Working Paper

R2006:002

Evaluation: definitions, methods and models

An ITPS framework

Federica Calidoni-Lundberg

SWEDISH INSTITUTE FOR GROWTH POLICY STUDIES

ITPS, Swedish Institute For Growth Policy Studies
Studentplan 3, SE-831 40 Östersund, Sweden
Telephone: +46 (0)63 16 66 00
Fax: +46 (0)63 16 66 01
E-mail info@itps.se
www.itps.se
ISSN 1652-0483

For further information, please contact Federica Calidoni-Lundberg
Telephone +46 063 166617
E-mail federica.calidonilundberg@itps.se

Foreword

In the last decades evaluation has became more and more of an independent science, that has its roots in many disciplines and turns out to be a useful tool for understanding and implementing policy studies, performance assessment, engineering design, investment portfolio and so on. Notwithstanding the large amount of definitions and objectives of evaluation programmes it is clear cut, and widely participate in the scientific community that evaluators strive to contribute to social betterment and this could only be achieved if evaluation findings are fed back to inform programme administrators, policymakers and other stakeholders and to improve the programme structure and operations. One of ITPS main tasks is to develop evaluation methods and use them.

This work shed light on the existing literature on evaluation research and it aims to classify and easily illustrate the most recent approaches to evaluation theory, underlying their advantages and disadvantages in different context. Looking at the latest contributions, the project will focus on (1) existing definition of evaluation, (2) qualitative and quantitative methods of evaluation and (3) the role of the evaluator.

The report summarizes and clarifies what is *evaluation* in a governmental agency, which functions and values it serves, and which issues, models and methods of the existing literature can be useful instruments for evaluators facing budget, time or data constraints.

The report was written by Federica Calidoni-Lundberg, ITPS Östersund.

Östersund, July 2006

Håkan Gadd
Head of the Evaluation department

Table of Content

Summary

Evaluation has acquired in recent years the status of independent science in many countries and has recently been object of increased interest from the point of view of different disciplines. This report aims to highlight what are the most discussed topics in the current literature in order to have a clearer picture of the actual state of the debate.

To do this, it first concentrates on the historical evolution of the concept of evaluation, its philosophical roots and the reasons why it has different shades according to different point of views, and then it sketches a picture of the current state of the literature. If the core development of evaluation practice can be found in the post-war era for most of the developed countries, the way it has grown and the methods and procedures applied in different fields have had a heterogeneous growth and program evaluation has not been seen in the same manner in different countries: if in the United States evaluation practice has its heredity in social sciences studies, in Canada and in Scandinavia the basic purpose of evaluation is to provide information to clients and users.

The report presents a number of definitions in order to shed light on the historical upgrading of the term itself. The main reason for the existence of such a number of definitions is the large number of actors involved in the evaluation process, with different aims and objectives, different methods applied, diverse priorities and so on. Notwithstanding the great deal of actors, methods and procedures involved, some core ideas can be found in most of the definitions mentioned in the report, such as: the common attempt of evaluators to attribute observed outcomes to known inputs, and vice versa; the importance of efficiency, accountability and resource allocation; and the attention towards programs' implementation and delivery.

In the light of these definitions, the report stress how the practice of evaluation at ITPS is very much coherent with Vedung's idea of evaluation as a *"tool to determine the worth and value of public programs, with the purpose of providing information to decision-makers and improve institutional performance in the spirit of looking backwards to improve forward directions"* (Vedung, 1997).

With this in mind the report wants to illustrate the existing debate on evaluation models and methods, in order to make clear for those that are in charge of the evaluation practice at ITPS, or elsewhere, that every program is a nexus in a set of political and social relationships and it involves different actors that the

evaluators can not ignore because they affect the choice of models and methods.

From the philosophical roots of positivism, constructivism and realism, different schools of thought have led the development of evaluation: on one side positivist theorists aimed to discover regularities and laws for social sciences as it is done for natural ones, opposed to authors that deny the possibility of objective knowledge and assert that the world can be understood only through theorization of the observer (constructivism); and on the other hand the recognition that programs are embedded in multi-layered societies which should be taken into account for the different ways in which they affect different context. These different traditions are very helpful in explaining why approaches to evaluation can be so different from each other, but the belonging to different theoretical streams is not too strict and it is easy to find overlapping of theories, models and methods.

The mainstay of this work focuses on practical aspects and tries to answer the question *"what does it mean to evaluate?"*, bearing in mind the ITPS perspective of evaluation as a tool to look and study backward events to improve forward directions. To do this the evaluator has to engage in four work steps: defining the purpose of the evaluation, the issue, the model and the method. Most likely for ITPS evaluation work the definition of the purpose and the issue is easy task given the common focus on evaluation of Swedish national and regional policies for industrial development and innovation. When it comes to define model and methods there are, however, many aspects to keep into account and the evaluators' choice could be constraint by financial resources, data and time availability. The report presents different models (Result-oriented, Actor-oriented and Economic models) and sheds light on their advantages and disadvantages, focusing in particular on economic and result-oriented models. Moreover, each model can be applied either through quantitative or qualitative methods, that are briefly summarised in the work.

Finally, given the emphasis recently given to the role of the evaluator in ethical terms, the focus shifts to the effects that different perceptions, values and background of the evaluator can have on the results of evaluation: it is natural that evaluators, as human beings, might feel strongly for certain issues and promote them in their work, and the evaluator's job of ensuring the highest quality evaluation might be under constraints of limited budget, time or data availability.

The report concludes that evaluation, from ITPS point of view, can not ignore the limits due to method's choice, time and budget constraints and the influence of different stakeholder points of view. It stresses the importance of well-defined purposes and issues, and the need of transparency in motivating the choice of a given method over others in relation to the evaluation's aim. Notwithstanding the personal experiences, interests and abilities of the evaluators, ITPS reports and analysis should clearly inform clients and stakeholder of possible bias and constraints that could affect the quality of the evaluation, choice of models and methods must be justified from the beginning and the evaluator's point of view should be clearly stated as such.

1 Where does evaluation come from? History

In the last decades the definition of evaluation has been subject of intense debate and has therefore undergone deep changes, leading to a great deal of works on the topic. The intense debate on evaluation is mainly due to the fact that it is, first of all, a transversal discipline that crosses many other fields of science, has many different purposes, perspectives and uses.

1.1 Historical background

For a better understanding of what it means *to do evaluation* it is useful to briefly recall the historical background of the concept. Evaluation has a long history, dating back four thousands years to China where it was used to assess public programs, but it emerged as a distinct area of professional practice only in the post-war period. It is in the Unites States that for the first time we find the term *evaluation* applied to three important social topics: educational innovation, resource allocation and anti-poverty programs. Since then, the concept of evaluation has conventionally been linked to social science studies with its traditions of quantitative and experimental studies, economic appraisal methods and participatory methods involving the beneficiaries of the programs in the evaluation process.

A first wave of evaluation development starts in the 60s, when many countries, such as the United States, Canada, Germany and the United Kingdom, start feeling the need of monitoring the progress of programs, evaluate the effectiveness of the operations and assess the performance of government activity. The 1960s were also a very successful period for the natural sciences, and this led to an almost unshakable faith in the natural sciences and their methods, which, by consequence, where adopted by social scientists to tackle society's ills. Patton (1997) refers to this as "a new order of rationality in government – a rationality underbinded by social scientists" (p.7). With the application of scientific methods to program evaluations, traditional evaluation was born.

Traditional evaluation is characterized by its emphasis on scientific methods. Reliability and validity of the collected data are key, while the main criterion for a quality evaluation is methodological rigor. Traditional evaluation requires the evaluator to be objective and neutral and to be outcome-focused (Fine, Thayer, & Coghlan, 2000; Torres & Preskill, 2001).

It is only from the 1970s onwards that evaluation has begun to take root in different European countries, with different traditions and emphasis. Program evaluation has not been seen in the same manner in the different countries for many reasons. In the States, for example, evaluation is a significant part of social science research and it has its origin in the massive initiative supported under the banner of the "War to Poverty" by the presidents Kennedy and Johnson. The development represents the melting and maturing of two streams of intellectual inquiry during this time: efforts by government managers to get control of their programs and resources through a new rationalization of the process and the development of applied social sciences, particularly the strategies of survey research and large scale statistical analysis. The movement of such techniques into the public sector did not occur concurrently across government: it came first under the Defense Department under Robert MacNamara in the early 60s and shortly after President Johnson ordered its implementation in all executive-branch agencies. By 1970, however, it was largely abandoned because the measures needed to justify the Planning, Programming and Budgeting System did not exist at the time and it was not clear how to measure inputs and outputs. But while fascination with the hyper rationalization offered by economic models has waned considerably, the broader concern with policy evaluation has diffused into other parts of the federal government and policy evaluation is now seen as a necessary tool for good management.

In the same years the concept of evaluation develops in Canada in a different way: there is less of a history of evaluation as a significant part of social science research and Canadian government has taken more of a leadership role than has been the case in the United States. In Canada the basic purpose of program evaluation is, therefore, to provide clients and users with relevant and timely information and analysis to assist them to make better decisions on resource allocation and program improvements, less emphasis is placed on truly scientific studies that attempt to provide definitive statements about program outcomes, effects and impact.

In Scandinavia, and in particular in Sweden, evaluation was at first seen to be an open process of inquiry, producing information close to users and with a focus on being useful. Nowadays while the international debate calls for a return to evaluation systems that are closer to monitoring and performance management, Sweden has developed a renewed interest in the institutional issues and the whole field is moving towards more centralized and institutionally independent forms of evaluation. In France evaluation has, until recently, mirrored the characteristics of the French state with a formal

structured approach at a central government level and a more dynamic practice at a regional and local level. Evaluation has however not been static in any of these countries, and in many of them the focus and scale of evaluative activity has reflected the changing policies of the different governments.

This very brief depiction of the historical context of evaluation practice is intended to provide a backdrop against which recent developments can be assessed. In short, the 1970s were characterized by a predominantly social-scientific approach to program evaluation. Other approaches were not generally accepted as valid or scientific, so the variety of methods at the evaluator's disposal was limited. The 1980s and 1990s were characterized by a host of developments both in the political realm and in the academic realm. As a consequence of the increased emphasis on accountability, nonprofits and government agencies start facing pressure to demonstrate results, be held accountable, show high performance, and to behave like business generally (Fine et al., 2000; Bozzo, 2000; Hoefer, 2000; Renz, 2001; Lindenberg, 2001; Poole et al., 2000; Love, 2001; Wholey, 2001). The underlying assumption appears to be that agencies and nonprofits can and should be run the way businesses are run.

1.2 Philosophical roots

In order to understand the development of different evaluation methods, their interactions and the reasons for choosing one method instead of the other according to the object of evaluation, it is useful to briefly recall the philosophical tradition underpinning methodological approaches to evaluation.

Three philosophical traditions underpin the broad methodological approaches to evaluation that are used in socio-economic development programs.

From the 18th century onwards *positivism* has provided the philosophical underpinning of mainstream science from the 18th century onwards. The word *positivism* in social science and philosophy means the application of scientific methods to social phenomena. This tradition believes that observation is the instrument to obtain objective knowledge, so that different researchers applying the same observation instruments should obtain the same findings. These results, when analyzed by objective techniques, should lead to the same outcome whoever applied the technique. Positivist traditions aim to discover regularities and 'laws', applying natural sciences rules to social sciences. Explanations rest on the aggregation of individual elements and their behaviors and interactions, the whole is understood by looking at the parts, the basis for survey methods and econometric models used in evaluation. At best these methods can provide quantifiable evidence on the relationships between the inputs of interventions and their outcomes.

When it comes to the application of this tradition to evaluation of socio economic development, the limitations stem from the difficulties of measuring many of the outcomes that are of interest, the complexity of interactions between the interventions and other factors and the resulting absence of insights into 'what works'. Nowadays these limitations are well recognized, in particular the fact that what can be observed is usually incomplete and needs to be interpreted by frameworks or theories; it is always mediated, simplified or even distorted by the tools and techniques we use to collect data; the difficulty in most human settings to expect to find regularities and 'laws' that do not vary across 'local' contexts; problems of complexity where phenomena themselves change as they interact – often in unpredictable ways.

These limitations of positivism have led to the emergence of various *post-positivist* schools. The most radical, rejecting most of the assumptions of positivism, is *constructivism* which denies the possibility of objective knowledge and contends that it is only through the theorizations of the observer that the world can be understood. Facts are, therefore, always theory laden and facts and values are interdependent. In this tradition evaluators and stakeholders are at the centre of the enquiry process. The evaluator is likely to assume a responsive, interactive and orchestrating role bringing together different groups of stakeholders with divergent views for mutual exploration and to generate consensus.

In 1962 Kuhn argued that scientific knowledge is not "discovered" but "constructed" in a social context. The knowledge "constructed" depended on the particular "paradigm" within which the research was situated. And this was the first serious challenge to the supposed universality of truth from within the natural sciences and led to a long-standing debate in scientific circles. Lincoln and Guba (1985) brought the debate to the field of evaluation, launching what has often been referred to as the "paradigm wars" (Caracelli, 2000) and challenging the privileged status of the traditional evaluation over alternative approaches. Essentially, all disagreement boils down to a philosophical argument: whether or not the world is ultimately knowable, and whether or not there is such a thing as objectivity. If, as constructivists Lincoln and Guba (1985) argue, each "truth" is socially constructed, then whose truth matters most? The one of the evaluator?

Realism, on the other hand, concentrates on understanding different contexts and seeks to open up the 'black box' within policies and programmes to uncover the mechanisms that account for change. In doing so the tradition recognises that programmes and policies are embedded in multi layered social and organisational processes and that account should be taken of the influence of these different 'layers' as well as different contexts. Emphasis is placed on

social inquiry explaining interesting regularities in 'context-mechanism-outcome' patterns. The systems under investigation are viewed as open and the focus of evaluators is on the underlying causal mechanisms and on explaining why things work in different contexts.

In practice evaluators are unlikely to see themselves as operating exclusively within any one of these philosophical traditions but will tend towards one or another depending on the circumstances of the evaluation. In general terms evaluation tools applied in the tradition of positivism will be helpful for the purposes of scrutiny and realist approaches are likely to generate formative insights especially where the evaluation work takes place within a context where policy is being developed. Constructivist approaches can be particularly helpful in 'putting programmes right' but are especially dependent upon the trust between the evaluator and stakeholders. It is important to recognise these different traditions, if only because they help explain why approaches to evaluation can be so different from each other; and it is certainly useful for evaluators to be explicit about their philosophical traditions and preferences.

1.3 Theoretical developments

After a period of relative inactivity in the 1950s, several events and developments sparked an increased interest in evaluation in the 1960s, when achievements in natural sciences create an almost unshakable faith in their methods, immediately adopted by social scientists to tackle society's ills. With the application of scientific methods to program evaluations, traditional evaluation was born.

Competing approaches have since been developed, mostly in response to one of the most serious drawbacks of traditional evaluation: the fact that many reports are not used or even read (Torres and Preskill, 2001; Fetterman, 2001; Patton, 1997). One of the earliest alternatives to traditional evaluation is what is known as *Responsive Evaluation* an approach to evaluation that is less objective and more tailored to the needs of those running the program. In Stake's words, responsive evaluation *"sacrifices some precision in measurement, hopefully to increase the usefulness of the findings to persons in and around the program"* (Stake, 1973). In short, the 1970s were characterized by a predominantly social-scientific approach to program evaluation. Other approaches were not generally accepted as valid or scientific, so the variety of methods at the evaluator's disposal was limited.

The 1980s and 1990s were characterized by a host of developments both in the political realm and in the academic realm. Increased scrutiny, increased competition for decreased levels of funding, increased demand to demonstrate results and increased emphasis on accountability led nonprofits and govern-

ment agencies to face higher pressure to demonstrate results, be held accountable, show high performance, and to behave like business generally. In this period the practices identified in the literature can be divided into three broad categories: strategic analysis/alignment and organizational effectiveness; impact evaluation; and performance management.

In sum, the net result of the increased popularity of business practices, methodological innovations like theory-based evaluation, and improved technology could be called a consolidation of the traditional evaluation. Still concerned with numbers, objectivity, and rigor, traditional evaluation has shifted its attention from activities (Sawhill & Williamson, 2001) and indicators such as operating expense ratios (Kaplan, 2001) to outcomes or impacts. In the same years there is, similarly to the drive toward accountability, a drive toward democratization due to a set of separate but related forces. While the accountability drive seems to come from government and business, the democratization drive appears to originate in the academic world, as a consequence to the increased attention to "constructed" knowledge (Khun, 1962) and the serious implication that the constructivists' argument that the cultural context of research is an important determinant of its outcomes has for program evaluation.

Moves toward alternative approaches to evaluation are not only driven by academic or philosophical debates. It also flows from a major shortcoming of the traditional evaluation: its lack of use. Taken together, the challenges to the legitimacy of traditional research methods, the recognition that language in itself is not neutral, the acknowledgment that the aims of evaluation may vary, and under-utilization of traditional evaluation have significantly undermined its authority in the 80s and can be identified as the driving forces behind the second trend in evaluation practice: increased popularity of more participative approaches in program evaluation from the middle of the 1990s.

Ryan (1998) argues that such approaches improve decision-making, are more credible, and consistent with evaluation's overall goal of being democratic and inclusive. Different participative strategies call for different levels of stakeholder involvement and, by extension, different roles for the evaluator. The three main categories of participative approaches that are found in the literature are stakeholder-based evaluation, empowerment evaluation, and self-evaluation.

In sum, it seems fair to say that while traditional evaluation has "hardened" because of its shift in emphasis from activities and outputs to outcomes and results, the competing approaches have "softened" because of the evaluator's gradual move from content expert to methodological expert and, finally, coach and mentor.

1.4 Where is evaluation theory now?

From the brief summary of the previous paragraph it is clear that the field of evaluation practice has diversified, evaluators have a more diverse set of tools to tackle evaluations, and the days of the one-type-fits-all approach to evaluation are past. Moreover the role of the evaluator, as well as other variables, changes according to the evaluation approach. However the "paradigm war" is, nowadays, still open and the debate has not yet been settled. Smith (2001) agrees, saying that the debate *"is and was about differences in philosophy and "world view" […] No sooner is it put to bed under one guise than to raise its ugly head under another"*

If anything, the distance is greater, at least in terms of articulated positions, between those who see evaluation as a quest for social justice which requires advocacy for the disenfranchised and those who see evaluation as the most nonpartisan, fair search we can mount for understanding what is happening and why, and for reaching judgments on merit, worth, and value.

In conclusion, while the argument originally revolved around incompatible philosophical positions on objectivity, it now focuses on the espoused purpose of program evaluation. Those who argue for social justice are the former constructivists and those who still subscribe to the assessment of value or worth generally fall into the objectivist camp.

In spite of the continued paradigm war, which tends to polarize the field between two alternatives (objectivist or constructivist assumptions; quantitative or qualitative methods; summative or formative purpose; etc.), the literature shows an increase in popularity of pragmatic approaches (Bengston & Fan, 1999; Mohr, 1999; Pratt et al., 2000). These approaches essentially ignore the paradigm debate and show no hesitation to mix approaches in ways that loyalists to either paradigm would never do out of fear of compromising their findings.

One might even speculate that these pragmatic approaches are appearing because of the persistence of the paradigm war. Possibly the best justification for calling the advent of mixed-method approaches a trend is the work by Henry, Julnes, and Mark (1997) and Mark, Henry and Julnes (2000). These authors attempt to give the pragmatic approach more legitimacy by providing a theoretical basis for it, called *emergent realism*. Thus far, there are no articles reporting on an application of this philosophy to program evaluation. If this trend continues, it may have profound implications for program evaluation as an emerging field of practice.

2 Definition

A lot is written about evaluation, a great deal of which is misleading and confused. Many informal educators are suspicious of evaluation because they see it as something that is imposed from outside. It is a thing that we are asked to do; or that people impose on us. As Gitlin and Smyth (1989) comment, from its Latin origin meaning 'to strengthen' or to empower, the term evaluation has taken a numerical turn - it is now largely about the measurement of things - and in the process it can easily slip into becoming an end rather than a means.

One of the main reasons why it has been difficult to reach a unique definition of evaluation can be found in the great number of actors involved, with four main groups whose interests sometimes compete with each other in defining evaluation priorities: policy makers, professionals and specialists, managers and administrators, citizens and those affected by public policies.

Each of these groups makes assumptions about how evaluation can help them. For example, policy makers tend to see evaluation as a tool to ensure the accountability and justification for policy decisions; citizens are more likely to regard it as an instrument for democratic accountability and an opportunity to shape public intervention to their needs; managers and administrators are often concerned with the delivery of policies and programs; while professionals often regard evaluation as an opportunity to improve the quality of their work or the autonomy of their own professional group. This review will therefore try to summarize the mainstream ideas and explain the historical development and the theoretical and philosophical background of the different definitions bearing in mind the different interests and pressures in action.

Table 2.1 reports some of the most common definitions of evaluation. Reading these few definitions it is clear that evaluation has varied roots and it is not an unified practice derived from a single set of tradition; notwithstanding the different origins it is, however, evident the existence of some core ideas concerning:

1 *Scientific research and methods.* Many of the basic ideas and methods used in evaluation are shared with the wider research community in the social sciences and economics. Even thought in complex socio-economic programs explanations are rarely straight forward, independently of the definition of evaluation taken into account much of the work of evaluators is an attempt to attribute observed outcomes to known inputs, and vice versa.

2 *Economic theory and public choices.* Economic thinking is present within evaluation at different levels: such as notions of efficiency and resource allocation, institutional incentives and behavior, and macro-economic studies that seek to identify aggregate effects of policy interventions.

3 *Organization and management theory.* This has begun to feature more prominently in evaluation in recent years as the focus has increasingly shifted to implementation and delivery of programs and policies.

4 *Political and administrative sciences.* As public programs and their managers address issues of the policy process and public sector reform they increasingly draw on ideas concerned with governance, accountability and citizenship.

Table 2.1 Definition of evaluation

Source	**Definition**
Stufflebeam (2000)	Evaluation is a study designed and conducted to assist some audience to assess an object's merit or worth
Vedung (1997)	Evaluation is a careful retrospective assessment of the merit, worth and value of administration, output and outcome of government intervention, which is intended to play a role in future practical situations.
Scriven (1991)	Evaluation is the process of determining the merit, worth and value of things and evaluations are the products of that process. Evaluation is not the mere accumulation and summarizing of data that are clearly relevant for decision making…gathering and analyzing the data that are needed for decision making comprise only one of the two key components in evaluation, a second element is required to get to conclusions about merit or net benefits: evaluative premises or standards. Evaluation has two arms: one in engaged in data gathering, the other collects, clarifies and verifies relevant values and standards.
Centre for Program Evaluation–Government of the United States	Evaluation (1) assesses the effectiveness of an ongoing program in achieving its objectives, (2) relies on the standards of project design to distinguish a program's effects from those of other forces, and (3) aims at program improvement through a modification of current operations.
COBUILD English Language Dictionary-Collins	Evaluation is a decision about significance, value, or quality of something, based on careful study of its good and bad features.
ASEAN Australia Development Cooperation Program	The assessment of how well a project/activity achieved its objectives. Ongoing evaluation (during project implementation) is referred to as 'review' and is linked closely with monitoring.
Australian Government	The process of reviewing the overall efficiency (did we do the right thing?), effectiveness (did we do the best possible way?) and economy (did we get the best possible value for what we invested?) of a project. Evaluation also considers the alignment of a project's outcomes to the program's objective(s).
www.evaluateit.org	Assessment at a point in time of the value, worth or impact of a project or program.
Eurydice–The information Network on Education in Europe	The forming of a judgment based on the collection of data with a view to determining the quality of one or more (educational or administrative) tasks and improving the way they are performed.

From the above it follows that evaluators are similarly diverse: they might be economists concerned with efficiency and costs; or managers' consultants interested in the smooth running of the organization; policy analysts with a commitment to public sector reforms and transparency; or scientists concerned to establish truth, generate new knowledge and confirm or disconfirm hypothesis.

Given these basic homogenous characteristics three are the aspects that lead to the flourishing of different evaluation frameworks: the object of evaluation, the purpose of evaluation and the methods used (mainly dictated by the evaluator's background and the benchmark philosophical theory).

But, what does it mean to do evaluation in a governmental agency? How can we define ITPS main purpose for evaluative activity? And what is the definition that better fits it?

Lately the need for evaluation has been felt more urgently given the increasing number of global political forces that are transforming societies and governmental agencies are called to understand and explain how societies are changing, to strengthen institutions, to improve their performance and to help governments to effectively react to global changing such as aging population, immigration, increased number of new technologies, globalization and so on. Chelimsky (1997) and Cronbach (1980) clearly express the importance of evaluation in any democratic society with the following words: *"the ability of evaluators to serve policy depends on what they know about how politics work"* Chelimsky and *"a theory of evaluation has to be a theory of political interactions as well as a theory of how to determine facts or how knowledge is constructe"* (Cronbach).

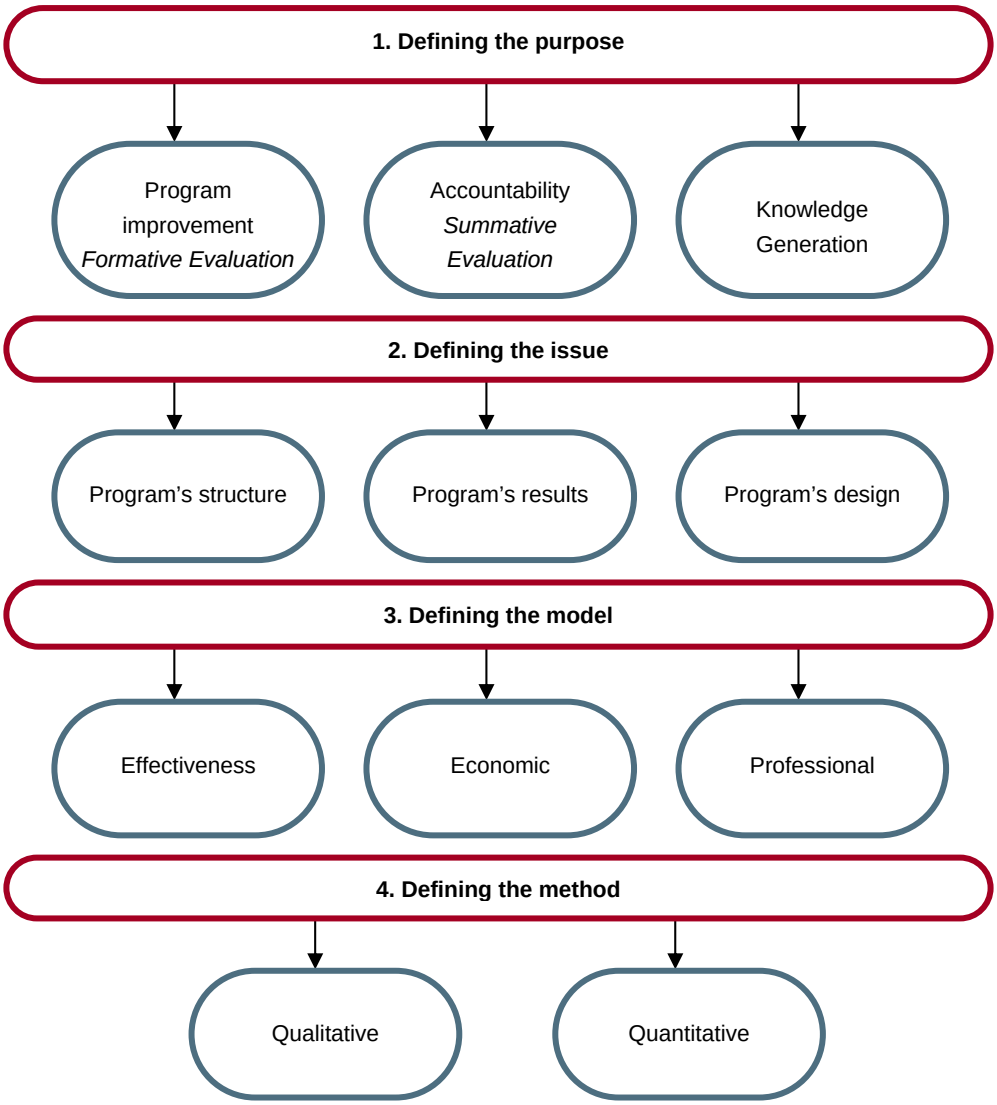
ITPS is therefore called to determine the worth and value of public programs, both finished and ongoing, with the main purposes of providing information to decision makers and improve institutional performance, in the spirit of *"looking backwards to improve forward directions"* (Vedung, 1997). ITPS evaluation concerns government intervention, it is focused on administrative outputs and outcomes and it is called to play a role in future practical action. In Sweden, where the interest in evaluation grew as a result of the wish to expand central government as a instrument to achieve special goals, evaluators should bear in mind that every program is a nexus in a set of political and social relationships and there is a great number of stakeholders the evaluator may need to consider, among which: policymakers and decisionmakers, program sponsors and target participants.

3 What does it mean *to evaluate*? Procedure

Evaluation operates within multiple domains and serves a variety of functions at the same time. Moreover it is subject to budget, time and data constraints that may force the evaluator to sacrifice many of the basic principles of impact evaluation design.

Before entering into the details of evaluation methods it is important for the reader to have a clear picture of the way an evaluation procedure works. The phases of the evaluation process are schematized in Figure 3.1.

Figure 3.1 Evaluation process



It is obvious that all the possibilities presented for the four steps of the evaluation procedures can be combined in different ways leading, therefore to a vast number of possible evaluation techniques.

We start from the point of view of ITPS and we try to keep in mind that the present work is focused on program's evaluation and that ITPS is most of the time called to do program's evaluation and design program alternatives at three different points in time: ex-ante, mid-term and ex-post evaluation. In this discussion of evaluation we will be focusing on how we can bring questions of value (rather than only numerical worth) back into the centre of the process.

This report explores some important dimensions of this process; the theories involved; the significance of viewing ourselves as action researchers; and some issues and possibilities around evaluation at ITPS.

First, it is helpful to define program and project evaluation from ITPS perspective, as opposite to practice evaluation. Program and project evaluation is typically concerned with making judgements about the effectiveness, efficiency and sustainability of pieces of work, it is a management tool to provide feedback so that future work can be improved or altered. While practice evaluation is directed at the enhancement of work undertaken with particular individuals and groups, and to the development of participants, and it tends to be an integral part of the working process.

The matter of program objectives and purposes is quite broad and many can be considered to be intermediate objective of an evaluation program, such as: measure and account for the results of public policies, determine the efficiency of program and projects, gain explanatory insights in problems, understand how organizations learn, strengthen institutions and improve performance, reform government, expand efficiency measurements and results, ensuring that there is justification for a policy or program, and so on. However the program objectives provide guidance for achieving the program's purposes that can be classified in three main groups:

- Evaluation for development: aimed to improve institutional performance.
- Evaluation for accountability: aimed to provide information to decision makers.
- Evaluation for knowledge: aimed to generate understanding and explanation.

Notwithstanding the importance of such classification it is relevant to keep in mind that there is not clear cut between these purposes, they have multiple methodological interactions and unavoidable overlapping points. However, evaluation at ITPS tests, in the majority of the cases, the worth and value of ongoing or finished programs with improvement or accountability purposes.

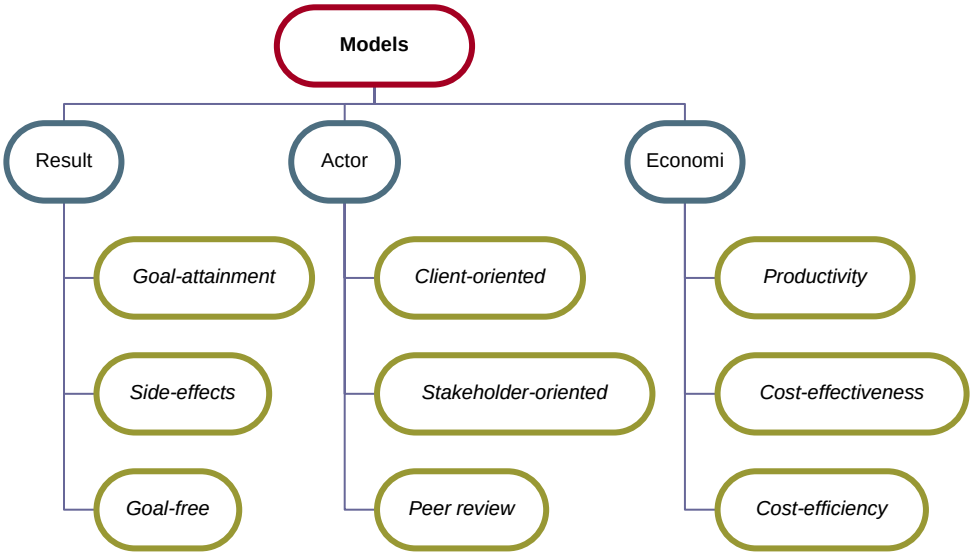
3.1 Models of Evaluation

The present paragraph summarizes the existing classification of evaluation models, which is broadly accepted in evaluation theory. It will, furthermore, highlight advantages and disadvantages and the possible use of different models by evaluators at ITPS according to their purposes.

The evaluation literature presents a large amount of alternative approaches to evaluation and in recent years the number of possible models has increased systematically. However, most authors agree on the basic classification of models in three categories: effectiveness models, economic models and professional models.

Following Vedung (1997) and Foss Hansen (2005) we can schematize the theoretical mainstream in the following way:

Figure 3.2 Evaluation Models



Source: Vedung (1997), "Public Policy and Program Evaluation", Transaction Publisher.

Each one of these has, obviously, different purpose and present advantages and disadvantages according to the object of evaluation.

Result models focus on the results of a given performance, program or organization and they inform on whether the goals have been realized or not and on all the possible effects of the program, both foreseen and unforeseen. There are at least two distinct methodologies, reflecting distinct methodological principles: goal-bound and goal-free procedures. Broadly speaking, goal-bound

evaluation[1] is focused on the relative degree to which a given product effectively meets a previously specified goal, while goal-free evaluation[2] measures the effectiveness of a given product exclusively in terms of its actual effects - the goals and motivations of the producer are ignored. Each approach has relative advantages and disadvantages. On the one hand, goal-bound evaluation is ordinarily more cost-effective than goal-free evaluation; on the other hand, measuring effectiveness entirely in terms of the degree to which stated goals are met can have at least two undesirable consequences: (a) since effectiveness is, on this model, inversely proportional to expectations, effectiveness can be raised simply by lowering expectations, and (b) deleterious or otherwise unwanted effects, if any, are left out of account, while unintended benefits, if any, go unnoticed.

Economic models, on the other hand, test whether program's productivity, effectiveness and utility have been satisfactory in terms of expenses. Cost analysis is currently a somewhat controversial set of methods in program evaluation. One reason for the controversy is that these terms cover a wide range of methods, but are often used interchangeably. Whatever position an evaluator takes in this controversy, it is good to have some understanding of the concepts involved, because the cost and effort involved in producing change is a concern in most impact evaluations (Rossi & Freeman, 1993).

- *Cost allocation* is a simpler concept than either cost-benefit analysis or cost-effectiveness analysis. At the program or agency level, it basically means setting up budgeting and accounting systems in a way that allows program managers to determine a unit cost or cost per unit of service. This information is primarily a management tool. However, if the units measured are also outcomes of interest to evaluators, cost allocation provides some of the basic information needed to conduct more ambitious cost analyses such as cost-benefit analysis or cost-effectiveness analysis;
- *Cost-effectiveness* and *cost-benefit* studies are often used to make broad policy decisions, the terms might be used interchangeably, but there are important differences between them: by definition, cost-effectiveness analysis is comparative, while cost-benefit analysis usually considers only one program at a time. Another important difference is that while cost-

[1] Tyler (1949) is the first one to propose goal-oriented and objectives-based models in order to describe whether students have met their goals or not, underlying how this model can be very useful in providing information about how to handle new strategies and reform old ones

[2] For an introduction to the distinction between goal-free and goal-bound evaluation methods as well as a thorough review of their respective strengths and weaknesses, see Michael Scriven, "Evaluation Perspectives and Procedures," in W. James Popham, ed., Evaluation in Education: Current Applications (Berkeley, CA: McCutchan Publishing Corporation, 1974).

> benefit analysis always compares the monetary costs and benefits of a program, cost-effectiveness studies often compare programs on the basis of some other common scale for measuring outcomes. The idea behind cost-benefit analysis is simple: if all inputs and outcomes of a proposed alternative can be reduced to a common unit of impact, they can be aggregated and compared. If people would be willing to pay money to have something, presumably it is a benefit; if they would pay to avoid it, it is cost. In practice, however, assigning monetary values to inputs and outcomes in social programs is rarely so simple, and it is not always appropriate to do so (Weimer & Vining, 1992; Thompson, 1980; Zeckhauser, 1975).

Economic models therefore can provide estimates of what a program's costs and benefits are likely to be, before it is implemented; they may improve understanding of program operation, and tell what levels of intervention are most cost-effective and they might reveal unexpected costs. But surely they are not free from drawbacks such as not being able to tell whether or not the program is having a significant net effect on the desired outcomes and whether the least expensive alternative is always the best alternative.

Finally *actors' models* are based upon the actors' own criteria for assessment. As the term suggests, they emphasize the central importance of the evaluation participants, especially clients and users of the program or technology. Client-centered and stakeholder approaches are examples of participant-oriented models, as are consumer-oriented evaluation systems.

With all of these strategies to choose from, how can an evaluator decide? Debates that rage within the evaluation profession are generally battles between these different strategists, with each claiming the superiority of their position; but most of the recent development in the debate have focused on the recognition that there is no inherent incompatibility between these broad strategies and each of them brings something valuable to the evaluation table, attention has therefore increasingly turned to how one might integrate results from evaluations that use different strategies, carried out from different perspectives, and using different methods. Clearly, there are no simple answers here. The problems are complex and the methodologies needed will and should be varied.

It is in the last couple of years that the core of the debate as moved towards the need of "defining evaluation model according to the purpose of evaluation, its object or the problem to be solved by the evaluated program" (Foss Hansen, 2005). Three are, so far, the existing school of thought. On one side authors that claim that the choice of models must be based on the purpose of the evaluation (formative evaluation and stakeholder models if the evaluation is

intended to create learning, summative and goal attainment model if it is planned to control performance); on the other side the advocates of the choice of different combination of evaluation models due to the characteristics of the object to be evaluated; and finally those that argue that evaluation design should be determined on the basis of an analysis of the problem that the object of evaluation is meant to solve.

3.2 Methods of Evaluation

Once and evaluation program has been designed according to its purpose, object or problem to be solved it is time to choose which method (or methods) to apply in order to estimate the worth and value of a program.

One of the major controversies in evaluation since its dawn and ongoing challenge that will determine the future of evaluation is the so-called *quantitative/qualitative debate*. Even thought some authors such as Greene, House and Newcomer suggest that the qualitative/quantitative debate is no longer an issue for evaluators, others, among which Smith and Worthen, claim that the debate still hold the interest of many evaluators. If House in 1994 stated that *"the debate between qualitative and quantitative methods is the most enduring schism in the field of evaluation…but it will recede in importance and mixed method studies will become the norm in the future"* Worthen in 2001 assert *"the qualitative/quantitative debate is a 'mutant' and is still very much alive in our midst"*.

This part of the work summarizes qualitative and quantitative evaluation methods but, given the impossibility to deeply present every method, methods will briefly be listed with their advantages and disadvantage. It is therefore a picture of the current state of the debate on qualitative and quantitative methods and it will shed light on the way to combine different analysis techniques in order to ensure the highest quality of evaluation under constraint.

In most sciences, as mentioned, the use of either qualitative or quantitative methods has become a matter of controversy and even ideology, with particular schools of thought within each discipline favouring one type of method and pouring scorn on the other. Advocates of quantitative methods argue that only by using such methods can the social sciences become truly scientific; advocates of qualititative methods argue that quantitative methods tend to obscure the reality of the social phenomena under study because they underestimate or neglect the non-measurable factors, which may be the most important. The modern tendency (and in reality the majority tendency throughout the history of social science) is to use eclectic approaches: quantitative methods might be used with a global qualitative frame and qualitative methods might be used to understand the meaning of the numbers produced by quantitative methods.

But even this practice has risen a debate over whether quantitative and qualitative research methods can be complementary: some researchers argue that combining the two approaches into a comparative research method is beneficial and helps to build a more complete picture of the social world, while other researchers believe that the epistemologies that underpin each of the approaches are so divergent that they cannot be reconciled within a research project.

Generally, qualitative research studies rely on three basic data gathering techniques: participant observation, interviews, and documents or artifact analysis (Wolcott, 1995, 1999). Some examples of qualitative methods:

- *Analytic induction* refers to a systematic examination of similarities between various social phenomena in order to develop concepts or ideas. Social scientists doing social research use analytic induction to search for those similarities in broad categories and then develop subcategories.
- *Focus groups* are a form of qualitative research in which a group of people are asked about their attitude towards a product, concept, advertisement, idea, or packaging. Questions are asked in an interactive group setting where participants are free to talk with other group members.
- *Ethnography* includes direct observation of daily behavior, conversations, genealogical methods, in-depth interviews and longitudinal research.
- *Participant observation* is a major research strategy which aims to gain a close and intimate familiarity with a given group of individuals and their practices through an intensive involvement with people in their natural environment. Such research usually involves a range of methods: informal interviews; direct observation; participation in the life of the group; collective discussions; analyses of the personal documents produced within the group; self-analysis; and life-histories. Thus, although the method is usually characterized as qualitative research, it often includes quantitative dimensions.

Other qualitative methods known in the literature are: *Semi-structured interviews, Unstructured interviews, Textual analysis* and *Theoretical sampling*.

Shortly some examples of quantitative methods:

- *Statistical surveys* are used to collect useful information in many fields. When the questioned are administrated by the researcher the survey is called *structured interview* (essentially aimed to ensure that each interviewee is presented with exactly the same questions and answers can be reliably aggregated) while when the questions are administered by the respondent

the survey is referred to as *questionnaire* or *self-administered survey*. Surveys present a number of advantages: they are efficient in collecting information from a large number of respondents; they are flexible and allow gathering a wide range of information; they are standardized and easy to administer; they re focused and do not spend time and money in tangential questions. However they are not free from disadvantages: they are not appropriate for studying complex social phenomena; they depend on subjects' motivation, honesty, memory and ability to respond; it is difficult to create random samples and respondents are usually self-selected.

- Content or textual analysis defined by Holsti (1969) as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages". The method of content analysis enables the researcher to include large amounts of textual information and identify systematically its properties by detecting the more important structures of its communication content; such amounts of textual information must be categorized according to a certain theoretical framework, which will inform the data analysis, providing at the end a meaningful reading of content under scrutiny.
- *Statistical descriptive techniques*, the most common include: graphical description (histograms, scatter-grams, bar chart, …); tabular description (frequency distribution, cross tabs, …); parametric description (mean, median, mode, standard deviation, skewness, kurtosis, …)
- *Statistical inferential techniques* which involve generalizing from a sample to the whole population and testing hypothesis. Hypothesis are stated in mathematical or statistical terms and tested through two or one-tailed tests (t-test, chi-square, Pearson correlation, F-statistic, …)

Notwithstanding the importance and utility of both quantitative and qualitative methods, the evaluators are most of time called to work in natural settings, where history and context matter and experiences are shaped by relationships and institutions and where the complexity can be studied only with a marshalling of all the ways of understanding. The practice of studying an issue using several different methods, as if you're seeing it from different angles, is called *triangulation* and has ruled over the quantitative/qualitative debate in the last years.

The advocates of triangulation support the idea that, though different methods come up with different results, the results should be similar enough that they might be plotted on a graph as a small triangle. Somewhere inside that triangle is the "real truth". Triangulation offers, therefore the prospect of enhanced confidence, because as Webb et al. (1966) suggested *"once a proposition has*

been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced".

There are different types of triangulation, among which: data triangulation (involving time, space, and persons); investigator triangulation (which consist of the use of multiple observers); theory triangulation (which consists of using more than one theoretical scheme in the interpretation of the phenomenon); methodological triangulation (which involves using more than one method and may consist of within-method or between-method strategies); and multiple triangulation (when the researcher combines in one investigation multiple observers, theoretical perspectives, sources of data, and methodologies).

The main advantages of evaluation are, as mentioned, related to the improved reliability of the results and in particular: it can be employed in both quantitative (validation) and qualitative (inquiry) studies; it is a method-appropriate strategy of founding the credibility of qualitative analyses; and it is the preferred line in the social sciences because, by combining multiple observers, theories, methods, and empirical materials, researchers can hope to overcome the weakness, intrinsic biases or problems coming from single method, single-observer, single-theory studies.

Nonetheless, the idea of triangulation has been criticized on several grounds. First, it is sometimes accused of subscribing to a naive realism that implies that there can be a single definitive account of the social world, and this have come under attack from theorist of constructivism, who argue that research findings should be seen as just one among many possible renditions of social life. On the other hand, writers working within a constructionist framework do not deny the potential of triangulation; instead, they underline the utility of triangulation in terms of adding richness and complexity to an inquiry. A second criticism is that triangulation assumes that sets of data deriving from different research methods can be unambiguously compared and regarded as equivalent in terms of their capacity to address a research question, and this does not take into account the different social circumstances associated with the administration of different research methods.

Triangulation and mixed-methods evaluation is therefore the new frontier of evaluators' work because it offers much for increased understanding of programs. However the evaluator is called to act with increased reflexivity and responsiveness, more openness to diversity, acceptance of differences and tolerance of diversity to skirt the eventuality of doing triangulation merely choosing from a smorgasbord of available methods.

4 What is the role of evaluators? Ethics

Despite using similar methods, professional evaluators are not the same, they come from many different backgrounds, they have different aspirations, perceptions and values that might influence the interpretation they make, and all these characteristics must be kept in mind when considering the results of an evaluation. This last part of the analysis investigates to which extent conflict or confluence of interest can be considered a problem and what the standards say about advocacy and constraints to evaluation, bearing in mind that evaluations are powerful and influential instruments in democracies.

As already mentioned, the main purpose of evaluation is social betterment and improvement. However, an evaluation is most of all required to be influential in the sense of setting off changes of various kind, such as persuading others, justifying policies, changing attitudes or behaviors, placing an item in the public agenda or substantiating public expenditure. And the role of evaluator in improving the influence of evaluation is essential.

One of the hot topics of today's debate is how the power that evaluation has in leading to changes in behavior can be affected by sponsorship and advocacy. Most evaluators claim to make unemotional searches for quality and speak scornfully of advocacy and promotion, yet it is clear that evaluators, as human beings, might feel strongly for certain matters and promote them in their work.

In 1994 the *Joint Committee of Standards* made a first attempt to verge on ethics for evaluators with a list of ethical guidelines including the call for frank and full disclosure and for balanced and objective reporting, with strong implication to the fact that evaluators should stick to the job of finding merit and worth of a program. Shortly after, in 1995, the *American Evaluation Society* has somehow tried to state some *Guiding principles* (www.eval.org) to solve the increasing debate on conflict and confluence of interests. According to these *Guiding principles "the evaluator should inform a client if there is reason to believe he or she might be object to a particular value commitment"*; moreover *"Evaluators should explore with the client the shortcomings and strengths of the various evaluation questions, it might be productive to discuss the various approaches that might be used in answering these questions"* and *"evaluators should seek to determine, and where appropriate to be explicit about, their own, their clients' and other stakeholders' interests concerning the conduct and outcomes of an evaluation"*. Since then the debate on ethical principles has not gone much further and most of the guiding principles are still suggestions and acknowledgements rather than constraints or support.

The situation nowadays remains subject of debate: evaluators can not help but see some things differently, evaluation practice is still, and probably will always be, influenced by the value commitments of the evaluator and the challenge nowadays is that of developing standards and commitments in order to deal with the uncertainty and individuality of evaluating, to do so *"the full use of validation, triangulation and meta-evaluation is essential, but it will not eliminate uncertainty in the evaluation findings"* (Stake, 2004).

Moreover, beside an evaluator's own values and commitments there other aspects that constraint evaluation practices: budget, time and data constraints. Evaluators are therefore called to face not only their own limits in making influential evaluation but external limitations as well which might compromise many of the basic principles of sound evaluation. It is therefore an evaluator's job to ensure the highest quality evaluation under constraints of limited budget, time or data availability and to inform the client of the drawbacks due to such limits.

Evaluators, including those working at ITPS must therefore be aware of the limits due to method's choice, time and budget constraints and the influence of different stakeholder points of view. Each analysis and report must clearly state purposes and issues, and be evaluators should always motivate the choice of a given method over others in relation to the evaluation's aim. Notwithstanding the personal experiences, interests and abilities of the evaluators, ITPS reports and analysis should clearly inform clients and stakeholder of possible bias and constraints that could affect the quality of the evaluation, choice of models and methods must be justified from the beginning and the evaluator's point of view should be clearly stated as such.

Conclusions

The present report illustrates the most discussed topics in the current literature on evaluation. Starting from a panoramic of the historical evolution of the concept of evaluation, its philosophical roots and the reasons why it has different shades according to different point of views, it moves to analysis of existing definitions, the large number of actors involved in the evaluation process, the different aims and objectives, different models and methods applied, diverse priorities and so on.

In the light of these definitions, the report stress how the practice of evaluation at ITPS is very much coherent with Vedung's idea of evaluation as a *"tool to determine the worth and value of public programs, with the purpose of providing information to decision-makers and improve institutional performance in the spirit of looking backwards to improve forward directions"* (Vedung, 1997).

Moreover it becomes clearer and clearer while reading this paper that in the last decades much progress has been made towards clarifying the essential and core nature of evaluation, which has become itself a discipline.

The mainstay of this work focuses on practical aspects and tries to answer the question *"what does it mean to evaluate?"*, bearing in mind the ITPS perspective of evaluation as a tool to look and study backward events to improve forward directions, but the decisive question remains *"what does it mean to evaluate in a Swedish governmental agency such as ITPS?"*. As mentioned, different perceptions, values and background of the evaluator can have different effects on the results of evaluation: it is natural that evaluators, as human beings, might feel strongly for certain issues and promote them in their work, and the evaluator's job of ensuring the highest quality evaluation might be under constraints of limited budget, time or data availability.

Some question, however, still remain unanswered. What should be done to correspond to what is actually done at ITPS? Shall ITPS engage in judgement of program's implementation and results or shall evaluators limit themselves to simple measurement? Which evaluation methods can be better applied to public programs?

The report concludes that evaluation, from ITPS point of view, can not ignore the limits due to method's choice, time and budget constraints and the influence of different stakeholder points of view. It stresses the importance of well-defined purposes and issues, and the need of transparency in motivating the

choice of a given method over others in relation to the evaluation's aim. In spite of the personal knowledge, interests and aptitudes of the evaluators, ITPS reports and analysis should clearly notify clients and stakeholder of possible bias and restraints that might impinge on evaluation's quality, choice of models and methods must be justified from the beginning and the evaluator's point of view should be clearly stated as such.

References

Bamberger, Rugh, Church and Fort (2004). Shoestring evaluation: designing impact evaluation under budget, time and data constraints, *American Journal of Evaluation*, Vol.25(1), 5-37.

Bengston and Fan (1999). An innovative method for evaluating strategic goals in a public agency: Conservation leadership. *Evaluation Review*, 23(1), 77-10.

Berg (1989). *Qualitative research methods for the social sciences*. Allyn and Bacon.

Bozzo (2000). Evaluation resources for non-profit organizations. *Non-profit Management and Leadership*, 10(4), 463-472.

Caracelli (2000). Evaluation *use at the threshold of the twenty-first century,* in V. J. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use* (pp. 99-111). *New Directions for Evaluation*, no. 88. San Francisco: Jossey-Bass.

Chelimsky (1997). *The Coming Transformations in Evaluation*, in Chelimsky & Shadish (eds.) *Evaluation for the 21st century. A Handbook*. Thousand Oaks: Sage Publications.

Cobb-Clark and Crossley (2003). Econometrics for evaluations: an introduction to recent developments, *The Economic Record*, Vol.78 (247), 491-511.

Creswell (1998). *Qualitative inquiry and research design: choosing among five traditions*. Thousand Oaks, Sage Publications.

Cronbach (1980). *Toward reform of program evaluation: aims, methods and institutional arrangements*. San Francisco, CA, Jossey-Bass.

Denzin (1970). *The research act in sociology*. Chicago, Aldine.

Denzin and Lincoln (2000). *Handbook for qualitative research*. Thousand Oaks, Sage Publications.

Feller and Ruegg (2003). *A Toolkit for Evaluating Public R&D Investment Models, Methods, and Findings from ATP's First Decade* (www.atp.nist.gov/eao/gcr03-857/contents.htm).

Fetterman (2001). *Foundations of empowerment evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Fine, Thayer, and Coghlan, A. T. (2000). Program evaluation practice in the non-profit sector. *Non-profit Management and Leadership*, 10(3), 331-339.

Foss Hansen (2005). Choosing evaluation models. A discussion on evaluation design, *Evaluation*, Vol 11(4), 447-462.

Greene, Benjamin and Goodyear (2001). The merit of mixing methods in evaluation, *Evaluation*, Vol.7(1), 25-44.

Henkel (1976). *Tests of significance*. Series: quantitative applications in the social sciences, Sage University Paper.

Henry (2003). Influential evaluations, *American Journal of Evaluation*, Vol.24(4), 515-524.

Henry, Julnes and Mark (Eds.). (1997). Realist evaluation: An emerging theory in support of practice. *New Directions for Evaluation*, no. 78. San Francisco: Jossey-Bass

Hoefer (2000). Accountability in action? Program evaluation in non-profit human service agencies. *Non-profit Management and Leadership*, 11(2), 167-177.

Holsti (1969). *Content analysis for the social sciences and humanities*, Reading, MA, Addison-Wesley Publishing Company.

House (1989). *Assumptions Underlying Evaluation Models*. In Madaus G.F., 1989 (Eds.), *The courts, validity, and minimum competency testing*, Boston, Kluwer-Nijhoff Publishing.

House (1994). The future perfect of evaluation, *Evaluation Practice*, Vol. 15, 239-247.

House (2001). Unfinished business: causes and values, *American Journal of Evaluation*, Vol. 22, 309 - 315.

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards: how to assess evaluations of educational programs?*. Newbury Park, CA, Sage Publications.

Kaplan (2001). Strategic performance measurement and management in non-profit organizations. *Non-profit Management and Leadership*, 11(3), 353-370.

Kuhn (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Leviton(2003). Evaluation use: advances, challenges and applications, *American Journal of Evaluation,* Vol.24(4), 525-535.

Lincoln and Guba (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.

Lindenberg (2001). Are we at the cutting edge or the blunt edge? Improving NGO organizational performance with private and public sector strategic management frameworks. *Non-profit Management and Leadership*, 11(3), 247-270.

Love (2001). The future of evaluation: Catching rocks with cauldrons. *American Journal of Evaluation*, 22(3), 437-444.

Mark and Henry (2004). The mechanisms and outcome of evaluation influence, *Evaluation*, Vol.10(1), 35-57.

Mark, Henry and Julnes (2000). *Evaluation: An integrated framework for understanding, guiding, and improving public and non-profit policies and programs.* San Francisco: Jossey-Bass.

Mohr (1999). The qualitative method of impact analysis. *American Journal of Evaluation*, 20(1), 69-84.

Morris (2001). Ethical challenges, *American Journal of Evaluation,* Vol.22, 105 - 106.

Newcomer (2001). Tracking and probing program performance: fruitful path or blind alley for evaluation professionals?, *American Journal of Evaluation*, Vol.22, 337 - 341.

Patton (1982). *Practical Evaluation*. Newbury Park: Sage Publications.

Patton (1997). *Utilization-focused evaluation: The new century text (3rd ed.)*. Thousand Oaks, CA: Sage Publications.

Poole, Nelson, Carnahan, Chepenik and Tubiak (2000). Evaluating performance measurement systems in non-profit agencies: The program accountability quality scale (PAQS). *American Journal of Evaluation*, 21(1), 15-26.

Pratt, McGuigan and Katzev (2000). Measuring program outcomes: Retrospective pretest methodology. *American Journal of Evaluation*, 21(3), 341-349.

Renz (2001). Changing the face of nonprofit management. *Nonprofit Management and Leadership*, 11(3), 387-396.

Rist (1990). *Program evaluation and the management of government*. Transaction Publishers.

Rossi and Freeman (1993). *Evaluation. A Systematic Approach*. Newbury Park: Sage Publications.

Rossi, Lipsey and Freeman (2004). *Evaluation: a systematic approach*. Thousand Oaks, Sage Publications.

Ryan (1998). Advantages and challenges of using inclusive evaluation approaches in evaluation practice. *American Journal of Evaluation*, 19(1), 101-122.

Sawhill and Williamson (2001). Mission impossible? Measuring success in nonprofit organizations. *Nonprofit Management and Leadership*, 11(3), 371-386.

Scriven (1972). Pros and Cons about Goal-free Evaluation, *The Journal of Educational Evaluation*, Vol. 4, 1–7.

Scriven (1991). *Evaluation thesaurus*. Beverly Hills, CA, SAGE Publications.

Shadish, Cook and Leviton (1991).*Foundations of Program Evaluation: Theories and Practice*. Newbury Park, SAGE Publications.

Shadish, Newman, Scheirer and Wye (1995). *Guiding principles for evaluators*, New Directions for Program Evaluation, N.66, San Francisco, CA, Jossey-Bass.

Smith (2001). Evaluation: Preview of the future #2. *American Journal of Evaluation*, 22(3), 281-300.

Smith (2001). Evaluation: preview of the future 2, *American Journal of Evaluation*, Vol.22(3), 281-300.

Sonnichsen (2000). *High impact internal evaluation*. Thousand Oaks, sage Publications.

Stake (1973). *Program evaluation, particularly responsive evaluation*. Keynote address at the conference "New trends in evaluation," Institute of Education, University of Goteborg, Sweden, in G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff, 1987.

Stake (2004). How far dare an evaluator go towards saving the worlds´?, *American Journal of Evaluation*, Vol.25(1), 103-107.

Stufflebeam (2000). *Checklists development checklist* (www.wmich.edu/evalctr/checklists/)

Stufflebeam (2001). The metaevaluation imperative, *American Journal of Evaluation*, Vol.22, 183-209.

Thompson (1980). *Benefit-cost analysis for program evaluation.* Beverly Hills, Sage Publications.

Torres and Preskill (2001). Evaluation and organizational learning: Past, present, and future. *American Journal of Evaluation*, 22 (3), 387-395.

Tyler (1949). *Basic Principles of Curriculum and Instruction*. Chicago, MA, University Chicago Press.

Van der Knaap (1995). Policy Evaluation and Learning. Feedback, Enlightenment or Augmentation?, *Evaluation,* Vol. 1 (2), 189-216.

Vedung (1997). *Public policy and program evaluation*, London, Transaction Publishers.

Webb, Campbell, Schwartz and Sechrest (1966). *Unobstrusive measures: nonreactive measures in the social sciences*. Chicago, Rand McNally.

Weimer and Vining (1992). *Policy Analysis: Concepts and Practice*. Englewood Cliffs, New Jersey, Prentice Hall.

Wholey (2001). Managing for results: Roles for evaluators in a new management era. *American Journal of Evaluation*, 22(3), 343-347.

Wolcott (1995). *The art of fieldwork*. Walnut Creek, CA, Altamira Press.

Wolcott (1999). *Ethnography: a way of seeing*. Walnut Creek, CA, Altamira Press.

Worthen (2001). Whither evaluation? That all depends, *American Journal of Evaluation*, Vohl.22, 409-418.

Zeckhanser (1975). Procedures for valuing lives, *Public Policy*, Vol.23 (4), 419-464.