Measuring the unmeasurable: Dealing with complex and complicated evaluations

Australasian Evaluation Society, Hobart, 16 July 2010

Presented by Dr John Guenther

1 Introduction

The purpose of this paper is to offer a counter to the argument that changes emerging from social programs are often unmeasurable. I offer a critique of the kinds of indicators used to evaluate outcomes of different types of programs and asks the question: 'Are these indicators a true reflection of outcomes?' The critique is based on a number of evaluations carried out by the author and his colleagues 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 Tasmanian Government departmental annual reports. Some of the author's recent evaluation experience is then described as a backdrop to a discussion about the use of appropriate tools and indicators for complex evaluations. The author concludes by suggesting that, provided outcomes are correctly attributed to program activities, there is no reason why outcomes should be unmeasurable.

I acknowledge the support of others I have worked with on evaluations discussed here: Dr Allan Arnott of Charles Darwin University, Darwin and Ms Emma Williams of Maburra Consulting, based in Canberra. The paper is based on an earlier presentation built more specifically around complex evaluations in the Northern Territory (see Guenther et al. 2009).
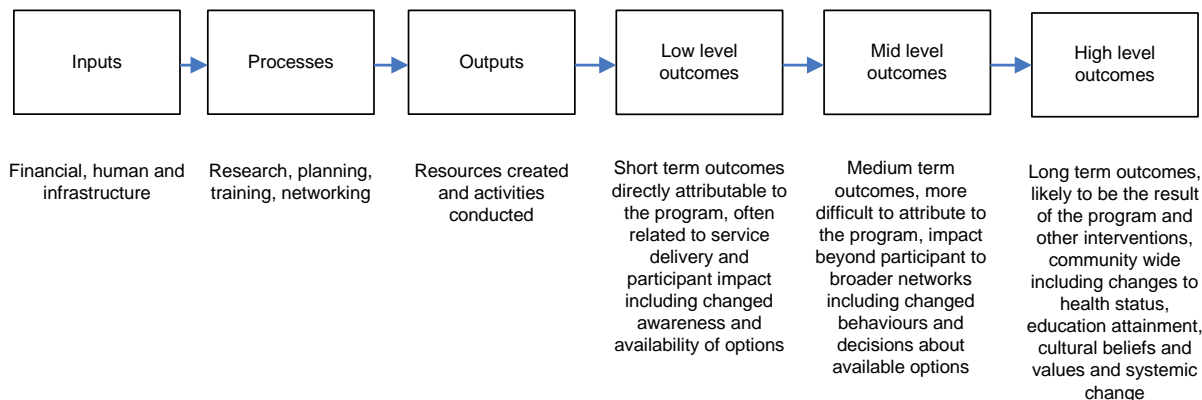
2 Evaluation methodologies

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

2.1 *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 2002:162-164; 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

2.2   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 numbers and percentages. 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)*

The task of 'quantitizing' (Tashakkori and Teddlie 1998) qualitative data may at first seem abhorrent to qualitative researchers, but it is now an accepted way of increasing the validity and improving the generalizability of data (Falk and Guenther 2007; Franzosi 2010).

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 the translation of 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.

2.3   *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 argument 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 being required for agreement on evaluation parameters. There are multiple and simultaneous causal strands where different causal mechanisms occur in different contexts. Complex evaluations on the other hand are those where outcomes are achieved through non-linear feedback loops 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 the 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. I would argue that almost all evaluations carried out in an Indigenous context (as are some 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 those being evaluated, the funding bodies, the evaluators and 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)

I would argue that application of this understanding goes well beyond the need for 'cultural competence' in evaluation (Botcheva et al. 2009). I would also suggest that the sub-cultural differences that may be observed within a culture are also significant for evaluation and add a degree of complexity to the process. Cultural competence in this context means that the subtle differences between the culture of the evaluator and the evaluand must not be assumed to be negligible. Indeed evaluators need to recognise the distinctives of their own culture before making assumptions about another's.

3   Programs and outcomes measures in government funded programs

The paper now turns its 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 Tasmanian Government departments: The Department of Health and Human Services (DHHS), the Department of Justice (DoJ), the Department of Education (DoE) and the Department of Police and Emergency Management (DPEM). These are selected because of work carried out 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.

I 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. I 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.

3.1   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/output group aims**

Department	Selected divisional/output group aims	Nature of reported performance measures
DHHS	Developing responsive, accessible and sustainable services.	Service delivery and service use measures related to hospital admissions, waiting times, length of stay. Some quality indicators and one cost measure.
DHHS	Supporting individuals, families and communities to have more control over what matters to them	Service delivery measures related to children and family services as well as some outcome measures related to quality service delivery
DoJ	…to contribute to a safer Tasmania by ensuring the safe, secure containment of prisoners and to provide them with opportunities for rehabilitation, personal development and community engagement	Measures related to operational capacity, escape rates, cost and service uptake while prisoners are in custody. Outcomes reported in terms of prisoner employment only.
DPEM	…a reduction in crime.	Measures of operational hours delivered plus measures of perceived safety and satisfaction with police service
DPEM	…a community where people are safe.	Measures of operational hours delivered plus measures of perceived satisfaction with police service and timeliness of response
DoE	…every student has the opportunity to learn and achieve his or her potential	Service delivery measures related to schools and enrolments, cost of education and national benchmark achievements
DoE	The state is committed to providing meaningful pathways and enhanced facilities to encourage Year 10 students to stay on to further education and training.	Service delivery measures related to hours of training provided, cost of training and conformance to QA standards

3.2   Some observations about performance indicators

I make a few brief observations about the data presented in Table 1.

- Performance indicators in annual reports are frequently summarised in terms of service delivery and in terms of capacity to deliver and/or uptake of service.
- The indicators are almost always described in numerical terms (in the form of an amount or percentage).
- The majority of performance indicators are related directly to service delivery measures (as they relate to quantity and cost).
- 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 can see that the DPEM 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 other cases the objective statements used to articulate divisional or output group aims are somewhat ambiguous and somewhat difficult to measure. In fact, in some cases a clear articulation of goals and objectives is so obfuscated by the rhetoric of 'directions' or 'strategy' that it is almost impossible to see from the annual reports, exactly what the objectives are. A case in point is the DoE Annual Report for 2008-2009 which does not mention output group objectives for pre-, compulsory or post-compulsory education, except in relation to Tasmania Together goals.

The reports tend to intersperse internal performance indicators with Tasmania Together goals. Readers are therefore led to believe that divisional activities are aligned to Tasmania Together goals. The DHHS report for example goes to great lengths to demonstrate the apparent links between the goals and divisional activities. The problem though is that other Tasmanian and Australian Government programs (some of which are explicitly described) are also arguably responsible for outcomes related to the goals. This makes it nigh on impossible to attribute success to individual initiatives. The assumptions behind the linkages are rarely explicitly stated.

The lack of internal 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 the DHHS '805 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. A case in point is the DoJ indicator relating to the cost of prison services. What does it mean when the cost of prison services is $263 per day? Does an increase mean that the goal of achieving a safer Tasmania is being achieved or does a decrease mean simply that the quality of service is getting worse?

3.3 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? My colleagues and I 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 DHHS goals where the mission is to 'design and implement a sustainable, people-focused health and human services system which supports individuals and communities to be active partners in the management of their own health and wellbeing'. (Department of Health and Human Services 2009b:3)

The strategic objectives that support this goal are:

1. Supporting individuals, families and communities to have more control over what matters to them.

2. Promoting health and wellbeing and intervening early when needed.

3. Developing responsive, accessible and sustainable services.

4. Creating collaborative partnerships to support the development of healthier communities.

5. Shaping our workforce to be capable of meeting changing needs and future requirements.

Taking the first of the DHHS examples shown in Table 2, which is drawn from the Acute Hospitals 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 DHHS Acute Hospitals activities and objectives**
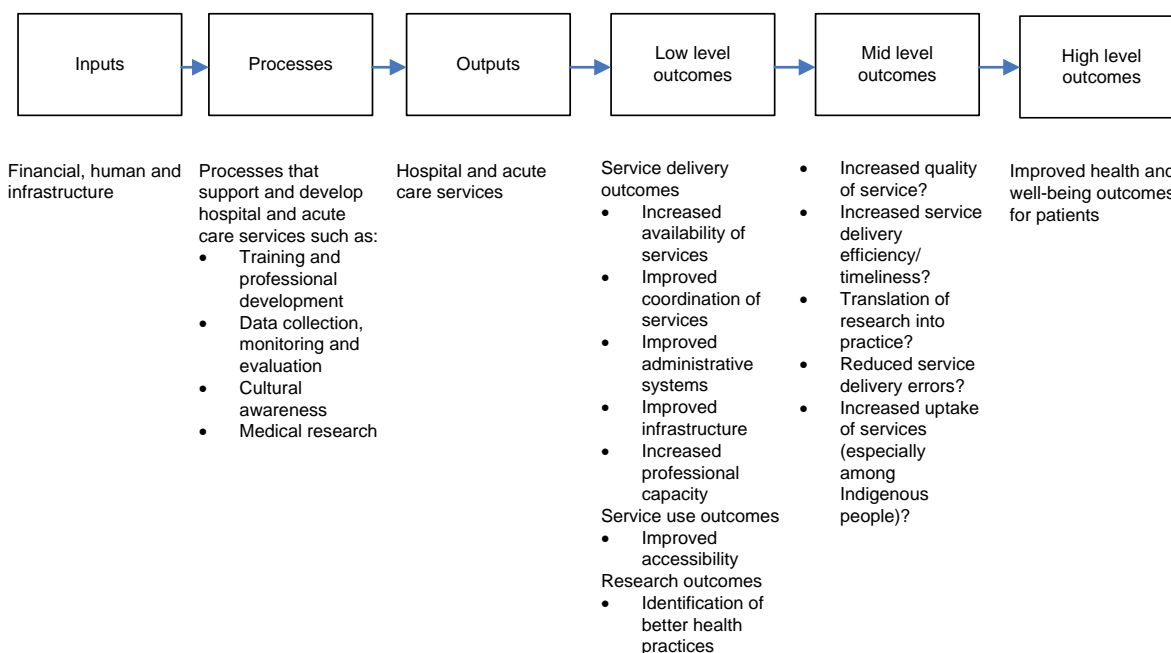
From here we could then consider indicators of the low, mid and high level outcomes. However, it is first important to identify the definitional assumptions associated with outcomes. For example, what constitutes improved health? What constitutes well-being? If patients are the object of these outcomes, 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 it is suggested 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 it could be helpful to:

- Measure the perceptions (in terms of satisfaction with level of care and awareness of available treatment options) of service users—and particularly more vulnerable service users;
- Report service failures (not just service achievements);
- Measure 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
- Identify and recording measures of quality of care, from a patient perspective (particularly from a vulnerable 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 my colleagues and I have conducted complex evaluations in the Northern Territory and Tasmania, 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.

4 Findings from evaluations of complex programs

This section very briefly describes an abridged list of evaluations conducted by my colleagues and I over the last three to four years. Each evaluation has elements of complexity that were discussed earlier (see Evaluations: what makes them complicated or complex?, page 3).

4.1 *Family Violence Partnership Program evaluations*

Between 2006 and 2009 a series of program evaluations were conducted in conjunction with CDU 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 CDU in an overarching scope of work. The evaluations are briefly described in a paper presented to the the Australian National University's North Australian Research Unit 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 Northern Territory Government data sources such as PROMIS (Police), IJIS (Justice), CCIS (Health and Families);
- 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. This suite of evaluations was inherently complex, partly because of its multifaceted nature with learnings from nine individual activities being synthesised into one final report. The formative nature of the evaluations also contributed to complexity because of the uncertainty of outcomes. Most of the activities were conducted in remote and/or cross cultural contexts.

4.2   Evaluation of Centacare's Weathering Family Separation program

Centacare, in northwest Tasmania, have a focus on addressing the needs of families and children. One program designed with the needs of children in mind is called 'Weathering Family Separation'. It is designed to help separating parents focus on the needs of their children. The program itself is fairly straightforward, and includes one to one support and a group workshop, with follow-up as required. Conventional approaches to evaluation would suggest that a simple summative evaluation of the program and its elements would be required to assess outcomes, and indeed in an earlier version of the program this is exactly what was done (Webb 2007). With the Centacare evaluation it was recognised early on that the single instrument approach would not yield the kind of responses expected. Further, assent to true/false questions to determine responses that were more or less conducive to a child focussed outcome would not yield the kind of data required. Instead, a pre- and post- instrument was designed for use at the point of enrolment and after completion of the workshop. The responses sought from participants were designed to be affective rather than cognitive. In other words, we wanted to know how the program impacted on the feelings of individuals rather than their thoughts. While the responses sought were qualitative (with respondents able to choose from a range of emotions) the analysis was quantitative.

In conjunction with this, a follow-up telephone interview was planned at six months after the program to identify what changes had taken place for individuals as a result of the program. The analysis of this data required a combination of quantitative and qualitative techniques which informed the findings of the pre-post survey. The complexity in this evaluation lies in the uncertainty of outcomes—as the findings emerge so does the delivery of the program.

4.3   Evaluation of Burnie Communities for Children program

Communities for Children is a national program funded by the Department of Families, Housing, Communities and Indigenous Affairs. The program is designed to meet the needs of vulnerable families with children up to the age of 12, and deliver on outcomes in four prescribed outcome areas. The Burnie Communities for Children program has been running for five years and has had many activities built in. Each element is required to include a 'local evaluation' which is designed to produce evidence to support the achievement of the prescribed outcomes. Bringing these elements together into a coherent evaluation is in itself a challenge and one of the aspects of complexity in the program is that the integrated nature of the program makes it difficult to attribute outcomes to a single activity. A case in point is the activity called the Mobile Family Resource Van, which was essentially designed as a health promotion activity and designed to connect with other activities. It did so very effectively but evaluating it according to the originally anticipated outcomes then became problematic. Rather than focussing on tangible outcomes of changed behaviour the final evaluation report concluded that the real benefit of the program lay in the networks that were developed (Guenther and Arvier 2009). Here, we see that expected outcomes were replaced with alternatives.

4.4   Evaluation of Akeyulerre Healing Centre

Akeyulerre is an Aboriginal healing centre based in Alice Springs, built on an understanding of Arrernte traditional healing. 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. An evaluation team was built around academic researchers from Charles Darwin University (of which the author was part) and the Tangentyere Research Group (TRG) to conduct the evaluation. TRG collected data from local Arrernte people using storytelling and interviewing techniques in language. The CDU team gathered data from mainstream service providers. This evaluation is a good example of complexity created by cross-cultural context (Guenther et al. 2010).

4.5 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 (Communities and Families Clearinghouse Australia 2009; Guenther and Boonstra 2009). 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 of remote communities 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 of evaluators worked for 18 months to develop an alternative set of tools for use in remote sites. The complexity in this evaluation relates both to the cross-cultural context and to the uncertainty of outcomes.

5 Discussion

5.1 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 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? I 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, when a performance indicator is used to 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.

—then it is not a valid indicator of performance. As evaluators, my colleagues and I found that sometimes indicators need to be revisited to make them more appropriate for the program. In some cases we have managed to do this effectively early on in the life of the program before evaluation took place.

5.2   *Evaluation tools to fit the complex evaluation context*

The evaluation teams I have been involved with 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. Tools described here are drawn from an array of recent program evaluations, some of which are not explicitly described above.

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. The tool itself can take many forms and is context dependent.

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.

Tools that measure affective rather than cognitive response

In the Centacare evaluation discussed earlier, the need was to measure affective as well as cognitive response to the program. The tools used have effectively 'quantitized' a qualitative emotional response to assess the extent to which a program has affected participant feelings as they relate to their children and the child's other parent. Initial findings suggest that the program has indeed influenced a range of feelings, though we have picked up a difference between males and females.

Culturally sensitive visual tools

In a project sponsored by Families and Schools Together (FAST) a consultancy project has developed an instrument for use in remote Indigenous contexts. The instrument is designed to 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 is being used before and after an eight week program to determine the extent to which the program has resulted in change in participants' feelings as they relate to five domains associated with the program: social relationships; child strengths, family environment, educational involvement, and parenting capacity. One of the reasons for using the tool is to engage the participants in a meaningful conversation so that they express their views, not the views that they might be expected of them. The process has been documented and is publicly available (see Guenther and Boonstra 2009).

Community of practice approaches

In two recent evaluation projects I have worked on a 'community of practice' (Wenger 1998) approach has been employed. 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 what they describe as a 'discovery education' process (Trudgen 2000). The other was for an evaluation of the Northern Territory Department of Health and Families (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 recently completed with colleagues from CDU is an evaluation of the Akeyulerre Healing Centre in Alice Springs. The Centre is built around traditional healing for Arrernte families. The evaluation was 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. The CDU team's role in the evaluation was to gather stories from mainstream service providers. The evaluation outputs were produced as a partnership between TRG and CDU teams.

Customised databases

In several projects I have worked on the evaluation has supported program staff with the development of databases. An important aspect of these databases is that they are not for the evaluation team to use. Rather, they are designed for use by program staff. Training and support is provided to ensure that local staff could use the tools. While program staff are responsible for data entry, the evaluation team has had an independent role validating and analysing the data.

Summary

The purpose of the foregoing discussion about tools and approaches has been to demonstrate the wide variety of instruments available for use in complex evaluation contexts. As evaluators my colleagues and I 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.

6   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 author and his colleagues 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 Tasmanian 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.

The evaluations of complex programs discussed here 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 often 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.

7 Appendix 1: Selected performance indicators and measures from Tasmanian Government departmental annual reports

Table 2.   Selected performance indicators and their measures, taken from Tasmanian Government department annual reports (2008-2009)

Tasmanian Government Department	Division/ Output group	Aim/mission	Selected performance indicators	Measure 2008-2009
Department of Health and Human Services	Acute Care	"Promoting health and wellbeing and intervening early when needed." (Department of Health and Human Services 2009b)**	Admitted patients: weighted hospital separations	100 957
			Average length of stay	6.2 days
			Elective surgery waiting list	7 742
			Emergency Department presentations	125 594
			Percentage of elective surgery (multi-day stay) patients admitted on the day of surgery	87.3%
			Urgent patients admitted within 30 days for elective surgery	73%
Department of Health and Human Services	Disability, Child, Youth and Family Services	"Supporting individuals, families and communities to have more control over what matters to them". (Department of Health and Human Services 2009b)	Proportion of child protection notifications referred for further investigation	24.1%
			Proportion of finalised child protection investigations that were substantiated	57.2%
			Clients accessing crisis support services	
			Number of children in out-of-home care at 30 June	807
			Child abuse or neglect: Number of unallocated cases	26
			Children on Care and Protection Orders (per 1 000 children)	8.8
			Supported accommodation waiting list	33
			Community Youth Justice: Average daily number of active clients	662
			Supported accommodation clients	n.a.
Department of Health and Human Services	Housing Tasmania	"Developing responsive, accessible and sustainable services" (Department of Health and Human Services 2009b).	Public Housing Occupancy Rate	99.0
			Number of applicants housed	883
			Average wait time for people who are housed	29
			Number of applicants on the wait list	3039

Tasmanian Government Department	Division/ Output group	Aim/mission	Selected performance indicators	Measure 2008-2009
Department of Justice	Corrective Services: Prison services	"…to contribute to a safer Tasmania by ensuring the safe, secure containment of prisoners and to provide them with opportunities for rehabilitation, personal development and community engagement." (Department of Justice 2009:40)	Prison capacity utilisation rates Escape rate of prisoners Total escapes Daily average prisoner population Cost per prisoner per day Proportion of eligible prisoners enrolled in education and training Prison capacity utilisation rates	82.0% 0 per 100 prisoner years 0 522 $263 33.9% 82.0%
	Corrective Services: Community Corrections*	"…committed to working with offenders on community based orders in ways that aim to reduce re-offending and contribute to a safer society." (Department of Justice 2009:48)	Net operating expenditure per offender per day (2008-09 $) Average number of persons per day Number of Probation Orders imposed Number of Probation Orders revoked/cancelled Number of Parole Orders imposed Number of Parole Orders completed Number of Parole Orders revoked Return to corrections, Proportion returned within two years 2006-07	12.00 1 177 683 21 118 81 21 23.2%
Department of Police and Emergency Management	Community Safety	"…a community where people are safe" (Department of Police and Emergency Management 2009:28)	People aged 15 years or over who felt "safe" at home alone during the day People aged 15 years or over who felt "safe" at home alone after dark People aged 15 years or over who said they were "satisfied" with police services	96% 87% 71%
Department of Police and Emergency Management	Crime	"…a reduction in crime." (Department of Police and Emergency Management 2009:42)	Offences recorded Offences cleared Change in Offences Against the Person (2009/2008) Change in Offences Against Property (2009/2008)	31615 14478 -6.3% -5.2%

Tasmanian Government Department	Division/ Output group	Aim/mission	Selected performance indicators	Measure 2008-2009
Department of Police and Emergency Management	Traffic	"…improved driver behaviour through traffic law enforcement." (Department of Police and Emergency Management 2009:54)	Fatalities and serious injuries/per cent change on previous year On the spot notices/per cent change on previous year Per cent of vehicles detected speeding by road safety cameras/last year Random breath tests per cent charged/last year	366/+15% 53604/-5% 0.91%/1.24% 0.67%/0.72%
Department of Education	Pre-compulsory and compulsory education	"…every student has the opportunity to learn and achieve his or her potential."(Department of Education 2009:15)	Percentage of Kindergarten students achieving expected outcomes Percentage of Prep students achieving expected literacy outcomes Percentage of Prep students achieving expected numeracy outcomes Percentage gap in Indigenous students achieving expected outcomes NAPLAN results are included in the annual report but are not described as 'key performance measures'	76.8% 85.1% 84.6% 7.4%
Department of Education	Post-Compulsory Education and Skills Development	"The state is committed to providing meaningful pathways and enhanced facilities to encourage Year 10 students to stay on to further education and training." (Department of Education 2009:30)	Apparent retention rate Years 7–12 Apparent retention rate Years 10–12 Percentage of 15 to 19 year-olds not employed, not at school and not in fulltime tertiary education (Vocational) Student satisfaction – Graduates	61.3% 61.8% 10.8% 91.6%

* The Community Corrections section of the DoJ Annual report provides a vision, mission and objective statement. Prison services only provide a mission statement. The mission statement is shown here

** Note that the DHHS objectives listed are not tied to output groups. The 2008/09 Annual Report does not indicate that individual output groups have specific objectives/aims.

8   Appendix 2: Selected Tasmanian Government department mission statements

8.1   *Tasmanian Department of Health and Human Services*

A key aim of the DHHS is to improve the overall health and wellbeing of Tasmanians. (Department of Health and Human Services 2009a:18).

Vision: High quality, safe services for the people of Tasmania when they need them, so they can live well and live longer.

Mission: To design and implement a sustainable, people-focused health and human services system which supports individuals and communities to be active partners in the management of their own health and wellbeing. (Department of Health and Human Services 2009b:3)

Strategic objectives:

1. Supporting individuals, families and communities to have more control over what matters to them.

2. Promoting health and wellbeing and intervening early when needed.

3. Developing responsive, accessible and sustainable services.

4. Creating collaborative partnerships to support the development of healthier communities.

5. Shaping our workforce to be capable of meeting changing needs and future requirements.

8.2   *Tasmanian Department of Justice Mission Statement*

The Agency, in partnership with others, will work towards achieving a safe, fair and sustainable Tasmanian community where the diversity and rights of individuals are respected (Department of Justice 2009:1).

8.3   *Tasmanian Department of Police and Emergency Management*

Vision: A safe and secure Tasmania

Mission: To deliver quality policing and emergency management services to the people of Tasmania (Department of Police and Emergency Management 2009:7)

8.4   *Tasmanian Department of Education*

The activities of the Department of Education are aimed at enabling every Tasmanian to reach their potential, at all stages of life and to nurture, a culturally rich, socially cohesive and economically productive community (Department of Education 2009:4).

9 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 Learning 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.

Communities and Families Clearinghouse Australia 2009. FAST (Families and Schools Together), Darwin. Promising Practice Profiles. Australian Institute for Family Studies, Retrieved June 2009, from http://www.aifs.gov.au/cafca/ppp/profiles/cfc_fast.html.

Department of Education 2009. Annual Report 2008-2009, Retrieved June 2010, from http://www.education.tas.gov.au/annualreport/07-08/complete.pdf.

Department of Health and Human Services 2009a. Annual Report 2008–2009. Hobart, Retrieved June 2010, from http://www.dhhs.tas.gov.au/__data/assets/pdf_file/0010/48277/DHHS_Annual_Report_08-09_Parts1-and-2.pdf.

Department of Health and Human Services 2009b. Strategic Directions 09-12: putting our patients and clients at the centre of all we do. Hobart, April 2009, Retrieved June 2010, from http://www.dhhs.tas.gov.au/__data/assets/pdf_file/0017/42137/Strategic_Directions_2009-2012_FINAL.pdf.

Department of Justice 2009. Annual Report 2008-2009, Retrieved June 2010, from http://www.justice.tas.gov.au/annual_reports/2008-09.pdf.

Department of Police and Emergency Management 2009. Annual Report 2008-2009, Retrieved June 2010, from http://www.police.tas.gov.au/__data/assets/pdf_file/0014/30551/DPEM_Annual_Report_2008-09.pdf.

Falk, I. and Guenther, J. 2007. Generalising from Qualitative Research: Case studies from VET in Contexts. Australian Vocational Education and Training Research Association n200 Annual Conference, Victoria University, from http://www.avetra.org.au/publications/10-Guenther.pdf.

Franzosi, R. 2010. Quantitative Narrative Analysis, Thousand Oaks, Sage Publications.

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, Darwin, North Australia Research Unit, 27 March 2008, Retrieved June 2009, from http://www.catconatus.com.au/docs/080327_NARU_role_of_evaluator.pdf.

Guenther, J., Arnott, A. and WIlliams, E. 2009. Measuring the unmeasurable: Complex evaluations in the Northern Territory. NARU Seminar Series, Darwin, covaluator.net, 8 October 2009, Retrieved January 2010, from http://www.catconatus.com.au/docs/091008_NARU_measuring_unmeasurable.pdf.

Guenther, J. and Arvier, M. 2009. "Assessing the impact of a mobile family resource service for parents and young children." International Journal of Child Health and Human Development 2(2): 135-141.

Guenther, J. and Boonstra, M. 2009. Adapting Evaluation Materials for Remote Indigenous Communities and Low-Literacy Participants. 12th Australasian Conference on Child Abuse and Neglect, Perth, Western Australia, 15-18 November 2009, Retrieved January 2010, from http://www.catconatus.com.au/docs/091117_APCCAN_FAST.pdf.

Guenther, J., Davis, V., Foster, D. and Arnott, A. 2010. Two knowledges working together. AUCEA 7th Annual Conference, Launceston, Australian University Community Engagement Alliance, 5-7 July 2010.

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#.

Patton, M. 2002. Qualitative research and evaluation methods. Third edition, 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.

Tashakkori, A. and Teddlie, C. 1998. Mixed Methodology: Combining Qualitative and Quantitative Approaches. Applied Social Research Methods Series Volume 46, Thousand Oaks, CA., Sage.

Trudgen, R. 2000. Why Warriors Lie Down and Die, Darwin, Aboriginal Resource and Development Services.

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

Webb, N. 2007. Building Connections pilot program: Evaluation Report. Interrelate Family Centres, March 2007, Retrieved January 2010, from http://www.ag.gov.au/www/agd/rwpattach.nsf/VAP/(756EDFD270AD704EF00C15CF396D6111)~Building+Connections+-+EvaluationReport+-+March+2007.pdf/$file/Building+Connections+-+EvaluationReport+-+March+2007.pdf.

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