ELSEVIER

Evaluation and Program Planning 28 (2005) 1–14

EVALUATION and PROGRAM PLANNING

www.elsevier.com/locate/evalprogplan

Assuring the quality of evaluative information: theory and practice

Robert Schwartz[a,*], John Mayne[b,1]

[a]Department of Political Science, University of Haifa, Mount Carmel, Haifa 31905, Israel
[b]654 Sherbourne Rd, Ottawa, Ont., Canada K2A 3H3

Abstract

There is now a large supply of evaluative information in the forms of evaluation, performance reporting and performance auditing. Relatively little attention has been paid to assuring the quality of this information. The article explores the origins, practice and consequences of evaluative information quality assurance in light of the political and organizational environments within which it occurs. Information was collected from nine countries and two international organizations. While these jurisdictions practice a wide variety of structural, formative, summative and systemic quality assurance approaches, routine active metaevaluation tends to be a sporadic and spotty undertaking. The prevalence of quality assurance initiative varies across types and jurisdictions. Performance audit leads the pack, followed by program evaluation and performance reporting. There is considerable incidence of unintended consequences including decoupling and colonization. The risks of these phenomena increase when quality assurance is cast upon organizations from the outside.

Keywords: quality assurance; evaluation; performance audit; performance reports; metaevaluation; evaluation audit

There is now a plethora of evaluative information available to public sector managers, senior government officials, legislators and the public at large. Evaluation, performance reporting and performance auditing provide considerable amounts of evaluative information about government activities. While evaluative information has become widely available, relatively less attention has been paid to issues of quality including reliability, validity, credibility, legitimacy, functionality, timeliness and relevance. Yet evaluative information that lacks these characteristics stands little chance of legitimately enhancing performance, accountability and democratic governance.

Scholars and practitioners have not ignored quality issues of evaluative information. They devote considerable attention to developing standards for program evaluation and for performance auditing and have made some headway in developing standards for performance reporting. The evaluation literature also includes substantial descriptive and prescriptive writing about metaevaluation.

Yet little has been reported about the practice of quality assurance across these three different types of evaluative information. Even less has been written about the consequences (positive and negative) of quality assurance practices on evaluative information or about obstacles to its successful implementation. The article explores how various national and international organizations assure the quality of evaluative information. The purpose is to describe and compare quality assurance approaches for evaluative information. We generate hypotheses concerning the impacts of quality assurance and about organizational and political impediments to successful implementation. And we use illustrations to provide preliminary support for these hypotheses.

1. A need for quality assurance

Program evaluation activities flourish in a variety of settings, producing thousands of reports each year for scores of local, regional and national governments throughout the world and for a slew of international organizations, notably the World Bank and the European Union. New public management or reinvention initiatives in a large number of

* Corresponding author. Tel.: +972-4-8240599; fax: +972-4-8257785.
E-mail addresses: robsch@poli.haifa.ac.il (R. Schwartz), john.mayne@rogers.com (J. Mayne).
[1] Tel.: +1 613 729 9877; fax: +1 613 729 4161.

0149-7189/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.evalprogplan.2004.10.001

countries, regional governments and municipalities now require government organizations to produce performance reports that are used, inter-alia, in the annual budgeting process. The mandates of state (national) auditors in a wide range of settings have been expanded to include performance auditing with a gradual transfer of audit resources from traditional concerns of legality, proper management and financial management to issues of economy, efficiency and effectiveness.

The success of the current boom in the use of evaluative information will remain largely dependent on its credibility. Program evaluations, performance reports and performance audits all claim to provide objective representations of the reality of program outputs and outcomes, economy, efficiency and effectiveness. Perceptions that evaluative information misrepresents reality (intentionally or not) are likely to render it useless—other than as a tactical weapon in political and bureaucratic skirmishes. There is some evidence suggesting the risk of a credibility crisis regarding much evaluative information.

One threat to the credibility of evaluative information stems from political and organizational pressures. Observers of program evaluation practice have long warned that political and commercial pressures on evaluation clients and on evaluators lead to a priori bias in evaluation reports (Chelimsky, 1987; Palumbo, 1987; Schwartz, 1998; Weiss, 1973; Wildavsky, 1972). Administrators' interests in organizational stability, budget maximization and the promotion of a favorable image, contribute to a general desire to prefer evaluations and performance reports that do not cast programs in a bad light. Smith (1995) describes similar reasons for unintended consequences of performance reporting.

A second threat to the credibility of much evaluative information comes from the apparent pervasiveness of shoddy practice. Unlike professions such as accounting, law, medicine and architecture, neither performance measurement nor evaluation has accreditation, certification or licensing systems. Anybody can call themselves an evaluator and bid for evaluation contracts. Purchasers of evaluations and performance reports often lack the expertise to distinguish professional evaluators and competent performance measurers from well-intentioned amateurs or charlatans. They tend to lack the skills to determine whether evaluation and performance measurement products constitute solid work or worthless words and data.

Where evaluation findings and performance information are used in decision-making this can have grave consequences. Muir (1999) provides evidence to this effect on the use of evaluation findings for education reform policy-making. One hundred and sixteen evaluation studies which constituted the evaluative support base for 24 common school reform programs were assessed on the basis of: scope, objectivity of measurement instruments, construct validity, internal validity, sample bias, use of appropriate statistical technique, and external validity. 'Out of the two dozen programs examined, only three had both adequate research base and strong evidence of success.' The experience of a former editor of a prominent evaluation journal lends further support to concerns with evaluation quality: "My own experience leads me to believe that such protection is needed. Many of the manuscripts submitted to me during the seven years I was editor of Evaluation Practice caused me to believe that clients often pay for evaluations that could lead to unsubstantiated conclusions and to unwarranted changes in programs." (Smith 1999).

Studies of the use of performance measurement systems raise similar credibility concerns. Bouckaert (1993) ascertains a growing need to monitor the quality of performance measurement systems as they are applied to more 'intangible services' (education, medical treatment, care for children) and 'services that involve the processing of ideas' (think tanks, policy staffs, people who prepare legal work). In a classic article, Smith (1995) outlines eight 'unintended consequences of performance measurement', many of which express problems with the reliability of data and validity of measures. Empirical assessments of the quality of performance measurement systems lend support to these claims. For example, a recent study of the use of performance measurement systems in 695 American municipalities demonstrates considerable weaknesses of validity, legitimacy and functionality concerns (Streib & Poister, 1999).

The third type of evaluative information—performance audit reports—presents a somewhat different picture on quality. Audit reports are widely considered by legislatures, the public and many others as the epitome of credibility. State auditors pride themselves on their independence, objectivity, neutrality and professionalism. Yet performance audits are not without their criticism. Some scholars have begun to question the credibility of evaluative information found in performance audit reports that deign to measure program effectiveness.

An empirical study of the quality of effectiveness evaluation audits in six countries demonstrates that some state auditors have not met the challenge of providing non-politicized, professional and objective reports (Schwartz, 1999). Six out of 13 audit reports that examined outcome effectiveness were found deficient in dealing with causality, failing to utilize standard social science techniques for measuring change and for attributing change to program interventions rather than to intervening variables.

Despite such misgivings, performance audits attract much less critical comment about their quality than either evaluations or performance reports. And as we shall see, performance audit practice often has in place an impressive array of quality assurance mechanisms.

2. Recognizing the need

Evaluators and auditors have long recognized the need for quality assurance. This is not surprising coming from

analytical professions whose work is largely about assuring the quality of program provision. The development of standards for evaluation, performance reporting and performance auditing reflect first steps in quality assurance practices for evaluative information. Some evaluation thinkers went further in devising and promoting both formative and summative models of meta-evaluation or evaluation audit (Chelimsky, 1983; Schwandt & Halpern, 1988; Stufflebeam, 1974, 2001a). Indeed Standard A12 of the JC Program Evaluation Standards stipulates that, "the evaluation itself should be formatively and summatively evaluated…" (Joint Committee on Standards for Educational Evaluation, 1994, 185) The title of a recent article, 'The Metaevaluation Imperative,' in the American Journal of Evaluation reflects the continued, or perhaps increased, importance of this activity (Stufflebeam, 2001b). And the Evaluation Center at Western Michigan University has developed checklists and guidelines for use in metaevaluation work (Stufflebeam, 2000) and posted them on its website (www.wmich.edu/evalctr/). Similar models have been developed in the 1990s by practitioners of performance measurement and by public audit institutions. Several SAIs have also developed formative and summative models of assuring the quality of their own performance audit work.

A few publications report on practical experiences with metaevaluation. For example, in 1992, *Evaluation and Program Planning* published a symposium on the description and analysis of a case study of a formative metaevaluation (or evaluation audit) (of the East Central Education and Training Center). More recently, in 2002, a special issue of the same journal devoted to evaluation in Africa included a metaevaluation, based on the African Evaluation Guidelines, of 14 evaluations (Patel, 2002). Stufflebeam (2001b) notes that while evaluators are making progress in conducting metaevaluations, "sustaining and increasing efforts to systematize and increase the rigor, relevance, and contributions of metaevaluations are in the interest of professionalizing the field".

While considerable effort has been expended on establishing quality assurance models—at least for program evaluation and performance auditing—little has been written about the actual practice of assuring the quality of evaluation, performance reporting and performance auditing. Certainly, there is no comparative literature on this. There has been little investigation of the conditions that promote the application of quality assurance, about the extent to which it occurs, how it is done, its accomplishments, and the obstacles it faces. Power's (1997) investigation of the consequences of auditing in the UK is an exception. He illustrates numerous examples where various types of audits have become 'rituals of verification' that provide illusions of quality assurance, but do little to improve practice.

Power demonstrates that quality assurance may produce deleterious side effects that he labels *decoupling and colonization*. In Power's terms, decoupling refers to a state in which audit—or more generally, quality assurance activities—activities are kept separate from the real 'organizational activities', making the assurance activity irrelevant at best. When decoupling occurs, there is ritualistic compliance with quality assurance processes, but little actual impact on organizational activity. Colonization refers to the 'ingraining' of audit—quality assurance—values and practices into the 'core of organizational operations'. Dysfunctions of colonization include adherence to orthodox practice, inhibiting of innovation and tunnel vision in which the organization strives to excel in measurable activities included in the audit purview to the detriment of less measurable (and less audited) pursuits. This latter problem has received some attention in the American literature as well. Schwandt (1992, 99) followed House (1987) in opposing formalized quality control, as in the establishment of 'some kind of evaluation institute that would monitor quality', for fear that this would impose 'an orthodox point of view about evaluation work'.

The present study aims to cast some light on the origins, practice and consequences of quality assurance practices for evaluative information in light of the political and organizational environments within which it occurs. While methodological limitations prevent conclusive answers to questions of influence and causality, the findings do provide preliminary evidence for further testing of hypotheses associated with questions such as: To what extent do political and organizational considerations prevent the conduct of quality assurance, impede its success and result in decoupling and colonization? What conditions enhance the possibilities for successful quality assurance?

3. Methodology for this study

Data for this article were collected as part of a study conducted predominantly by members of the INTEVAL group.[2] In soliciting contributions to the study, we sought

[2] The International Working Group on Policy and Program Evaluation addresses issues of evaluation theory and practice in a cross national perspective. Research results usually take the form of co-edited books, most of them being part of the Comparative Policy Analysis Series (Transaction Publishers). We draw on the following chapters prepared for an upcoming book. Barbier, Jean-Claude. Devising and Using Evaluation Standards: The French Paradox; Boyle, Richard. Assessment of Performance Reports: A Comparative Perspective; Ginsburg, A. and Pane, N. Decentralization Does Not Mean Poor Data Quality: A Case Study from the US Department of Education; Grasso, Patrick. Quality Of Evaluative Information At The World Bank; Kraan, A. and van Adrichem, H. The Netherlands Court of Audit and Meta-research: Principles and Practice; Lonsdale, J. and Mayne, J. Neat And Tidy…And 100% Correct: Assuring The Quality Of Supreme Audit Institution Performance Audit Work; Mayne, J. and Wilkins, P. Believe it or not? The Emergence of Performance Information Auditing; Toulemonde, J., Summa-Pollitt, H. and Usher, N. Triple Check for Top Quality or Triple Burden? Assessing EU Evaluations; Widmer, T. Instruments and Procedures for Assuring Evaluation Quality: A Swiss Perspective.

Table 1
Approaches to assuring quality of evaluative information

Approaches used	Types of evaluative information		
	Evaluations	Performance reports	Performance audits
Structural approaches: setting guidelines and standards	Professional evaluation society standards	Professional practice guidance (e.g. text books, academic writing)	Professional audit/accounting society standards
	Professional practice guidance (e.g. text books, academic writing)	Organizational/governmental guidance and standards	SAI manuals
	Organizational/governmental guidance and standards		Training and capacity development
	Training and capacity development		Certification
Formative approaches: real-time assessments of individual reports	Advisory committees	Advisory committees	Advisory committees
	Internal quality control procedures	Internal quality control procedures	Internal quality control procedures
Summative approaches: ex-post assessments of individual reports	Independent assessments, such as SAI audits	SAI assurance	External review of audit reports
	Semi-independent assessments	Self-assessments	
	Self-assessments	Internal audit	
Systemic approaches :assessments of systems and procedures for producing evaluative information	SAI audits	SAI audits	Independent review of SAI audit practices
		Internal audit	

examples for each type of evaluative information and each different type of approach to quality assurance. Table 1 sets out this matrix. We sought (and received) information about both positive and negative quality assurance experiences. Nevertheless, the experiences reported here are self-selected to an extent. We do not pretend to represent the worldwide state of quality assurance practices for evaluative information.

For each quality assurance experience we collected information about the following questions:

1. What were the reasons for establishing assurance systems? Who initiated them? Who supported their development? Who opposed? Why is more attention paid to providing assurance of evaluative information in some cases and not in others?
2. How is the quality of evaluative information enhanced or assessed? Who finances and administers the establishment of standards and procedures of assessment? Who performs the assessments?
3. Are assessors members of the evaluation, performance measurement and audit communities? How are problems of friendship, connections, rivalry and competition managed?
4. What standards are in use? If evaluative information is distinct from scientific research, to what extent can standards of reliability, internal validity, external validity and causality be applied? Are additional standards in use—such as relevance, timeliness and use?
5. How do assurance systems manage disagreement amongst judges, a common occurrence in academic peer review? What is the trend for the future?
6. To what extent do a range of stakeholders take part in assessment?
7. What can be said about the efficacy of assurance approaches in contributing to better practice of evaluation, performance measurement and effectiveness auditing—perhaps in the eyes of practitioners and users? Do these assurance approaches contribute to filling the gap between practice and standards?
8. What kind of evaluation would be needed to assess whether the various approaches to enhancing quality are working? To what extent do assurance systems serve primarily to create an image of credibility? Do the benefits of image management justify the costs?

For the purposes of this article, responses to these questions were grouped by categories according to type of evaluative information, assurance approach, dysfunctions, impacts and political and organizational impediments The article covers quality assurance experiences across three types of evaluative information: evaluations, performance reports and performance audits (Table 1). We analyze evaluation quality assurance in two international organizations—the European Union and the World Bank—and in four countries: Canada, France, the Netherlands and Switzerland. Data about performance reporting quality assurance practices come from Australia, Canada, New Zealand, the United Kingdom and the United States. And quality assurance of performance audits is based on Supreme Audit Institution (SAI) experiences in Australia, Canada, the Netherlands, New Zealand, Sweden, and the United Kingdom. Table 2 outlines the external and internal quality assurance agents that were studied by evaluative information type and jurisdiction.

This is not a representative sample of jurisdictions. The countries and international organizations included are either widely considered to be leaders in an evaluative information field or have quality assurance experiences that are useful for learning. In the evaluation area: Canada and the Netherlands are two of a handful of countries that have made program evaluation mandatory and the European Union and World Bank are considered international

Table 2
External and internal quality assurance agents by jurisdiction

Jurisdiction	External	Internal
Evaluation		
European union	DG Audit (1997–9) European Court of Audit	DG agriculture
World Bank		Quality Assurance Group Operations Evaluation Department
Canada	Office of the Auditor General	
France	Conseil scientifique de l'evaluation Conseil national de l'évaluation	
Netherlands	Netherlands Court of Audit	
Switzerland	Swiss evaluation society Academics	Health Office
Performance reporting		
Australia	Australian National Audit Office Institute of Public Administration Australia Auditor General of Western Australia	
Canada	Office of the Auditor General Auditor General of Alberta	
United Kingdom	Audit Commission National Audit Office	
United States	US General Accounting Office	Planning and Evaluation Service, Department of Education
	Mercatus Center	Internal auditor of Maricopa County, AZ International City/ County Management Association
Performance auditing		
Australia	New Zealand's Office of the Controller and Auditor General Victoria Audit Office ANAO's external auditor	Australian National Audit Office
Canada		Office of the Auditor General
Netherlands		Netherlands Court of Audit
New Zealand	Australian National Audit Office	Office of the Controller and Auditor General
Sweden		Riksrevisonsverket
United Kingdom	London School of Economics; Ad hoc committees	National Audit Office

evaluation leaders (Furburo, Rist, & Sandahl, 2002); Switzerland has developed highly advanced evaluation standards and France has implemented, largely unsuccessfully, a national system for assuring evaluation quality. The performance reporting area is covered with experiences from leading New Public Management countries, expected to be most developed in quality assurance. And the SAIs are amongst those with the longest experience in conducting performance auditing. The data likely represents leading edge practice.

Quality assurance practices were classified as belonging to one of four approaches: structural, formative, summative, or systemic (see Table 1). *Structural* approaches refer to efforts to develop an infrastructure that makes quality work possible, including the development and promulgation of standards, training of individuals and capacity development of organizations. *Formative* approaches attempt to assure quality of specific evaluations, performance reports or performance audits during the course of conducting the work. *Summative* assessments of individual evaluative information reports. The *systemic* approach to quality assesses the extent to which systems for producing credible evaluative information function successfully.

We start with a look at the extent and characteristics of quality assurance practices for evaluative information. The following section examines conditions that have led to the advent of quality assurance measures. We then explore the consequences of quality assurance, including the emergence of *decoupling* and *colonization* behaviors. We continue with a discussion of impediments and politics of quality assurance and conclude by drawing out some 'best practices' from the experiences reported.

4. The extent and nature of quality assurance practice

Table 1 lists the types of quality assurance approaches we found for each of the three types of evaluative information examined. A glance at Table 1 may leave one with the impression that quality assurance of evaluative information is pervasive. This is not so. There are significant differences amongst assurance approaches and amongst types of evaluative information. What is pervasive, across all types of evaluative information in our sample, are structural approaches—particularly standards, albeit with varying levels of authority. Formative approaches are highly prevalent in performance auditing, but appear to be used routinely in few jurisdictions for assuring quality in either program evaluation or performance reporting. And while routine summative assessments of performance reports are common, few departments and agencies apply summative assessments systematically for program evaluation work. Summative assessments of performance audit are conducted routinely by only one SAI. Finally, SAIs have conducted systemic assessments of program evaluation and of performance reporting, but have been much less frequently subject to external review of their own performance audit work. Exceptions are a report on the United States GAO by the National Academy of Public Administration (1994) and a legislative review of Australia's National Audit Office (Joint Committee of Public Accounts, 1989). What emerges then is more a potpourri of quality assurance practice which when summed up still portrays a lot of quality assurance activity.

Below we discuss some of the specific examples that make up the entries in Table 1.

4.1. Structural approaches

Standards are the basic building block upon which all other quality assurance approaches rest. They exist in some form across jurisdictions and across types of evaluative information, yet with considerable variance in sources of authority, objectives, specificity, content and weighting. Standards are used also in training and capacity development to develop structure that enables high quality work.

Sources of authority. The source of authority for standards varies amongst types of evaluative information. For program evaluation, the US Joint Committee standards (Joint Committee, 1994) have served as a basis for the development of professional standards in several jurisdictions (Africa, European Union (EU), Germany, and Switzerland), giving them some international acceptance. In other countries, the evaluation professional association has worked towards developing their own standards. For example, the Canadian Evaluation Society is exploring certification of its member (Stierhoff, 1999).

There are no international professional performance reporting organization societies. Standards for performance reporting have been developed initially by central government organizations and by supreme audit institutions. Now, however, in countries such as the USA and Canada, performance reporting standards are being developed by standard or quasi-standard setting bodies (Canadian Comprehensive Auditing Foundation 2002; Canadian Institute of Chartered Accountants 2002; Government Accounting Standards Board 2003). Standards for performance auditing have been developed both at the international level by INTOSAI and at the national level by accounting standard-setting bodies and by SAIs (Lonsdale & Mayne, 2004).

Objectives and specificity. Standards differ in objectives. The European Union's MEANs standards (European Commission, 1999a) for evaluations, for example, are minimalist in that all evaluations are expected to achieve at least minimal compliance with each criterion. The Swiss Evaluation Society's SEVAL (Widmer, Landert, & Bachmann, 2000) and French Conseil scientifique de l'évaluation (CSE) (Conseil scientifique de l'évaluation 1996) set out maximal quality standards to which all evaluations should strive.

In the jurisdictions studied, the level of detail of program evaluation standards ranges from skeletal (World Bank) to intricate (European Union and Switzerland). While the World Bank makes do with five concisely stated general good practice guidelines, Europe's MEANs Grid includes nine detailed criteria and the Swiss SEVAL specifies 27 individual standards across four broad categories. The Office of the Auditor General of Canada uses a comprehensive set of 19 criteria from five broad categories to assess the quality of performance reports. And Lonsdale and Mayne's (2004) synthesis of quality standards used by various SAIs includes six process criteria and eight product criteria.

Content. We identify three types of standards: product quality—technical quality of the information produced; process quality—quality of the process used to obtain the information; and usefulness—the usefulness of the information produced.

Criteria for *product quality* are similar across the three types of evaluative information. Although different terms are used in different jurisdictions there is a high degree of uniformity in standards of good quality evaluative information, namely:

- *Well-defined scope*. The objectives of the information, the purposes to be served and the range of coverage should be clearly set out.
- *Accurate data*. The data collected should be valid and reliable.
- *Sound analysis*. The analysis of the data collected should be based on robust methodology.
- *Substantiated and impartial/objective findings/conclusions*. The findings and conclusions presented should be supported by the evidence gathered (data and analysis) and should be presented in an impartial (objective) manner.

Similarity here is perhaps not surprising. All three types of evaluative information have their roots in the social sciences and hence what quality information entails ought to be similar. Indeed, it would be surprising if there were significant differences here. On the other hand, there are rather significant differences among the three when we examine *process quality* and *usefulness*.

Evaluation *process quality* criteria, focus more on stakeholder involvement and consideration than do criteria for performance audit. Evaluation quality standards quite explicitly address the concerns and interest of those being evaluated and those affected by the evaluation. These concerns are considerably less evident in performance audit standards. This distinction probably reflects the different roles usually played by evaluation and audit. One gets audited whether you like or not and legislative auditors are often required to report weaknesses. So while auditors would like to get buy in for and acceptance of their findings and recommendations, auditees are usually legally required to cooperate. Evaluation often—although not always—undertaken from quite a different perspective (such as a research perspective), might need to convince those evaluated to assist in the evaluation process.

Criteria for performance audits stress the importance of the independence of the auditors. Both evaluation and audit call for objectivity, but only performance audit makes independence a key element of the audit process.

Three essential *Usefulness* criteria are common to all three types of information:

- *Timeliness*. The information is produced at a time when it can make a difference in improving the performance of the program reported on,
- *The 'right' scope*. The information produced is relevant to the issues of the day, and
- *Clarity*. The information is understandable by the intended audience.

Other Usefulness criteria vary across evaluation types. Quality criteria for performance reports stress communication and accessibility of the information more than do those for evaluations and performance audits. Performance reports serve their purpose to the extent that legislators and the public read them. Evaluations and audits have a predetermined audience (those who commissioned the evaluation or legislative committees). While both are written with an eye to the more general public, they are useful if these predetermined 'clients' use them.

Quality criteria for evaluation stress the need to explain methodology more than do criteria for performance audits. This is perhaps a question of degree, but while performance audits are expected to explain how they arrived at their findings and conclusions, they usually do not provide the kind of detail found in evaluations. The research roots of evaluation stress the importance of careful explaining of the methodology used so that others can repeat the approach.

Weighting. There is a longstanding debate in evaluation literature about the extent to which quality standards should emphasize technical matters as opposed to use (Greene, 1990; Patton, 2001). At the heart of this debate is the contention that preoccupation with technical issues such as reliability and validity may reduce the chances that findings will be utilized because of time, expense and complexity of reporting. On the other hand, focus on use risks overlooking vital aspects of technical quality leaving the work open to credibility challenges.

Almost all of the standards reviewed, across types of evaluative information, included both technical (product quality) and use standards, with no predetermined suggestions of their relative weight. Directions for the application of standards generally recognize the need to consider contextual factors in weighting criteria. The SEVAL standards explicitly note that the relative importance of criteria varies amongst evaluation projects. SEVAL recommends a functional approach in which only some of its 27 standards apply to any given evaluation. While all nine MEANs standards are designed to be applicable to all evaluations their relative weight may vary depending on contextual factors of specific evaluations. Lonsdale and Mayne (2004) suggest that the technical—use tension may be particularly sharp for performance auditing, where, "highly technical reports might satisfy the professional pride of auditors and appeal to a specialist audience, but may be of limited interest to parliamentarians."

Training and capacity development. Several jurisdictions promote the quality of evaluative information by using standards as educational tools. When standards are widely accepted and promulgated, they can serve to improve the understanding of commissioners, producers of information and other stakeholders as to the desired characteristics of evaluative information. The Swiss SEVAL standards stand out in their educational function, having become central components of various training programs. The cross-national and cross-field span of European Union evaluation requirements means that the MEANS grid is accepted as representing evaluation canon by an increasing circle of commissioners, evaluators and users. The educational function of standards thus presents considerable opportunity for evaluation standards to have an enlightenment effect. Similarly, audit standards are the basis for training in performance auditing.

4.2. Formative approaches

We found that formative assessments are applied routinely within SAIs, and for performance reporting in a few jurisdictions. There is no requirement that all evaluations be subject to formative assessment in any country or organization we examined.

Prevalent amongst the formative approaches to quality assurance, in our sample, is the assessment of evaluation proposals and interim reports. The French CSE (Conseil scientifique de l'évaluation 1996) in-depth reviews of evaluation proposals, for example, commonly revealed significant flaws in evaluation approaches and methods. Toulemonde, Usher, and Summa-Pollitt (2004) describe a step-by-step formative quality assurance process implemented in the European Union DG Agriculture in which evaluators are briefed about the MEANs grid criteria and evaluation steering groups subject successive interim reports to quality assessments. And the Swiss Federal Office of Public Health Office generally submits both evaluation proposals and interim evaluation reports for review by independent experts in a preventive mode of quality assurance (Widmer, 2004). Toulemonde et al. report that, in DG Agriculture of the European Union, formative assessments of proposals have been a key tool for improving the quality of evaluation work. French CSE proposal reviews, by contrast, were generally ignored.

Formative quality assurance is highly developed in several SAIs (Lonsdale & Mayne, 2004). It is common practice to subject draft performance audit reports to various forms of peer review, clearance with audited agencies and stakeholder response. A number of SAIs also make extensive use of external experts in reviewing drafts of audit criteria, plans and reports. Often this is done in the framework of advisory committees which also include internal experts. The National Audit Office (UK) sometimes contracts with external experts for external 'hot reviews' of reports as they are being prepared.

Formative assessment of performance reporting is less developed. Yet a number of innovative practices were

identified, including benchmarking networks, working groups and technical support groups in Norway, the United States and the United Kingdom. For example, in the United States, the International City/Council Management Association established the Performance Measurement Association in which local government staff, independent researchers and a consulting firm work together in developing performance reports that enable comparisons of performance (Coe, 1999). A byproduct of this benchmarking exercise is the improvement of the measurement systems.

4.3. Summative approaches

In our sample, summative approaches are routine for performance reporting in some jurisdictions and a frequent practice in several SAIs for performance reporting by their governments. While several instances of summative metaevaluation were found, it is rarely a routine practice for evaluations.

The range of practice of summative metaevaluation of evaluations is wide, spanning mandatory external reviews in the now defunct French CSE, steering committee reviews in the European Union, and sporadic reviews conducted by academics (Switzerland) and by state audit institutions (the Netherlands and Canada). While there is occasional use of summative reviews to prevent poor quality evaluations from being used, their main function is to provide lessons to be learned for future evaluation work.

The World Bank routinely subjects internal evaluations conducted by operational staff to independent summative assessments by its Operations Evaluation Department (OED) staff (Grasso, 2004). OED follows up these desk reviews with full independent field assessments of one-fourth of completed projects. One of the most common reasons for conducting a field review is that there are reasons to believe that the information available for the internal evaluation and for the OED desk review was insufficient, invalid or unreliable. This is the only case that we know of where quality assurance is achieved by a sort of 'replication' of the original evaluation work.

Evaluation synthesis can be used as a summative approach to assessing the coverage and quality of evaluation work regarding a whole policy area. Evaluation synthesis, as practiced by the Netherlands Court of Audit, assesses the quality of individual evaluation reports and examines the extent to which existing evaluations provide sufficient evidence, in terms of coverage and quality of studies, to support policymaking. The conduct of evaluation synthesis thus assesses not only the work of evaluators, but also the success of government officials and commissioners of evaluations.

Some jurisdictions require that their SAI annually conduct summative reviews of annual performance reports produced by government bodies (Sweden, New Zealand, Western Australia and Canada for three agencies).

Summative assessments of performance audit reports are less common in the jurisdictions studied. Some SAIs (Canada, the Netherlands and United Kingdom) conduct internal 'post-mortem' reviews. Others (Australia and New Zealand) have established a reciprocal quality assurance programme in which each SAI reviews samples of performance audits from other SAIs. And one SAI (United Kingdom) subjects every performance audit report to external 'cold review' by a team of academic specialists.

4.4. Systemic approaches

SAIs dominate systems approaches to assuring the quality of both program evaluation and performance reporting. There has been little systems oriented effort for assuring the quality of SAI performance audit work.

Several SAIs and audit units have conducted one or more large scale audits of the systems that public organizations have in place for conducting evaluations (Auditor General of Canada, 1983, 1986, 1993, 1996; Australian National Audit Office, 1991, 1992a,b, 1993, 1997; European Commission, 1999a,b, 2000; New Zealand Controller and Auditor General, 2000; Algemene, 1990–91). These system audits reveal weaknesses in evaluation planning, commissioning and attention to evaluation content and method. For example three reviews conducted by European Union's Directorate General Audit (Toulemonde et al., 2004), revealed that most departments lacked evaluation standards, did not systematically apply evaluation findings in decision-making, and did a poor job of monitoring the quality of evaluation processes. The World Bank's annual review of the quality of its internal evaluations is the only non-audit based systemic evaluation quality assurance reported (World Bank, 2002).

SAI system audits of performance reporting have been conducted in the UK (National Audit Office, 2001), Canada (Auditor General of Canada, 1997, 2000), United States (General Accounting Office, 1999) and New Zealand (New Zealand Controller and Auditor General, 2001). These audits provide an overview of the current state and progress of performance reporting. They identify common problems and highlight good practice for purposes of accountability and to provide lessons learned for guiding future quality improvement. The UK Audit Commission has, in the past, had an unusual mandate to specify the performance indicators that are reported as well as conduct reviews to check that the systems used by local government authorities are able to produce accurate information.

Summing up this section, the jurisdictions covered in this study use a large variety of quality assurance approaches. Patterns of quality assurance vary across types of evaluative information and jurisdictions. In many jurisdictions, SAIs play a central role. The following sections assess the variance in quality assurance practice by looking at insitutional and environmental drivers and impediments.

5. Institutional contexts: why initiated, by whom

In the jurisdictions studied, there appear to be three broad sources for initiatives to develop systems for assuring the quality of evaluative information: legislative/central government demand; management self-initiative; and pressure from audit organizations.

5.1. Legislative and central government demand

Governance changes that place evaluative information at the center of much public sector reform have spurred legislatures, central governments and/or SAIs in some countries to stipulate quality assurance measures. This is particularly evident in jurisdictions that have adopted new pubic management type reforms in which agencies are granted considerable autonomy, but in turn are required to report on their performance. Information quality is of particular concern when evaluative information is expected to play a role in budget allocation decisions. European Union interest in quality assurance for evaluation stems largely from its turn of the century governance changes which, amongst other things, placed evaluation at the center of accountability improvement. Similarly, the French CSE evaluation quality assurance system was established in the context of management reform in which evaluation was one of four central principals. Management reform in Canada, Australia, New Zealand, Sweden and the United States spurred central governments and legislatures to set out quality assurance standards and to request that SAIs assure the quality of annual performance reports. When evaluation became an important performance improvement tool in the World Bank, it too stipulated quality assurance measures in a top-down fashion.

5.2. Management self-initiative

Management evaluation occurs when program managers have a sincere interest in learning about the performance of their programs and policies. Managers order such evaluations for use by managers. They tend to ask questions concerning ongoing implementation and efficient management rather than questions about the overall effectiveness of a program (Auditor General of Canada, 1983; Schwartz, 1998, 306–7). The cases we examined describe a number of instances of quality assurance initiated by agency level management with a genuine interest in making sure that evaluation work conducted for them is credible and useful. Examples include the extensive use of metaevaluation by the Swiss Federal Office of Public Health and by European Union's DG Agriculture. In the absence of 'top-down' requirements, these agencies developed evaluation quality assurance systems from the 'bottom-up'.

Lonsdale and Mayne (2004) note that SAI performance audit quality assurance has also been self-initiated. They attribute auditors' interest in quality assurance, "partly partly (due) to self-interest, partly professional pride and the self-critical nature of those involved, and partly a result of pressure from outside." These pressures stem from the higher profile of performance auditing and the desire of performance auditors to comply with quality requirements accepted in the academic research community.

5.3. SAI initiative

In many jurisdictions, SAIs play a prominent role in assuring the quality of both evaluation and performance reporting. SAI quality assurance work usually complies with legislative stipulations for the SAI to examine the quality of the information going to its legislature. Quality assurance of evaluation and performance reporting is viewed as part of the 'value-for-money' mandate many SAIs now have; it is seen as as a natural extension of their financial and performance auditing roles. Indeed the classical function of auditors is to assess systems of control.

6. Positive impacts of quality assurance practices

Here we meet what might be called the paradox of evaluating the evaluation of evaluative information. While there is an expectation that systems for assuring the quality of evaluative information include a look at usability and impacts, it is a challenge to measure the usability and impact of quality assurance systems themselves. Power (1997, 28) notes this issue regarding auditing: "It is in this sense that auditing has a 'weak' knowledge base; there is no way of specifying the assurance production function independently of a practitioner's own qualitative opinion process…there is something unspecifiable about its output." Moreover an evidence-based approach to results measurement would require a research design that could not only describe changes in the quality of evaluative information, but also attribute these changes to the application of quality assurance mechanisms while controlling for the influence of other variables.

While such an evidence-based approach is beyond the scope of the present undertaking, we have assembled some empirical and some more anecdotal indications about the effects of quality assurance practices for evaluative information. These reports should be interpreted cautiously. They should be seen as leading to hypotheses for further research rather than as definitive conclusions about the efficacy of quality assurance.

The data gathered provide illustrations from all three types of evaluative information where quality assurance does appear to contribute to improvement. For example, widespread indoctrination of SEVAL standards through education and training has made good evaluation practice

part of the common knowledge of Swiss evaluation commissioners and evaluators thereby making a significant contribution to overall evaluation quality (Widmer, 2004). European Union DG Agriculture experience shows that assessments of draft interim and final reports enable considerable improvements in the final evaluation product (Toulemonde et al., 2004).

A number of reports indicate that quality assurance for performance reporting has had some positive impacts: The Institute of Public Administration Australia (2001) attributes improvements in performance reporting to the existence of common standards. The UK National Audit Office (2001) notes that significant increases in outcome targets stems from the use of assurance-based critiques. And Law (2001) attributes improvements in UK chief constable of police reports to national initiatives aimed at improving the quality of information.

There is a general feeling amongst audit offices that attest auditing of performance reports has made significant contributions to improving quality. One SAI—the Swedish National Audit Office—provides empirical evidence of the correction of deficiencies.

'Harder' evidence of positive impacts is available for performance auditing. The UK NAO has submitted some 400 performance audits to independent external scrutiny over the past ten years. The grades obtained by reports improved steadily over the years. And Lonsdale (2000) found, in focus groups with NAO auditors, that the external review was 'a useful stimulus to thinking about new methods, and comments and criticisms acted as a direct spur to improve'.

7. Dysfunctions

Engaging in quality assurance activities is not without risks, despite whatever benefits might accrue. We identify several types of dysfunctions.

7.1. Wasted resources

The quality assurance activities may not work, with the result that considerable resources may have been used with no noticeable improvement in the quality of the evaluative information assessed. Even if some improvement in quality can be identified, the often substantial resources used in a quality assurance effort, may be difficult to justify in terms of the improvements realized.

7.2. Decoupling

Quality assurance efforts are sometimes ignored, or only ritualistically adhered to, because they lacked credibility, were seen as demanding too high a standard, or were not perceived as significant to core organizational operation. Segsworth and Volpe (2004) partially attribute the limited impact of the Auditor General of Canada's systemic audits of the evaluation function in the 1980s and 1990s to decoupling. There has been an extreme case of ineffective quality assurance and decoupling in France. CSE imposed an orthodox, foreign, academically-oriented perspective that was not acceptable among other evaluation milieus nor amongst evaluation stakeholders. The result was that while CSE engaged in formal metaevaluations, evaluation practice continued with little attention paid to its findings and recommendations. Grasso (2004) observes decoupling in the World Bank where operational staff has weak incentives to be concerned about the quality of monitoring and evaluation activities.

7.3. Colonization

Too much attention to adhering to quality assurance dictates can distort evaluative information. Power (1997) refers to this as colonization, whereby the 'ingraining' of audit values and practices into the 'core' of organizational operations' occurs. Boyle's (2004) comparative review of quality assurance suggests that performance reporting can be prone to colonization when reports are rated in 'name, shame and blame' style. Smith (1995) notes that this type of reporting generates such risks as tunnel vision, convergence, gaming and misrepresentation, all of which may have an adverse effect on the quality of data."

8. Impediments: the politics of quality assurance

The production of evaluative information often occurs in politicized contexts where stakeholders act to limit risks of unflattering reports. What is the likelihood that under these circumstances, political considerations will affect the establishment and operation of quality assurance systems? The extent to which a jurisdiction has quality assurance mechanisms in place is one possible indicator of the significance it attaches to evaluative information. While many jurisdictions now make annual performance reporting and periodic program evaluation mandatory (Derlien & Rist, 2002), the data collected for this study indicate that routine active quality assurance tends to be a sporadic and spotty undertaking. Political considerations may also affect the operation of established quality assurance systems. We have examples from both evaluation and performance reporting experience.

Political considerations appear particularly strong when the initiative for quality assurance originates in top-down, especially external, initiatives to institutionalize quality assurance of evaluative information reports. In its extreme form, institutionalization might entail the establishment of an external 'meta-evaluation' institute that would routinely monitor the quality of evaluative information produced for government use. In favor of institutionalization is

the potential to provide independent and comprehensive quality assurance. Problems associated with institutionalization include: exacerbating problems of decoupling and colonization; financial and time costs; and increased bureaucratization of the process.

External institutionalization of quality assurance occurs in the work of SAIs, often seen as an extension of their traditional role in providing external quality assurance on financial statements. In their systemic assessments of evaluative information produced by government organizations, SAIs provide quality assurance. More directly, in SAI assurance audits of annual performance reports there is real time assessment of the evaluative information produced. Boyle's (2004) indicates that SAI activity in promoting and auditing performance reporting systems presents a risk of decoupling and colonization. The Auditor General of Canada (2000, 19–27) is surprisingly upfront about political impediments to the success of its own efforts to improve the quality of performance reporting, noting that 'performance reporting has political dimensions; and there are few incentives for good reporting or sanctions for poor reporting'.

In the evaluation domain, the French experience is a prominent example where an attempt to impose metaevaluation from the top down encountered organizational and political pressures. The result was extreme decoupling which rendered CSE work largely irrelevant. This example shows the potential dangers of institutionalizing external quality assurance practices, perhaps illustrating some of the concerns of House (1987) and Schwandt (1992).

The idea of an external independent meta-evaluation institute to monitor the quality of performance audits is anathematic to state auditors who guard their own independence with a vengeance. Such an institution risks tarnishing the image of SAIs as guarantors of the quality of financial and evaluative information produced by audited agencies. It is not surprising therefore that no such institution exists. The closest thing to it are the external reviews carried out on behalf of a few SAIs, and the external 'hot' reviews of UK NAO performance audit reports carried out by the London School of Economics. Until now these 'hot' reviews have been a strictly self-initiated and internally reported exercise. It will be interesting to see the consequences of a recent decision to publish results of these reviews.

Self-initiated management quality assurance appears to hold promise for avoiding internal political and organizational impediments. Yet it too may be subject to external political use or perhaps abuse. Toulemonde et al. (2004), for example, illustrate attempts by external stakeholders to undermine the credibility of DG Agriculture metaevaluations.

There have been cases where the quality of a report has been assessed 'poor' while the report concluded negatively on the policy under evaluation. In such cases, DG Agriculture has been exposed to criticism like: 'they rated the quality of the report poorly because they did not like the conclusions'

.

Overall, most organizations that we studied, other than SAIs, seem not to worry about the quality of their evaluative information. In the European Union for example, DG Agriculture's quality assurance system is the exception rather than the rule, other DGs having opted not to invest much in this exercise. Similarly, routine application of formative and summative self-assurance is reported to exist in only a handful of individual departments or agencies in the various jurisdictions covered in this study. SAIs are a notable exception, perhaps because producing reports is their raison d'etre. For most other government bodies, the production of evaluative information is a byproduct, often conducted in order to comply with stipulations from above and not intrinsic to operational or management needs. Moreover, many government bodies believe that the evaluative information they produce receives little attention by policymakers. The credibility of evaluative information is not likely then perceived as being critical to the success of most government bodies.

Our preliminary evidence of impacts, dysfunctions and impediments leads to the following hypotheses for testing in further research:

(a) The promulgation and inculcation of clear standards will have independent positive impacts on the quality of evaluative information;
(b) A greater incidence of formative and summative assessments will be accompanied by the following factors: production of evaluative information is a core organizational mission; high management commitment to the evaluative information function; perceived high use of the evaluative information by policy makers.
(c) A higher impact of quality assessments will be accompanied by the following factors: high management commitment to the evaluation function; strong incentives to staff for investing in quality;
(d) Routine summative attest auditing of performance reports will lead to improvements in the quality of subsequent performance reports.
(e) Routine external summative assessments of performance audit reports will lead to improvements in the quality of subsequent performance audit reports.

9. Cross-learning and innovative practices

There is some evidence that quality assurance stands the most chance of avoiding decoupling and colonization

pitfalls when there is an organizational need for it. The rationale is the desires of managers to assure that their investment in evaluative information results in a credible and reliable product for use in improving program operation. Alternatively, if evaluative information produced by an organization receives significant external attention managers may wish to avoid potential embarrassment by implementing quality assurance systems.

The case of performance audits carried out by SAIs exemplifies a situation where an organization feels a real need to pay attention to the quality of evaluative information. Such a focus on quality evaluative information can be found in other organizations, but much more selectively. The US Department of Education with regards to its performance measurement and reporting efforts and the EU DG Agriculture for its evaluation work illustrate such cases.

In the absence of strong organizational need for quality assurance, a wide variety of approaches are practiced in the various jurisdictions. Here we summarize some of the more innovative practices and identify opportunities for learning across jurisdictions and across evaluative information types.

Innovative structural approaches aim to enlighten producers, commissioners and users in the practice of evaluation. Examples include standards based training (SEVAL), capacity development (World Bank), promulgation of standards and 'best practice' guidelines (EU). Professional standards are the bedrock of performance audit practice. And in the less developed performance reporting area, efforts are underway in a number of jurisdictions to develop and promulgate reporting standards, often using the financial statements reporting standards model as the basis for the development.

Formative approaches assist in the production of high quality work. Steering groups represent a now accepted practice across evaluative types. In one novel approach a consortium of American local authorities work together to develop and improve measures for benchmarking purposes. A more centralized approach is taken in the UK, where Treasury based 'technical' panels advise departments on the development of performance measures. Uncharacteristically, even some SAIs take a formative approach to assuring the quality of performance reporting, preferring to educate rather than blame. SAIs themselves have what is probably the most developed and longstanding internal formative approach to quality assurance. Formative quality assurance appears least practiced in the evaluation domain despite the extensive prescriptive literature on this. Evaluation may have something to learn from the positive formative quality assurance experience of DG Agriculture along with the experiences in performance reporting and auditing.

We identify two innovative summative approaches, both of which might be called 'friendly'. Auditors in Australia and New Zealand review samples of one another's performance audit reports in a collegial and constructive environment. And The US Department of Education provides incentives for high quality, by publicizing exemplary practice. Less friendly summative approaches use sticks rather than carrots as for example in publishing rankings of the quality of performance reports (United Kingdom).

Our study also allowed us to compare the quality assurance approaches across the three type of evaluative information examined. The fairly extensive focus on quality assurance in the performance audit field stands in contrast to the modest efforts in evaluation and the still more limited attention paid to quality assurance in performance reporting, an albeit much less developed field. To our knowledge, evaluation practice has not to date paid much attention to performance auditing practices.

Performance auditing could also learn from evaluation in this area. Auditing tends to prefer things, including quality, to be black or white. Evaluation practice is quite comfortable with the idea that good quality is not absolute but varies depending on circumstances. Quality that is 'fit for purpose' is an idea that is not readily adopted by performance auditing.

The data collected for this study is insufficient to tell us when to use what type of quality assurance approach. The study sample and methodology have enabled us to describe the state of evaluation practice in some of the world leaders in the practice of evaluation, performance reporting and performance auditing. Their experience with quality assurance has led to the development of some preliminary ideas about the origins, effects and dysfunctions of quality assurance practices for evaluative information. More conclusive findings about these relationships require further research.

What is clear is that assuring the quality of evaluative information often involves fundamental questions about organizations and their programs and policies, and as a result can be subject to self-interest pressures. It is likely that the risks of decoupling and colonization are greater when quality assurance is cast upon organizations from the outside. Those interested in imposing quality assurance from above, or from the outside, might consider what might be called a developmental approach: start with the building blocks of standards, capacity building and education; continue with formative approaches and then friendly summative approaches; conduct periodic system checks, supplemented if necessary with less friendly summative approaches.

References

Algemene, R. (1990–91). *Verslag 1990, vergaderjaar 1990-1001, 22 032, Nos. 1–2*. The Hague: Staatsuitgeverij.

Auditor General of Canada (1983). *Report of the auditor general to the House of Commons*. Ottawa: Supply and Services (Chapter 3).

Auditor general of Canada (1986). *Report of the auditor general to the House of Commons*. Ottawa: Supply and Services (Chapter 15).

Auditor General of Canada (1993). *Report of the auditor general to the House of Commons*. Ottawa: Supply and Services (Chapters 8–10).

Auditor General of Canada (1996). *Report of the auditor general to the House of Commons*. Ottawa: Public Works and Government Services Canada (Chapter 3).

Auditor General of Canada (1997). *Reporting performance in the expenditure management system Report of the auditor general of Canada to the House of Commons. Ottawa* (Chapter 5).

Auditor General of Canada (2000). *Reporting performance to parliament: Progress too slow Report of the Auditor General of Canada to the House of Commons. Ottawa* (Chapter 19).

Australian National Audit Office (1991). *The auditor-general audit report no. 23*. Canberra, ACT: Australian Government Publishing Office.

Australian National Audit Office (1992a). *The auditor-general audit report no. 13*. Canberra: Australian Government Publishing Office.

Australian National Audit Office (1992b). *The auditor-general audit report no. 26*. Canberra: Australian Government Publishing Office.

Australian National Audit Office (1993). *The auditor-general audit report no. 35*. Canberra: Australian Government Publishing Office.

Australian National Audit Office (1997). *The auditor-general audit report no. 3*. Canberra: Australian Government Publishing Office.

Bouckaert, G. (1993). Measurement and meaningful management. *Public Productivity and Management Review*, *17*, 1.

Boyle, R. (2004). Assessment of performance reports: A comparative perspective. In R. Schwartz, & J. Mayne (Eds.), *Quality matters: Seeking confidence in evaluation, auditing and performance reporting*. New Brunswick NJ: Transaction.

Canadian Comprehensive Auditing Foundation. (2002). *Reporting principles—Taking public performance reporting to a new level*. CCAF-FCVI. Ottawa. Retrieved 14 July 2003 from http://www.ccaf-fcvi.com/english/reporting_principles_entry.html

Canadian Institute of Chartered Accountants. (2002). *Management discussion and analysis: Guidance on preparation and disclosure*. Review Draft. Retrieved 14 July 2003 from http://www.cica.ca/index.cfm/ci_id/10383/la_id/1.htm

Chelimsky, E. (1983). The definition and measurement of evaluation quality as a management tool. In R. G. St Pierre, *Management and organization of program evaluation*. *New directions for program evaluation, no. 18* (pp. 113–126). San Francisco: Jossey-Bass, 113–126.

Chelimsky, E. (1987). The politics of program evaluation. *Social Science and Modern Society*, *25*, 24–32.

Coe, C. (1999). Local government benchmarking: lessons from two major multigovernment efforts. *Public Administration Review*, *59*(2), 110–123.

CSE (Conseil scientifique de l'évaluation) (1996). *Petit Guide de l'évaluation des politiques publiques*. Paris: La Documentation française.

Derlien, H., & Rist, R. C. (2002). Policy evaluation in international comparison. In J. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 439–456). New Brunswick, NJ: Transaction, 439–456.

European Commission (1999a). *Evaluation socio-economic programmes*, Vol. 1. Luxembourg: Office des Publications.

European Commission (1999b). *SEC(1999)69/4—Communication from Mrs Gradin and Mr Liikanen in agreement with the President Spending more wisely: Implementation of the Commission's evaluation policy*.

European Commission (2000). *SEC (2000)1051—Communication to the commission from Mrs Schreyer in agreement with Mr Kinnock and the President Focus on results: Strengthening evaluation of commission activities*. Brussels: EC.

Furubo, J. E., Rist, R. C., & Sandahl, R. (Eds.). (2002). *International atlas of evaluation*. New Brunswick and London: Transaction Publishers.

General Accounting Office (1999). *Managing for results: Opportunities for continued improvements in agencies' performance plans. Washington DC*.

Government Accounting Standards Board. (2003). Reporting performance information: Suggested criteria for effective communication. Retrieved 30 January 2004 from http://www.gasb.org

Grasso, P. (2004). Quality of evaluative information at the World Bank. In R. Schwartz, & J. Mayne (Eds.), *Quality matters: Seeking confidence in evaluation, auditing and performance reporting*. New Brunswick NJ: Transaction.

Greene, J. (1990). Technical quality versus user responsiveness in evaluation practice. *Evaluation and Program Planning*, *13*, 267–274.

House, E. (1987). The evaluation audit. *Evaluation Practice*, *8*(2), 52–56.

Institute of Public Administration Australia. (2001). *The Judging Criteria*. www.wa.ipaa.org.au/lonnie/criteria.html

Joint Committee of Public Accounts of the Parliament of the Commonwealth of Australia. (1989). Report 296, *The auditor general: Ally of the people and parliament; Reform of the Australian Audit Office*, Canberra: Australian Government Publishing Service.

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed.). Thousand Oaks: Sage.

Law, J. (2001). Accountability and annual reports: the case of policing. *Public Policy and Administration*, *16*(1), 75–90.

Lonsdale, J. (2000). *Advancing beyond regularity: Development in value for money methods at the National Audit Office 1984–1999* (unpublished PhD thesis, Brunel University).

Lonsdale, J., & Mayne, J. (2004). Neat and tidy…and 100% correct: Assuring the quality of SAI performance audit work. In R. Schwartz, & J. Mayne, J. (Eds.), *Quality matters: Seeking confidence in evaluation, auditing and performance reporting*, New Brunswick NJ: Transaction.

Muir, E. (1999). They blinded me with political science: On the use of nonpeer reviewed research in education policy. PS. *Political Science and Politics*, *32*(4), 762–764.

National Academy of Public Administration. (1994). *The roles, mission and operation of the US General Accounting Office*. Report prepared for the Committee on Governmental Affairs, United States Senate.

National Audit Office (2001). *Measuring the performance of government departments. HC301*. London: The Stationery Office.

New Zealand Controller and Auditor General. (2000). *First Report for 2000*. Wellington.

New Zealand Controller and Auditor General. (2001). *Reporting public sector performance*. Wellington.

Palumbo, D. J. (Ed.). (1987). *The politics of program evaluation*. Newbury Park, CA: Sage.

Patel, M. (2002). A meta-evaluation, or quality assessment, of the evaluations in this issue, based on the African evaluation guidelines: 2002. *Evaluation and Program Planning*, *25*, 329–332.

Patton, M. Q. (2001). Use as a criterion of quality in evaluation. In A. Benson, D. M. Hinn, & C. Lloyd, *Visions of quality: How evaluators define, understand and represent program quality*. *Advances in program evaluation* (Vol. 7) (pp. 155–180). Amsterdam: JAI, 155–180.

Power, M. (1997). *The audit society*. Oxford University Press.

Schwandt, T. A. (1992). Constructing appropriate and useful metaevaluative frameworks: Further reflections on the ECAETC audit experience. *Evaluation and Program Planning*, *15*, 95–100.

Schwandt, T. A., & Halpern, E. S. (1988). *Linking auditing and metaevaluation: Enhancing quality in applied research*. Beverly Hills, CA: Sage.

Schwartz, R. (1998). The politics of evaluation reconsidered: A comparative study of Israeli programs. *Evaluation*, *4*(3), 294–309.

Schwartz, R. (1999). Coping with the effectiveness dilemma: Strategies adopted by state auditors. *International Review of Administrative Sciences*, *65*(4), 511–526.

Segsworth, B., & Volpe, S. (2004). Auditing the evaluation function in Canada. In R. Schwartz, & J. Mayne (Eds.), *Quality matters: seeking confidence in evaluation, auditing performance reporting*. NewBrunswich NJ: Transaction.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, *18*(2 and 3), 277–310.

Smith, M. F. (1999). Should AEA begin a process for restricting membership in the profession of evaluation. *American Journal of Evaluation*, *20*(3), 521–531.

Stierhoff, K. 1999. *The certification of program evaluators: A pilot survey of clients and employers*. Retrieved 8 July 2003 from http://www.evaluationcanada.ca/txt/certification_survey_sep99.pdf

Streib, G. D., & Poister, T. H. (1999). Assessing the validity, legitimacy, and functionality of performance measurement systems in municipal governments. *American Review of Public Administration*, *29*(2), 107–123.

Stufflebeam, D. L. (1974). *Meta-evaluation*. Kalamazoo, MI: Western Michigan University Evaluation Center. Occasional Paper Series #3.

Stufflebeam, D. L. (2000). *Guidelines for developing evaluation checklists* [On-line]. Available: www.wmich.edu/evalctr/checklists/

Stufflebeam, D. L. (2001a). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation*, *22*, 71–79.

Stufflebeam, D. L. (2001b). The metaevaluation imperative. *American Journal of Evaluation*, *22*, 183–209.

Toulemonde, J., Usher, N. & Summa-Pollitt, H. (2004). Triple check for top quality or triple burden?: Assessing EU evaluations. In R. Schwartz & J. Mayne (Eds.), *Quality matters: Seeking confidence in evaluation, auditing and performance reporting*. New Brunswick NJ: Transaction.

Weiss, C. (1973). Where politics and evaluation research meet. *Evaluation*, *1*, 37–45.

Widmer, T., Landert, C., & Bachmann, N. (2000). Evaluations—Standards der Schweizerischen Evaluations gesellschaft (SEVAL-Standards). Bern/Genève: SEVAL.

Widmer, T. (2004). Instruments and procedures for assuring evaluation quality: A Swiss perspective. In R. Schwartz & J. Mayne (Eds.), Quality matters: Seeking confidence in evaluation, auditing and performance reporting. New Brunswick NJ: Transaction.

Wildavsky, A. (1972). The self-evaluating organization. *Public Administration Review*, *32*, 509–520.

World Bank (2002). *2002 Annual report on operations evaluation. Operations evaluation department*. Washington: World Bank.