crime prevention programs delivered People aged 15 years or over who felt "safe" or "very safe" at home alone during the day People aged 15 years or over who felt "safe" or "very safe" at home alone after dark People aged 15 years or over who said they were "satisfied" or "very satisfied" with police services	597 765 5091 90% 75% 59%
Northern Territory Department of Police, Fire and Emergency Services	Response and Recovery Services	"Effective and efficient response services" (NT Police Fire and Emergency Services 2008:24)	Police hours People who were 'satisfied' or 'very satisfied' with police in their most recent contact. Proportion of 000 calls answered within 10 seconds Proportion of other general calls answered within 20 seconds Proportion of incidents where police are dispatched within 10 minutes	344 170 79% 80.5% 77.3% 82%
Northern Territory Department of Police, Fire and Emergency Services	Investigations	"Effective and efficient investigation services" (NT Police Fire and Emergency Services 2008:)	Quantity: Police hours Quality/Timeliness: Outcomes of investigations within 30 days Property crime Crimes against the person	526056 22% 61%

Northern Territory Government Department	Division	Aim	Selected performance indicators	Measure 2007-2008
Northern Territory Department of Police, Fire and Emergency Services	Road Safety Services	"Environment that encourages road users to behave safely and lawfully" (NT Police Fire and Emergency Services 2008:36)	Police hours Vehicles passing a speed camera checkpoint Drivers breath tested People who had driven in the previous 6 months without wearing a seatbelt People who had driven in the previous 6 months when possibly over the 0.05 alcohol limit People had driven in the previous 6 months more than 10km/h above the speed limit Proportion of infringements detected by speed cameras Proportion of drivers breath tested who were detected for drink driving offences Proportion of incidents where police are dispatched within 10 minutes	125261 892227 98035 13% 16% 60% 1.7% 3.7% 81%
Department of Employment, Education and Training	Government Education	"Improved educational outcomes for all students, particularly Indigenous students, in all key learning areas" (Northern Territory Department of Employment Education and Training 2008:25).	Quantity: Total primary school student enrolments Quantity: Indigenous primary student enrolments Quantity: Schools providing primary education Timeliness: Primary education delivered during four terms Cost: Average cost per student Quality: Non-Indigenous students achieving national numeracy benchmark: Year 3 Year 5 Year 7 Quality: Indigenous students achieving national numeracy benchmark: Year 3 Year 5 Year 7	20 445 9 504 136 100% $14 499 95% 91% 87% 53% 34% 29%
Department of Employment, Education and Training	Training	"Maximising training opportunities for Territorians" (Northern Territory Department of Employment Education and Training 2008:25).	Quantity: Annual hours of curriculum delivered (million) Quality: Registered training organisations' compliance with Australian Quality Training Framework (audit) Quality: Level of invalid student unit enrolments (audit) Quality: Successful unit enrolment completions Timeliness: Resource agreements issued and monitored within appropriate timeframe Timeliness: Agreed timeframes met for submission of information to national agencies Cost: Average cost per hour of annual hours of curriculum	3.8 100% 2.38% 73% 95% 100% $16.00

Appendix 2: Selected Northern Territory Government department mission statements

Northern Territory Department of Health and Families Mission Statement

We promote, protect and improve the health and well-being of all Territorians in partnership with individuals and the community (Department of Health and Families 2008:12).

Northern Territory Department of Justice Objectives

- Informed government decision making
- A workforce that is representative of the community
- Timely and effective implementation of the Government's policy agenda
- Improved understanding of the justice system and people's rights and responsibilities
- A society that is non-discriminatory and respects people's rights
- A reduction in alcohol and substance abuse
- A reduction in anti-social behaviour and violence
- An effective criminal justice system and civil dispute resolution system
- A safe, humane and responsible corrections system that reduces re-offending
- High level of community confidence in regulatory, registration and trustee services

(Department of Justice 2008:10)

Northern Territory Police, Fires and Emergency Services Mission

Working in partnership to reduce crime, protect the community from fire and other emergencies and enhance community confidence. (NT Police Fire and Emergency Services 2008:7)

Northern Territory Department of Employment, Education and Training Objective statement

Develop employment and training initiatives to enhance the social and economic prosperity of the Northern Territory, develop Northern Territory students through preschool, primary, secondary and vocational education and training (VET) programs and manage work health programs to ensure that people in the Territory have a safe working environment (Northern Territory Department of Employment Education and Training 2008:25).

References

Arney, F. M., Bromfield, L. M., Lewig, K. and Holzer, P. 2009. "Integrating strategies for delivering evidence-informed practice." *Evidence & Policy: A Journal of Research, Debate and Practice* **5**: 179-191.

Arnott, A., Guenther, J. and Williams, E. 2009. *Towards an Investment Framework to reduce Family Violence in the Northern Territory*. Social Partnerships in Larning Consortium, Charles Darwin University. Darwin.

Botcheva, L., Shih, J. and Huffman, L. C. 2009. "Emphasizing Cultural Competence in Evaluation: A Process-Oriented Approach." *American Journal of Evaluation* **30**(2): 176-188.

Briskman, L. 2007. *Social work with Indigenous communities*. Sydney, The Federation Press.

Burton, P., Goodlad, R. and Croft, J. 2006. "How Would We Know What Works?: Context and Complexity in the Evaluation of Community Involvement." *Evaluation* **12**(3): 294-312.

Department of Health and Families 2008. *Annual Report 2007-2008*. DHF, Retrieved September 2009, from http://www.nt.gov.au/health/Publications/Corporate_Publication.

Department of Justice 2008. *Annual Report 2007-2008*. DoJ, Retrieved September 2009, from http://www.nt.gov.au/justice/docs/depart/annualreports/ADC%20Annual%20Rpt%2007-08.pdf.

Glasby, J., Walshe, K. and Harvey, G. 2007. "Making evidence fit for purpose in decision making: a case study of the hospital discharge of older people." *Evidence & Policy: A Journal of Research, Debate and Practice* **3**: 425-437.

Glouberman, S. and Zimmerman, B. 2002. *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?* Romanow Papers: Changing health care in Canada. Commission on the Future of Health Care in Canada Discussion Paper 8, July 2002, Retrieved October 2009, from http://www.healthandeverything.org/files/Glouberman_E.pdf.

Guenther, J. 2008. *The role of the evaluator in remote contexts*. Public Seminar Series. North Australia Research Unit. Darwin, 27 March 2008, Retrieved June 2009, from http://www.catconatus.com.au/docs/080327_NARU_role_of_evaluator.pdf.

Hatry, H. and Newcomer, K. 2004. Pitfalls of evaluation. *Handbook of Practical Program Evaluation*. J. S. Wholey, H. P. Hatry and K. E. Newcomer. San Francisco, Jossey-Bass**:** 547-570.

Mason, P. and Barnes, M. 2007. "Constructing Theories of Change: Methods and Sources." *Evaluation* **13**(2): 151-170.

Mitchell, P. 2000. *Building capacity for life promotion: Technical report volume 1*. Australian Institute of Family Studies, Evaluation of the National Youth Suicide Prevention Strategy, Retrieved March 2006, from http://www.aifs.gov.au/ysp/yspevaluation/evalrep2chap1.pdf.

Ngoc Nguyen, T., Kagawa-Singer, M. and Kar, S. 2003. *Multicultural Health Evaluation: Literature Review and Critique*. UCLA School of Public Health, Social Policy Research Associates,. Los Angeles, Fall 2003, Retrieved October 2009, from http://www.calendow.org/Collection_Publications.aspx?coll_id=74&ItemID=416#.

Northern Territory Department of Employment Education and Training 2008. *Annual Report 2007-2008*, Retrieved October 2009, from http://www.det.nt.gov.au/__data/assets/pdf_file/0018/4635/AnnualReport.pdf.

NT Police Fire and Emergency Services 2008. *2008 Annual Report*. NTPFES, Retrieved September 2009, from http://www.nt.gov.au/pfes/documents/File/police/publications/annrep/AR_0708.pdf.

Patton, M. 2002. *Qualitative research and evaluation methods*. Thousand Oaks, CA., Sage.

Patton, M. 2008. *Utilization-Focused Evaluation*. Thousand Oaks, Sage Publications.

Pawson, R. 2002. "Evidence-based Policy: The Promise of 'Realist Synthesis'." *Evaluation* **8**(3): 340-358.

Rogers, P. J. 2008. "Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions." *Evaluation* **14**(1): 29-48.

Stake, R. and Schwandt, T. 2006. On Discerning Quality in Evaluation. *The Sage Handbook of Evaluation*. I. Shaw, J. Greene and M. Mark. London, Sage Publications Ltd.**:** 404-418.

Stufflebeam, D. and Shinkfield, A. 2007. *Evaluation Theory, Models & Applications*. San Francisco, Jossey-Bass.

W.K. Kellogg Foundation. 2004. *Logic Model Development Guide*.   Retrieved August 2005, from http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf.

Wenger, E. 1998. *Communities of Practice. Learning, meaning and identity*. Cambridge, Cambridge University Press.

Wild, R. and Anderson, P. 2007. *Ampe Akelyernemane Meke Mekarle: "Little Children are Sacred"*. Board of Inquiry into the Protection of Aboriginal Children from Sexual Abuse. Darwin, Report of the Northern Territory Board of Inquiry into the Protection of Aboriginal Children from Sexual Abuse, from http://www.nt.gov.au/dcm/inquirysaac/pdf/bipacsa_final_report.pdf.