Monitoring: The slow cousin of evaluation or an equal partner?

Marc de boer wrote this one. I don't see his name on it anywhere and there may be other authors. Liz.

Abstract

Evaluators often regard monitoring as playing a secondary and relatively simple role compared to evaluation proper. This paper argues that this underestimates the potential of monitoring information in enhancing the value of evaluative work, in particular to increase the 'half life' of evaluation findings. Moreover, it suggests the possibility of a more dynamic interchange between monitoring and evaluation. Specifically, monitoring complements the fragmented and *ad hoc* nature of evaluation work, so that the process of monitoring presents an opportunity to develop a framework within which individual evaluations can exist. Drawing on programme logic and CMO theory, the key to this interchange is to build of systematic body of knowledge and theory that drives, and is in turn informed by, evaluation and monitoring. The work within the Department of Work and Income New Zealand provides an example of an attempt at implementing such a framework.

Introduction

Over the past 25 years two approaches have emerged to provide empirical information to decision makers on the effectiveness of social programmes. One is evaluation, which uses principles of social science research to assess the concept, design and implementation of programmes (Rossi & Freeman, 1994). The other is monitoring,[1] which provides decision-makers with timely information on a programme's progress, often against set goals or benchmarks.

Several authors within the discipline of evaluation have raised concerns over the apparent independence of these two approaches and how they often come into conflict within the decision making process (Bernstein, 1999; Blalock, 1999). The challenge they pose is how to marry evaluation and monitoring information approaches. This paper shows one attempt, focusing particularly on the contribution that monitoring can make in the generation of organisational knowledge.

---

[1] The literature discuses monitoring under a number of headings, most often as performance measurement and management information systems. The term used in this paper primarily refers to any regular source of information or data on programmes rather than the management systems that they support.

Criticism of evaluation and monitoring

Before outlining how monitoring and evaluation might work together, it is useful to first examine the weaknesses of each.

Among evaluators at least, evaluation is regarded as the best means to judge and understand the impact of programmes on outcomes. But this strength is also its Achilles heel, in that good evaluations need expertise, resources and, above all, time. This often leads to a lagged cycle of commissioning evaluations to address policy questions, only to have evaluations reported well after the necessary decisions have been made. Conversely, decision-makers often view earlier evaluations as out of date; and, rightly or wrongly, irrelevant to current policy questions.

This game of "catch up" also produces an incoherent body of work. The limited time-frame of individual evaluations often precludes review of previous evaluation or research findings. By acting in isolation, such evaluations have limited opportunity to contribute new insights into the policy or programme under review (Lipsey, 2000; Anderson, 1998). This failure to accumulate evaluative knowledge is ironic, as this is one of the cornerstones of scientific inquiry upon which evaluation basis its legitimacy.

In contrast, the perception of monitoring information, especially among evaluators themselves, is that it is second-best to evaluation proper (Blalock, 1999; Davies, 1999). Monitoring information is overwhelmingly quantitative of what can be measured easily, often leaving important aspects under-represented.[2] Further, monitoring analysis often comprises only simple descriptions of programme operation and outcomes, reflecting its audit role for public accountability and the focus on the "what" rather than "why" questions (Newcomer, 1997). Evaluators see particular risk in the use of monitoring information in assessing programme impact, in that monitoring can fail to properly address, or even acknowledge, issues of causality; misleading decision-makers over the effectiveness of programmes (Mayne, 1999; Blalock, 1999).[3]

The positive conclusion of this rather negative introduction is the weaknesses and, by implication, the strengths of evaluation and monitoring are complementary (Perrin, 1999). Synthesis of these two sources of information serves to increase the value of both

---

[2] As Perrin (1999) points out, 'many activities in the public policy realm, by their very nature, are complex and intangible and cannot be reduced to a numerical figure ... What is measured or even measurable, often bears little resemblance to what is relevant."

[3] This does not even begin to address the more serious issues with monitoring when linked to unrealistic performance targets producing goal displacement and other perverse behaviours, which can also seriously undermine the veracity of the information gained through these systems. In this respect, Maoist China and Stalinist Russia are perhaps some of the most instructive examples of the misuse of performance management.

in the decision making process (Davies, 1999). The question is how this can be best achieved.

Role of theory

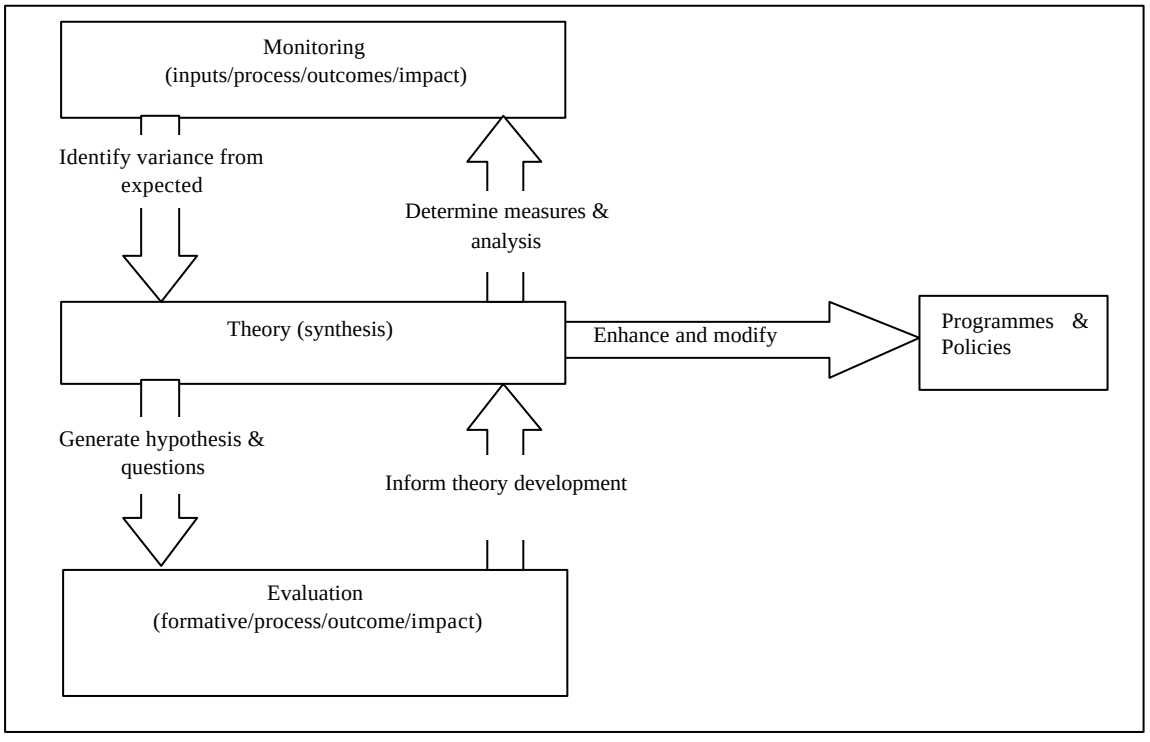
It is argued that grounded theory is the key to linking evaluation and monitoring. Moreover, such theory operates at two levels, the first being at the individual programme level, while the second places programmes within the broader setting of the problem or issue being addressed.

At the programme level, good programme design will clearly state how it is to be implemented (inputs, processes and outputs) and the intervention logic (and assumptions) of how the programme's outputs will influence outcomes in the desired way. It is also important to understand the context within which programmes and policies operate (e.g. other policies/programmes, resource availability) and to deal with the competing theories of how programmes influence outcomes (Wholey, 1977; Pawson and Tilley, 1999; Funnel, 1997; Sheirer, 2000).

The second level of theory positions the intervention logic of programmes within midlevel theories (Pawson and Tilley, 1999) of the social or economic problem being addressed. This provides an opportunity to be able to trace how the outputs of specific programmes influence the broader outcomes and to contrast this with alternative programmes and initiatives. This helps to anchor programme evaluation and monitoring within a broader body of research; enabling analysts to use this framework to better understand the processes that contribute to the observed impact of individual programmes.

Accordingly, monitoring and evaluation play central roles in testing and expanding

Figure 1: Schematic of the proposed relationship between evaluation and monitoring



programme theory. Put simply, monitoring describes 'what is', while evaluation seeks to understand 'why this is so'. The ideal would be that after construction of a programme's intervention logic a monitoring framework would be implemented to measure critical aspects of a programme's intervention logic. Sufficient care needs to be taken to ensure the framework is credible to external stakeholders and decision makers understand the validity, reliability and limitations of the measures used (Blalock, 1999).[4] Where monitoring signals the programme is not working as intended or having its expected impact or where competing causal theories cannot be ruled out, then evaluation can be applied (Perrin, 2000 & 1998). Such evaluation will either confirm existing theory or produce new understanding, which in turn can redirect monitoring effort (Perrin, 1999). Therefore, a cycle emerges of continuing monitoring with periodic evaluation (Mayne, 1999), each enhancing the robustness of the programme(s) theory (Figure 1).

Such a monitoring/theory/evaluation cycle increases the value of both monitoring and evaluation information for decision-making. Firstly it allows evaluators to better respond to the information needs of policy makers, lessening the need for *ad hoc* or piecemeal evaluative research. In particular, it can help assure control agencies of the effectiveness of policies or programmes by using monitoring information for accountability as well as programme improvement (Mayne, 1999; Scheirer, 2000; Bernstein, 1999). Secondly, it enables identification of those areas where organisational understanding is limited and properly direct scarce evaluation resources.

Evaluation by this logic is necessarily intensive and targeted, while monitoring needs to be broad in coverage, and occurring in an ongoing and timely manner. This does suggest the methods of monitoring are less sophisticated and rigorous than for evaluation. However, this does not imply limiting monitoring information to simple counts of inputs, processes, outputs and outcomes (Scheirer, 2000). Rather monitoring information needs to be sufficient to provide stakeholders assurance the programme is working as intended and provides credible evidence of the effect it has on outcomes (Mayne, 1999). The key aim of monitoring, therefore, is to minimise the resources needed to provide credible and reliable information to inform decision-makers of the operation of programmes. This equation will be a function of the policy/programme being evaluated, available information and the technical expertise of the evaluator.

The following section provides an example of a first attempt at setting up this framework and the role that monitoring information plays within it. Moreover, while theoretical

---

[4] Perrin (1999) makes the useful distinction between "bad" and "dirty" data. Dirty data occurs with nearly all evaluation methods, "bad" data is a specific risk with monitoring data generated for performance management purposes where results or methods have been "engineered" to reach set targets.

frameworks strive for simplicity and elegance, practice is always a messier affair. Nevertheless, the example that follows will hopefully illustrate two points:

- The iterative nature of building organisational knowledge through evaluation and monitoring.
- The importance of theory in understanding the information produced.

Monitoring of employment programmes

The Department of Work and Income (DWI) is responsible for administering of income support and providing employment assistance, and was created through the merger of Income Support Service and New Zealand Employment Service in 1998.[5] This merger provided the opportunity for the integration of administrative data on income support and employment information. This, coupled with the rapid advancement in micro processing power and analytical applications like SAS, has enabled internal evaluators to have direct access to administrative data for evaluation and monitoring.

Outcomes and impact of employment programmes

One of the key questions for government is whether the employment assistance DWI provides is effective in helping disadvantaged job seekers. As a result, government directed DWI to review the effectiveness of 9 of its largest programmes. The internal evaluation team saw this as an opportunity to implement a monitoring framework focused on the outcomes and impact of employment programmes provided through the Department.

The two most significant challenges in setting up this monitoring framework was a reliable outcome measure and estimating the counterfactual (generative impact). In both cases, the original outcome measure and the estimation technique used were challenged by external agencies. Because the monitoring framework had to be credible to these agencies, DWI with these agencies had to work through the issues raised. This work and its acceptance has improved both the utility and robustness of the analysis.

In the end a simple "Independence of DWI" measure was outcome chosen, which reflected whether a job seeker was receiving either income support or employment assistance from the Department. The counterfactual was estimated using a quasi-experimental design. This involved constructing propensity weighted comparison groups, which produces a group of non-participants who have the same characteristic profile as that of the participants. The similarities between the two are strongest for those characteristics that distinguish participants from the average job seeker population (de

[5] As of October 2001, DWI was merged with the Ministry of Social Policy to form the Ministry of Social Development.

Boer, 2001b). The top-line results of the analysis are summarised in Table 1 below (de Boer, 2001a, 2001c).

Table 1: Proportion of participants Independent of DWI and impact ratio by programme at 12 and 24 months after participation start.

Type	Programme	Participants Independent of DWI [1]		Adjusted impact ratio [2]	
		12 months	24 months	12 months	24 months
Wage Subsidy	Job Plus	68%	70%	1.64	1.29
	Job Connection	41%	43%	1.99	1.46
On-the-job training	Job Plus Training	52%	58%	1.27	1.16
Work Experience	CTF/Community Work	26%	41%	0.83	0.94
	Task Force Green	47%	56%	1.31	1.18
	Job Plus Maori Assets	65%	61%	2.05	1.33
Self-Employment Assistance	Enterprise Allowance and Capitalisation	76%	70%	1.92	1.26
	Business Advice and Training Grant	55%	64%	1.40	1.21
Into Work Support	Work Start Grant	67%	65%	0.92	0.97

1: Independence of DWI is where a job seeker is no longer receiving a core benefit or participation in employment programmes.

2: Impact ratio: estimated using propensity weighted regression and is the ratio between the proportions of participants and non-participants Independent of DWI, controlling for other observable job seeker characteristics.

Base: Includes all programme participants who started between 1 January 1998 and 1 July 2000.

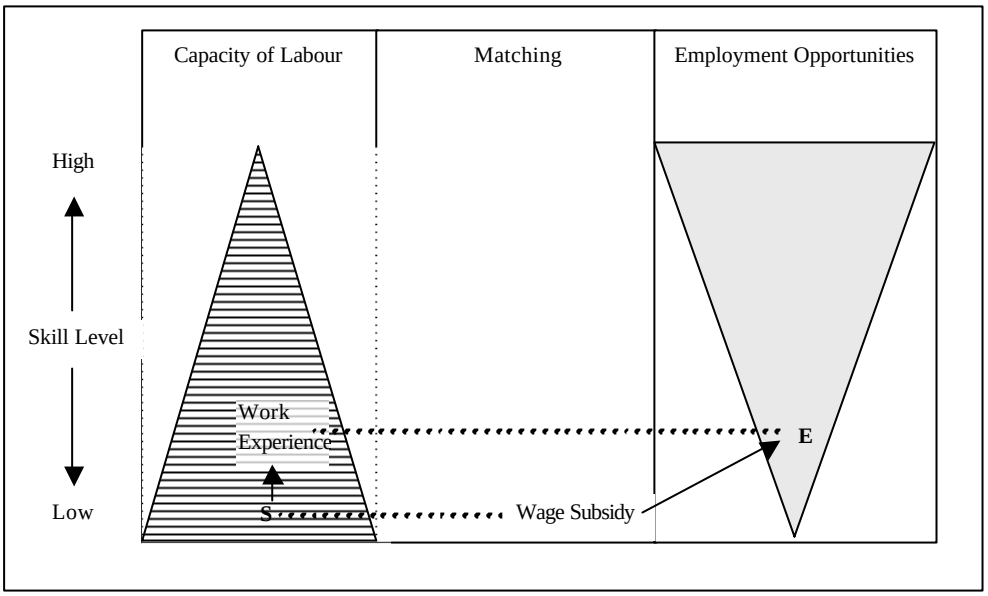
It is not the purpose of this paper to go into detail over the results of the findings, albeit to say that they illustrate the risk of interpreting gross outcomes as a proxy of the programme's impact. For example, Job Connection (wage subsidy targeted at the long-term unemployed) has low outcomes but has the greatest impact.

Theoretical model of the labour market

Having developed consistent measures of programme outcomes and impact, this immediately focuses attention on why differences exist. The first high-level attempt at explaining this is placing these programmes within a simple model of the labour market (Figure 2 below). Figure 2 divides the labour market into three parts Capacity of Labour, Matching, and Employment Opportunities, and assumes that unemployment is largely structural. That is, there is a miss-match between the supply and demand of labour at each skill level (represented by the two triangles). Different types of programmes can be placed within the model according to how they are intended to address unemployment within the labour market.

Although simple, the model is a powerful way of illustrating how employment programmes attempt to address unemployment and more importantly draws the links between the observed microeconomic impact of programmes to their potential macroeconomic affects. For example, wage subsidises are located close to the matching | opportunity boundary and therefore they are expected to result in high outcomes for participants and thus have a significant impact. On the other hand, work experience programmes seek to develop the capacity of labour, and therefore, the link to opportunities is more tenuous. Accordingly, the outcomes and impact of these programmes would be more modest. Both these conclusions are borne out by the monitoring data presented above.

Figure 2: Conceptual framework of the labour market with structural unemployment



However, the model points out that assisting participants into employment, is only part of the picture. What the monitoring information does not provide is information on the quality of the outcomes achieved (represented as an increase in the skills of job seekers and the types of work they move into). Whether programmes assist job seekers into higher skilled employment has implications for the programme addressing structural unemployment and the associated risk of displacing or substituting people who are equally disadvantaged in the labour market as those assisted. This applies particularly to wage subsidies where the apparent high microeconomic impact can be offset through unobserved displacement of disadvantaged workers.

This represents the first phase of the monitoring/theory cycle (see Figure 1). From this several new evaluation and monitoring initiatives have been proposed. The analysis of macroeconomic risks signals the Department needs to monitor the involvement of individual employer's use of wage subsidies and work experience programmes to reduce displacement risk. This work also identified a large gap in the Department's knowledge about the role that wage subsides play in persuading employers to hire disadvantaged job seekers. This places the research and evaluation team in a strong position in recommending that subsequent evaluation effort should be directed into these areas.

Conclusions

The work done so far within DWI represents the first iteration of our application of the monitoring/theory/evaluation cycle. However, what the paper hopes to show is the value of developing sophisticated monitoring techniques with theoretical frameworks to support and direct evaluation effort. To this end monitoring and evaluation become equal partners in developing and synthesising knowledge at all levels of the organisation.

References

Anderson, Diane., 1998, *What have we learned and where to next? A review of evaluations of employment programmes, 1994-1997*, Labour Market Policy Group, Department of Labour, Wellington.

Bernstein, David, J., 1999, "Comments on Perrin's 'Effective use and misuse of performance management'", *American Journal of Evaluation*, Vol 20, No 1, pp. 85-93.

Blalock, Ann Bonar., 1999, "Evaluation Research and the Performance Management movement", *Evaluation*, vol. 5, no. 2, pp. 245-258.

Davies, Ian C., 1999, "Evaluation and performance management in government", *Evaluation*, vol. 5, no. 2, pp. 150-159.

de Boer, Marc., 2001a, *Review of the Subsidised Work appropriation*, Department of Work and Income, Wellington, New Zealand.

de Boer, Marc., 2001b, *Methodology of the Review of the Subsidised Work appropriation*, Department of Work and Income, Wellington, New Zealand.

de Boer, Marc., 2001c, *August 2001 Six monthly review of DWI employment programmes*, Department of Work and Income, Wellington, New Zealand.

Funnel, Sue., 1997, "Program logic: an adaptable tool for designing and evaluating programs", Evaluation News and Comment 5 (July).

Lipsey, Mark., 2000, "Meta-Analysis and the Learning Curve in Evaluation Practice", *American Journal of Evaluation*, Vol 21, No 2, pp. 207-212.

Martin, Steve. & Sanderson, Ian., 1999, "Evaluating public policy experiments: measuring outcomes, monitoring processes or managing pilots?", *Evaluation*, vol. 5, no. 3, pp. 245-258.

Newcomer, Kathryn, E., 1997, "Using performance measurement to improve programmes", in *Using performance measurement to improve public and non-profit programmes*, Newcomer, Kathryn, E. (ed), New Directions for Evaluation, Number 75, San Francisco: Jossey-Bass.

Pawson, Ray. and Tilley, Nick., 1997, *Realistic Evaluation*, Sage Publications, London.

Perrin, Burt., 1999, "Effective use and misuse of performance measurement", *American Journal of Evaluation*, Vol 21, No 2.

Rossi, P. & Freeman, H., 1994, *Evaluation: A systematic approach*, Thousand Oaks, CA: Sage Publications.

Scheirer, Mary Ann., 2000, "Getting more 'bang' for your performance 'buck'", *American Journal of Evaluation*, Vol 21, No 2, pp. 139-149.

Wholey, J. S., 1997, "Evaluability Assessment", in *Evaluation research methods: a basic guide*, Rutmna, L. (Ed), Thousand Oaks, CA: Sage Publications.