Evaluation
http://evi.sagepub.com

Evaluation Research and the Performance Management Movement: From Estrangement to Useful Integration?

Ann Bonar Blalock

Evaluation 1999; 5; 117
DOI: 10.1177/13563899922208887

Evaluation
Copyright © 1999
SAGE Publications (London,
Thousand Oaks and New Delhi)
[1356–3890 (199904)5:2; 117–149; 008794]
Vol 5(2): 117–149

Evaluation Research and the Performance Management Movement

From Estrangement to Useful Integration?

ANN BONAR BLALOCK
Editor of Evaluation Forum, Department of Labor, Washington DC, USA

The burgeoning performance management movement, with its emphasis on social program 'results' measured typically by a limited set of quantitative indicators, has developed a life of its own largely apart from the evaluation research movement. Reflecting the differences in the professional history, interests and training underlying the two movements, the relationship between these disparate approaches to establishing public accountability has lacked coordination and defied integration. This article discusses the basic concepts guiding the evolution of these movements in the context of the goals of information production, and explores the major conceptual, measurement and methodological problems resulting from the lack of accommodation between them. It also provides suggestions about how these two important approaches can be better integrated, both professionally and organizationally, for the purpose of enhancing the reliability and validity of social program assessments, and therefore for improving policy development and program management.

Two 'movements' dedicated to improving government policies and programs and increasing government accountability have come together in the 1990s: the *performance management movement*, and the *evaluation research movement*.[1] Although both have been given greater governmental attention across a large number of developed countries in this decade, their developmental histories differ and their relationship has been problematic. This article examines the nature of this relationship, explores why the information base for judging the value of social policies and programs can benefit from complementary information from both movements, and proposes some ways to begin increasing the integration of evaluation research within performance management systems. A major question is whether the current, nearly exclusive emphasis in the performance management movement on the outcomes or *results* of programs, frequently in the absence of a commitment to collect information about why and how those results occurred, may be leading to flawed social policy and misguided judgments of programs.

The concept of performance management was persuasively seeded in the postindustrial world in the late 1980s and blossomed in the 1990s. It incorporates key features of past efforts to reform the management of social service systems and programs. What is new is the *context* in which this concept has flourished. Government deficits have reduced social program funding, leading to greater selectivity in what programs are to be given continuing support and what new initiatives are to be launched. The decentralization of program authority has resulted in the need for central governments to retrieve some level of control over the outcomes of devolved programs, as a trade-off for their loss of power over the way programs are being implemented at subnational and local levels. In this sense, the performance management movement has been strongly influenced by global economic change. Economic realities have generated new central government and public demands for evidence of social program *accountability.*[2] In the process, the definition of accountability has shifted from a previous emphasis on program *processes* to a more singular focus on program *results.*

Interestingly, in many of the postindustrial countries the *performance management movement* has developed alongside an already evolved *evaluation research movement,* the evaluation movement in the United States having had the longest and most scientifically elaborated history. The general purposes of these two movements are similar, to base judgments of the effectiveness of social program efforts on more appropriate and trustworthy information, and to improve those efforts. Given a common *general* purpose, it is rational to propose that a better articulation between these two significant professional phenomena would enrich them both. Unfortunately, these movements have failed to come together meaningfully to serve emerging information production and accountability needs.[3]

While the performance management movement has mediated the previous preoccupation with the characteristics of the 'customers' of social programs and the 'services' they receive ('process' issues), the heavy reliance on the collection, analysis and use of a restricted array of short-term outcome measures has the disturbing potential for leading to faulty judgments of the inherent value of social programs, and of the new coordinated 'systems' of related programs being created at subnational levels in response to program proliferation, fragmentation and excessive costs.

The possibility of producing fatal social remedies in the contemporary rush to 'manage performance' is particularly great if those designing and directing performance management systems, and the users of information flowing from such systems, are not careful about:

1. Distinguishing between results that can be attributed exclusively to the unique interventions of these programs, and those that may be due to a variety of influences both within and *outside* these programs, or which are occurring simply by *chance*;
2. Obtaining sufficient information about the way in which program implementation may be influencing program results. Of concern as well is the comparative influence of the two movements in the 1990s. As performance

management has grown in importance, there has been a tendency to undervalue the critical role of evaluation research.

The Nature Of The Two Movements

Despite commonalities, there are important differences between the two movements. The disparities have arisen out of different professional disciplines and bureaucratic environments, and have been shaped by different levels of public acceptance. Performance management is a blend of public/private planning and management ideas – particularly private sector ideas about quality assurance, customer satisfaction, and continuous improvement. Evaluation research is an applied offshoot of basic social science research. As significant elements in an *ideal-type* policy process, strategic planning, visionary management and evaluation research would seem to be integral and equally important parts. All yield significant benefits and usefully inform one another. But an integration of these three elements has been tentative, and often absent, in the real world of policymaking.

The Goals of Information Production

In analyzing differences between the performance management and evaluation research movements, it is important to define what policy-makers, administrators, program managers, stakeholders, elected officials and customers of the program *need to know*. Both performance management systems and evaluation research activities should respond to the collective information needs of these constituencies. So, what are the major goals of information production for policy-making, program improvement and accountability purposes? Several goals can be suggested:

- To determine whether a program's *interventions* (its unique combination of services, subsidies and/or activities) are those intended – and to increase their quality, as well as the quality and appropriateness of their delivery to a program's target population (its 'customers');
- To determine whether a program's interventions are being delivered to the right *target population*;
- To determine whether a program is being *implemented* as intended (whether its customers are being exposed to its interventions in the ways intended) – and to improve its implementation;
- To determine whether a program's *general outcomes* for customers, its *net impact*, and its *ratio of costs to benefits* are consistent with the outcomes desired – and to improve program results;
- To make a judgment about what *major influences are shaping a program's outcomes*: regarding the nature of its target population, its mode of implementation, its interventions, and the geographic, socio-economic and cultural environment in which it operates;
- To make an assessment of the appropriateness, utility and societal value of the *policies* on the basis of which a program is designed, in terms of the target

> populations selected to receive the program's interventions, the interventions proposed, the mode of program implementation recommended, the outcomes desired, and the characteristics of the program's environment which likely influence its outcomes.

Performance management systems are not intended, nor are they so designed, to respond to this full menu of information production objectives. Although the emphasis of such systems on *outcomes* represents a significant improvement in information production, these new systems tend to neglect the full range of program issues.

The assessment of this broader range of program issues has been encompassed traditionally in the academic training of evaluation researchers, as well as a professional commitment to utilizing research methodologies that reduce potential bias. This does not mean, of course, that all evaluation researchers are experts on comprehensive evaluations – that is, evaluations that combine process and outcomes studies – or that all are committed to professional standards.

Significant Differences Between Performance Management and Evaluation Research

Much of the current tension between the two movements appears to stem from a disturbing confusion about, and often vexing controversy over what *role* each should play in the policy process. In performance systems, the monitoring (or tracking) of program outcomes using *performance measures* or indicators, is viewed frequently as a substitute for science-based evaluations in judging program value. Outside performance systems, evaluation research may be used as the sole source of such judgment. This all too prevalent dichotomy masks substantial opportunities. However, to understand the lack of cohesion between these movements, one must appreciate that performance management is a *planning and managerial* tool; evaluation research is a *research* tool. The two movements' different purposes tend to condition the types of activities conducted – primarily *managerial monitoring* activities for performance management systems; primarily *scientific evaluative* activities for evaluation research.

It is true that there are gray areas of some consequence between performance monitoring and evaluation research. Monitoring information is appropriate to managers' need to obtain quick and continuous feedback on a limited number of outcome measures. And if performance measures have high utility and validity (that is, if they adequately represent the variables they measure), and are being collected reliably, monitoring information can be extremely useful to evaluators as part of their own data collection effort.[4] However, most monitoring responsibilities in performance management systems involve the collection of a circumscribed cluster of quantitative measures, and the purpose of most monitoring staffs is to report simple statistics about a program rather than to conduct more complex data analyses to answer critical planning and managerial questions.

Evaluation Research The purpose of evaluations is to increase our understanding of the major relationships imbedded in the *design* of social programs –

that is, in the set of propositions that describe them. In this context, Peter Rossi and Howard Freeman define evaluation research as follows:

> Evaluation research is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs. (Rossi and Freeman, 1994)

Scientific evaluations are frequently classified based on the questions they seek to answer about a program's design.

Process evaluations seek an answer to the following questions: 'Is the program being implemented as intended?' 'What structures, policies and practices are in place or are occurring, in the context of the desired mode of implementation?' 'How is the nature of program implementation affecting program outcomes?'

Gross outcome evaluations seek to answer the question, 'Are the program's customers experiencing the kinds of outcomes intended, as well as other outcomes of interest – in the short term or longer term – irrespective of whether the program is responsible for them?'

Net impact evaluations address the most critical policy question, 'Did the program's unique implementation mode and interventions make a real difference in the outcomes, independent of other influences?' Net impact studies therefore seek to determine which program results, or outcomes, can be attributed exclusively to the program rather than to other influences or to chance. It is important to distinguish net impact (net effect) studies from gross outcome evaluations that simply tell managers what changes may be occurring in the outcomes of customers, or changes in these outcomes between the pre-program and post-program periods (Heckman, 1993).

Cost/benefit evaluations attempt to answer a final and very difficult policy question, 'Does the program's net impact justify its costs?' Even if the program has a positive net impact, it does not necessarily mean that the program is worth continuing.

The multiple goals of information production require answers to all these questions. In answering such critical questions, a *set of logical scientific steps* are involved in planning and conducting evaluations. These steps assure that evaluations are competently conducted, are appropriate to the issues of greatest concern to those sponsoring and using the results of evaluations, and make sense in terms of the practical setting in which these evaluations are to occur. Three major steps are identifiable: *conceptualization*, *measurement*, and *methodology.* In the conceptualization step, the major variables and relationships of interest are identified and defined. In the measurement step, measures are developed for these variables (for intended influences and desired effects) – either qualitative or quantitative measures or both. In the methodological step, *research designs* and *methods* for *sampling*, *collecting data*, and *data analysis* are selected, consistent with the elements of the program to be evaluated as well as the scientific, political, organizational, socio-cultural and ethical constraints which frame what the evaluator can accomplish. Table 1 illustrates the complexity of the methodological step.

Exploratory, *descriptive, quasi-experimental* and *experimental* research designs lie on a continuum from least rigorous scientifically to most rigorous – if rigor is viewed as a function of the extent to which a design controls for potential biases in information production. Evaluators have recourse to strategies for reducing bias in exploratory and descriptive studies, and rich insights about the intricacies of causal relationships between interventions and outcomes can result from such studies. But quasi-experimental designs come closer to revealing the causal nature of these relationships, through the construction of a comparison group whose characteristics match the key attributes of the units being exposed to program interventions, and by using sophisticated statistical techniques to reduce selection bias. However, a continued and often intense debate exists, particularly in the US, about the extent to which these artificial controls are sufficient, and these designs have been criticized for their complexity and cost.

Experimental designs are viewed as the most rigorous option because they adhere most closely to scientific principles and methods, and therefore control for bias most effectively. As Gunther Schmid has commented in an article in *Evaluation*, they may be the design of choice for studying variations in single programs in relatively stable environments (Schmid, 1997). However, they are not appropriate for evaluating program implementation and its potential effect on outcomes. Also, the emphasis in field experiments is frequently on the measurement of outcomes to the neglect of careful definition and measurement of the interventions, leaving net impact results difficult to interpret – that is, we have to wonder what *elements of the program* caused the measured effects to happen. And experimental designs are rarely appropriate for evaluating programs that involve units other than individuals, since it is difficult if not impossible to randomly assign a set of complex interventions, such as multiple-treatment economic development strategies, to a universe of related programs, whole communities or different regions of a country.

Furthermore, evaluators must make the assumption (sometimes unwarranted) that the restricted cluster of key variables selected for study in experiments are, in fact, the most significant ones to look at. In addition, random assignment interferes with traditional service delivery functions, which may have unmeasured effects on outcomes, and can raise serious ethical issues in ongoing programs. Timeliness and cost are issues also. Finally, experiments trade realism for precision (greater control over bias), often making it difficult to generalize their results.

Despite these imperfections, the evaluation research profession's principles, methods and standards place an important scientific framework around the production of information, which distinguishes evaluation research from performance management.

Emphasis in the 1960s and 1970s in the US, and in some other countries until the late 1980s, has been on process evaluations focusing on program interventions, or outputs – who and how many have received what kinds of services, for example – and on gross outcome evaluations. In the 1980s a trend toward the use of quasi-experimental and classic experimental designs for conducting net impact evaluations took hold in the US, supported by econometricians convinced that the use of such designs was the best way of studying cause–effect relationships to

Table 1. EVALUATION METHODOLOGIES from which evaluators choose those most appropriate to program settings and the issues to be studied

Research Design	*Primary Purpose*	*Kinds of Issues for which design is most appropriate*	*Research Methods**	
			Data Collection Methods	*Predominant Kind of Data Collected*
I. Nonexperimental				
A. Exploratory	To identify the main variables (influences) of interest, in order to conduct more rigorous evaluations in the future.	The study of different aspects of program implementation whose nature and influence have not been systematically studied.	Interviews with or surveys of: staff, customers, relevant others. Review of program records. Collection of administrative (MIS) data. Case studies.	Qualitative
		The study of significant outcomes that have been neglected in previous evaluations.	Collection of gross outcome data via administrative data systems and/or thru fresh surveys.	Quantitative
B. Descriptive	To conduct a more rigorous evaluation, having identified, defined and measured key program variables, but without using a comparison or control group.	Program implementation.	Same as in exploratory.	Qualitative
		Short term/longer term gross outcomes for program participants. *Pre-post program-comparative approach is most rigorous option.*	Same as in exploratory studies.	Quantitative
II. Quasi-Experimental	To conduct a more rigorous evaluation, to determine whether a program 'made a	Short term/longerterm net outcomes (net effects or net impacts).	Use of MIS and/or other administrative data + fresh surveys.	Quantitative

Table 1. continued

Research Design	*Primary Purpose*	*Kinds of Issues for which design is most appropriate*	*Research Methods**	
			Data Collection Methods	*Predominant Kind of Data Collected*
	difference' – that is, to determine if the outcomes studied can be attributed to a program's influence rather than to other factors or to chance. This design utilizes a matched comparison group.	*Note*: certain aspects of program implementation need to be evaluated, particularly the selection and assignment of program participants, as a control for selection bias.		Qualitative
III. Experimental	Same as above, but utilizing random assignment of the interventions, producing equivalent 'treated' and 'control' groups.	Same as above.	Same as above.	Quantitative
IV. Mixed-Method	To study the influence of program implementation and/or interventions on outcomes,	Program implementation and gross outcomes.	A mix of methods.	Mix of quantitative & qualitative.

Table 1. continued.

Research Design	*Primary Purpose*	*Kinds of Issues for which design is most appropriate*	*Research Methods**	
			Data Collection Methods	*Predominant Kind of Data Collected*
	using a mixture of research designs and/or sets of methods, appropriate to the ongoing information needs of decisionmakers and sensitive to constraints on rigorous evaluations in the setting in which studies are to occur.			

* Data analysis methods typically increase in statistical sophistication as one moves from exploratory to experimental designs.

answer policy questions (Heckman, 1989/1993; Hotz, 1992). This trend was aided and abetted by the accumulation of evidence in government circles that less rigorous research designs were yielding ambiguous results at best, and fatally misleading ones at worst.

By the late 1980s, these more rigorous designs were the methodology of choice in the US for evaluating new initiatives (demonstration projects), since the random assignment of interventions in small-scale pilot projects designed to test new ideas posed fewer ethical, political and organizational problems (Gueron and Pauly, 1991). This translated later to increased use of experimental designs in evaluating ongoing, large-scale national programs (Bloom, 1993). A similar trend occurred in Sweden and Canada.

At the same time, American evaluators were moving toward more *comprehensive* evaluations. These combined well-designed process studies and experimental net impact studies within a single large-scale evaluation (Blalock, 1990). The recommendations for the US *School-To-Work Evaluation* followed this model (US Department of Labor, 1997). In the European Union countries, experimental approaches have been recommended, in addition to other approaches, in the new official government evaluation guides of the 1990s, because of a growing belief that this strategy can produce more valid information. But resistance to their use remains high in Europe because of considerations of cost, timeliness, ethics, and realism (see OECD, 1991; European Commission, 1995, 1997).

Despite the growing prestige of evaluation research in developed countries (apart from vigorous debates over methodology), much has been written about the lack of training of researchers in the realities of planning and management in large-scale organizations – that is, in understanding the needs and constraints of planners and managers, planning and management principles and methods, and the intricacies of organizational negotiation, compromise and consensus building. Even though this knowledge gap is slowly closing as researchers accumulate more experience in collaborative partnership settings, it remains a deterrent to the integration of evaluation research within performance management systems.

Performance Management The drive to develop performance measures consistent with program goals has supported more logical and strategic thinking at all levels of government. This approach builds on theory and practice in strategic planning, and on formal management principles and practices. In the US, the recently developed federal *National Performance Review* (NPR) has defined performance management as:

> The use of performance measurement information to help set agreed-upon performance goals, allocate and prioritize resources, inform managers to either confirm or change current policy or program directions to meet these goals, and report on the success in meeting those goals. (National Performance Review, 1995)

In this context, a *set of logical steps* is usually proposed in designing a performance management system, based on classic planning concepts:

- The definition of a *vision*, or mission, for the system;
- The development of system *goals*, consistent with the vision;
- The definition of *objectives*, consistent with the goals;
- The development of a limited set of key *performance measures* for monitoring progress in achieving the objectives;
- The development of *performance standards*, *short-term targets*, *and/or longer-term benchmarks* that will reinforce the system's commitment to meeting performance expectations;
- The development of *incentives and sanctions* for rewarding the meeting of standards, and penalizing failure to meet them.

It is intriguing, however, that the NPR states also that performance management is 'A process of assessing progress toward achieving predetermined goals, including information on the efficiency with which resources are transformed into goods and services (outputs); the quality of those outputs (how well they are delivered to clients and the extent to which clients are satisfied); outcomes (results of a program activity compared to its intended purpose); and the effectiveness of government operations in terms of their specific contributions to program objectives.' This emphasis on *efficiency and effectiveness* is reminiscent of standard American evaluation terms, implying the need for research expertise. In practice, however, performance management systems have not risen to that level of complexity. Typically they have viewed program monitoring, using quick-turn-around data in internal automated Management Information Systems (MISs) as the primary vehicle for determining progress in meeting performance expectations, rather than evaluation research. Analyses of MIS data have involved relatively simple comparisons of gross outcomes against performance goals, targets and/or standards.

The major activity in designing performance management systems is the development of *performance measures.* Measures can be developed to represent any type of variable that is of interest to those setting policies for a performance management system: for (1) *program inputs*, such as customers' ethnic status and work history; (2) *processes*, such as the way new educational strategies are being delivered; (3) the program's major *interventions*, such as the use of educational vouchers; (4) *interim outcomes*, such as an increase in educational skills in a program designed to increase college enrollment; (5) *customers' short-term* outcomes, such as attainment of a certificate at program completion and (6) *customer's longer-term outcomes*, such as enrollment in a college program. However, the primary focus of performance measurement tends to be on customers' *short-term gross outcomes.* This is a serious concern, since performance management systems tend not to distinguish between gross outcomes and outcomes that can be attributed directly to the program itself (net outcomes/net impact).

In addition, these systems rarely measure implementation processes and often fail to develop precise measures of program interventions. The nature of measurement in performance management systems is, understandably, influenced by the frequency with which data are to be collected, as well as by organizational support. System MISs collect on a continuous basis while evaluations are typically conducted only periodically. Therefore the cost of the MIS data collection effort is an important issue. Performance management systems consequently tend to rely heavily on easy-to-obtain quantitative measures that are conditioned by cost and political acceptance, whereas evaluations frequently seek answers to broader implementation and effectiveness questions that require a richer data base.

The incorporation of incentive/sanction systems to ensure compliance with performance expectations, which has occurred in a number of national programs in the US and projects in other countries, is another characteristic of performance systems that raises concerns. These incentive systems often tie performance management to budgeting. Rewards and penalities can range from simple recognition of a program's ability to meet performance standards to significant monetary rewards, and from reduced funding to the loss of the right to operate the program.

For example, the American Job Training Partnership Act (JTPA) programs have included incentives in the form of increased funding with fewer strings attached for good performance, and penalties in the form of loss of the right for existing administrative units to operate the program if poor performance continues for two consecutive years.

The trend toward linking performance monitoring directly to budgeting appears to be a global phenomenon. It helps explain the recent creation of 'performance auditing' functions in central governments' Inspector General's Offices (OICs). These audits are viewed frequently as quick-turn-around evaluations which have immediate implications for budgeting decisions, even though not conducted by those with research training (Chelimsky, 1985; Moran, 1990; Hatry and Fountain, 1990). This spin-off from the performance management and evaluation research movements is now competing with both for governmental attention, making the separate professional worlds of planners, managers, accountants and evaluators a serious barrier to coordinating performance monitoring and evaluation research.

Major Problems Associated With The Lack Of Integration Of Performance Management And Evaluation Research

Having discussed the nature of the two movements and their main differences in terms of professional development and evolution, it is important to identify the most critical areas in which a lack of commitment to view these two movements as potentially complementary, and to use them in tandem, creates significant problems.

Conceptualization

In focusing largely on outcomes, performance management systems tend to ignore key elements of programs that need to be given attention by policy-makers and program managers.

The description of social programs in legislation or in administrative directives, referred to as *program designs*, essentially expresses *theories of change*. These change theories involve a set of hypotheses that describe causal relationships. As with all theories, in proposing such relationships program designers make certain *assumptions* – about the needs and conditions of the units to be changed (the *target group* of the program), the ability of change agents (*interventions)* to produce the desired changes, and the nature of the *environment* in which a program seeks to produce these changes. Box 1 summarizes the major relationships of interest in program designs. Table 2 provides an example of the more specific elements in program design.

A significant first step in making an assessment of a program's outcomes is a conceptual one: clarification of a program's particular theory of change and the assumptions underlying it. Judgments of a program's value require a *test* of program theory. The quintessential question is 'Do the propositions in the theory have validity when applied in a real-life program environment?'

In most state-level work/welfare programs in the US, for example, a general hypothesis is that a mix and sequence of basic education, language proficiency, employment, occupational training and social services, delivered in a particular way, will increase the skill level, employability and ultimate earnings of welfare clients, therefore reducing welfare system costs. Some programs, however, may propose only that job placement assistance alone will lead to timely employment and economic self-sufficiency. Within each of these general hypotheses are more specific sets of quite complex causal relationships. Underlying them are assumptions about the innate capabilities of people and the nature of organizations and/or communities, the value of varying the way programs are implemented, and the viability of particular outcomes in the context of the severity of the social problem involved.

In outlining earlier the goals of information production, it was apparent that

Box 1. PROGRAM THEORY: Major Hypotheses or Propositions Expressed in a Program's Design

General Hypothesis Involved:

A set of planned influences (*interventions*) unique to the program will produce a number of desired changes (*outcomes* or *effects*) in those exposed to these program interventions (*the target group or unit*).

Specific Hypotheses:

Two major causal relationships are proposed in program theory:

1. The causal relationship given most policy attention: the relationship between the program's intended set of *interventions* and the intermediate, short-term and longer-term outcomes desired.

Example: Community college occupational programs will reskill dislocated workers and reduce their period of unemployment.

2. The causal relationship between the program's intended *implementation mode* (the way the program is expected to be implemented, including the way in which the interventions are to be delivered and the target group exposed to them), and the intermediate, short-term and longer-term outcomes desired.

Example: Performance-based contracts with providers of services will produce better outcomes for the target group.

Table 2. PROGRAM DESIGN: Key Elements or Components

ELEMENT Requiring Clear Definition	*EXAMPLES: using an employment and training program as the general illustration.*
Nature of a program's intended *resources*	*Program funding and in-kind assistance, personnel, access to other resources.*
Characteristics of a program's intended *target group* or unit (a program's customers)	*Demographic characteristics, work history, income history, history in other programs.*
Nature of a program's intended *implementation mode*	
Intended *organizational structures*	
Administrative/management units	*A central body that contracts with other entities for the delivery of services.*
Service delivery units	*'One-stop' multi-service centers.*
Monitoring units	*In-program units operating Management Information Systems, performing limited data analyses on data in these systems.*
Boards, councils, committees under a program, or with which a program articulates, or under which a program is operated	*Workforce Development Boards or Councils coordinating multiple related programs.*
Structures enabling liaison with other programs, organizations, stakeholders, communities	*Interagency committees.*
Intended *organizational functions*	
Administrative/management	*Developing performance-based contracts with service providers.*
Needs assessment, planning and budgeting	*Assessing customers' needs, developing strategic plans and budgets.*
Staff recruitment and training	*Developing hiring criteria, selecting staff.*
Service delivery:	*Service delivery functions separated into 1) units that specialize in client recruitment, eligibility determination,*
Outreach and recruitment of target group	

Table 2. continued

ELEMENT Requiring Clear Definition	*EXAMPLES: using an employment and training program as the general illustration*
Eligibility determination	*assessment and service assignment, and 2) units that focus exclusively on service provision.*
Appraisal of program participants' needs and statuses vis-a-vis the problem condition being addressed by a program	
Assignment of customers to a mix and sequence of interventions	
Case management re customer's progress through a program	
Follow-up with customers after their involvement with a program	
Development of relationships with other entities	*Creation of employer advisory committees.*
Monitoring of customers' progress through a program, customers' outcomes, and additional outcomes.	*Collection/analysis of MIS data.*
Reporting of information about implementation and/or outcomes to funders, political decisionmakers, and others invested in the program (stakeholders)	*Development of periodic reports.*
Characteristics of a program's intended *official interventions*	
A program's services: primary and supportive	*Primary: work orientation and experience, on the job and/or classroom training. Supportive: counseling.*
A program's subsidies	*Child care, medical care, transportation, training vouchers.*
Program activities designed to influence the target group's outcomes, and other outcomes	*Use of input from employer advisory groups.*
Nature of a program's intended *outcomes*	
Intermediate outcomes: *outcomes intended to be achieved as*	*Achievement of particular skill levels.*

Table 2. continued

ELEMENT Requiring Clear Definition	*EXAMPLES: using an employment and training program as the general illustration.*
a prerequisite for achieving the outcomes intended for program completers	
Short-term outcomes: *outcomes intended to be achieved at program completion*	*Employment in unsubsidized job, increased earnings from employment.*
Longer-term outcomes: *outcomes at a designated follow-up point beyond program completion, or over time*	*Retention in employment, earnings gains.*
Net outcomes or effects: *short-term or longer-term effects that can be attributed more exclusively to a program's interventions*	*Unsubsidized employment at higher earnings and with greater retention.*
Characteristics of a program's *environment*	*Aspects of the environment that are most important to control for, in attributing outcomes to a program's interventions.*
Economic	
Demographic	
Socio-cultural	
Historical	
Geographical	
Nature of human services organizational network and program's status within this network	

the full complexity of a program (or system of programs) needs to be understood and described by planners, managers, and evaluators. A clear sorting out and definition of the major variables or influences within the program design can allow these professionals to develop a *set of complementary performance management and evaluation research questions.* The answers to this larger spectrum of questions can be pursued through a *combination* of performance monitoring and competent external scientific evaluations. These complementary approaches can yield information useful to decisionmakers about a program's efficiency, effectiveness, and policy utility; about 'best practices' in implementing programs; and about the value of different program components. It is this professional interconnection that should guide first the selection of variables to be measured, and secondly the set of measures to be used as proxies for these variables, to support both performance management and evaluation purposes.

However, performance management systems tend to be so focused on the measurement of a limited set of outcomes that the true complexity of a program's design is frequently ignored in the information production process. Consequently, too little information is collected about important elements of program implementation, of the interventions considered unique to a program, or of a richer array of outcomes that may be very significant (see Funnel's 1997 article on developing a 'program logic matrix and model' for evaluating social programs in Australia, which supports the need to identify all the key elements of program design, as a basis for performance monitoring and evaluation.)

In the US, the State of Oregon has one of the best developed performance management systems, encompassing all of the State's workforce education, training and employment programs. Oregon's performance plan involves ten goals, seven strategies for achieving the goals, ten interim outcome measures and benchmarks, and five short-term outcome measures and associated benchmarks. Agreement on these core elements across a large number of individual programs has been a difficult even though impressive process, forming the basis for a collective MIS for monitoring the entire system. However, none of the measures provide information about the way programs are implemented or the system is organized, despite process goals, or about the characteristics of the interventions or the quality of the outcomes. Although 'cost effectiveness' is to be measured, only internal staff surveys and assessments are suggested. No role is defined for evaluation research, even to validate the results of performance monitoring (Office of the Governor of Oregon, 1998). The State of North Carolina's workforce plan moves in the direction of the Oregon model, again with an abstract commitment to analyzing effectiveness and cost/benefit tradeoffs but without appreciating the need for process measures or evaluation expertise (Office of the Governor of North Carolina, 1998). In both cases, assessments depend exclusively on simple internal analysis of a small number of gross outcome measures collected in system MISs.

Clearly the use of performance measures and standards is appropriate for *monitoring compliance* with governmental regulations regarding a program, for *comparing program realities against formal program plans*, and for *judging the level of outcomes achieved for a program or system's customers.* But performance management is of far less use in understanding how the interventions or the implementation process may or may not have caused or influenced these outcomes. Therefore, lacking periodic evaluations of implementation and net impact, performance management systems are not capable of *testing program theory* (Wirt, 1995; Pawson and Tilley, 1997). Yet such tests are the logical basis for making policy and program modifications.

A serious problem related to these conceptualization issues is the tendency for those responsible for monitoring and reporting results in performance management systems to interpret data on performance measures *as if* they represented the direct and exclusive effects of program interventions, when these outcomes may be only weakly correlated with them. In many cases, it is the *combination* of a program's interventions and other 'treatments' to which a program's customers are exposed that are responsible for the effects observed. These innuendos are missed by most performance management systems. Periodic scientific evaluations could provide more accurate and comprehensive information to decisionmakers, and fill critical information gaps left by traditional performance management approaches.

Measurement

Tending to rely on a narrow set of quantitative gross outcome measures accessible through Management Information Systems, performance management systems have been slow to recognize and address data validity, reliability, comparability, diversity, and analysis issues that can affect judgments of programs.

General Problems In the rush to develop performance measures, targets and standards emphasizing gross outcomes, there are substantial potential risks that information production will be insufficient or misleading:

- The measures selected may not be highly correlated with the system's goals and objectives;
- The measures selected for key outcome variables may lack validity – that is, they may not be the best proxies for these variables;
- Too few measures of each of the key variables may have been included in MISs;
- The measures selected may not yield sufficient information about the characteristics of target groups, the nature of service delivery, and the services or activities (outputs) that are expected to directly influence outcomes (Funnel, 1997). *Examples are the income and work histories of those receiving employment and training interventions, the practices of program staff regarding service assignment, the quality and duration of services;*
- Key outcomes that require qualitative measurement may not be incorporated within the monitoring mandate (that is, *outcomes such as increased partnership development, organizational coordination or social cohesion*);
- The measures may not include program costs, as a basis for gaining insights about the tradeoff between costs and the gross outcomes realized;
- The measures may be applied equally to diverse program environments and

situations, when in fact a given set of measures may be more useful in one environment or context than in another;
- The performance measurement process may become an end in itself (Kettl, 1997).

Most states in the US have confined their set of measures for federally funded, state-operated employment and training programs to job placement, earnings, and retention in employment. These quantitative measures can be accessed from an ongoing national database. Typically no measures for intermediate outcomes are collected, such as the acquisition of basic educational or occupational skills, nor are measures of longer-term outcomes collected, such as access to career development and promotion opportunities, or increased work benefits. More important, missing from the database is information which could provide insights about why the outcomes collected are occurring. The Advisory Panel of the National Academy of Public Administration in the US has, itself, recommended that richer, more valid, more reliable, and more timely measures be used in performance management systems (National Academy, 1997). Cautions have been raised in Australia about the decline in support for evaluation research in the context of the performance management trend, with the fear that critical linkages between inputs, processes, outputs and impacts will be neglected (Funnel, 1997).

Data Validity Measurement 'validity' reflects the extent to which a data element truly represents a more abstract variable. For example, grades on an educational achievement test may or may not be a valid measure of educational performance. In the performance systems of many employment and training programs, 'obtained a job following program termination' represents employment success, even if the job is tenuous, of low quality, has few benefits and exhibits poor working conditions.

More seriously, in most of the American states participating in the National Governors' Association's project on performance management, certain outcome measures were taken for granted as *representing* program or system goals (National Governors' Association, 1994). This illustrates the tendency for performance management systems to leap from abstractions to specific measures without first clearly defining goals, operationalizing objectives under these goals, identifying key variables involved in the objectives, and then selecting or developing measures that more closely represent these variables. Such a broad leap reduces the validity of the measures. The problem is compounded by selecting measures mainly on the basis of their availability in pre-existing information systems.

Evaluation researchers have sometimes been naive about measurement issues as well, but the professional commitment to developing valid measures, and to testing their validity formally, has been greater in the evaluation research movement, consistent with its emphasis on scientific standards.

Data Reliability Evaluation researchers utilizing MIS monitoring data in addition to freshly collected information usually must deal with significant reliability problems. Often these data sets must be 'cleaned' and statistical strategies used to compensate for missing data, both sizable tasks. Data quality and data collection reliability are frequently ignored (Decker, 1989; Stevens, 1989)

A 1994 Office of Technology Assessment project studied numerous efforts to use administrative data for performance management purposes and identified a number of data reliability problems. The Northeast/Midwest Institute, the National Commission for Employment Policy, and Mathematica Policy Research, sources of competent research at the national level in the US, all found problems with the data collected in MISs. Yet most states' performance management systems rely exclusively on these administrative data (US Congress, 1994).

Data Comparability In the US, states have been encouraged by federal legislation to develop performance management systems for statewide multi-program human service councils. In 1996, a federal interagency task force developed a core set of measures for monitoring and evaluating labor market programs, which was to form the basis for state-level MISs supporting performance monitoring by workforce councils (US Department of Labor, 1997). Some individual programs under these councils utilized this core set, but no state is operating a fully *integrated* data system serving multiple programs. Consequently the choice and definition of performance measures differ substantially from one program to another, even though under the rubric of a single performance management system.

Performance management systems that lack technical cross-walks from one data set to another, or lack a single automated system for a cluster of related programs, must use genuine caution in interpreting the meaning and significance of the outcomes that may be compared across programs. However, few performance systems have used such caution. As a result, often evaluators have had to develop their own set of comparable measures.

Data Diversity Evaluation researchers must frequently collect both qualitative and quantitative measures of the major variables, since performance management systems tend to ignore the need to develop qualitative measures. Frequently variables that are best measured qualitatively are left out of performance monitoring efforts.

Data Analysis Important insights can be gained from studying simple statistical tables reporting the results of performance monitoring. However, performance management systems tend not to apply the more sophisticated statistical strategies in analyzing monitoring data to answer managerial questions. This can result in misleading interpretations of program outcomes which then affect decisions about improving the program. Evaluations using state-of-the-art statistical methods can correct for interpretation errors.

Methodology

Performance management systems usually do not seek to isolate the net impact of a program – that is, to distinguish between outcomes that can be attributed to the *program rather than to other influences. Therefore, one cannot make trustworthy inferences about the nature of the relationship between program interventions and outcomes, or about the relative effects of variations in elements of a program's design, on the basis of performance monitoring alone.*

Performance management systems pose a major methodological problem. There is a tendency to fail to distinguish between (1) *gross outcomes*, using MIS measures collected for performance monitoring purposes, which may be due to influences other than the program being monitored, and (2) *net impacts* that are more directly attributable to the unique nature of that program. *This is a significant methodological weakness.* Gross outcomes may, in some cases, constitute reasonable proxies for net effects. But net effects may differ in important ways from gross outcomes, leading to quite different policy and managerial decisions. If cautions are not forthcoming about the potential misinterpretation of gross outcomes, faulty decisions can be made about modifying or eliminating programs.

A cogent example is the European Commission's ESF workforce program. Gross outcomes for the hard-to-serve population, which were collected for performance monitoring purposes, proved to be negative, but a net impact evaluation yielded positive effects. Similarly, performance monitoring data tracking the gross outcomes of students participating in occupational programs in two-year colleges in Washington State in the US suggested that these programs were having positive employment effects. But a subsequent net impact evaluation yielded negative results. If decisions about these programs were to have been made solely on the basis of performance monitoring, they could have led to flawed decisions about their continuation.

Because a performance management system is expected to provide data quickly and inexpensively, it is sometimes rationalized that the interpretation of the data need not be as accurate or qualified as in evaluations. In too many instances performance monitoring has been *defined* as 'evaluating', revealing a lack of knowledge of the benefit of scientific procedures. In such cases the need for periodic pre/post gross outcome, net impact, and cost/benefit studies of individual programs is particularly great.

The importance of combining performance management with evaluation research was recognized by the US Department of Labor in the 1990s. A performance management system was developed for all workforce training and employment programs, using a range of performance measures. In addition, the Department developed a plan for systematic evaluations of workforce programs: both process and net impact evaluations. Although evaluations have not been closely coordinated with performance reporting to serve key decision points in policy-making or management, the Department now views the two forms of assessment as complementary information tools.

The difference between gross and net outcomes has implications for cost/benefit analysis, since the calculation of the cost/benefit ratio is dependent on the estimation of net impacts. However, many performance management systems are expected to feed gross outcome information directly into budgeting decisions, without the benefit of cost/benefit evaluations. Evaluation research can serve as an essential check on judgments based only on performance data, prior to linking judgments of programs to budgeting policies.

For example, in the US the National Performance Review requires each federal agency to produce 'results-oriented financial statements' based on performance agreements using performance measures. These statements essentially link policy priorities, performance results, and funding – potentially setting a dangerous precedent (National Performance Review, 1994/1995).

Performance Standards

An additional issue is the use of *performance standards,* and *incentive/sanction systems* to assure compliance with them. Frequently standards are developed prematurely, prior to a careful definition of program design and the selection of the main variables to measure within it. Even though the measures chosen for the standards may not be the best outcome measures to use, programs are held accountable if these imperfect standards are not met (Hatry and Fountain, 1990). Incentive/sanction systems supporting standards pose their own problems. Together, standards and incentives have been shown to redefine program goals away from their original intent.

For example, a series of studies conducted in 1995 analyzed data from the American Job Training Partnership Act Experiment to determine the influence performance standards and incentives had on managerial behavior. They provided evidence that JTPA standards and incentives led to a reorientation of goals in the direction of meeting or exceeding standards, away from the pursuit of program goals. The studies also indicated that the short-term measures in the standards were only weakly (and often negatively) correlated with longer-term employment and earnings effects, the intended goals of these job training programs (Heckman, 1999).

A National Governors' Association report on the use of skills standards suggested that setting absolute standards was directing behavior in the wrong direction (National Governors' Association, 1994). A 1988 SRI International study of JTPA concluded that performance standards had decreased services to the 'hard to serve'. A Manpower Demonstration Research Corporation study suggested that the outcomes collected in welfare-to-work programs were only weakly correlated with program goals and sometimes undercut them. A 1991 report by the US General Accounting Office suggested that JTPA performance standards were resulting in serious inequities in service provision (US General Accounting Office, 1992).

These studies reveal a major methodological issue facing performance management systems, since programs with high performance against standards may be shown by rigorous evaluations to be failing to produce the outcomes intended for them by their designers.

Internal Self-Assessment vs *External Evaluations*

In addition, there is the issue of *internal self-monitoring* vs *external evaluation* activities. Internal assessments cover studies carried out by staff within a performance system or a program, usually within the system's or program's monitoring unit, which has minimal training in research methods and an understandable investment in the success of the program being monitored. Scientific evaluations are expected to be carried out by researchers independent of (external to) the program being studied – that is, by professionals that do not have such an investment. Many performance management systems view internal self-assessments as sufficient to satisfy information production and accountability requirements. However, these self-evaluations can be highly vulnerable to the agendas of those managing the performance system, introducing considerable bias. (This position is taken in the European Commission's MEANS Handbook No. 1, 1995, the first in a series of evaluation guides for Monitoring Committees overseeing the Commission's Structural Funds projects in the member countries of the European Union.)

Given that the major purpose of performance management systems is to develop indicators that can be used as short-run indicators of long-term goal achievement, the potential for managerial manipulation is quite serious. If meeting standards is made the basis for budget decisions, programs that are ineffective in achieving program goals but successful in meeting standards may be continued, while programs having difficulty meeting standards perhaps due to their efforts to meet program goals may be eliminated.

Overflow audiences for sessions on performance management at the Canadian Evaluation Society's 1997 international evaluation conference, which revealed the global influence of the performance measurement movement, nevertheless identified similar concerns to those voiced here. Conference attenders from Australia took the position that the 'performance indicator craze' was repeating the mistakes of the 1970s, and predicted that a singular focus on a limited set of outcome measures would be abandoned soon in Australia. A report on the Netherlands' use of performance management criticized the system for its minimal information base, its resistance to integrating evaluation research, and its overreliance on quantitative measures and standards (see Blalock, 1997).

Recommendations for Integrating Performance Management Systems and Evaluation Research

Hopefully the preceding discussion has generated an interest in securing the potential benefits of creating better articulation between the performance management and evaluation movements. Since performance management systems appear to be the preferred postmodern strategic planning and quality management model for ongoing goal-directed activity, and for assuring accountability in terms of goal achievement, the best strategy for maximizing the complementary benefits of the two movements would seem to be a consideration of ways to integrate evaluation research *within* performance management systems. Integrating so that the differing purposes, perspectives, priorities and expertise of the two movements *complement* one another would be a major task. Easier said than done, given the current institutional status of both professional territories.

However, there are two directions one might take: one is to address some of the conceptual, measurement and methodological concerns raised in the previous discussion; another is to consider pragmatic ways to accomplish integration.

Conceptualization

- Bring researchers into the strategic planning process as partners in producing critical information, so that this planning process is infused with competent advice about the kind of evaluations that will be most feasible and useful to decisionmakers, and at what decision points.
- Encourage evaluation researchers and performance management staff to familiarize themselves more thoroughly with the theories on which programs are based.
- Include within a performance management system a set of both qualitative and quantitative measures describing the key variables involved in the theory of a program. This database can assist evaluators and reduce evaluation costs.
- Integrate periodic comprehensive evaluations, *at key decision points* within the strategic planning process, to complement the information base for decisionmaking and to act as a check on the accuracy of information gathered through system MISs.

Measurement

- Involve evaluation researchers, performance management staff, customers and stakeholders in the development of a range of process and outcome measures reflecting the complexities of program design, that can enrich a simpler core set of performance monitoring measures. This would support both performance management and evaluation research activities.
- Increase the sophistication of MISs in terms of their technical data analysis capabilities, and of MIS staff in terms of their data analysis training.
- Give more attention to data quality and reliability.
- Build in organizational incentives for managers to actively *utilize* the complementary information available to them.

Methodology

- To reduce information bias, enhance professional integration, and secure the commitment of planners/managers and evaluators to the collection and use of complementary information, provide training seminars to evaluators on the principles and practices of performance management, and to performance management staff on basic research principles and methods.
- Mesh evaluation research activities with performance management activities through a *formal monitoring/evaluation planning process.*

Performance Standards

- Clearly define the purpose of performance standards and incentive systems.
- Base the development of standards on a comprehensive set of measures that serve as valid proxies for key variables describing a program.
- Develop performance standards only after a broad range of relevant program information has been collected over a reasonable length of time,

so that analysts can determine which measures are the most valid, reliable, appropriate and useful for incorporation in standards. Develop incentive/sanction systems only very gradually, drawing insights from the operation of a program, the apparent effects of using standards, and the results of evaluations, being aware of the potential negative impact of incentive/sanction systems on management decisions.

- Make available to funders and stakeholders information that will help *explain* the meaning, significance and limitations of the gross outcomes collected in performance management systems, using analyses of MIS and evaluation research data.
- Adjust performance standards using regression models, assuring that the adjustments are fair and create the right incentives. Adjustments can be used to encourage equity as well as to correct for different environments, and can increase the positive correlation between standards and program goals.

Ways to Accomplish Integration

Two beginning questions may be helpful in addressing the integration issue: (1) *at what key decision points in the strategic planning/managerial process would evaluation expertise and/or activities be most valuable? and (2) what kind of division of labour between planners/managers and researchers would be most useful at these decision points?*

Answering these beginning questions will require changes both within and outside a performance management system.

Internal cultural change: a new perspective and commitment on the part of planners and managers that supports the complementary use of performance management and evaluation research.

External cultural change: new expectations on the part of those funding and overseeing performance systems which mandate useful integration between performance monitoring and evaluation research. (This may require new legislation or new priorities for the use of administrative funds.) Funders need to view the start-up costs of integration as a cost-effective *investment* in the context of the longer-term benefits of obtaining more accurate, in-depth, varied and useful information.

Technical change: revised expectations for information production, dedicated to: (1) identifying what questions need to be answered given the information needs, interests and concerns of planners and managers, and in the context of a system's mission, goals and desired outcomes (it will be critical to distinguish between questions that primarily satisfy a *monitoring* purpose, those that primarily satisfy an *evaluation* purpose, and those that satisfy both purposes); (2) developing a timetable for introducing performance standards and associated incentive sanction systems, based on a study of monitoring and evaluation information over time; and (3) increasing the validity and reliability of MIS data and MIS analysis capability, so that MIS information is more useful for evaluation purposes, can reduce evaluation costs, and can enable clearer, more technically sophisticated and 'researchable' requests for independent (external) evaluations.

These improvements require supportive organizational structures and clearly defined professional roles, which can create and sustain a cooperative, coordinated division of labor between planners/managers and evaluators throughout system activities – a division that reduces the perception that established professional territories are being compromised.

A general organizational model is dependent on the creation of four organizational structures additional to those operating in most performance management systems: a small, carefully selected, interdisciplinary *research advisory group* that provides assistance throughout the planning and management process; the creation of a special position for an *oversite coordinator* responsible for assuring the complementary use of monitoring and evaluation research information; a *dissemination/utilization unit* dedicated to preparing information for a variety of users; and a *cross-training effort* that provides monitoring staff with a familiarity with research principles and methods, and gives researchers advising the system or performing external evaluations for the system an orientation to the tasks managers and monitoring staff perform. Table 3 suggests a division of labor regarding the research advisory group.

In the US, the proposal developed for the State of Texas's performance system for workforce education and training is an example of such an integrated system. The proposal distinguishes three kinds of accountability: fiscal, compliance-oriented, and program-oriented. The monitoring of performance measures against goals, targets, standards and/or benchmarks on an ongoing basis illustrates compliance accountability, requiring managerial expertise and the development of a common automated information system collecting a core set of performance measures. Program accountability is concerned with whether or not a program within the system 'made a difference', requiring periodic evaluations of gross or net outcomes using evaluation research expertise. All three acountability functions are to provide complementary information for use by planners, managers and stakeholders in making system decisions (King and McPherson, 1997).

Conclusions

As a monitoring tool for program managers, performance management can provide important short-term, quick-turn-around information for tracking progress against stated goals, focusing on outcome measures. As a research tool, evaluation research can provide the broad range of information needed to make relatively unbiased judgments of program or system efficiency and effectiveness. Each approach has its own benefits in terms of increasing the accountability of programs and human service systems to their customers and stakeholders, and enhancing policymakers' ability to make choices among a range of program options.

Used as *complementary* tools, the two movements can offer vastly more valid and reliable information to decisionmakers, and can more accurately guide improvements in programs. Performance management contributes information

Table 3. INTEGRATING EVALUATION RESEARCH WITHIN PERFORMANCE MANAGEMENT SYSTEMS

Major Tasks of Performance Management Systems	*Roles Evaluation Researchers Can Play in Performance Management Systems, As Members of A Research Advisory Group*
I. Strategic Planning Tasks	
Development of mission, goals and objectives for system.	Researchers can play a participant-observation role, to absorb accurate information about planning decisions and provide appropriate input.
Identification of key variables (influences) reflected in the objectives, for which measures need to be developed: both process and outcome measures.	Researchers can contribute their expertise to the identification of key variables in program designs.
Development of a core set of measures to be collected in a Management Information System for *monitoring* implementation and performance, and additional measures that could be helpful to evaluators in *evaluating* the system and/or programs within the system.	Researchers can contribute to the selection of core measures and additional measures likely to be useful for evaluations.
Development of performance priorities, targets, standards, and incentive systems.	Researchers can provide advice regarding the appropriate time to introduce standards and incentives.
Development of a joint *monitoring and evaluation plan* for the system: issues to be addressed, kinds of activities to be conducted, budgets, timetables. (This would include the purpose of oversight activities, the activities to be engaged in to achieve that purpose, the resources needed, and the timing of process, gross outcome and/or net impact evaluations to provide information at key decision points.)	Researchers can help develop an annual and multi-year *monitoring/evaluation plan* for the system and its programs.
II. Managerial Tasks	
Organizational and fiscal management of system and its programs.	Researchers have no special role.
Development of structures usefully linking the system and its programs with funders, stakeholders, other public/private entities important to the system.	Researchers can help identify potential external research firms and individual evaluation researchers in academia for contracting for evaluations, and review their qualifications for conducting the kinds of evaluations identified in the monitoring/evaluation plan.

Table 3. continued

Major Tasks of Performance Management Systems	*Roles Evaluation Researchers Can Play in Performance Management Systems, As Members of A Research Advisory Group*
Construction of an MIS capable of collecting, storing, and analyzing information collected using the measures developed.	Researchers can provide guidance.
Development of a Monitoring/MIS Unit capable of analyzing and reporting MIS information for monitoring performance and assisting evaluators.	Researchers can provide advice regarding what level of research familiarity is needed.
Performing analyses of ongoing MIS data, and conducting internal *self-evaluations* by the Monitoring Unit.	Researchers can advise Monitoring Unit staff about strategies for reducing information bias.
Development of *Requests for Proposals* for evaluations to be conducted by external researchers.	Researchers can assist in developing such requests.
Selection of external research firms/evaluators, and development of research contracts.	Researchers can assist in the selection.
Coordination of oversite activities between the Monitoring Unit and external evaluators.	Researchers can suggest ways to relate effectively to external evaluators.
Organization of the service delivery systems for programs within the system: recruitment, assessment, service assignment, customer monitoring, follow-up.	Researchers can help interpret evaluation activities to service delivery staff, particularly in cases where the design of the research alters or interferes with some aspect of service delivery.
Review of the progress and results of activities intended to achieve the goals of the monitoring and evaluation plan.	Researchers can contribute their review.
Presenting and distributing oversite information for different user groups.	Researchers can assist in explaining the methodologies used and review the cautions surrounding the interpretation of findings.

for strategic planning, monitoring and operational efficiency. Evaluation research contributes information about the causal processes involved in programs, and provides a check on the validity of shorter-term performance monitoring strategies. Evaluations can be used in selecting performance measures and developing strategic plans, and the strategic planning process can help agencies determine the outcomes against which programs should be evaluated.

Consequently, a major direction for the future should be to more fully coordinate evaluation research within performance management systems, moving toward full integration. This will require a new interdisciplinary culture – a belief in the utility of meshing different kinds of professional knowledge and expertise to accomplish common goals, and an organizational commitment to do so.

Notes

1. Clearly some readers will quarrel with the designation of performance management and evaluation research as 'movements'. However, movements are characterized by the gradual development and fluorescence of certain patterns of thought and action over time, and both performance management, as a strategic planning and management approach, and evaluation research, as a new direction in applied social research, come close to qualifying as bona fide movements within the life of public and private bureaucracies. This conclusion assigns special importance to the influence of each movement on planning, managerial and research perspectives.
2. In response to a fiscal crisis in the early 1980s, New Zealand applied a set of principles and approaches across all government agencies. Some of the key elements were a reorganization of government administration through clarification of goals and the collection of information on performance, a new management framework for the delivery of services based on performance objectives, and a distinction between outcomes and outputs. Scott et al. (1997) suggest that the New Zealand model has drawn insights from the American economic literature, and from the cyclical planning/programming, monitoring, reporting, budgeting process pursued by the US General Accounting Office. But the process follows a classic planning chronology.
3. The disparate history of the performance management and evaluation research movements reveals their different professional roots and courses of development. In the US, the evaluation research movement was launched in the 1960s and has been at its developmental peak in the 1990s, emphasizing the importance of experimental methodologies. In Europe and other postindustrial countries, evaluation research developed much later, receiving public administration attention only in the late 1980s and 1990s, and emphasizing, in most cases, nonexperimental research designs.

   Performance management had many precursors, such as 'management by objectives' concepts in the US, and the New Zealand model in a significant number of other countries. In all cases, new approaches for establishing accountability were spawned by budget deficits, the decentralization of social programs, and strong governmental efforts to determine the results of social expenditures. By the 1990s, most developed countries had applied performance management ideas to government activities, as well as increasing mandates for scientific evaluation. But little effort has been made to coordinate the two movements.

   For information on the history of these movements, as a context for understanding their current statuses, please see the following references: Demings,1982; Osborne and

Gaebler, 1992; US General Accounting Office, 1993a, 1993b, 1994, 1995a, 1995b, 1997a, 1997b, 1997c, 1997d; Rossi and Freeman, 1994; Sanders, 1994; Friedlander and Burless,1994; National Governors' Association, 1994; Hatry, 1994; Leeuw and Rozendal, 1994; National Performance Review, 1994/1995; US Department of Labor, 1994–1996; Duran et al., 1995; Muller-Clemm, 1995; European Commission, 1995a, 1995b, 1997; Organization for Economic Cooperation and Development, 1995, 1997; Pawson and Tilley, 1997; Blalock, 1997; Scott, et al., 1997; Kettl, 1997; Pollit, 1998; Toulemonde et al., 1998.

In a paper given at the International Evaluation Association conference in Vancouver, BC in 1994, Sue Funnel commented that Australia used a hierarchy of six stages of organization as a framework for studying continuous improvement, which also formed the basis for implementation studies. And she reported that Australia has a formal mandate and timetable for evaluation activities, a cycle of 3–10 years, to determine the appropriateness, efficiency and effectiveness of programs. Evaluations are expected to be an ongoing managerial responsibility, an integral part of planning and budgeting. However, evaluation funding has permitted mainly internal self-evaluations, with consequent problems with their objectivity (Blalock, 1997).

4. Both Sue Funnel (1997) and Michael Patton (1990) refer to program designs as 'theories of action'. Funnel shares the view that a program's theory (what she calls its 'program logic') is a critical tool for conceptualizing and measuring the performance of a program. She proposes a model for use in performance management/evaluation systems. Her 'hierarchy of outcomes' in the model mirror, to some extent, those suggested in Table 2.

References

Blalock, A. B. (ed.) (1990) *Evaluating Social Programs.* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Blalock, A. B. (1997) 'The European Commission's Effort to More Effectively Structure Evaluation Activities', *Evaluation Forum 12.*

Bloom, H. (1993) 'The National JTPA Study: Origins, Objectives and Issues of Interpretation', *Evaluation Forum* 9.

Chelimsky, E. (1985) 'Comparing and Contrasting Auditing and Evaluation', *Evaluation Review* 9.

Decker, P. (1989) *Systematic Bias in Earnings Data Derived from Unemployment Insurance Wage Records and Implications for Evaluating the Impact of Unemployment Insurance Policy on Earnings*. Princeton, NJ: Mathematica Policy Research.

Demings, W. E. (1982) *Out of the Crisis.* Cambridge, MA: Massachusetts Institute of Technology's Center for Advanced Engineering Study.

Duran, P., E. Monnier and A. Smith (1995) 'Evaluation a la Française', *Evaluation* 1(1): 45–63.

European Commission (1995a) *Common Guide for Monitoring and Interim Evaluation.* Brussels: EC Structural Funds.

European Commission (1995b) *MEANS Handbook No.1: Organising Intermediate Evaluation in the Context of Partnerships.* Brussels: European Commission.

European Commission (1997) *Evaluating European Expenditure Programs: A Guide.* Brussels: European Commission.

Friedlander, D. and G. Burtless (1994) *Five Years After: The Long-Term Effects of Welfare-to-Work Programs.* New York: Russell Sage Foundation.

Funnel, Sue (1997) 'Program Logic: An Adaptable Tool for Designing and Evaluating Programs', *Evaluation News and Comment* 5(July).

Gueron, J. and E. Pauly (1991) *From Welfare to Work*. New York: Russell Sage Foundation.

Hatry, H. and D. Fountain (1990) *Service Efforts and Accomplishments Information: Its Time Has Come.* Norwalk: Governmental Accounting Standards Board.

Hatry, H. and J. Wholey (1994) *Toward Useful Performance Measurement: Lessons Learned from Initial Pilot Performance Plans Prepared Under the Government Performance and Results Act.* Washington, DC: National Academy of Public Administration.

Heckman, J. (1993) 'Randomization and Social Program Evaluation', in C. Manski and G. Garfinkel (eds) *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.

Heckman, J. (1999) *Performance Standards in a Government Bureaucracy*. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Heckman, J. and V. J. Hotz (1989) 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs', *Journal of the American Statistical Association* 84.

Hotz, V. J. (1992) 'Designing an Evaluation of the Job Training Partnership Act', in C. Manski and G. Garfinkel (eds) *Evaluating Welfare and Training Programs.* Cambridge, MA: Harvard University Press.

Kettl, D. (1997) *The Global Revolution in Public Management.* Washington, DC: The Brookings Institution.

Kettl, D., P. Ingraham, R. Sanders and C. Horner (1996) *Civil Service Reform: Building a Government That Works.* Washington, DC: The Brookings Institution.

King, C. T. and R. E. McPherson (1997) *Evaluation Plan for the Texas Workforce Development System*. Austin, TX: Center for the Study of Human Resources, University of Texas.

Leeuw, F. L. and P. J. Rozendal (1994) 'Policy Evaluation and the Netherlands Government: Scope, Utilization and Organizational Learning', in F. L. Leeuw (ed.) *Can Governments Learn?* New Brunswick, NJ: Transaction Publishers.

Moran, William C. (1990) 'Evaluation Within the Federal Offices of Inspectors General', *New Directions in Program Evaluation* 48.

Muller-Clemm, W. J. (1995) 'A Historical Perspective on Federal Program Evaluation in Canada'. Paper given at the International Evaluation Research conference in Vancouver, British Columbia.

National Academy of Public Administration (1997) *Improving Performance/Improving Government.* Washington, DC: National Academy of Public Administration.

National Governors' Association (1994) *Building State Workforce Development Systems: The Critical Roles of Policy Coordination and Quality Assurance.* Washington, DC.

National Performance Review (1994/1995) *From Red Tape to Results: Creating a Government that Works Better and Costs Less*; *Putting Customers First: Standards for Serving the American People; and Common Sense Government*. Washington, DC.

Office of the Governor of North Carolina (1998) Report of the Governor's Commission on Workforce Preparedness. Raleigh, NC: Office of the Governor of North Carolina.

Office of the Governor of Oregon (1998) *Report of the Governor's Office of Education and Workforce Policy*. Salem, OR: Office of the Governor of Oregon.

Organization for Economic Cooperation and Development (1991) *Evaluating Labour Market and Social Programmes: The State of the Art.* Paris: OECD.

Organization for Economic Cooperation and Development (1995) *Governance in Transition: Public Management Reforms in OECD Countries.* Paris: OECD.

Organization for Economic Cooperation and Development (1997) *Benchmarking, Evaluation and Strategic Management in the Public Sector*. Paris: OECD.

Osborne, D. and T. Gaebler (1992) *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector.* Reading, MA: Addison-Wesley Press.

Patton, M. (1990) *Qualitative Evaluation and Research Methods.* Thousand Oaks, CA: Sage Publications.

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

Pollit, C. (1998) 'Evaluation in Europe', *Evaluation* 4(2): 214–24.

Rahn, M., E. Hoachlander and K. Levesque (1992) *State Systems for Accountability in Vocational Education.* Washington, DC: National Center for Research in Vocational Education, US Department of Education.

Rossi, P. and H. Freeman (1994) *Evaluation: A Systematic Approach.* Thousand Oaks, CA: Sage Publications.

Sanders, J. (1994) *The Program Evaluation Standards.* Thousand Oaks, CA: Sage Publications.

Schmid, Gunther (1997) 'The Evaluation of Labour Market Policy', *Evaluation* 3(4): 409–34.

Scott, G., I. Bell and T. Dale (1997) 'New Zealand's Public Sector Management Reform: Implications for the US', *Journal of Policy Analysis and Management* 16(3).

Smith, P. (1995) 'On the Unintended Consequences of Publishing Performance Data in the Public Sector', *International Journal of Public Administration* 18.

Stevens, D. (1989) *Using State Unemployment Insurance Wage Records to Trace the Subsequent Labor Market Experiences of Vocational Education Program Leavers. National Assessment of Vocational Education*. Washington, DC: US Department of Education.

Toulemonde, J., C. Fontaine, E. Landren and P. Vinche (1998) 'Evaluation in Partnership', *Evaluation* 4(2): 171–88.

US Congress (1994) *Performance Standards for the Food Stamp Employment and Training Program.* Washington, DC: Office of Technology Assessment.

US Department of Labor (1997) *Core Data Elements and Common Definitions for Employment and Training Programs*, Washington, DC: US Department of Labor.

US General Accounting Office (1992) *Program Performance Measures.* Washington, DC: US Congress.

US General Accounting Office (1993a) *Using Performance Measures in the Federal Budget Process.* Washington, DC: US Congress.

US General Accounting Office (1993b) *Measuring Performance and Acting on Proposals for Change.* Washington, DC: US Congress.

US General Accounting Office (1994) *Managing for Results: State Experiences Provide Insights for Federal Management Reforms*. Washington, DC: US Congress.

US General Accounting Office (1995a) *Government Reform: Goal-Setting and Performance.* Washington, DC: US Congress.

US General Accounting Office (1995b) *Managing for Results: Experiences Abroad Suggest Insights for Federal Management Reforms.* Washington, DC: US Congress.

US General Accounting Office (1997a) *Measuring Performance: Strengths and Limitations of Research Indicators*. Washington, DC: US Congress.

US General Accounting Office: (1997b) *GPRA – Managerial Accountability and Flexibility Pilot Did Not Work As Intended*. Washington, DC: US Congress.

US General Accounting Office (1997c) *Managing for Results: Analytic Challenges in Measuring Performance.* Washington, DC: US Congress.

US General Accounting Office: (1997d) *Agencies' Strategic Plans Under GPRA: Key Questions to Facilitate Congressional Review.* Washington, DC: US Congress.

Volcker, P. (1989) *Leadership for America: Rebuilding the Public Service.* Washington, DC: The Federal Reserve.

Wholey, J. K. (1989) *Improving Government Performance.* San Francisco, CA: Jossey-Bass.

Wirt, J. G. (1995) *Performance Assessment Systems: Implications for a National System of Skill Standards.* Washington, DC: National Governors' Association.

Zornitsky, J. and M. Rubin (1995) *Establishing a Performance Management System for Targeted Welfare Programs.* Washington, DC: National Commission for Employment Policy.

ANN BONAR BLALOCK is an evaluation research consultant and editor of the US Department of Labor's research journal, *Evaluation Forum*. She has written and edited three books and numerous articles on research. She has been consultant to state governments in the US, as well as a presenter in OECD and European Commission research seminars. Please address correspondence to Admiralty Inlet Consulting, PO Box 409, Hansville, WA 98340, USA. [email: AnnBlenski@aol.com]