IDPM - UA discussion paper

Dealing with Stakeholder Values in the Evaluation of Development Programs

A Methodological Framework for Mid Term Evaluation

Jos Vaessen

2005 - 1

UNIVERSITY OF ANTWERP

INSTITUTE OF DEVELOPMENT POLICY AND MANAGEMENT

Comments on this Discussion Paper are invited.
Please contact the author at <jos.vaessen@ua.ac.be>

Instituut voor Ontwikkelingsbeleid en -Beheer
Institute of Development Policy and Management
Institut de Politique et de Gestion du Développement
Instituto de Política y Gestión del Desarrollo

Venusstraat 35, B-2000 Antwerpen
België - Belgium - Belgique - Bélgica

Tel: +32 (0)3 220 49 98
Fax: +32 (0)3 220 44 81
e-mail: dev@ua.ac.be

http://www.ua.ac.be/dev

Dealing with Stakeholder Values in the Evaluation of Development Programs

A Methodological Framework for Mid Term Evaluation

Based on a paper presented at the 6th European Evaluation Society Conference: Governance, Democracy and Evaluation Berlin, September 30 - October 2, 2004

Jos Vaessen[1]

Institute of Development Policy and Management
University of Antwerp

January 2005

[1]Jos Vaessen is a research and teaching assistant at the Institute of Development Policy and Management at the University of Antwerp.

Contents

Abstract

In mid term program evaluations evaluators are often confronted with the double task of retrospectively judging the program's merit and worth while at the same time advising decision makers concerning future adjustments in courses of action. In such cases, it can be argued that it is particularly important that evaluators take into account the divergent views and needs of different stakeholder groups. In principle, program theory evaluation can constitute a sound basis for dealing with the double objective of retrospective judgment and proactive program improvement. However, as argued in the paper, current approaches in program theory evaluation may not be sufficiently equipped to systematically deal with divergent stakeholder values. Taking into account lessons from the literature on stakeholder values in evaluation, an alternative methodological framework is presented. The framework combines program theory evaluation with elements of multicriteria decision aid. An example is used to illustrate the framework.

Résumé

.Pour les évaluations à mi-parcours les chercheurs sont souvent confrontés à la double tâche d'évaluer rétrospectivement le bien fondé et la valeur d'un programme et en même temps de conseiller les décideurs en ce qui concerne les ajustements futurs dans le cours même de leur mise en acte. Dans de tels cas, on peut avancer qu'il est important que les évaluateurs tiennent compte des points de vue divergents et des besoins des différents groupes qui détiennent les enjeux dans le domaine concerné. En principe, l'approche d'évaluation basée sur les « théories du programme » peut constituer une bonne base pour appliquer le double objectif de jugement rétrospectif et d'amélioration pro-active de ce programme. Pourtant, il est affirmé dans l'article que les méthodes courantes concernant cette approche pourraient ne pas être suffisamment outillées pour traiter systématiquement les valeurs divergentes des « stakeholders ». Tenant compte des leçons qu'on peut retirer de la littérature sur les valeurs des « stakeholders » en ce qui concerne l'évaluation, un cadre méthodologique alternatif est proposé. Ce cadre combine l'approche d'évaluation basée sur les « théories du programme » avec des éléments de l'aide multicritère à la décision. Un exemple est présenté pour illustrer le cadre méthodologique.

1 Introduction[1]

Over the past decades, evaluators of social programs have developed a number of evaluation approaches which start out from some kind of 'theory' of how a program works or should work. Nowadays, for many in the evaluation community, as phrased by Pawson and Tilley (1997: 56-57), "the careful enunciation of program theory is [a] prerequisite to sound evaluation". In this sense, evaluators have used terms like policy theories (e.g. Majone, 1980), program theories (e.g. Bickman, 1987), intervention theories (e.g.Vedung, 1997) or theories of change (e.g. Weiss, 1997). The common element that unites these 'theory-oriented' approaches (Stame, 2004)[2] is the reconstruction of a causal model (the program theory) on the basis of different sources of information in order to arrive at an understanding of how programs intend to bring about a number of intended and unintended outcomes. Program theory evaluation (PTE) as used in this paper refers to this process of reconstruction of the theory as well as an assessment of the validity of the reconstructed theory (vis-à-vis multiple benchmarks).

In a recent article, Leeuw (2003) purports that, notwithstanding the renewed attention in the literature for program theory in evaluation, there continues to be a lack of systematic methods for reconstructing program theories. An important element that is not sufficiently addressed in most of the methodological discussions on program theory reconstruction and evaluation is the question of how to deal with multiple stakeholder perspectives on what a program is (or should be) about. While this question has received little attention in the literature on PTE, elsewhere the significance of the topic is reflected by the myriad of approaches in evaluation dealing with the issue of stakeholder participation (e.g. Cousins and Whitmore, 1998; Stufflebeam, 2001). The importance of incorporating the diversity in stakeholder values in the design and implementation of evaluation studies (i.e. an element of stakeholder participation) is also explicitly expressed in the guiding principles for evaluators (Shadish et al., 1995).

The arguments in favor of stakeholder participation in evaluation are the following (see Greene, 2000; Cousins and Whitmore, 1998). First of all, stakeholder participation can improve stakeholder ownership and hence the relevance and utilization of evaluation findings. Second, the fact that stakeholders are actively involved in the design and implementation of evaluations can have an empowering effect. Indeed, some evaluation approaches (e.g. empowerment evaluation (Fetterman, 1994)) are specifically geared to serve this purpose. Finally, one finds the argument that evaluations of public programs or policies in a democratic society should be based on democratic values and therefore should include the viewpoints of all major stakeholder groups in a fair and transparent manner.

In view of the above, a number of questions come to mind. First of all, how can evaluators operating from the perspective of PTE deal with multiple

[1] The author would like to thank Frans Leeuw and Robrecht Renard for their comments on the paper and Johan Springael for the interesting discussions on multicriteria techniques. In addition, the author would like to thank participants at the 6th EES Conference "Governance, Democracy and Evaluation" for their useful feedback. Any remaining errors are the responsibility of the author.

[2] The three most important evaluation approaches that fall under the banner of theory-oriented evaluation are: theory-driven evaluation (Chen and Rossi), theory-based evaluation (Weiss) and realistic evaluation (Pawson and Tilley).

stakeholder values? A number of sub questions can be helpful in guiding the discussion on this principal question:

- How can PTE benefit from other evaluation approaches specifically dealing with stakeholder participation?
- Which stakeholders are to participate in PTE?
- In what phase of the PTE process should they participate (e.g. the reconstruction of the program theory, the assessment of the validity of the theory)?
- In what manner should they participate in PTE (e.g. consultation, joint deliberation)?

These questions constitute the main focus of this paper. However, the answers to these questions are not unequivocal but depend on the specific context and objectives of the program evaluation study. Therefore, rather than attempting to provide a comprehensive treatment of these questions under all circumstances (a rather daunting task), in this paper we focus on a specific type of evaluation study. The type of study we focus on falls in the category of what Stufflebeam (2001) calls decision/accountability-oriented studies. This type of study combines the objective of retrospective judgment of merit and worth of a program with the objective of proactively improving the program strategy. Many midterm program evaluations can be classified under this type of evaluation study. PTE is especially useful in midterm evaluations since it helps stipulating the consecutive assumptions underlying a program and consequently provides a structure to judge the validity of the program using multiple sources of information. These information sources include among other things tentative expert judgments or research evidence from elsewhere in order to provide arguments for or against a certain assumption. The use of this type of preliminary or more tentative evidence is very important in midterm evaluations where outcome and impact effects are not yet manifest and therefore difficult if not impossible to measure. Judging the 'solidity' of the theory underlying a program is fundamental to redefining and improving the program strategy. Evidently, in such moments of evaluation the legitimate question arises how the different stakeholders' priorities could be taken into account, especially in the redefinition and adjustment of the program strategy towards the future.

In order to provide a methodological/procedural response to these issues, an alternative methodological framework is presented. From the above it follows that this paper does not simply seek a methodological framework that integrates stakeholder participation into PTE. Stakeholder participation in evaluation, though important for several reasons, does not necessarily lead to the most correct or complete assessment of a program. Instead, on the face of it, the contrary might be the case. Moreover, stakeholder participation itself is neither straightforward nor unproblematic (Cooke and Kothari, 2001). Therefore, our methodological quest would necessarily focus on a combination of participatory evaluation and the need for analytical and scientific rigor.

The methodological framework presented in this paper combines elements of existing approaches to PTE and multicriteria decision aid (MCDA). The core of MCDA constitutes the evaluation of different alternatives (e.g. program strategy scenarios) based on a number of criteria, each receiving a specific weight in the evaluation process (Belton, 1990). MCDA concerns a broad range of techniques with a wide field of application in both public and corporate environments, mostly in contexts of ex ante appraisal of policies, programs, processes or products (see Belton and Stewart (2002) for an overview of different schools in MCDA). In this paper we will discuss the rationale and illustrate the potential of the proposed framework for dealing with the double objective of retrospective judgment of merit and worth and proactive program strategy refinement while systematically taking into account divergent stakeholder perspectives on the program.

The structure of the paper is the following. In the next section we will briefly review some of the principal elements in the general debate on stakeholder values in evaluation. Subsequently, we will take up the issue of stakeholder values within the realm of PTE. A number of PTE approaches will be compared regarding the treatment of stakeholder values. It is concluded that in evaluation contexts that fit Stufflebeam's (2001) category of decision/accountability-oriented studies, current PTE approaches may not be sufficiently equipped to deal with the double objective of retrospective assessment of merit and worth and proactive program improvement, in such a way that multiple stakeholder perspectives are systematically included in the analysis. To remedy this situation, MCDA is introduced in the fourth section as a potential enrichment of existing PTE approaches. Section five illustrates the proposed methodological framework by means of an example. The last section summarizes and suggests some elements for further debate.

2 Stakeholder Values and Program Evaluation

In a comprehensive review of participatory approaches in evaluation, Cousins and Whitmore (1998) distinguish between transformative participation and practical participation. In the latter approach the main principle underlying participation is that it enhances stakeholder ownership and therefore the relevance of the study and utilization of the findings. In contrast, transformative participatory evaluation aims at strengthening the capacities of stakeholder groups by enhancing their control over the programs that affect them. Cousins and Whitmore (1998) compare these (and other approaches) on the basis of three dimensions: stakeholder selection, control of the evaluation process, and depth of participation. While differing in the first two dimensions, both approaches are quite similar with respect to the dimension of depth of participation. Both can be characterized by a rather intensive form of participation (i.e. dialogue and joint deliberation among evaluators and stakeholders as opposed to 'mere' consultation) as well as a comprehensive form (i.e. in the sense of including stakeholder participation in all phases of the evaluation study).

A recent approach called deliberative democratic evaluation (House and Howe, 1999) is rather similar to the participatory approaches just mentioned in the sense that it presupposes an intensive interaction between evaluators and different stakeholder groups. In contrast to the other two types of approaches, however, the elicitation of stakeholder values regarding a program rather than the involvement of stakeholders in issues of design and implementation of program evaluations is the central element of this approach. House and Howe (1999) argue that the elicitation of the interests of different stakeholders is not an easy and straightforward task. Stakeholders do not have a clear viewpoint about a given program but need to be involved in a process of extensive dialogue and deliberation in order to be able to clarify and express their views.

The authors advocate a three-tiered approach for program evaluation: inclusion of all major stakeholder groups, extensive dialogue, and finally a phase of deliberation in order to arrive at a judgment of a program's merit and worth. With respect to the third element, the authors claim that the value claims held by different stakeholder groups can and should be subjected to rational analysis in a similar way as one would deliberate about factual claims. They argue that although facts and values are two different things, the differences are much narrower than thought. Many claims encompass both factual and value elements. Therefore, under the condition that all major stakeholder groups' perspectives are taken into account and sufficient efforts have been undertaken to clarify their views, the evaluator (in a continuous reflection with stakeholders) is in the position to arrive at a judgment about the merit of the different (mixed value and factual) claims and ultimately arrive at a balanced judgment on the program. This implies a judgment about which group was 'correct' about a certain aspect of a program.[3]

[3] See also House (1995) for a thorough discussion on the synthesis of different stakeholder group perspectives in order to arrive at a final judgment.

House and Howe acknowledge that their model for evaluation in a democratic society is rather ideal typical. More specifically, there are a number of constraints to a successful implementation of this approach in practice.[4] First of all, there is the difficulty and high cost of physically incorporating all major stakeholder groups in a process of dialogue and deliberation. Second, even if one might be able to organize such processes (e.g. in smaller programs), power imbalances and other features of group dynamics might crowd out the viewpoints of less powerful and underrepresented groups.

[4] The same aspects apply (to a certain extent) to the two above-mentioned approaches of practical and transformative participatory evaluation.

One can also question the rationale behind a single balanced judgment as the outcome of a deliberative democratic evaluation process. House (1995) asks himself the question, "[s]hould one construct multiple syntheses based on each group's perceptions, values, and interests?" (ibid.: 43). He then argues that, "evaluators should try to determine which group was correct about the services, and that should be the basis for the synthesis" (ibid.: 43). In other words, after deliberation on the different factual and value claims put forward by different stakeholders, the evaluator arrives at a final verdict about the program. However, having said the foregoing he contends that, "[i]f the evaluator could not make this determination [regarding which group was correct], then multiple syntheses might be appropriate" (ibid.: 43).

The latter statement which House considers a 'contingency solution' is in fact in line with the viewpoint held by other authors like Shadish, Cook and Leviton (1991). They argue that in her final judgment, the evaluator should include different value positions and illustrate how different value positions would affect evaluative conclusions. "One could construct several value summaries, each of the form 'If X is important to you, then evaluand Y is good for the following reasons,' where X is drawn from the interests of different stakeholders or from prescriptive theories" (Shadish et al., 1991: 101).

This kind of reasoning is in many aspects more compatible with the program reality. In practice, program decision makers are often reluctant to share their decision-making power with other stakeholders. The creation of an 'artificial' decision-making forum with multiple stakeholders to decide on a program's worth, risks ending up being disarticulated from the decision makers' real concerns. As an alternative, if evaluators contrast the evaluative conclusions drawn from the point of view of decision makers' values with the conclusions drawn on the basis of other stakeholders' values, this would probably have quite an illuminating effect on decision makers. In practice, they often do not have at their disposal the information concerning how different value positions (e.g. of program participants) would affect evaluative outcomes.

Recently, an approach called values inquiry (Mark et al., 1999) has taken up the task to fill this information gap. "Values inquiry refers to a variety of methods that can be applied to the systematic assessment of the value positions surrounding the existence, activities, and outcomes of a social pol-

icy and program" (Mark et al., 1999: 183). Values inquiry is in fact a rather generic approach to incorporate stakeholder values into the evaluation process. In contrast to the deliberative democratic evaluation approach which also focuses on the elicitation of stakeholder values, the depth of participation is in general much lower. In principle it is a method-free type of inquiry. Henry (2002) for example used surveys to elicit the values of four different stakeholder groups in the evaluation of a preschool program. Alternatively, focus groups or other more qualitative techniques of inquiry can be used to capture stakeholder values.

A central element of values inquiry is to determine which criteria (according to different stakeholder groups) should be used to determine a program's worth and merit.[5] A core premise hereby is that "the choice of criteria for program success should be justified by the process used to obtain them" (Henry, 2002: 183). In fact, one of the basic principles of values inquiry, i.e. the elicitation of criteria and the relative importance of criteria that reflect a stakeholder's perspective on what is important in a program, is also one of the central elements of multicriteria decision aid (MCDA). Indeed, some of the interview and survey techniques underlying different MCDA techniques could be applied as part of a values inquiry exercise. In values inquiry, the mapped value positions are important findings that can be presented to different evaluation audiences (Mark et al., 1999). MCDA is a particular technique to take the analysis a step further by showing in a systematic manner how different value positions would affect evaluative conclusions and subsequent policy choices.

Before we take a closer look at MCDA, let us first recapitulate the discussion on stakeholder values within PTE.

[5] In this sense, values inquiry can be perceived as being an elaboration of the earlier approach of stakeholder-based evaluation (Bryk, 1983).

3 Stakeholder Values and Program Theory Evaluation

Program theories are not neutral models of the programs they represent. Their construction is the product of a complex social reality, partly shaped by conscious choices about which methodology to follow and who to involve in the process. The evaluator should be aware that her reconstruction of the program is but one possible reconstruction among many (see Dahler-Larsen, 2001). There is nothing wrong about this realization. However, the final reconstruction of the program should be defendable from the point of view of the objective of the program theory evaluation and the methodological choices made in the process. In this sense, the final reconstruction is more than just another construction which according to social constructivists in principle would be 'equal' to any other (see Guba and Lincoln, 1989).

The incorporation of stakeholder values in PTE is only one of the elements which affect the 'neutrality' of program theory reconstruction and subsequent evaluation.[6] Nevertheless, the manner in which they are dealt with in the evaluation process is especially relevant in the current debate and practice of public program evaluation. By exposing the methodological choices the evaluator makes with regard to dealing with different stakeholder values, she reinforces her position and the scope for making subsequent judgments about the program's merit and worth.

[6] See Dahler-Larsen (2001) for a wider discussion on constructivism and program theory evaluation.

We will briefly discuss a number of aspects regarding the incorporation of stakeholder values in PTE. With the risk of some oversimplification, table 1 compares a number of PTE approaches on the basis of six aspects. We have chosen five different approaches from recent publications in the American Journal of Evaluation. The list is far from exhaustive but probably represents a significant part of the variety in methodological approaches in PTE that are used today.

In PTE much of the discussion on values has focused on the question of whose assumptions make up the program theory (Chen, 1990; Weiss, 1998). This discussion has largely revolved on the relative roles of the evaluator and program staff in determining the program theory. Who is/are the principal determinant(s) of the program theory? Some authors (e.g. Wholey, 1987) have argued for a focus on program staff and other key stakeholders (e.g. policy makers, interest groups) as principal determinants of the program theory on the grounds that they are the major actors who continuously shape the implementation and outcome of the program. Others have emphasized the role of the evaluator backed by her experience and knowledge of social science theory (e.g. Chen and Rossi, 1980). In more recent work, authors tend to favor an integrative approach (e.g. Chen, 1990; Pawson and Tilley, 1997).

Most of the approaches in table 1 are examples of integrative approaches to PTE. All five approaches rely on program staff as the main information

sources for the reconstruction of the theory. Moreover, in at least three of the five approaches (2, 3 and 4 and to a lesser extent approach number 1) the evaluator consults research evidence from the social sciences either as an inspirational basis for the reconstruction of the theory or as a benchmark for assessing the theory. The most obvious element that binds all five approaches is the fact that only program staff members[7] are included in the core exercises of reconstructing and evaluating the theory (though in the margin other stakeholders might be consulted as well). Renger and Bourdeau (2004) admit that the inclusion of other stakeholder groups (e.g. program beneficiaries) might have added important extra information to the theory. Indeed, Rossi et al. (1999) among others have argued for the incorporation of a wider group of stakeholders in PTE. Not only would this generate additional knowledge, in the spirit of both pragmatic reasons (stakeholder ownership, accountability) and possibly stakeholder empowerment it is important for a wider group of stakeholders to be involved

[7] Or, in one case, the somewhat broader group of primary users. Primary users concern those stakeholders who have a principal role in decision-making and are in the position to utilize the results (Christie and Alkin, 2003).

Table 1. Comparison of program theory evaluation approaches on six aspects

approach	which stakeholders	method of reconstruction (main elements)	aggregation of the theory	method of evaluation (main elements)	stakeholder participation in the process	main objective evaluation
1) strategic assessment approach (Leeuw, 2003)	program staff	four-tiered approach based on group dynamics and dialogue (incorporating existing research evidence)	group consensus mediated by evaluator	deliberation without external validation (reconstruction and evaluation are not separate processes)	high	program and organizational improvement (strategy clarification)
2) policy-scientific approach (Leeuw, 2003)	program staff	interviews, documents, argumentational analysis	evaluator determines final theory without much stakeholder intervention	assessment of the theory against three pre-determined criteria (e.g. on the basis of scientific research evidence)	low	ex post evaluation of program's merit and worth
3) elicitation approach (Leeuw, 2003)	program staff	interviews, creating dialogical tension, iterative process of reconstructing cognitive maps	evaluator determines final "collective cognitive map"	confrontation of the theory with relevant scientific research	medium	program and organizational improvement (strategy clarification)
4) utilization-focused program theory evaluation (Christie and Alkin, 2003)	primary users, program staff	grounded in social science theory, interviews, iterative process of Delphi inquiry (questionnaire-based)	group consensus determined by the evaluator on the basis of iterative Delphi process	not specified	medium	program and organizational improvement (strategy clarification)
5) values inquiry and program theory evaluation (Renger and Bourdeau, 2004)	program staff	three-tiered approach based on interviews and group meetings	group voting and consensus building mediated by evaluator	not specified	high	ex post evaluation of program's merit and worth

To what extent would the incorporation of a wider group of stakeholders be feasible under the different methodological processes? The second, third and fifth aspect in table 1 will provide some insights on this question. Looking at the fifth aspect, the degree of stakeholder participation differs from low to high. In the policy scientific approach (2), the degree of participation is low. The evaluator consults with program staff without relying on systematic feedback or debate for aggregating the information into the final reconstruction. In approaches 3 and 4, some kind of iterative process of consultation and feedback with program staff [8] constitutes the basis for the theory. However, in both approaches the evaluator determines the final "group consensus", either on the basis of different individual and/or collective mental maps (3) or anonymous individual response forms (4). In approaches 1 and 5, stakeholders play a direct active role in the determination of the final picture (resulting from discussions and/or voting), characteristic of a high degree of stakeholder participation in the process. In fact, the latter two approaches most adequately reflect earlier opinions by among others Chen (1990) and Vedung (1997) who talk about negotiation and consensus building between stakeholders (including the evaluators) as a main engine for deriving the final theory.

[8] This idea of iterative theory reconstruction has also been described vividly by Pawson and Tilley (1997) who envision a relationship of student and teacher, where the stakeholder teaches the theory to the student (the evaluator). The evaluator uses the ideas from her 'teachers' to define a theory which is subsequently fed back to them for further refinement.

Rephrasing our question from the previous section, do group processes as under approaches 1 and 5, which in practice require the physical presence of (representatives from) stakeholder groups present potentially feasible and viable options for wider application to stakeholder groups 'outside' the program organization (e.g. program beneficiaries)? In our discussion on the deliberative democratic evaluation approach we provided two major critical arguments also pertinent in this case which would justify a cautious 'no'. In contrast, both the elicitation of individual mental maps (3) as well as the individual Delphi method (4) are more easily applicable to wider groups. However, especially the construction of individual mental maps (and the iterative feedback to refine the maps) involves a rather cumbersome exercise and therefore is more difficult to apply on a larger scale. The same is valid for the policy-scientific approach (2) which in case of each respondent requires a rather elaborate process of argumentational analysis (see Leeuw, 1991, 2003) in order to uncover the strings of causal logic that underlie a program. A less elaborate values inquiry exercise using for example surveys (e.g. Henry, 2002) would perhaps constitute a more feasible method of inquiry to capture stakeholder values among wider groups of stakeholders (outside the program organization) which are affected by or in any other way involved in the program.

A final important element in our succinct comparison of the five methodological approaches concerns the fact that all approaches arrive at one single synthesis, one single program theory, for evaluation. Let us briefly review this issue from the point of view of the objective of the program evaluation. Mark et al. (2000) discern four primary purposes of evaluation studies: assessment of merit and worth, oversight and compliance, program and organizational improvement, and knowledge development. In our case, the five approaches either serve the objective of ex interim/ex post assessment of

merit and worth or ex interim program and organizational improvement. In both cases, a single overall theory, comprising the different value and factual claims offered by stakeholders, makes sense. In ex post assessment one needs a clear idea of 'the program' in order to be able to judge whether it has merit or not. Similarly, in ex interim (or ex ante) strategy building, it is important that all program staff members share a joint view on what should be done and how (Weiss, 1998).

Nevertheless, considering the diversity in perspectives ('theories') of different stakeholders related to a program, does a single program theory do justice to this diversity? Consensus-seeking group processes, group voting or an evaluator's discretionary decisions to present an 'average' or 'decision maker biased' overall perspective can easily spirit away quite substantial differences in priorities between stakeholders. By failing to address explicitly these differences, the evaluation process, rather than contributing to a stronger stakeholder commitment and the program's accountability towards stakeholders, might risk alienating stakeholders from the program. Hence, what is needed is a PTE approach that captures these differences and systematically evaluates how each of these different value positions (reflecting alternative theories about how a program should work) would affect evaluative outcomes and subsequent policy choices. Only after such a systematic treatment of different perspectives should a consensus or a well-considered choice in the form of a single program strategy be sought. Evidently, this line of argument is especially relevant in evaluation contexts where (among other objectives) the objective of program improvement towards the future is important. More specifically, such an approach to PTE is relevant in cases when:

- accountability of the program and the relevance of the (findings of the) program evaluation to stakeholders outside the program organization warrant explicit attention;
- there are substantially divergent needs and views of different groups of stakeholders (both within as well as outside the program organization);
- both the objectives of ex interim assessment of merit and worth as well as program improvement towards the future are (equally) important.

In the next section we will illustrate how PTE can be enriched by a systematic approach that involves the definition and evaluation of alternative program strategies from the point of view of multiple stakeholder value positions.

4 Enter… Multicriteria Decision Aid

Multicriteria decision aid (MCDA) is a decision support technique that has been widely used in ex ante product, project and program evaluations in both corporate and public domains of decision-making. Nowadays, there are many different methods in MCDA and the number is growing (see for example Belton and Stewart, 2002; Roy, 1996). The essence of a MCDA approach is to support decision makers in making informed choices regarding a number of alternatives based on a number of criteria. In addition, the relative importance of the different evaluation criteria in the final decision can vary according to the preferences of different stakeholders involved the decision-making process (Belton, 1990).

In this section, we will illustrate the logic and utility of the link between PTE and MCDA. In fact, the potential role of PTE in decision support has already been acknowledged quite some time ago by for example McClintock who asserted that "[p]rogram theory […] plays an important role in decision-making, since it can be used to both expand conceptions of problems and solutions and to narrow attention on a manageable set of action alternatives" (McClintock, 1987: 43). MCDA offers a framework to systematically evaluate such alternatives and thereby enhances the utility of program theory in supporting decision-making processes.

The compatibility between MCDA and PTE can be logically established by conceiving the former method as an extension of the latter. A MCDA extension to PTE would comprise the following elements (see for example Dodgson et al., 2001):

1 the elicitation of evaluation criteria and the relative importance of evaluation criteria
2 the definition of alternative program strategies
3 the evaluation of the alternatives on the basis of the criteria
4 the application of a MCDA technique to derive a global ranking of alternatives as a basic structure for a process of deliberation among decision makers (and possible other stakeholders)

A methodological framework for midterm program evaluation comprising elements of PTE and MCDA can be usefully divided into two phases. The first phase serves the main purpose of assessing the strengths and weaknesses of the program under review. The elicitation of evaluation criteria, the first element in the abovementioned list on MCDA, is integrated into this phase. On the basis of the findings of the first phase, the second phase, encompassing elements two to four from the above-mentioned list, is designed to support decision makers in defining an improved program strategy. The main participatory element in the framework is the systematic elicitation and incorporation of stakeholder values in the evaluation process. The final deliberation process may or may not be participatory (in the sense of including multiple stakeholders in the discussion) depending on the characteristics of the decision-making

process (willingness and feasibility). Considering the costs and constraints surrounding stakeholder participation on the one hand and the purpose of each step of the framework on the other, many aspects of the methodology are unilaterally determined by the evaluator only. The basic underlying thought is to generate a particular balance between stakeholder participation and analytical rigor. The following framework reflects our interpretation of this balance given the features of the particular evaluation context sketched earlier.

Description of the methodological framework

First phase: PTE

In the first phase of the program evaluation the evaluator relies on one of the existing approaches to PTE to reconstruct a single program theory which according to her interpretation of program processes and additional assumptions elicited from program staff best fits the program reality. Other stakeholders (outside the program organization) very probably have other 'theories' regarding how a program works and should work. These theories are captured by means of systematic values inquiry (e.g. Henry, 2002), i.e. by focusing on what stakeholders perceive to be important. Hence, rather than reconstructing multiple theories by means of, for example, an extensive 'constructivist' process of inquiry (Guba and Lincoln, 1989), the evaluator tries to capture the different stakeholder theories on the basis of the criteria which stakeholders deem important regarding a program. Differences in 'sets of criteria' are the basis for the definition of different value positions, different stakeholder theories that adequately represent the diversity in values existing among the groups of stakeholders. Depending on the type and size of the stakeholder group, interviews and short surveys can be used to elicit the most important criteria that stakeholders deem important and the relative importance of these criteria according to the preference patterns of the stakeholders.

Subsequently, the assumptions underlying the original program theory (which is reconstructed from the point of view of the program (staff)) are assessed on the basis of the elicited criteria (reflecting the values of all relevant stakeholder groups, including those outside the program organization). In the context of midterm evaluations empirical data collection is mostly restricted to inputs, processes and outputs, whereas the links with potential impacts are appraised on the basis of more tentative empirical evidence. In addition, in elaborating her judgment, the evaluator can make use of existing research evidence on similar programs and state of the art research in the social sciences (e.g. Leeuw, 2003).

Second phase: MCDA

The definition of alternative program strategies

On the basis of the first phase, the evaluator is able to identify a number of weaknesses in the existing program. These will provide the basis for:

- a set of general guidelines to improve the program;
- the definition of a number of alternative strategies that provide partial solutions to the detected weaknesses in the existing program.

Given the trade-offs between the different criteria it is highly unlikely that there will be one alternative that can (potentially) solve all the weaknesses in the original program, hence the definition of multiple alternatives. For example, there might be a trade-off between the costs of a program (which need to be controlled) and its potential impact (which should be optimized), i.e. a higher positive impact implying higher costs. In practice, these trade-offs are not always apparent. Systematic evaluation of the alternatives on each criterion is the enabling condition for a comprehensive comparative analysis of the trade-offs.

The evaluation of the alternatives on the basis of the criteria

The original program theory provides a useful guidance for appraising the different alternative strategies that have been defined, since all the alternatives are in essence (slightly) different versions of the existing program. In practice, the appraisal of the alternatives on the different criteria can be based on multiple sources of information and analysis. The evaluator's judgment is subsequently translated into an evaluation score.[9] MCDA techniques can deal with these different types of scores in order to derive a global comprehensive ranking of all the alternatives while taking into account all partial expected performances on all criteria. All the evaluation scores are summarized in a performance matrix of *m* alternatives by *n* criteria.

The application of a MCDA technique to derive a global ranking of alternatives as a basic structure for a process of deliberation among decision makers (and possible other stakeholders)

In the final step of the appraisal process the evaluator uses a (simple) formal algorithm[10] to produce a global ranking of the alternatives, based on the scores of the different partial evaluations on all the criteria and the relative weights of the criteria.[11] This kind of ranking (and the algorithm that produces it) should not be seen as a substitute for a more intuitive analysis of the performance matrix. Instead, intuitive analysis precedes and guides the procedure of generating a final ranking of the alternative strategies. MCDA is a tool that helps structuring the kind of intuitive deliberation that always takes place when people need to choose between alternative courses of action. It is especially useful when the number of alternatives and the number of criteria which count in the evaluation of the alternatives increase, since without such support the human mind alone cannot grasp such complexity.

An important condition for the practical application of MCDA in decision-making environments is simplicity. Although simple MCDA models are less 'realistic' (e.g. in terms of modeling of preferences) they are more easily understood by decision makers. Belton and Stewart (2002) argue that simple models (that are easily applicable in practice) very often generate the same global insights as more advanced models, if the assumptions of the model

[9] Such an evaluation score can be measured on a metric scale (e.g. expressing monetary costs) or on a non-metric scale (e.g. an ordinal semantic scale ranging from 'very low' to 'very high'). A special type of non-metric scores concerns the procedure of ranking alternatives according to their expected performance on a criterion. In such a procedure, the alternative with the best expected performance receives the rank 1, the second-best alternative receives rank 2, etc. This technique is often used in cases when the available information is imprecise and/or the expected effects are uncertain and depend on a complex interplay of (external) factors.

[10] A common problem in MCDA is the fact that it involves many arbitrary choices, hence creating the complication of method uncertainty. In practice, the choice of a particular MCDA technique should be based on a balance between realism (e.g. in terms of the assumptions regarding stakeholder preferences, being able to show 'real' trade-offs) and applicability, taking into account aspects like data scarcity, stakeholder willingness to cooperate, available computer hardware and software, time and resource constraints, conditions posed by program decision makers, etc.

[11] Our discussion on the intricacies of the MCDA process is restricted to a minimum. Important issues in the aggregation of scores and preferences and the choice of MCDA technique are for example: to what extent can a low performance on criterion A compensate for a high performance on criterion B; are preferences regarding the compensation between criteria linear or non-linear; are preferences regarding scores on a given criterion linear or non-linear; is there a minimum threshold level on a given criterion in order for an alternative to be accepted at all in further analysis. For these and other issues see for example Dodgson et al. (2001); Belton and Stewart (2002); Roy (1996); or more classical texts like Keeney and Raiffa (1976).

(e.g. stakeholder preferences regarding the choice and relative importance of criteria) are extensively discussed and varied (i.e. sensitivity analysis).

An attractive feature of MCDA concerns its potential for analyzing the possible 'optimal' choices regarding alternative program strategies from the point of view of different value positions. In MCDA the differences in priorities between stakeholders can be captured by distinct sets of criteria and relative preferences for the criteria (which as described earlier represent the essence of different stakeholder theories regarding the program). For example, whereas program staff might consider the criterion X and Y to be important in defining a new program, program beneficiaries might value X, Y and Z. Moreover, the relative importance of X and Y might be different for program beneficiaries in comparison to program staff.

In MCDA, the final ranking of alternatives resulting from the appraisal process is based on assumptions about the expected performance of alternatives on criteria and preferences regarding the criteria. It makes perfect sense that the evaluator produces multiple rankings of alternatives each of them representing a specific choice and relative importance of evaluation criteria, associated with a certain group of stakeholders. These multiple syntheses produce the basis for a deliberation process in order to define the new program strategy. In principle, there are three broad scenarios for organizing this type of deliberation process:

- the evaluator facilitates a discussion between representatives of different stakeholder groups where the initial viewpoint of each group is represented by a ranking of alternatives that is based on each group's value position;
- the evaluator facilitates a discussion between program decision makers only; it is the task of the evaluator to confront program decision makers with the consequences of different stakeholder values on the ranking of the program alternatives;
- the evaluator does not participate in the decision-making process but describes in her report how different value positions would affect choices between alternative strategies and where possible compromises between stakeholders might be found.

The characteristics of such a deliberation process, as well as other features of the methodological framework are illustrated in the next section.

5 An Example: The Evaluation of a Training Program in Organic Agriculture in Guatemala

Short description of the program and the program evaluation

In this section we will illustrate the potential of the methodological framework by means of an evaluation of a training program on organic agriculture in the Western Highlands of Guatemala (Vaessen and De Groot, 2004). The study was implemented in 2001[12] and served the main purpose of evaluating program outcome and impact. For the present, we have modified some of its features in order to increase its relevance for the discussion at hand. We assume that instead of being an ex post study, the evaluation study would serve the additional purpose of assisting decision makers in redefining the program for a second phase of implementation.

[12] The study also included a baseline study in 1998.

The three most important stakeholder groups involved in the training program are: indigenous (Mayan) farmers participating in the program, the managers and implementing staff from ORGANIC [13] (the organization which implemented the original program), and finally the managers of an integrated rural development program (IRDP) which finances ORGANIC. The training program involved two main components. The first component concerned the organization of training workshops of two to three days every two months (for a period of three years) on an experimental farm in the central part of the territory. Participants were selected on the basis of their commitment to attend the courses and apply the knowledge on their own farms. Traveling expenses to the farm and costs of food and lodging were covered by the program. The second component was the provision of technical assistance by ORGANIC staff to participant farmers in between the courses to assist them in the application of organic practices in the field.[14] Besides participating in the courses and applying the acquired knowledge in their own farms, farmers were stimulated to organize themselves in groups for future exchange of ideas and cooperation on farming techniques. In addition, they were stimulated to share their knowledge with neighboring farmers in their communities, thereby creating a diffusion effect of the innovations. The main objective of the program was to improve agricultural production and hence the living conditions of participating farm households. Annex 1 depicts the reconstructed program theory for the training program.

[13] ORGANIC is a fictitious name.

[14] The main topics that were covered by the program can be roughly divided into two categories: physical practices and cultural practices. Physical practices involve the use of knowledge, labor and sometimes capital in order to be implemented (and maintained). Examples are the construction of sties, latrines (for human manure collection), and soil conservation measures like windshields, ditches and terraces. Cultural practices basically concern a change of habit or technique. The only essential input is knowledge, though sometimes additional labor might be required. Examples include the substitution of organic 'homemade' fertilizers for 'chemical' purchased fertilizers, plowing along the contour lines of the plot, crop diversification, and collecting instead of burning crop residues. A large part of the practices imparted by the program is in fact based on traditional Mayan farming techniques that have been neglected or all but forgotten in the past. To some extent ORGANIC, being a Mayan organization, acts as a catalyst by collecting pieces of Mayan knowledge all over the country and bundling and adapting this knowledge to fit specific Mayan production systems.

1 The evaluation of the existing program (including the element of elicitation of evaluation criteria and the relative importance of evaluation criteria)

Per stakeholder group, the criteria for evaluating the existing program and subsequently the program alternatives are reported in table 2. In addition, table 2 shows the relative importance of the different criteria. To make things simple, we assume that there are no significant differences within the stakeholder groups with respect to how the program is perceived and valued.[15] Nevertheless, between the three groups there are clear differences. In the case of ORGANIC, being the organization that designed and implemented

[15] In reality, differences in values regarding a program determine the definition of groups rather than the generic categories of program beneficiaries, program managers, etc.

the training program, we assume that the original program represents the most ideal option for continuing operations in the second phase. Therefore, our subsequent analysis is restricted to the value positions of IRDP decision makers and farmers.

Table 2. Evaluation criteria and relative importance of evaluation criteria for two stakeholder groups: IRDP decision makers and farmers

	total costs	yields per unit land	soil quality	soil and water conservation	labor use	reliance on external inputs	conservation of Mayan farming tradition	working with farmer organizations
IRDP	1	1	2	2	3	2	2	1
farmers	n.r.	2	3	3	1	3	3	n.r.

Note. 1 = most important, 2 = relevant but not very important, 3 = least important, n.r. = not relevant (in the case of farmers).

Given the difficulties surrounding the determination of the relative importance of criteria a simple ranking approach supported by a simple semantic scale is used. As shown in the table, for IRDP decision makers, total costs, the effect on yields and the creation of farmer organizations are the most important criteria. In contrast, for farmers labor input is the primary criterion.[16] Program costs and the creation of farmer organizations are considered not to be important at all.

[16] The opportunity costs of labor are quite high in the region. Almost all farmers are part-time farmers dedicating a significant part of their time to non-agricultural activities.

2 The definition of alternative program theories

Annex 2 presents a summary of the most important findings of the evaluation of the existing program (taking into account the different criteria). These findings provide the basis for the definition of alternative program strategies.

The following alternatives represent (partial) remedies to the flaws detected in the original program:

1 original program
2 maintaining more or less the same management structure and content but with more emphasis on laborsaving techniques
3 abandoning the original management structure and content; the new program focuses on a balance between conventional (laborsaving) and organic techniques
4 3 plus credit provision
5 3 plus assistance in organization building
6 3 plus assistance in organization building and credit provision

3 The evaluation of the alternatives on the basis of the criteria

In order to evaluate the expected performance of the alternatives on the different criteria, the evaluator returns to the original material (interview transcripts from stakeholders and experts, survey data, academic literature, other written sources) collected for the assessment of the original program

theory. On the basis of these information sources, the original program theory (for guidance), and consultations with key stakeholders and experts, the evaluator is able to appraise the alternatives on the different criteria. Given the imprecise nature of many of the expected effects of the alternatives on the criteria, an ordinal ranking approach is used.

Table 3. Expected performance of the alternatives per criterion

Alternative	costs	yields	soil quality	soil and water conservation	labor use	reliance on external inputs	conservation of Mayan farming tradition	working with farmer organizations
1	1	4	5	6	3	2	5	4
2	2	3	4	5	2	1	4	4
3	3	2	3	4	1	3	3	3
4	5	1	2	2	2	4	2	3
5	4	2	3	3	1	3	3	2
6	6	1	1	1	2	4	1	1

Note. The alternatives are ranked according to their perceived performance on the different criteria. 1 = best performance, 2 = second-best, etc.
The rankings already take into account that some criteria (costs, labor use and reliance on external inputs) are criteria to be minimized (e.g. the alternative that implies the lowest costs receives the highest ranking = 1) while the rest of the criteria are to be maximized.

Table 3 presents the rankings of the alternatives per criterion. Quite elaborate reasoning (guided by the program theory) may lie behind a certain ranking. For example, the criterion 'conservation of the Mayan farming tradition' is evaluated at the level of the population of participating farmers and rests on explicit assumptions regarding expected adoption levels, expected diffusion levels and specific content of the training program.[17] This explains why the original program strategy has such a low ranking. Although the original program implies the richest variety in traditional Mayan techniques, expected adoption and diffusion levels are much lower than for the other alternatives.

[17] Alternatives 1 and 2 are not so different from each other in terms of content of the training program. In contrast, alternatives 3 to 6 are substantially different from the first two.

4 The application of a MCDA technique to derive a global ranking of alternatives as a basic structure for a process of deliberation among decision makers (and possible other stakeholders)

In this section we will illustrate how the differences in values between IRDP decision makers and farmers will lead to different rankings. Subsequently, these different rankings provide the basis for a process of reflection, either among IRDP decision makers or including representatives from the farmers.

First, let us briefly treat the issue of how to aggregate the evaluation scores (table 3) and weights (table 2) into an overall ranking of alternatives. In our example we applied a simple ordinal ranking approach as a basis for both the determination of the relative preferences regarding criteria as well as the expected performance of the alternatives on the different criteria. This approach is especially useful in cases when it is costly or difficult (e.g. in view

of the lack of clarity, complexity and/or uncertainty involved) to determine clear evaluation scores and preference patterns. There are different MCDA techniques available that allow for dealing with this type of data. In this example we used a simple intuitive approach for transforming ordinal rankings to cardinal scores in order to create a complete ranking.[18] As argued by Belton and Stewart, such a simple approach has its merits (e.g. stakeholders can more easily understand the procedure), but is not without problems.[19] Nowadays there are a number of techniques available that treat this type of imprecise information.[20]

Figure 1 shows the base ranking of alternatives for the two stakeholder groups. This ranking by no means represents the end of the evaluation process. In fact it is the beginning of a process of deliberation which involves a combination of human judgment and adaptations of the model. The ideal situation would involve some kind of workshop involving decision makers (and other stakeholders), guided by the evaluator in which the assumptions (e.g. regarding stakeholder preferences) underlying the model would change as the deliberation moves along. In this way, one would create a constructive process of deliberation, progressively moving towards a shared conviction of the best strategy to be undertaken. Alternatively, if the evaluator is barred from the decision-making process, or if for some other reason such a deliberation process cannot take place, it is the evaluator's task to inform decision makers as best as possible of the consequences of different assumptions regarding preferences (and expected performance) on the ranking of program strategies.

Let us briefly illustrate the kind of deliberation that could arise after presenting the base ranking (figure 1) and changing the underlying assumptions of the model (figure 2). In figure 1 we can see that for both groups the original strategy (alternative 1) represents the least desirable option.[21] For IRDP alternative 6 is the most desirable option given the high potential for working with farmer organizations, the high adoption rates and high yield effects. These benefits apparently sufficiently offset the high costs associated with alternative 6 (as long as the high costs do not pose a too high obstacle in the eyes of IRDP decision makers). For the farmers, alternatives 3 to 6 all represent a significant improvement in comparison to the original strategy or its close derivative (alternative 2). Alternatives 3 and 5 are the most desirable options. The provision of credit (implied in alternatives 4 and 6) tied to the adoption of (physical) practices only pays off if farmers are sufficiently willing to invest heavily in their farm. However, in reality this is in an important bottleneck. Given the overall importance of non-farm activities, often making up more than half of the household income, for the majority of farmers there are clear limits in terms of how much labor they are willing to invest in agriculture without endangering other income activities. Given current preferences and assumptions regarding expected performance perhaps alternative 5 would constitute a good compromise between farmers and IRDP.

[18] This procedure is treated in Belton and Stewart (2002). In the case of evaluation scores they argue "that if the number of alternatives is small, then it may be possible to rank order all alternatives in terms of the criterion under consideration. Each rank position might then in this case be represented as a 'category' …[on an ordinal scale], and the estimation of values corresponding to each category becomes in effect a direct rating of alternatives…" (ibid.: 168). In other words, further reflection will help to determine the 'distance' between the rank positions. In the case of weights we used the simple rank sum approach to create cardinal weights (Stillwell et al., 1987). Finally, a simple additive model was used to create an overall ranking (see for example, Keeney and Raiffa, 1976).

[19] The most extensive debate on the shortcomings of the linear additive model (as a simple MCDA technique) has been waged within the MCDA and wider operational research literature (and has led to the development of a wide variety of more sophisticated MCDA tools). In the evaluation literature, both arguments in favor (e.g. Shadish et al., 1991) as well as arguments against (e.g. Scriven, 1991) the linear additive model can be found. However, most of these critical arguments are related to the application of the linear additive model in function of creating a single (fact-value) synthesis or some of the methodological features of simple linear additive models. In this paper, we do not propagate the construction of a single (fact-value) synthesis nor the application of simple linear additive models.

[20] One of the approaches that deals systematically with the problem of transforming ordinal rankings into cardinal ratings is called MACBETH (see Bana e Costa and Vansnick, 1994). Another approach called ARGUS (De Keyser and Peeters, 1994) maintains the ordinal nature of both weights and evaluation scores up until the final ranking. Both approaches are quite accessible and the main ideas underlying these approaches can be quite easily explained to decision makers.

[21] The reader should consult the list of alternatives presented earlier in this section.

Figure 1. Base ranking

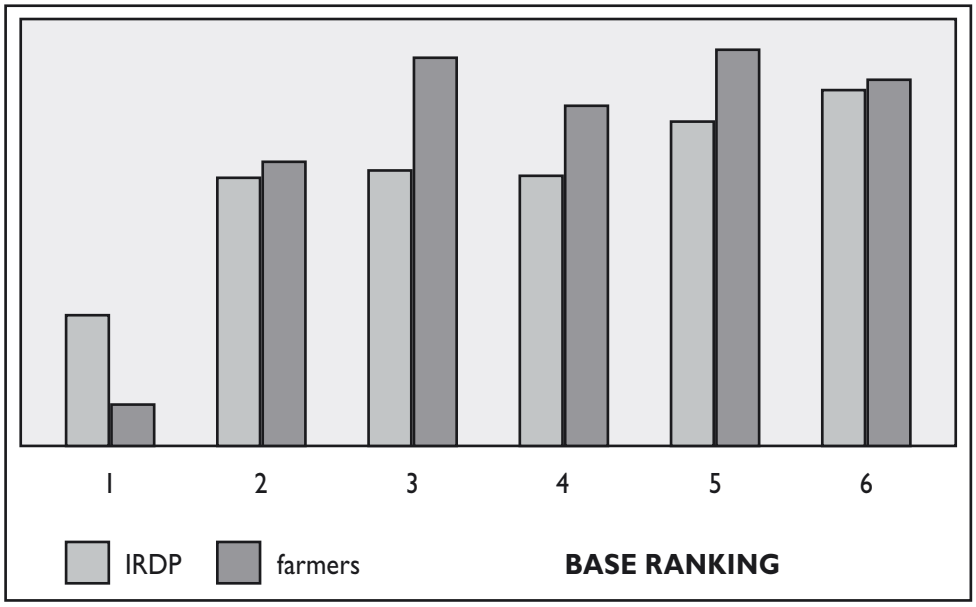
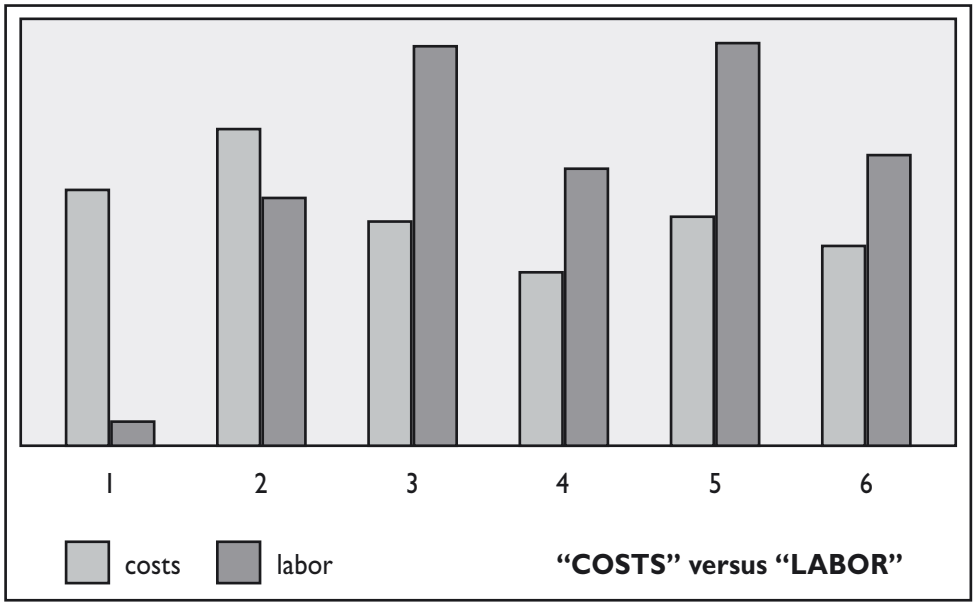


Figure 2. Sensitivity analysis: enhanced weight of costs (IRDP) and labor use (farmers)



What if we change the assumptions regarding stakeholder preferences? Let us suppose that the cost criterion for IRDP and the labor use criterion for the farmers become more important vis-à-vis other criteria. With respect to the farmers, figure 2 shows that the ranking does not change significantly but the preference for alternatives 3 and 5 is now more marked. In contrast, from the point of view of IRDP the ranking substantially changes. IRDP decision makers now face a different trade-off than before. Alternative 2 becomes the most attractive one, with alternative 1 (the original strategy and the cheapest

option) ending second. If IRDP chooses alternative 2, it is able to offer the farmers a substantially better alternative than the original strategy though not exactly the most attractive option from the farmers' point of view. On the other hand, alternative 1, the second most attractive strategy for IRDP, is out of the question since it is less attractive for both IRDP and the farmers in comparison to alternative 2. Perhaps alternative 2 would constitute the best compromise solution for the two parties. Choosing alternatives 3 or 5 would content the farmers but at a substantial cost for IRDP.

The above reasoning supported by the MCDA model in principle could be extended to include additional value positions. The discussion clearly illustrates the supportive nature of the MCDA model, not substituting but merely complementing in a constructive manner the deliberation process among decision makers.

6 Conclusions

The treatment of stakeholder values in evaluation (including PTE) remains "a fertile area for future work" (Mark, 2001: 467). Systematic values inquiry, a flexible mode of inquiry which allows for (among other things) more extensive and 'rapid' modes of stakeholder consultation and participation, appears to be promising, particularly because of its potential compatibility with real political realities surrounding programs and policies (ibid.: 469). In this sense, simple decision aid techniques articulated with systematic values inquiry may constitute useful methodological combinations to improve the links between governance, evaluation and stakeholder participation.

This paper has tried to clarify and emphasize two main elements. First and foremost, its aim has been to illustrate the utility of a specific methodological framework encompassing elements of PTE and MCDA. Second, it has delimited the evaluation context in which such a methodology might be most useful. The attractiveness of PTE approaches lies (among other things) in the aspect of making explicit the consecutive assumptions underlying social programs. Such a reconstruction of the program theory is quite useful to structure further evaluation activities in order to determine the program's merit and worth as well as organizational processes of strategy clarification, consensus building and improving the program towards the future. Nowadays, there are many methods in the field of PTE that (mainly) serve either of the two objectives in a more or less participatory manner. We have shown that in evaluation contexts (e.g. midterm program evaluations) where both the objectives of retrospective judgment and proactive program improvement are important, and where accountability towards stakeholders and the inclusion of different stakeholder priorities are important, our proposed framework displays a clear added value by illustrating how different stakeholder values would affect evaluative outcomes and subsequent policy choices.

In principle, the methodological framework presented can be easily adapted to the specific demands posed by different program evaluation contexts. For example, the elaborateness of the data collection and analysis underlying the evaluation of the original program and the alternatives derived from it depend essentially on the size and complexity of the program on the one hand and the budget for evaluation on the other. In addition, in some public programs the need for a systematic treatment of stakeholder values in view of greater transparency and accountability towards stakeholders and/or resolving potential conflicts, may be more pressing than in others. A final dimension concerns the specific balance between retrospective judgment and proactive program improvement. For example, the evaluator can choose to limit the depth of analysis in the evaluation of the original program and/or the appraisal of the different program alternatives, by relying only on imprecise qualitative appreciations for the different evaluation criteria. This might be desirable when the emphasis of the evaluation study lies on the second phase of the framework, the deliberation process and comprehensive analysis of program strategy alternatives.

References

Bana e Costa, C.A. and J.C. Vansnick (1994) *"MACBETH – An Interactive Path towards the Construction of Cardinal Value Functions"*, **International Transactions in Operational Research**, 1, pp. 489-500.

Belton, V. (1990) *"Multiple Criteria Decision Analysis - Practically the Only Way to Choose"*, in: L.C. Hendry and R.W. Eglese (eds.) **Operational Research Tutorial Papers 1990**, Operational Research Society, Birmingham.

Belton, V. and T.J. Stewart (2002) **Multiple Criteria Decision Analysis: An Integrated Approach**, Kluwer Academic Publishers, Dordrecht.

Bickman, L. (ed.) (1987) **Using Program Theory in Evaluation, New Directions for Program Evaluation**, 33, Jossey-Bass, San Francisco.

Bryk, A.S. (ed.) (1983) **Stakeholder-Based Evaluation, New Directions for Program Evaluation**, 17, Jossey-Bass, San Francisco.

Chen, H.T. (1990) **Theory-Driven Evaluation**, Sage Publications, Beverly Hills.

Chen, H.T. and P.H. Rossi (1980) *"The Multi-Goal, Theory-Driven Approach to Evaluation: A Model Linking Basic and Applied Social Science"*, **Social Forces**, 59, pp. 106-122.

Christie, C.A. and M.C. Alkin (2003) *"The User-Oriented Evaluator's Role in Formulating a Program Theory: Using a Theory-Driven Approach"*, **American Journal of Evaluation**, 24(3), pp. 373-385.

Cooke, B. and U. Kothari (eds.) (2001) **Participation: The New Tyranny?**, Zed Books, New York.

Cousins, J.B. and E. Whitmore (1998) *"Framing Participatory Evaluation"*, in: E. Whitmore (eds.) **Understanding and Practicing Participatory Evaluation, New Directions for Evaluation**, 80, Jossey-Bass, San Francisco.

Dahler-Larsen, P. (2001) *"From Programme Theory to Constructivism: On Tragic, Magic and Competing Programmes"*, **Evaluation**, 7(3), pp. 331-349.

De Keyser, W.S.M. and P.H.M. Peeters (1994) *"ARGUS – A New Multiple Criteria Method Based on the General Idea of Outranking"*, in: M. Paruccini (ed.) **Applying Multiple Criteria Aid for Decision to Environmental Management**, Kluwer Academic Publishers, Boston.

Dodgson, J., M. Spackman, A. Pearman and L. Phillips (2001) **Multi Criteria Analysis: A Manual**, Department for Transport, Local Government and the Regions, London.

Fetterman, D.M. (1994) *"Empowerment Evaluation"*, **Evaluation Practice**, 15(1), pp. 1-15.

Greene, J.C. (2000) *"Challenges in Practicing Deliberative Democratic Evaluation"*, in: K.E. Ryan and L. DeStefano (eds.) **Evaluation as a Democratic Process: Promoting Inclusion, Dialogue, and Deliberation, New Directions for Evaluation**, 85, Jossey-Bass, San Francisco.

Guba, G.E. and Y.S. Lincoln (1989) **Fourth Generation Evaluation**, Sage Publications, Newbury Park.

Henry, G.T. (2002) *"Choosing Criteria to Judge Program Success – A Values Inquiry"*, **Evaluation**, 8(2), pp. 182-204.

House, E.R. (1995) *"Putting Things Together Coherently: Logic and Justice"*, in: D.M. Fournier (ed.) **Reasoning in Evaluation: Inferential Links and Leaps, New Directions for Evaluation**, 68, Jossey-Bass, San Francisco.

House, E.R. and K.R. Howe (1999) **Values in Evaluation and Social Research**, Sage Publications, Thousand Oaks.

Keeney, R.L. and H. Raiffa (1976) **Decisions with Multiple Objectives – Preferences and Value Tradeoffs**, Cambridge University Press, Cambridge.

Leeuw, F.L. (1991) *"Policy Theories, Knowledge Utilization, and Evaluation"*, **Knowledge and Policy**, 4(3), 73-92.

Leeuw, F.L. (2003) *"Reconstructing Program Theories: Methods Available and Problems to be Solved"*, **American Journal of Evaluation**, 24(1), pp. 5-20.

Majone, G. (1980) *"Policies as Theories"*, **The International Journal of Management Science**, 8, pp. 151-162.

Mark, M.M. (2001) *"Evaluation's Future: Furor, Futile, or Fertile?"* **American Journal of Evaluation**, 22(3), pp. 457-479.

Mark, M.M., G.T. Henry and G. Julnes (1999) *"Toward an Integrative Framework for Evaluation Practice"*, **American Journal of Evaluation**, 20(2), pp. 177-198.

Mark, M.M., G.T. Henry and G. Julnes (2000) **Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs**, Jossey-Bass, San Francisco.

McClintock, C. (1987) *"Conceptual and Action Heuristics: Tools for the Evaluator"*, in: L. Bickman (ed.) **Using Program Theory in Evaluation, New Directions for Program Evaluation**, 33, Jossey-Bass, San Francisco.

Pawson, R. and N. Tilley (1997) **Realistic Evaluation**, Sage Publications, Thousand Oaks.

Renger, R. and B. Bourdeau (2004) *"Strategies for Values Inquiry: An Exploratory Case Study"*, **American Journal of Evaluation**, 25(1), 39-49.

Rossi, P.H., H.E. Freeman and M.W. Lipsey (1999) **Evaluation – A Systematic Approach**, Sage Publications, Thousand Oaks.

Roy, B. (1996) **Multicriteria Methodology for Decision Aiding**, Kluwer Academic Publishers, Dordrecht.

Scriven, M. (1991) **Evaluation Thesaurus**, Sage Publications, Newbury Park.

Shadish W.R., T.D. Cook and L.C. Leviton (1991) **Foundations of Program Evaluation: Theories of Practice**, Sage Publications, Newbury Park.

Shadish, W.R., D.L. Newman, M.A. Scheirer, and C. Wye (eds.) (1995) **Guiding principles for Evaluators, New Directions for Program Evaluation**, Jossey-Bass, San Francisco.

Stame, N. (2004) *"Theory-Based Evaluation and Types of Complexity"*, **Evaluation**, 10(1), pp. 58-76.

Stillwell, W.G., D. Von Winterfeldt and R.S. John (1987) *"Comparing Hierarchical and Nonhierarchical Weighting Methods for Eliciting Multiattribute Value Models"*, **Management Science**, 33, pp. 442-450.

Stufflebeam, D.L. (2001) **Evaluation Models, New Directions for Evaluation**, Jossey-Bass, San Francisco.

Vaessen, J. and J. De Groot (2004) *"Evaluating Training Projects on Low External Input Agriculture: Lessons from Guatemala"*, **Agricultural Research & Extension Network Papers**, 139, Overseas Development Institute, London.

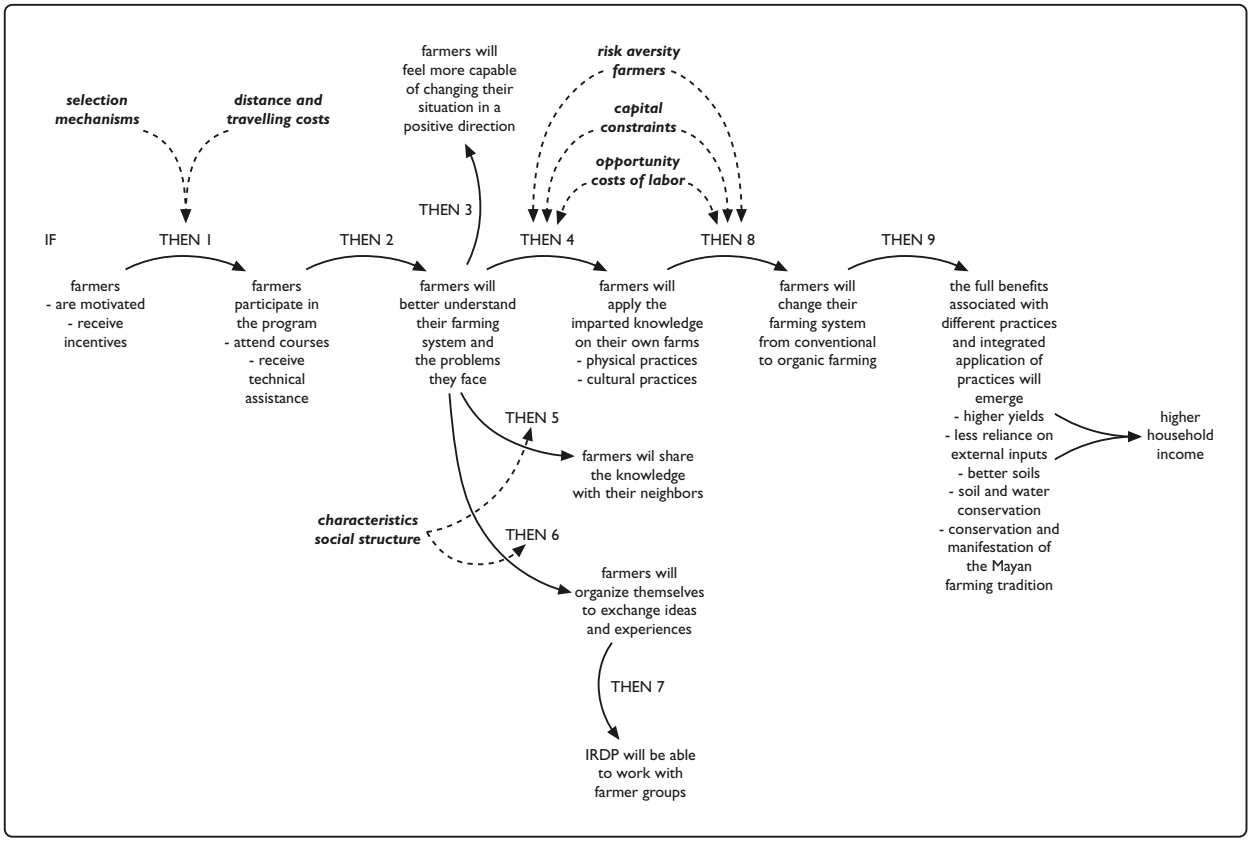
Vedung, E. (1997) **Public Policy and Program Evaluation**, Transaction Publishers, New Brunswick.

Weiss, C.H. (1997) *"Theory-Based Evaluation: Past, Present and Future"*, in: D.J. Rog and D. Fournier (eds.) **Progress and Future Directions in Evaluation: Perspectives on Theory, Practice and Methods, New Directions for Evaluation**, 76, Jossey-Bass, San Francisco.

Weiss, C.H. (1998) **Evaluation – Methods for Studying Programs and Policies**, Prentice Hall, Upper Saddle River.

Wholey, J.S. (1987) *"Evaluability Assessment: Developing Program Theory"*, in: L. Bickman (ed) **Using Program Theory in Evaluation, New Directions for Program Evaluation**, 33, Jossey-Bass, San Francisco.

Annex 1 Program Theory Training Program in Organic Agriculture



Note: Variables in bold italic and with dashed arrows are examples of external variables that influence specific links in the program theory.

Annex 2 Main findings of the assessment of the different links in the program theory

link	simplified assessment summary
THEN 1	Despite the fact that expenses for traveling, food and lodgings are reimbursed by the program, organizing the training program on a central experimental farm (or a limited number of locations) will discourage farmers from distant municipalities and communities to participate. This will lead to a concentration of participants from neighboring communities. In addition, clear and transparent information needs to be made available in the target area. If not, the program risks a selection bias in favor of those participants well-connected to program staff and past participants. Clear selection procedures respecting outreach (maximum number of participants per community, and minimum number of communities covered) are needed.
THEN 4 THEN 8	Most of the (poor) farmers in the region are risk-averse. They will carefully choose which practices to adopt and which not. Adoption will start on an experimental scale. Farmers in the region are mainly part-time farmers, most of the families being involved in trade, weaving and other artisanal activities. Competition for household labor is fierce and determined by the perceived return to labor. Capital is scarce. Very probably, the adoption of many physical practices demanding significant labor and capital input will be limited. Adoption will most likely be selective and partial (not on the whole farm). If adoption rates are to improve for those (physical) practices deemed important and desired by the farmers, then additional assistance perhaps in the form of credit or grants (under the right conditions) is needed.
THEN 5	Farmers will probably share their knowledge with other farmers. However, in some communities internal divisions (based on the legacy of the civil war, loyalty to different community leaders, kinship, religious practice) are high. In those communities, careful selection of participants would greatly enhance the potential diffusion effect. Other interventions in the territory have often been monopolized by certain factions in the community with limited or no spill-over effects to other factions.
THEN 6 THEN 7	Some participants and some of the ORGANIC staff members (from the region) are in fact local leaders in their communities. Incipient farmer association on the basis of shared interests in organic agriculture may be clustered around these persons. Given other experiences in the region with similar interventions, probably these groups will not become sustainable farmer associations. Because of the social divisions in the community and the role of leadership, these groups will not become groups of free access with equal membership rights. In any case, organizations of the type desired by IRDP will not come into existence without further external support.
THEN 9	Many organic practices once adopted will generate positive effects in different domains (e.g. agriculture, nutrition). However, field evidence and evidence in the literature suggest that a combination of conventional (chemical) inputs and organic inputs will generate significantly higher positive effects in the same variables.

Note 1: the findings are formulated in the present tense to serve as recommendations for alternative programs.

Note 2: this subset of findings is based on:
- empirical data collection and analysis
- literature on farmer adoption and innovation processes


INSTITUTE OF DEVELOPMENT
POLICY AND MANAGEMENT


UNIVERSITY
OF ANTWERP