"PARTICIPATORY EVALUATION AS A TOOL TO FOSTER ORGANIZATIONAL LEARNING: THEORY AND EVIDENCE FROM THE WORLD BANK"

Silvia Paruzzolo

Student ID 1003813
PhD in Business Administration and Management
XX cycle
Bocconi University
Tutor: Prof. Giovanni Fattore

ENHANCING LEARNING THROUGH PARTICIPATION IN EVALUATION PROCESSES

A CONCEPTUAL FRAMEWORK

Candidate: Silvia Paruzzolo
PhD in Business Administration and Management
XX cycle - Track: Public Management
Paper 1/3

OUTLINE

1. **Introduction**
2. **Theoretical Context**
3. **My Argument**
4. **Conceptual framework**
5. **Discussion and Conclusions**

References

1. Introduction

The growing importance of performance based management in government and non-profit operations has created a demanding environment for today's program managers and opportunities for evaluation professionals (Newcomer and Neill, 2001). Governments and organizations all over the world are grappling with internal and external demands and pressures for improvement and reforms in public management. Governments and organizations must be increasingly responsive to internal and external stakeholders to demonstrate tangible results. The cultural change required to achieve performance-based management in any public or non-profit agency presents an enormous challenge. Manager's thinking and action have to change from considering primarily what is needed to obtain resources and deliver programs, to concerns with achieving measurable results (Scheirer and Newcomer, 2001). Result-based monitoring and evaluation (M&E) is a powerful public management tool that can be used to help policymakers and decision makers track progress and demonstrate the impact of a given project, program or policy (Kusek and Rist, 2004).

Measuring results is extremely powerful (Osborne and Gaebler, 1992), and it can have a summative nature if it enables to tell success from failure or a formative nature if it provides "real-time" feedback to improve program implementation. Yet measuring does not necessary imply a direct effect in terms of action, i.e. rewarding success or correcting failure neither during the process nor after the evaluation are findings obtained. Time and effort is required in order to evaluate results and produce rigorous, useful and used evaluations, and useful and used findings. Nonetheless, their actual use has always been a concern for evaluation theorists.

In my opinion, one of the reasons why "neither learning nor utility are outstanding attributes of evaluations" is that "researchers and practitioners in evaluation setting are an "odd couple" (Myers-Wall, 2000). As Myers-Wall points out this conflict needs to managed, and one of the guidelines that the author proposes is "finding the intersection of interests that are shared by both evaluation researchers and practitioners". I argue that this intersection lies in the evaluation process and findings, and not only in the findings. To create alignment in the objectives of researchers and practitioners this intersection of interests needs to be made explicit and clear before the evaluation starts. For instance, managers will value participatory processes (Marra, 2004) only if can see the benefits deriving from the participation (given the time constraints and in a context of limited resources). It is logic to assume that if aware of the learning effects of being involved in the evaluation, managers would value evaluations more than if just seeing evaluations as a tool to provide information, no matter how "valuable".

As summarized by Forss and Reiben (2006), it has been said that "we have too much information – but not enough knowledge, that is, knowledge meaning information transformed into learning, action, and change. Evaluations are part of the information

industry. But evaluations do not necessarily produce knowledge". I maintain that the key factor for evaluations to successfully produce learning and change lies in the way the evaluation is managed and the absorptive capacity of individual and organizations exposed to a managed processes of evaluation. I concur with Perrin (2006) and Forss and Reiben (2006) in maintaining that evaluation is and - should be - used as a process rather than a product.

The present paper is structured as follows. I will first briefly describe the theoretical context by discussing the evolution of the evaluation theory and the contribution of different disciplines to the issue of evaluation use discussed hereafter. Next I will discuss my argument in favor of increasing awareness of process use to promote evaluation use in general, and process use, in particular. Finally I will present the conceptual framework on which I based my argument.

2. Theoretical Context

This conceptual framework presented here tries to link evaluation concepts and processes to theories and frameworks of public administration and management, including, for example, theories of organizational learning. As described in depth by Rossi and Wright (1984), evaluation research came into prominence as an applied social scientific activity. Policymakers and public administrators recognized that evaluation could be conducted systematically using social scientific research methods producing results which were considered to have more use and validity than the judgmental approaches used before. As explained by the authors, "the entire gamut of the social scientific disciplines was involved: economists, sociologists, psychologist, and educators..." The interdisciplinary character of this social scientific activity was and remains especially noteworthy. Initially, according to Alkin's (2004) work on evaluation roots, all evaluation was derived from social science research methodology and accountability. Accountability provided the rationale, but it was primarily from social inquiry that evaluation models have been derived. The framework developed hereafter draws on several social science disciplines: public policy and administration, policy analysis and evaluation, psychology and organizational learning, yet given the strong "results orientation" of contemporary public decision making, it offers a practical approach to evaluation suggesting a managerial approach designed to assist public officials in meeting these contemporary demands for "results."

The present paper focuses on the problem of use. According to Alkin's (2004) work evaluation use is one of the three major branches which developed from the accountability/social inquiry dual foundation (the other two being the other two "methods" and "valuing"). The use branch began its growth with what are often referred to as "decision-oriented theory". In fact, the work done by the theorists interested in use initially focused on an orientation toward evaluation and decision making, and expressed

concern for the way in which findings are used in these processes (Alkin, 2004). Alkin and Christie (2004) highlight that the first extension of this branch was made by theorists who were not primarily focused on decision makers' needs, but on emphasizing procedures that would enhance the use of evaluation to a broader spectrum of identified stakeholders. Moreover, the interpretation of the term "use" in itself has changed: from being initially narrowly limited to the use of evaluation findings/results, it has been broadened to include as use also what happens as a result of being involved in an evaluation, which can manifest both as evaluation capacity building (hereafter ECB) and as what Patton (1997) first defined as *process use* (hereafter PU).

To make the discussion more concrete, I will occasionally report examples from my personal professional experience in the context of impact evaluation work at the World Bank. Also in the table below I exemplify some of the changes I observed and attributed to participation in evaluation processes (distinguishing between what in my opinion is PU, and what is ECB).

CHANGES THAT CAN OCCUR through PARTICIPATION IN THE PROCESS OF EVALUATION

Changes in evaluand (PU)

- ...the evaluation helps question tacit assumptions about the program theory
- …the evaluation helps improving program designs as they are implemented

Changes in behavior (PU)

- …the evaluation stimulates team discussions on innovative ideas for project implementation
- …the evaluation increases team's shared understanding of the project
- …the evaluation increases the ability to solve problems during project implementation
- …the team has to pay more attention to changes in the external environment and contextual factors

Changes in thinking and skills (ECB)

- …team attitudes toward evaluation have become more positive after their involvement in the evaluation
- …evaluation develop team's systematic inquiry skills

The present paper generally focuses on the learning that occurs through participatory evaluation processes and addresses the question of whether there are ways to increase the contribution of evaluations towards result-based management and organizational learning. I argue that, in order to fully exploit the learning effects of

evaluation processes, more attention should be paid to awareness creation of and reflection on this kind of "use" of evaluations.

3. My argument

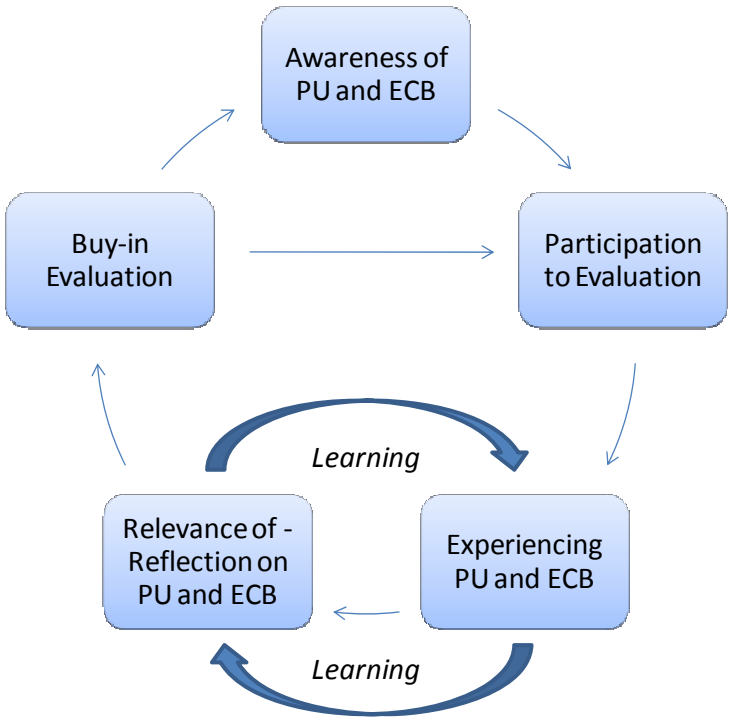
Assuming that a manager is willing to evaluate its program[1], designing and conducting a rigorous evaluation requires expertise, commitment, and follow-through. An evaluation also implies increased monitoring in general. Such monitoring can elicit higher effort from all of those involved in a project and reduce graft. It is believed that both donors and society benefit from the results of increased efforts, the measurement of impacts (Levine, 2006). Yet it is not always clear how. If one were to do a cost-benefit analysis of conducting an evaluation, I think that the benefit side of the equation should estimate all the possible learning effects of the evaluation. As highlighted by Patton (2008) the cost benefit ratio changes when the evaluation produces not only findings, but also serves immediate programmatic needs like staff and program development. Again, evaluations take a relatively long time to show end results: the focus on the short-term and non-awareness of possible immediate effects of process use leads to systematic under-investment in knowledge. From the point of view of a program manager, learning about a program's effectiveness retrospectively, once the program is over, might not justify the costs of an evaluation, especially if it requires substantial investments in terms of human and financial resources. In fact, the costs of evaluation activities shouldn't be justified only by the benefits obtained using the final outputs, but also by the effects (if any) of the evaluation process itself. Findings are ultimately the output of evaluation, and are crucial in informing management decision making processes in terms of improving, replicating, scaling up or rechanneling resources to different programs. Yet, awareness of the use of evaluation processes, can also contribute in convincing managers of the effectiveness of their investments in evaluations, and getting their buy-in. Stakeholder's awareness of process use activates a virtuous cycle (see Fig. 1). Being aware of the possible learning effects of being involved in an evaluation increases the likelihood of participation, which in turn creates the conditions to benefit and learn from the evaluation process (experiencing process use and ECB), which in turn increases the relevance attributed to evaluation use and the buy-in of the evaluation, increasing ownership, involvement, etc.. I suggest that the results of evaluation process use and ECB cannot be assessed adequately at the organizational level of analysis where important within-organization variance cannot be detected. Moreover, I suggest that understanding how different evaluation team members perceive differently the evaluation processes might shed light on how managers can direct those learning processes towards what is strategically desirable[2].

[1] As discussed in the literature there are many disincentives to conduct evaluations. See, for example "It pays to be ignorant" (Pritchett 2002).

[2] Here I would also cite paper 2 "Modeling Perceptions…" Paruzzolo (2009)

Based on this premises I developed a theoretical framework which discusses the learning effects of participatory evaluation processes.

Figure 1. The virtuous cycle triggered by awareness of PU and ECB



4. Conceptual framework

The theoretical and conceptual framework developed in this paper is strongly influenced by Cousins's participatory evaluation model, Patton's concept of process use, and Preskill's work on organizational learning and development. I next describe the main proposition of the framework providing a brief discussion of the relevant literature and examples from my own experience.

> *Proposition 1: Evaluation is a process of knowledge generation which is a catalyst for learning. Using an evaluation means learning from the evaluation by acquiring the knowledge generated through the process and the findings and acting upon it.*

According to Weiss and Bucuvalas (1980), much of the research on use since the 1960s has focused on the factors that affect use. And the concept of use (or utilization)

has been interpreted since then "as a learning process involving a mutually defined and interdependent system of participants, evaluation, and context". In fact, several authors investigated the relationship between evaluation and learning (to cite only few: Argyris and Schoen, 1978; Forss, et al., 1994; Forss et. al, 2002; Jenlik, 1994; Marra 2000; 2004; Preskill, 1994; 2008; Preskill and Torres, 1999; 2000). As discussed above, the focus has shifted in interpreting the use of evaluation beyond the intended use of findings, to the learning effects of being involved in an evaluation. The shift is discussed in the issue of the New Directions of Evaluation (2000) "The expanding scope of evaluation Use". I assume this was an inevitable path for theorists of evaluation use. In fact the concept of learning is often inherent in the same definition of evaluation, especially when discussing its use. Let's take into consideration a few definitions. The OECD (2002) defines evaluation as "the systematic and objective assessment of an on-going or completed project, program or policy, including its design, implementation, ad results… An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process if both recipients and donors. According to Owen (2004) evaluation is a knowledge production activity, bounded by the need to provide program-related information to identified audiences and program stakeholders". Evaluation can be defined even more broadly as analytical inquiry based in collecting and analyzing evidence, and drawing conclusions and recommendations from this evidence (Valovirta, 2002). Many other definitions exist, but essentially the common element is that evaluation is a form of systematic inquiry that generates information, evidence, and knowledge to provide feedback. Feedback affects individual knowledge structures (Chen, 2006; Forss et al., 1994) and stimulates learning. As will be discussed further, an evaluation generates knowledge beyond its findings, and the process whereby this knowledge it is acquired also represents a learning process. As pointed out by Preskill (2004) "evaluation is a learning process that should result in findings that are useful and used". Moreover "…evaluation in some way is about learning, learning about a program and its outcomes, learning from an evaluation process, learning about how to do an evaluation, or learning about evaluation effects on others…If learning is the act, process or experience of gaining knowledge or skills, then it is hard to imagine evaluation as anything other than a means for learning" (Preskill, 2007).

Proposition 2: Integrating evaluations with project operations requires collaboration of different stakeholders in the evaluation process, i.e. a participatory approach to evaluation, even when producing findings is the only purpose of the evaluation

Participatory evaluation means different things to different evaluators. There exists, mostly in the context of evaluation and monitoring in the developing world, a long-standing tradition of participatory inquiry variously labeled participatory evaluation,

participatory action-research, and the like. Such approaches are deeply rooted in principles of equity emancipation and transformation and are normative in form and function (Cousins, 2004)[3]. "Participatory evaluation" is also intended as a form of evaluation, or a method, as opposed to an approach to evaluation management. According to Owen (2004), for instance, a participatory evaluation takes place during the delivery of a program with the purpose of providing knowledge for decisions related to continuous improvement by involving program providers in the evaluation process, as opposed to impact evaluation which is used to assess impact of a settled program, to determine the effects of a program in terms of the criteria selected to judge its success.

To avoid confusion it is important to clarify that in the present study I refer to Cousins and Earl's (1995) definition of practical participatory evaluation as "applied social research that involves trained evaluation personnel and practice-based decision makers working in partnership". To be more specific, there are three fundamental dimensions of participation: (i) who controls the evaluation (on a researcher – practitioner continuum) (ii) stakeholder selection for participation (a continuum from all legitimate groups to just primary intended users) (iii) depth of participation (a continuum from consultation to deep participation) (Cousins and Whitmore, 2007).

As stated by King (2004), all evaluation is participatory at least to a certain extent. In fact an evaluator needs to interact with someone, at least minimally, to frame an evaluation task. In my view participatory approach is a practical approach and it is a means to undertaking a successful evaluation and obtaining rigorous and meaningful findings. High levels of stakeholder involvement in the planning and oversight aspects of evaluation work are linked to perceptions of greater usefulness of the evaluation data (Thayer et al., 2001).

But this definition raises a question: what participation in what? in what does the evaluation work or process consist? Regardless of the size, program type or methodology used for the impact evaluation, there are several steps to be carried out, which represent the evaluation process. Yet, given the high variability of interpretation of concepts such as participatory approach and process use (Forss et al, 2002), let's fix the method and type of evaluation and take the case of a prospective impact evaluation based on the estimation of a counterfactual (yet most of these steps and considerations apply to any evaluation). The steps in which high collaboration is especially needed are highlighted and discussed in more detail[4].

In a prospective evaluation all of the design work and initial data collection should be done during project identification and preparation, and the participation of program implementation team in the evaluation process is necessary in order to design a feasible evaluation strategy built into project operations. The first step consists in *establishing*

[3] Moreover, there is empowerment evaluation which is the use of evaluation concepts, techniques, and findings to foster self-determination (Fetterman, 2004).

[4] This paragraph draws heavily on Baker (2000) and World Bank (2006) to be consulted for a complete description of the steps to carry out an impact evaluation in the context of World Bank projects.

clear objectives and agreeing on the core issues that will be the focus of the evaluation up front. Clear objectives are essential to identifying information needs, setting output and impact indicators, and constructing a solid evaluation strategy to provide answer to the questions posed. The use of a logical framework approach provides a good and commonly used tool for identifying the goals of the project and the information needs around which the evaluation can be constructed. *Quantifiable measures* should then be identified for each link in the project cycle. Stating measurable objectives is not as common as one would think, and it is usually an iterative process which often requires the willingness of program managers to re-state program development objectives and set realistic targets under the supervision of the evaluation specialist who assure the actual feasibility of measuring those objectives. Also, evaluators can often serve as the "bridge" to other content area experts to help select the most "up to date" and valid measures to operationalize an intended construct (Scheirer and Newcomer, 2001). Following is the design phase of the impact evaluation study. The choice of the methodology will depend on the evaluation question, timing, budget constraints, and implementation capacity. The way in which the project is implemented will affect the ability of the evaluator to develop a valid identification strategy, i.e. how to identify the impact of the project separately from changes due to other causes. At this stage it is worth examining the possibility of project refinements that will enhance the evaluation – for example the way in which the project is phased in, clarification of eligibility criteria, and procedure for selection of pilot areas. The point here is not to change the development objectives of the project or to change the intended beneficiaries but to explore marginal changes to project design that will improve the quality of the evaluation. Optimally, identification strategy and project selection are developed in a way as to meet both project and evaluation objectives. Even after the evaluation design has been determined and built into the project, evaluators should be prepared to be flexible and make modifications to the design as the project is implemented. In addition, provisions should be made for tracking the project interventions if the evaluation includes baseline and follow-up data so that the evaluation effort is parallel with the actual pace of the project. By the very nature, evaluations are subject to the time frame established by the rest of the project. In defining the design it is also important to determine how the impact evaluation will fit into the broader *monitoring and evaluation strategy* applied to a project. Moreover the most critical *timing* issue is whether it is possible to begin the evaluation design before the project is implemented and when the results will be needed. There is the need to *integrate budgets* and insure flexibility of funding in order to coordinate the efforts and timing of program implementation and evaluation. Also during the *development of measurement instruments and data collection process* there is need of supervision from the program implementation team. Sector specialists and local staff who can provide knowledge of the program and country can be critical to the quality of the information collected and need to be involved in the development of the questions, the pilot test, and in the review of the data from the pilot test. Sampling is also best practiced by an experienced sampling

specialist. The sampling specialist should be incorporated I the evaluation process at the earliest stages to review the available information needed to select the sample and determine whether any enumeration work will be needed, which can be time consuming. The specialist will work in consultation with the other members of the evaluation team in obtaining the information, including data on the selected outcome indicators for the power calculations (an estimate of the sample size required to test for statistical significance between treatment and comparison groups), a list of the population of interest for the sample selection and details on the characteristics of the potential treatment and comparison groups important to the sample selection process. Since power calculations can be performed using only one outcome measure and evaluation often consider several, some strategic decisions will need to be made regarding which outcome indicator to use when designing the sample. Working with local staff who have extensive experience in collecting data similar to that needed for the evaluation can greatly facilitate fieldwork operations. Not only these staff can provide the required knowledge of the geographical territory to be covered but their knowledge can also be critical to developing the norms used in locating and approaching informants. Finally, as with other stages of the evaluation process, the analysis of the evaluation data requires substantial collaboration between the analysis, data producers, and policymakers to clarify questions and ensure timely, quality results. Problems with the cleaning and interpretation of the data will almost surely arise during analysis and require input from various team members. It appears clear that all these steps require the involvement of different team members with different skills and pieces of knowledge fundamental to the whole evaluation process to lead to successful results. Concluding the participatory approach might have no other purpose other than that allowing for i) the most tailored design to the project and specific context under evaluation and ii) the most useful results in terms of knowledge generation.

> *Proposition 3: Being directly involved in an evaluation process (or parts of it) has an impact in terms of learning at the individual level which can trigger changes in program and organizational procedures*

Knowledge creation and transformation in organizations results from individual's active participation and interactions with tasks, technologies, resources, and people within a particular organizational context (Bartel and Garud, 2003). Social interactions and the sharing of information contribute not only to individual knowledge but also to shared knowledge. Shared knowledge helps form the "unarticulated background" (Tsoukas, 1996) that enables coordinated action among people in a given work context. The process of evaluating can have effects that go beyond the use of evaluation findings (Weiss, 1998).

Beyond creating evaluation capacity (i.e. developing evaluation knowledge, skills and attitude), what happens to people and organizations as a result of being involved in an evaluation process has been firstly labeled as "process use" by Patton (1997, 1998, 2008). Theoretical and empirical research on this type of evaluation "use" has since then expanded substantially[5] (e.g. see Amo and Cousins 2007; Carden and Earl 2007; Cousins 2007; Cousins and Whitmore 1998; Harnar and Preskill 2007; King 2007; Patton 1997, 1998, 2004, 2007, 2008; Cousins and Shulha 2006; Preskill, et al., 2003; Fetterman 2003; Forss et al., 2002; Kirkhart 2000, Preskill and Caracelli 1997). Patton refers to process use as "the cognitive, attitudinal and behavioural changes in individuals, and program or organizational changes resulting from engagement in the evaluation process". Process use occurs when those involved in the evaluation learn from the evaluation process itself rather than just the evaluation findings (Patton, 1997, 1998, 2008). A very important clarification is made by Alkin and Taut (2003), who make the distinction between *unplanned instrumental use of a formative nature* and *process use*. According to the authors "the former is the use of interim evaluation findings produced and communicated by the evaluators to make program changes, the latter is the learning that takes place with stakeholders through their engagement in the evaluation process, as opposed to their reactions to evaluation findings produced and shared by the evaluators. To rightly talk about process use, the stimulus that causes instrumental or conceptual changes to take place must be the participation in the evaluation process, not the acknowledgement of evaluation findings."

Process use in its different dimensions brings changes at three levels: changes in thinking, behavior and in the evaluand itself. Let me exemplify these changes with what I've observed as recurrent patterns during my work experience at the World Bank.

The decision to design a prospective impact evaluation of a World Bank funded project usually stems from the necessity/willingness to prove with a rigorous methodology that the project can achieve the stated development objectives and is a cost/effective choice among alternative programs in a context of scarce resources. The World Bank is promoting evaluations that are aligned with project operations, which requires different stakeholders, such as the program manager, the project implementation team, the government counterpart where the project is actually implemented to collaborate with an evaluation specialist in the definition of a feasible evaluation design. Evaluations that can produce rigorous, credible and generalizable results generally require a substantial investment in terms of financial and human resources, especially in the case of big World Bank lending programs. Even when the request of support in the design of an impact evaluation spontaneously comes from the program manager, or the government, there is initial skepticism with respect to the methodology due to the fact

[5] Studies of the consequences of evaluation that are not a function of evaluation findings or recommendations were published even before Patton's definition of process use (Greene, 1988, Cousins and Earl. 1992 and Cousins and Leithwood, 1993)

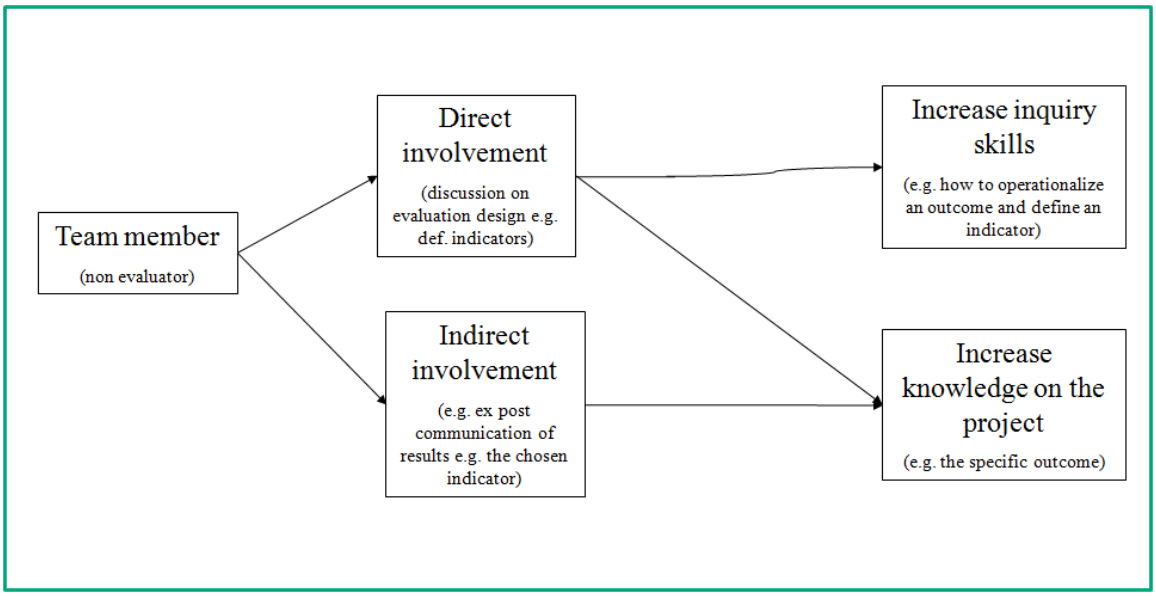
that the evaluation is seen as too expensive, too time-consuming, too academic and a limitation to program flexibility. The *change in thinking* occurs when the initial skepticism is substituted by total buy-in after many meetings between the evaluator and the program manager and team. In these occasions the evaluator explains the need to use a certain type of methodology to achieve a certain type of rigorous findings that would be accepted by the donor community to make funding decisions. Furthermore he assures the program team that the design wouldn't be imposed and that it actually needed to be the result of a partnership in order to fully grasp the dynamics of program implementation.

Having gained team's willingness to directly participate in the evaluation, time needs to be firstly devoted to a reality check. Especially when the evaluation is prospective when the *evaluability* assessment starts, there is not much clarity on how will the project be actually implemented. The existing program theory and log frame are generally vague and need to be re-assessed. They are usually prepared to request funds, not as actual implementation plans. This lack of clarity and specificity needs to be fixed in order to do an evaluation. The evaluator may or may not be involved in the actual process, but stimulates and supervises the theory-surfacing exercise of re-conceptualization of the program in a way that makes it evaluable. Objectives are redefined based on what could the program really achieve given the resources and time frame, the program theory and causal links between outputs/short and long term outcomes are carefully revisited. Generally speaking in order to answer the evaluator questions and allow him to think about the best design, the other stakeholders need to spend more time focusing on details of program design that they would have possibly overlooked before the actual implementation of the program *(change in behavior).*

Finally, changes in the program under evaluation *(change in the evaluand)* occur due to the necessity of aligning the evaluation to program operation, or due to the realization of possible improvements emerged by rethinking about the program through the analytical framework provided by the evaluation. For instance, changes could concern the roll-out plan when the design selected for an impact evaluation is experimental. In fact if one wants to achieve random selection of program beneficiaries, one solution can be that of phasing out the delivery in a random way instead of basing the decision of who gets the program first on other rules. Moreover, changes can occur when the necessity to have a big enough sample to obtain a statistically significant project impact can convince program managers to make a bigger investment in promotional activities in order to increase project take up.

All these changes would not occur without program management's understanding and buy-in of the impact evaluation logic. And if this understanding is the result of the evaluation process, I would argue that the involvement in the evaluation has also represented a form of ECB.

Figure 2. Effects of Direct (vs indirect) involvement in the evaluation process



As specified in the assumption, "being involved and directly experiencing the evaluation process" brings about the abovementioned changes. Process use and ECB both clearly occurs through active participation. Let's consider the following type of process use: the knowledge generated by discussing how to measure a particular outcome – e.g. youth risky behaviors. The fact that different stakeholders, such as the evaluation specialist, a youth development specialists, the program manager participate in the discussion of how to measure the outcome, not only generates the most appropriate indicator, but may also have other effects. Having to make explicit the definition of what is intended by risky behaviors might lead to a refined and common understanding of the concept itself whose tacit definition might have differed among different stakeholders. Moreover, the process of finding the correct indicator to measure a complex construct is something that requires to think empirically, to be realistic in what can/cannot be measured and quantified. To sum up, a discussion aimed at finding an outcome indicator is per se the generation of new skillful knowledge. Yet, if one of the possible stakeholders did not participate to the discussion on how to measure that outcome, they wouldn't have been involved in the learning process at all, and consequently would not have increased their evaluative inquiry skills by refining their capability to measure outcomes in general, nor would they have contributed to develop a common understanding of risky behaviors.

> *Proposition 4: The learning that occurs during the evaluation process is mostly individual and tacit, but can be enhanced, and be (partially) transferred to the team and organizational level by the intuition-reflection processes on the experience which generated the knowledge*

Quoting Preskill (2008): "learning from and about evaluation often requires us to change our mental models, to rethink our assumptions and beliefs and to develop new understandings about our programs and evaluation processes". So it is not a simple process. Learning is in the connection between action and reflection, in making knowledge actionable (implementation) and action knowledgeable (sensemaking) (Carroll et. al, 2003; Argyris et. al, 1985; Crossan et al., 1999). Even a well-used evaluation system does not generate learning automatically (Forss et al, 1994). Torres and Preskill (2001) list some specific catalyst that can help more organizations to adopt and implement an organization learning approach to evaluation, and suggest that the predominance and refinement of this approach will naturally occur when, as part of their ongoing work, evaluators and clients collaboratively *reflect* on the inquiry processes themselves. It is not particular tools such as root-cause analysis that lead to learning, but *rethinking* actions and assumptions in the context of new concepts that underlie the tools, such as data quality, rigorous cause-effect connections, systems thinking, mutual respect across groups, insight into personal and political relationships, and double-loop learning (Carroll et al., 2003). In the well-known distinction made by Argyris and Schoen (1978) between single-loop and double-loop learning, process use would qualify as deuteron-learning, a third level of learning which they call *deutero-learning*, that is, how to systematically generate and make use of single-and double-loop feedback mechanism. Especially if focusing on the learning by doing associated with process use, most of the knowledge acquired and learning occurred is tacit (as opposed to the knowledge generated by the evaluation findings which is mostly codified in order to be disseminated). Marra (2004) studied the contribution of evaluation to the socialization and externalization of tacit knowledge in line with Nonaka's theory (Nonaka, 1994; Nonaka et al. 2000). Marra's (2004) article contends that organizations always create new knowledge and that i) the process of creating evaluative knowledge takes place only when organizational members reflect on their actions, ii) participatory evaluation processes help share tacit knowledge through deep socialization. According to Nonaka and Takeuchi (1995) tacit knowledge is knowledge 'not yet articulated'. Yet, following Tsoukas (2003) I believe that not all tacit knowledge can be "articulated". Especially in the context of carrying out a specific task, we come to know a set of particulars without being able to identify them. In Polanyi's memorable phrase "we can know more than we can tell" (Polanyi, 1966 cited in Tsoukas). Yet, as explained by Tsoukas (2003) the ineffability of skillful knowing does not imply that we cannot speak about a practical activity at all. It is important to reflect on practical activities we engage in to draw the attention of those involved in it to certain unnoticed aspects of those activities, to see connections previously unnoticed. To sum up, tacit knowledge can be displayed, manifested, in what we do, and new knowledge comes about when our skilled

performance – our praxis – is punctuated in new ways through social interaction (Tsoukas 2001).

Continuing with the previous example on measuring risky behaviors, while the outcome indicator is tangible codified knowledge generated by the evaluation process, the refined understanding of the 'risky behavior construct' or the development of measurement skills are an example of Polanyi's ineffable skillful action. Most likely the individuals involved will not recognize it as learning, and they wouldn't be able to articulate the acquired knowledge on 'how to refine an impact indicator'. Moreover, even if they would recognize having learned a new skill, they might not attribute it to the fact of having participated in the evaluation discussion, but will rather credit it to their experience in dealing with youth issues. Yet, through *instructive forms of talk*, these practitioners would be moved to *re-view* their experience and possibly become more aware of the development of their evaluative skills. Reflection is an important human activity in which people recapture their experience, and think about it (Korthagen, 2005). In this context Shoen (1983) makes an important distinction between "reflection-in-action" (during the act) and "reflection-on-action" (after the act)

If the two conditions of participation, i.e. active involvement in the generation of the knowledge (i.e. intuition through participation) and recognition – awareness which lead to reflection-on-the-(skillful)-action and possible knowledge creation (new skills on how to measure an impact) and are not satisfied, this knowledge will remain individual and tacit. Moreover, individual learning is a necessary but not sufficient condition for organizational learning (Argyris and Schon, 1996, Senge, 1990). Outcomes of individual learning should be made comprehensible and available to others in order to have an impact of the core reflection process at the organizational level (Korthagen, 2005). According to Crossan et al. (1999), four social and psychological processes (intuiting, interpreting, integrating, and institutionalizing – 4Is) link the levels of organizational learning. What is particularly relevant here is that the first process of intuiting and interpreting requires awareness of the experiences. Intuition processes are difficult to describe in words (Korthagen, 2005). Motivation and willingness to reflect on these experiences are fundamental factors in activating the learning cycle (see Fig.3).

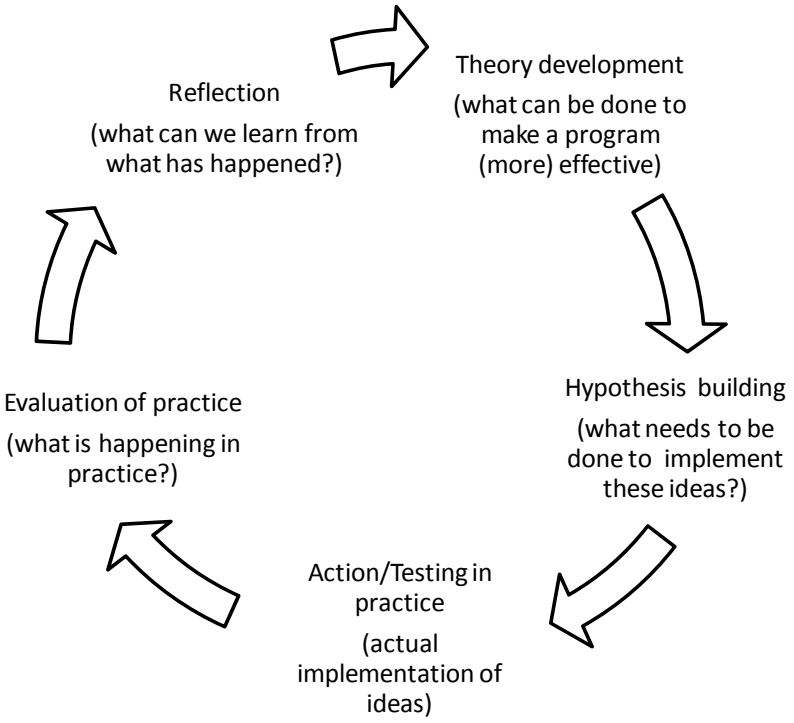
Moreover, the process of integration and institutionalization are respectively group-level and organizational level phenomena. In particular, integrating is the process of "developing shared understanding among individuals and taking coordinated action through mutual adjustment". If the coordinated action taken is recurring and significant, it will be institutionalized. Consequently, in order to stimulate organizational learning on the effects of being involved in an evaluation, integration and institutionalization also need to be stimulated.

This is not to say that with awareness all the learning will automatically occur. Two caveats. First, if the learning both at the individual and at the organizational level will

occur depends on the "absorptive capacity[6]". At the individual level, this ability depends on one's previous knowledge structures. I argue that the ability to absorb process use, as in, for example, strengthening the program or enhancing shared understanding, requires a very low level, or no level of expertise in evaluation. On the other side, evaluation capacity, as in the capability to conduct evaluation activities, requires a minimum level of training before the actual involvement in the evaluation for learning to occur.

With respect to the organizational level, there needs to be a specific organizational context for this learning to occur. Preskill and Torres (1999) have discussed this as the organization's infrastructure, and have identified the dimensions of culture, systems and structures, leadership, and communications as key factors that may enhance or inhibit learning from and about evaluation. I agree with the author's concern that if we fail to consider these various elements, sustainable evaluative thinking and practice will be severely compromised[7].

Figure 3. The learning cycle



Adapted from Osborne and Brown (2005)

Proposition 5: Stakeholder's buy into the evaluation, willingness and motivation to participate in the evaluation process and to re-view and re-

[6] The term coined by Cohen and Levinthal (1990) in context of the private sector as a firm's "ability to value assimilate and commercialize new knowledge".

[7] Preskill and Torres (2000) developed an instrument to measure Readiness for Organizational Learning and Evaluation that could be particularly useful in assessing organizational learning capacity

think actions and assumptions and learn from experience is key to fully stimulate learning

The complexity and pace of change of modern organizations requires more than a desire to learn. Special circumstances for learning end concepts and techniques that make learning more efficient are needed to break through long-held assumptions and cognitive habits (Carroll et al., 2003). This is true especially in large bureaucratic organizations. As discussed by Forss et. al (1994), in order to promote learning of a wide number of people in the organization, it is important to recognize the value of the learning process that do exist, such as process use. Yet, those new to evaluation might need help and facilitation in coming to view the experience as valuable (Patton, 2004). Perceptions, interpretation and motivating beliefs can impact individual learning (DeFilippi and Ornstein, 2003). Consequently, I suggest that even if the evaluators are the ones that, due to their experience, can draw other stakeholders' attention to certain evaluation processes make them "see connections" (Wittgenstein, 1958), there needs to be willingness and motivation to learn, in order to grasp the learning effects of process use.

5. Discussion and Conclusions

Levin (1993) distinguishes three purposes of collaborative research: increasing use; grounding data in practitioner's perspectives; and mobilizing for social action. Patton (2008) and Preskill (2008) identify a fourth objective: building evaluative capacity. Focusing on learning from the process as one (not the only) purpose of evaluation, by creating awareness and stimulating reflection on the learning and program development effects of stakeholder's involvement in an evaluation doesn't necessarily occur at the expenses of the resources devoted to generating the evaluation findings. This assumption is based on the following two points. First, in a continuum of investment of resources (financial, temporal and cognitive) directed towards the purpose of improving the collective understanding of action-performance linkages, which goes from the minimum required by the learning (by-doing) that happens in a semi-automatic fashion, to the efforts involved in the codification processes of the knowledge generated by experience, there is a middle way represented by knowledge articulation processes (de-briefing) (Zollo and Winter, 2003). Second, even if true that the evaluator should facilitate the process of awareness generation, I assume that program managers should be the ones responsible for the process of knowledge articulation and reflection which will lead to fully exploiting learning during the involvement in evaluation processes for the program implementation team. As opposed to Henry's (2000) view that "the potential of evaluation is more likely to be realized if informing rather than influencing policies and programs is the criterion for success", I don't think that one excludes the other. I actually

do not believe that it is possible to inform without influencing (Paruzzolo 2009)[8] ECB and PU are the result of making the "Virtue out of Necessity". Evaluations which aim at knowledge development by assessing the merit or worth of a program can achieve their purpose, while having an impact at the program and organizational level as well, and without diverting resources aimed at generating that knowledge. I don't agree with Henry's take that pursuing use generally will lead to an overemphasis on organizational and program-oriented evaluations, since I believe that most rigorous evaluation have by default an impact at the program and possibly organizational level. Even in the least participatory of the external evaluations, project stakeholders' involvement is critical for a) identifying the relevant policy questions, b) ensuring the internal integrity of the evaluation – helping to make sure that the evaluation keeps up with the reality of the implementation. Evaluation specialists need to work with stakeholders, on evaluation options, methods, and execution from the start. This represents an opportunity to strengthen the evaluand and further develop stakeholders' capacity for evidence-based policymaking. Why not taking this opportunity? Given the systematic time and budgetary constraints the question is how much human, organizational and financial resources to employ in order to perform planned and structured ECB activities. My opinion is that "none" is not an option. Yet I know from direct experience that it is mostly the case, especially in the international development arena. In that instance I would suggest to exploit the awareness and reflection processes on what occurs, by default in the evaluation processes. As discussed, the process of building evaluative thinking and learning from experience, takes place only when individuals reflect on their actions (Marra, 2004). Reflection is necessary to learn by making action knowledgeable, and may also activate knowledge transfers between the different levels of the organization. Reflecting, documenting and sharing experiences on evaluation processes is needed to transport learning within and across various organizational contexts.

The extent of the interaction and participation within evaluation process is a choice that has to be considered in an organization's advancement strategy. I argue that impact evaluations built into program operations requires a significant level of interaction of the different stakeholders, so the managerial choice is limited with respect to the extent of the participation within the evaluation process. Where the manager can make a difference with his/her strategic choice with respect to evaluation use, concerns the investments in sense-making and reflection on the evaluation process and the possible effects in terms of building evaluation capacity and an improved understanding of the program itself.

The skillful knowing acquired by the involvement in an evaluation is a skill that can have a more enduring value than evaluation findings (Patton, 2008). Also, evaluations take a relatively long time to show results. The average politician, bureaucrat, World Bank employee, or foundation program officer will not be in the same job in five years.

[8] Here I would cite paper 3 "The virtue of necessity" Paruzzolo (2009), the case study of the experimental design of the PBF evaluation

The resulting focus on the short run tends to lead to systematic under-investment in knowledge creation (Levine, 2006). I think that the focus on the short run is not something that will change easily and stakeholders need to perceive the usefulness of evaluation "as an opportunity to discuss problems openly, reflect critically and criticize constructively in order to learn what changes are needed to enhance impact (IFAD, 2002)". Raising awareness on PU and ECB as a product of evaluation can facilitate this vision. The fact that PU and ECB should be intentional has been discussed by different authors (Preskill et. al, 2003; Preskill and Boyle, 2008; Harnar and Preskill, 2007, Patton, 2007; 2008). I maintain that in an ideal world evaluation efforts should be always accompanied by ECB activities that would increase stakeholder's receptivity and "create a global cascade of evaluative thinking and practice" (Preskill and Boyle, 2008). Yet in a real world were evaluators due not always have enough resources to conduct primary activities such as data collection, i) it is still recommended, but is not strictly necessary to induce process use and ECB intentionally; but ii) it is necessary to intentionally create awareness and stimulate reflection in order to fully exploit the learning generated by the changes in thinking, behavior and the program under evaluation, that will most likely occur for stakeholders involved in an evaluation process.

References

Argyris C., Schon D.A. (1978), *Organizational learning: a theory of action perspective*. Reading. MA: Addison- Wesley.

Argyris, C. and Schön, D. (1996) *Organizational learning II: Theory, method and practice,* Reading, Mass: Addison Wesley

Argyris C., Putnam R., and McLain S. D. (1985), Action science: concepts, methods, and skills for research and intervention, San Francisco: Jossey-Bass

Alkin M.C. (2004), "Context adapted evaluation: a personal journey", in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Alkin M.C. and Christie C.A. (2004), "An evaluation Theory tree", in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Alkin M.C., Taut S.M. (2003), "Unbundling Evaluation Use", *Studies in Educational Evaluation*; 29: 1

Amo C. and Cousins J.B. (2007), "Going through the process: an examination of the operationalization of Process Use in Empirical Research on Evaluation", *New Directions for Evaluation*, 116: 5-26

Baker J. (2000), "Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners" Washington DC: The World Bank.

Bartel C.A. and Garud R. (2003), "Narrative Knowledge in action: adaptive abduction as a mechanism for knowledge creation and exchange in organization", in Easterby-Smith M. and Lyles M. A. (2003), *The Blackwell Handbook of Organizational Learning and Knowledge Management,* Blackwell, Oxford

Carroll J.S., Rudolph J.W., Hatakenaka S. (2003), "Learning from organizational experience" in Easterby-Smith M. and Lyles M. A. (2003), *The Blackwell Handbook of Organizational Learning and Knowledge Management,* Blackwell, Oxford

Chen, H. T. (1994). "Current trends and future directions in program evaluation". *Evaluation Practice*, 15: 229

Chen K-N (2006), "Library Evaluation and organizational learning. A questionnaire study", *Journal of Librarianship and Information Science*, 38(2): 93

Thayer C.E. and Fine A.H. (2001), "Evaluation and outcome measurement in the non-profit sector: stakeholder participation", *Evaluation and Program Planning,* 24: 61

Carden F. and Earl S. (2007), "Infusing Evaluative Thinking as Process Use: The Case of the International Development Research Centre (IDRC)",

Cousins J.B. (2004), "Crossing the bridge: toward understanding use through systematic inquiry", in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Cousins J.B. and Earl L. (1995),"Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning, London: Falmer

Cousins J.B. and Leithwood K.A. (1986), "Current Empirical Research on Evaluation Utilization", *Review of Educational Research*, 56(3):331

Cousins J.B. and Whitmore E. (1998), "Framing participatory evaluation", *New Directions in Evaluation*, 80: 5

Cousins, J. B., and Shulha, L. M. (2006). "A comparative analysis of evaluation utilization and its cognate fields" in I.F. Staw, M. M. Mark & J. Greene (Eds.), *International Handbook of Evaluation,* Thousand Oaks: Sage.

Crossan M., Lane H.W., and White R.E. (1999), An organizational Learning Framework: From Intuition to Institution. Academy of Management Review, 24(3): 522-37

DeFilippi R. and Ornstein S. (2003), "Psychological Perspectives underlying theories of organizational learning", in Easterby-Smith M. and Lyles M. A. (2003), *The Blackwell Handbook of Organizational Learning and Knowledge Management,* Blackwell, Oxford

Fetterman D.M. (2003), "Branching out or standing on a limb: looking to our roots for insight", in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Forss K., Cracknell B., Samset K. (1994), "Can Evaluation Help an Organization to Learn?" *Evaluation Review*, 18(5): 574

Forss K., Rebien C.C., and Carlsson J. (2002), "Process Use of Evaluations: Types of Use that Precede Lessons Learned and Feedback", *Evaluation*, 8(1): 29

Forss and Reiben (2006) "Evaluation, Knowledge Management and Learning: Caught between Order and Disorder", in Stame N., Rist R. (2006) "From Studies to Streams: Managing Evaluative Systems (Comparative Policy Evaluation)", Transaction Publisher

Harnar and Preskill (2007), "Evaluator's descriptions of process use: an exploratory study", *New Directions for Evaluation*, 116: 27-44

Henry G. (2000), "Why not Use?", New Directions for Evaluation, 88: 85

Jenlink P. M. (1994), "Using evaluation to understand the learning architecture of an organization", *Evaluation and Program Planning*, 17(3), 315

King J.A. (2004), "Tikkun Olam: the roots of participatory evaluation", in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

King J.A. (2007), "Developing Evaluation Capacity Through Process Use", *New Directions for Evaluation*, 116: 45

Kirkhart K. (2000), "Re-conceptualizing evaluation use: An integrated theory of use" *New Directions for Evaluation*, 88: 5

Korthagen F.A.J. (2005)," The Organization in Balance: Reflection and Intuition as complementary Processese", *Management Learning,* 36:371

Kusek J. Z. and R. C. Rist (2004), Ten steps to a result-based Monitoring and Evaluation System, Washington DC: The World Bank

IFAD, International Fund for Agricultural Development, (2002) *A guide for project M&E: Managing for Impact in Rural Development,* Rome: IFAD

Levine D. I. (2006), "Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry", Berkeley: Haas School of Business, University of California.

Marra M. (2000), "How Much Does Evaluation Matter? Some Examples of Utilization of the Evaluation of World Bank's Anticorruption Activities", *Evaluation* 6(1): 22

Marra M. (2003), "Dynamics of Evaluation Use as Organizational Knowledge - The Case of the World Bank", PhD Dissertation, George Washington University

Marra M. (2004), "Knowledge: The Case of the World Bank The Contribution of Evaluation to Socialization and Externalization of Tacit", *Evaluation* 2004; 10: 263

Myers-Walls J. A. (2000). "An odd couple with promise: Researchers and practitioners in evaluation settings", *Family Relations: Interdisciplinary Journal of Applied Family Studies, 49*, 341-347

Nonaka I. (1994), "A Dynamic Theory of Organizational Knowledge Creation", *Organization Science*, 5(1): 14

Nonaka I. and Takeuchi H. (1995), "The knowledge-Creating company", New York: Oxford University Press

Nonaka I., Von Krogh G. and Kazuo I. (2000) *Enabling Knowledge Creation. How to Unlock the Mystery of Tacit Knowledge and Release the Power of Innovation*. Oxford: Oxford University Press.

OECD, Organization for Economic Co-operation and Development (2002), *Glossary of key terms in evaluation and results-based management, Paris: OECD/DAC*

Osborne S.P. and Brown K. (2005), *Managing change and innovation in public service organizations*. London: Routdlege.

Osborne D. and Gaebler T. (1992), *Reinventing government. Boston, Mass.: Addison-Westley Publishing*

Owen (2004), "Evaluation forms: Towards an Inclusive Framework for Evaluation Practice" in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Patton M.Q. (1997), *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Patton M.Q. (1998), "Discovering process use", *Evaluation*; 4(2): 225

Patton M.Q. (1999), "Organizational development and evaluation", *Canadian Journal of Program Evaluation, Special Issue,* 93

Patton M.Q. (2004), "The roots of utilization-focused evaluation" in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Patton M.Q. (2007), "Process Use as Usefulism", New Directions for Evaluation, 116: 99-112

Patton M.Q. (2008), *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Perrin B. (2006) "How evaluation can help make knowledge management real" in Stame N., Rist R. (2006) "From Studies to Streams: Managing Evaluative Systems (Comparative Policy Evaluation)", Transaction Publisher

Preskill H. (1994), Evaluation's role in enhancing organizational learning: A model for practice. *Evaluation and Program Planning*, 14 (2), 291

Preskill H. and Boyle S. (2008), "A Multidisciplinary Model of Evaluation Capacity Building", *American Journal of Evaluation*, Vol. 29(4) 443-459

Preskill H. and Caracelli V. (1997), "Current and Developing Conceptions of Use: Evaluation Use TIG Survey Results", *American Journal of Evaluation;* 18: 209

Preskill H. and Torres R.T., (1999), "Building Capacity for Organizational Learning Through Evaluative Inquiry" *Evaluation*, 5(1): 42

Preskill, H. and Torres T. R. (2000), 'The Learning Dimension of Evaluation Use', *New Directions for Evaluation*, 88: 25

Preskill H., Zuckerman B. and Matthews B. (2003), "An Exploratory Study of Process Use: Findings and Implications for Future Research", *American Journal of Evaluation*; 24: 423

Preskill H. (2004), "The transformational power of evaluation: passion, purpose, and practice" in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Preskill H. (2007) as President of the AEA, Invitation letter to the Annual 2007 AEA Conference on "Evaluation and Learning".

Preskill H. (2008), "Evaluation's Second Act", *American Journal of Evaluation*, 29(2): 127

Preskill H., Zuckerman B., Matthews B. (2003), "An exploratory study of Process Use: Findings and Implications for Future Research", *American Journal of Evaluation*; 24: 423

Rossi P.H. and Wright J.D. (1984), "Evaluation Research: An Assessment", *Annual Review of Sociology,* Vol. 10

Senge P.M. (1990), *The Fifth Discipline: The art and Practice of the Learning Organization*. London: Century Business.

Scheirer M.A. and Newcomer K. (2001), "Opportunities for program evaluators to facilitate performance based management", *Evaluation and Program Planning*, 24: 63-71

Schön D. (1983) *The reflective practitioner*. Basic Books: New York

Torres R.T. and Preskill H. (2001), "Evaluation and Organizational Learning: Past, Present and Future", *American Journal of Evaluation*, 22: 387

Tsoukas H (1996), "The firm as a distributed knowledge system: A constructivist approach" *Strategic Management Journal*, Vol: 17 (Winter Special Issue): 11-25

Tsoukas H. (2001), "Where does new organizational knowledge come from?" Keynote address at the *International Conference Managing Knowledge: conversations and Critiques, Leicester University,* April 10-11, 2001

Tsoukas H. (2003), "Do we really understand tacit knowledge?" in Easterby-Smith M. and Lyles M. A. (2003), *The Blackwell Handbook of Organizational Learning and Knowledge Management,* Blackwell, Oxford

Valovirta V. (2002), "Evaluation Utilization as argumentation", *Evaluation*, 8:60

Weiss C. H., and Bucuvalas M. (1980*),* "Truth Tests and Utility Tests: Decision makers' Frames of Reference for social science research", *American Sociological Review*, Vol. 45(1)

Wittgenstein L. (1958), *Philosophical Investigations*. Oxford: Blackwell

Zollo M. and Winter S.G. (2003), "Deliberate learning and the Evolution of Dynamic Capabilities" in Easterby-Smith M. and Lyles M. A. (2003), *The Blackwell Handbook of Organizational Learning and Knowledge Management,* Blackwell, Oxford

MODELLING STAKEHOLDERS' PERCEPTIONS OF EVALUATION PURPOSE, APPROACH & USE

Candidate: Silvia Paruzzolo
PhD in Business Administration and Management
XX cycle - Track: Public Management
Paper 2/3

Outline

1. **Introduction**
2. **Brief review of the relevant literature**
3. **Analytical framework**
4. **Study Context**
5. **Research Methodology**
   - ***5.1 Data collection***
   - ***5.1.1 The instrument: questionnaire development and pilot testing***
   - ***5.1.2 The sampling strategy: participant recruitment***
   - ***3.2 Analysis***
   - ***5.2.1 Construct measurement***
   - ***5.2.2 Empirical specifications and results***
6. **Conclusion and Discussion**
7. **Annex**

1. Introduction

The main objective of the paper is to investigate how mental models can influence perceptions on program evaluation and how this understanding can contribute the research efforts on the relationship between program evaluation and learning. Both concepts of evaluation and learning have been extensively discussed in the literature, but many questions still have to be answered about how organizations create process and translate into action, new knowledge, and in particular the knowledge generated through evaluations. And, ultimately on whether and how does evaluation contribute to this process (e.g. Preskill, 2004; Preskill and Torres, 2000; Torres and Preskill, 2001; Preskill et. al, 2003; Forss et al, 1994; Marra, 2000, 2004). The writings on organizational learning affirm the idea that evaluation, being an approach to organizational inquiry, could help organizations resolve critical issues and grow from the collective experience of its members (Preskill, 2004). Attention to the 'positive externalities' of engaging in an evaluation led to an increasing interest in the notion of process use, that is, in understanding the ways in which individuals learn about the *evaluand*, evaluation practice, and each other, from their involvement in an evaluation study (Patton, 1994, 1997, 2008).

Empirical evidence supporting the power of evaluation to foster organizational learning is mixed (Shulha and Cousins, 1997). The theoretical and empirical literature is not currently aligned, and efforts to enhance the understanding of this complex organizational dynamic are highly recommended. If evaluations are meant to be an organizational learning 'tool', they need to be viewed as such by the primary users, i.e. the ones that initiate, and benefit from, the learning process. Consequently research should verify if program practitioners involved in an evaluation of their program as primary stakeholders agree with the belief that i) evaluation appears to be most useful when it is conducted using participatory approaches, ii) process use occurs in practice, and is a relevant aspect of the evaluation process. There is a need for continuous understanding and dialogue about how evaluations operate in practice, and program practitioners represent a unique source of information to unfold these processes. Adding to the existing body of literature this present study will provide evidence on the perspective of stakeholders - other than the evaluators - on process use and learning - building empirical evidence on what actually occurs in practice with a more in depth understanding of the influence of background factors in framing such perceptions. This study answers to the call for quantitative studies that capitalize on estimations of latent structures, by using logit regressions and structural modelling to move the research agenda forward (Cousins et al., 2004).

2. Brief review of the relevant literature

Participatory approach to evaluation and process use

Some of the more innovative thinking about use arises from efforts to understand the impact of the evaluation process quite apart from the nature of findings (Shuhla and Cousins, 1997). Greene (1988) was the first to develop a grounded conceptual framework for use that is very much tied to process, and other authors have discussed the effects in their research (for example, Cousins and Earl, 1992). Yet Patton (1997) was the one who coined the term "process use" to describe "individual changes in thinking and behaving that occur among those involved in evaluation as a result of the learning that occurs during the evaluation process and "changes in program or organizational procedures and culture that may also be manifestation of program impact". In particular, in his most recent edition of "Utilization-focused Evaluation", Patton (2008) lists six primary uses of the evaluation process: 1 Infusing evaluative thinking into the organizational culture 2 Enhancing shared understandings 3 Supporting and reinforcing the program intervention through intervention-oriented evaluation 4 Instrumentation effects (what gets measured gets done) 5 Increasing participants' engagement, self-determination, and sense of ownership 6 Program and organizational development. This list is used as a frame in the present research to operationalize the construct of process use and select the items for the survey instrument questionnaire (see Annex). The common and fundamental assumption is that these changes result from users' engagement in the evaluation (Patton, 1997, 1998, 2008)[9].

The term participatory evaluation often is used interchangeably with collaborative and/or empowerment evaluation (O'Sullivan and D'Agostino, 2002). Cousins, along with colleagues (1992; 1996; 1998, 2004), has done considerable work in the area of collaborative and participatory evaluation. He defines collaborative evaluation as 'any evaluation in which there is a significant degree of collaboration or cooperation between evaluators and stakeholders in planning and/or conducting the evaluation' (Cousins et al., 1996). Building on the existing literature and adapting the definitions to the specific study context, in the present study, an approach to evaluation is defined as participatory, i.e. the evaluation is conducted in a participatory way by engaging all relevant stakeholders (e.g. government counterparts, project management) when: i) the evaluation design is aligned with project operations (i.e. the evaluation strategy does not limit the flexibility of the project implementation) ii) the decisions about the evaluation activities (e.g. the design, the information to be collected, the timing of data collection) are made collaboratively by all team members iii) frequent consultation with project management and local counterparts take place in order to design the evaluation iv) attention to

[9] Different interpretation exists of what Patton defined as process use, but has been also referred to as influence (Kirkhart, 2000) or change (Weiss, 1998; Henry and Mark 2003).

contextual factors is made possible by the active involvement of different stakeholders (see Annex for more details on the operationalization).

Yet if we maintain that one possible goal of evaluation, through its process as well as its findings, is to contribute to learning and improvement – in other words, to support favourable changes in social settings, the psychological dimensions cannot be underestimated. A psychology of use undergirds and informs utilization-focused evaluation (Patton 2004). The theoretical analysis of social and (organizational) psychological concepts underlying evaluation use need to be explored in order to increase the current understanding of the complex dynamics of process use and its consequences. Similar to organizational development, programme evaluation involves people who are entangled in social and organizational networks; evaluation affects the individual in his or her social context. Social and organizational psychological theories and concepts are therefore considered especially useful in answering questions about evaluation's impact on organizational learning. Findings, such as "the perceived value of evaluation as a management tool emerged as a likely predictor of use" (Shulha and Cousins, 1997), as well as "effective" communication during the evaluation increases the likelihood of organizational learning through evaluation (e.g. Cousins and Leithwood, 1986; Forss, 1994; Preskill and Torres, 1999), point out the relevance to focus on psychological dynamics within the evaluation team. I argue that one fundamental step in this direction is to analyze stakeholder's take - and differences with the evaluators perspectives - on what occurs in practice and the value they attribute to the different purposes of evaluation, participatory approach and process use, in order to build a theory of how to stimulate use and learning.

Different cultures, different mental models

If that process use occurs when evaluations are carried out in a participatory way, it means that we are looking at team dynamics and we need to take into consideration the pluralistic values that different stakeholders bring to social programs and the necessarily political context in which both programs and evaluations are mounted (Lincoln and Guba, 2004). The idiosyncratic nature of evaluation teams calls for more attention, given the very heterogeneous range of skills is needed in evaluation work. Just thinking about evaluators and program managers, the skills required to implement programs are quite different from the skills required to design a rigorous evaluation (Levine, 2006). It has been widely discussed that the research-practice gap is partially due to the fact that researchers and managers have different frames of reference with respect to such things as the types of information believed to constitute valid bases for action, the way in which information is ordered and arranged for "sense-making", the past experience used to evaluate the validity of knowledge claims and the metaphors used to symbolically construct the world in meaningful ways (Rynes et al, 2001; Shrivastava and Mitroff, 1984). Not taking this into consideration can represent a great limitation in the implication of research findings

on their shared learning processes. The two groups, which need to collaborate in the evaluation process, represent different cultures, live in different worlds, with different traditions, rules and expectations. Cultural differences are seen in temporal orientation, cognitive resources or ways of "knowing", values, and definitions of excellence, patterns of communication, daily life-styles, and use of tools (Myers-Walls, 2000). These differences can shape the way evaluation purposes, and processes are interpreted. As highlighted by Marra (2000) and others, "there is disjuncture between the benefits desired by users in the short run and those promised by evaluators"; in her case study the following emerged: "managers asserted that by the time the report was finalized, the evaluation findings were already outdated because the program had undergone many changes". In the same way, it appears that both practitioner's and evaluator's value process use (cite articles) but it might be that each values different aspects of it. Also, the recognition that among other things, our background shapes what we know, how we know it and how we frame that knowledge when we share it with others is part and parcel of the experience of a more realistic program evaluation theory (Lincoln and Guba, 2004).

Mental models in evaluation teams

I argue that these differences between evaluators and non (evaluators) might reflect different mental models. The term *mental model* has been used as an explanatory mechanism in a variety of disciplines over the years (see Wilson & Rutherford, 1989). Essentially, mental models are organized knowledge structures that allow individuals to interact with their environment. Specifically, mental models allow people to predict and explain the behavior of the world around them, to recognize and remember relationships among components of the environment, and to construct expectations for what is likely to occur next (see Rouse & Morris, 1986). Furthermore, mental models allow people to draw inferences, make predictions, understand phenomena, decide which actions to take, and experience events vicariously (Johnson-Laird, 1983). Shared mental models allow team members to draw on their own well-structured knowledge as a basis for electing actions that are consistent and coordinated with those of their teammates (Mathieu et al. 2000). Especially under conditions in which communication is difficult--because of excessive workload, time pressure, and inherent conflicts of interest that necessary emerge when evaluating a program (see Pritchett, 2002), teams might not able to engage in necessary strategizing. In this case, shared mental models become crucial to team functioning because they allow members to predict the information and resource requirements of their teammates. Hence, members are able to act on the basis of their understanding of the task demands and how these will affect their team's response. Widely different mental models suggest that teammates work toward different objectives and predict different future system states, and therefore have difficulty coordinating their efforts. Yet teammates tend to exhibit greater sharedness over time as they gain

experience both with the task and with each other. Rentsch et al. (1994) found that more experienced team members conceptualized teamwork more concisely and coherently than did less experienced members. Dyer (1984) advanced similar conclusions and added that with increased experience teams are able to better coordinate their efforts and therefore exhibit better team processes and performance.

Concluding, the literature on mental models suggests that i) if team members do not share the same mental models they might interpret things differently, and ii) the accumulation of experience increases sharedness.

Empirical research on process use

Amo and Cousins (2007) conducted an extensive review of the empirical literature on process use and highlighted different limitations and direction for further research. In the present study I intend to contribute to this effort by: i) providing a measure of process use which limits the ambiguity of the analysis due to the variability of understanding of the concept of process use documented by Harnar and Preskill (2007) ii) documenting not only the occurrence, but also the intensity of process use iii) linking process use and the level of involvement of stakeholders in the evaluation process iv) offering a innovative research design by surveying a larger sample and not focusing only on case studies v) presenting multiple interpretations surveying perceptions of evaluators and non in light of the same operationalization of complex concepts such as process use and participatory approach. Studying process use empirically presents substantial methodological challenges (see Cousins et al., 2004; Preskill et al, 2003; Amo and Cousins, 2007) that I will address in the methods section.

3. Analytical framework

The analytical framework tested in the present study is based on the following assumption:

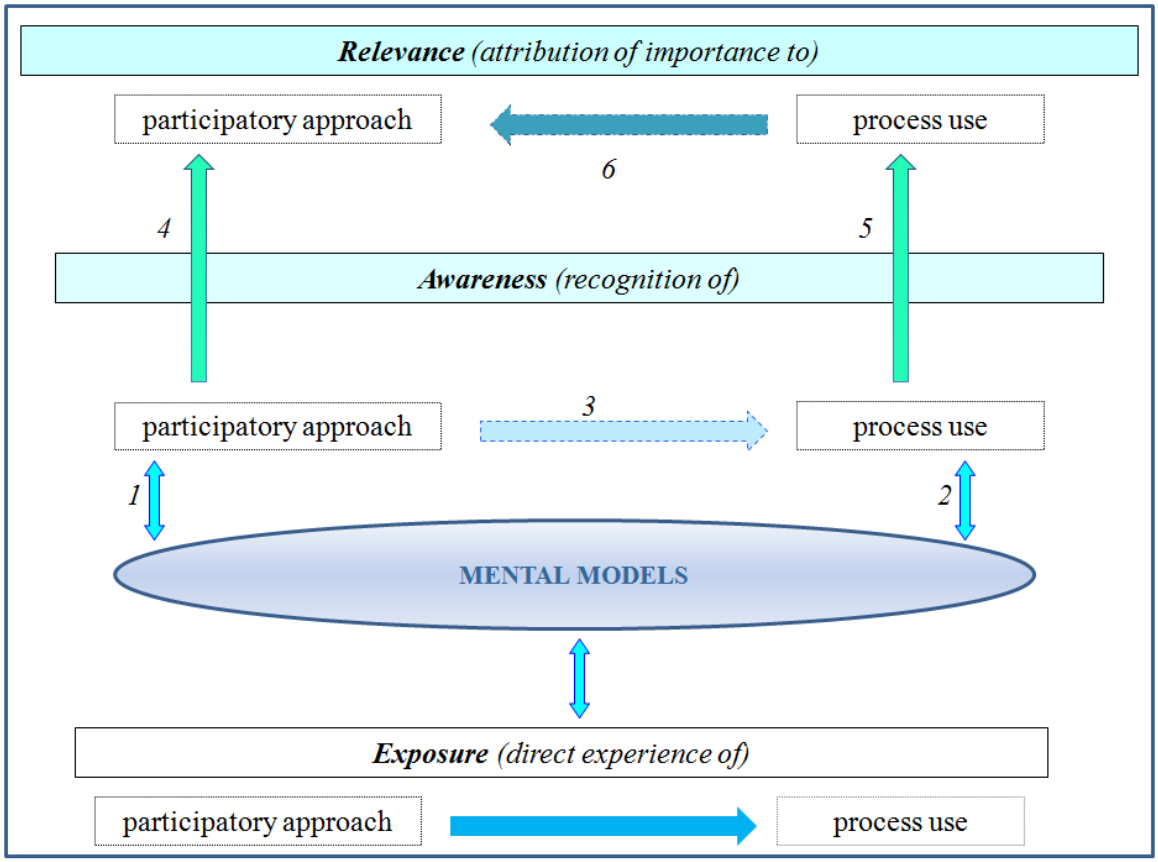
> *A range of specialized skills is employed in the evaluation team work. Members with different roles and different levels of experience with evaluations might not share the same mental models, in which case they will draw on their own knowledge structures in interpreting the evaluation process and selecting actions to carry out the evaluation task.*

The guiding question of this work is: *Can perceptions on participatory approach to evaluation, process use and use of findings be explained by one's role in the evaluation team, and previous experiences with evaluations?*

The framework is depicted in Figure 1 below. I hypothesize that perceptions on the evaluation process can be explained by member's mental model which is framed by the role in the evaluation (particularly evaluators versus non) and level of experience with

evaluation, participatory approaches, and process use, controlling for academic background. Moreover, awareness of, and the relevance attributed to, certain phenomena depends on having directly experienced them. In other words, as depicted in the figure below, the level of awareness depends on having actually experienced a certain event, and it is influences by each person mental models. In other words, one would recognize a certain event only after having experienced it, but might not be able to recognize it as such even after having experienced it, due to his/her own frame of reference (arrows 1-2). The level of awareness of process use depends on the level of awareness of participatory approaches (arrow 3): in other words, having experienced a participatory approach is a necessary condition to having experienced and being aware of process use. Moreover, if one has experienced a participatory approach and process use, he or she is more likely to attribute relevance to both (arrows 4-5). Finally, assuming that one would recognize the relevance of process use, more importance would be attributed also to participatory approaches, as the necessary condition for process use to occur (arrow 6). This framework will be tested in different steps and ultimately the path analysis will attribute coefficients to the arrows in the analytical framework (the graphical representation will be different given that there is a standard for describing path analysis results).

Figure 1. The Analytical Framework



In particular, the following set of hypothesis will be tested:

> *HYPOTHESIS #1*
> *Evaluators' perceptions on the purposes of evaluation differ from those of non-evaluator in the following manner:*

Proposition 1: Use of findings is generally more valued than process use as a purpose of evaluation
Proposition 2: Non-Evaluators attribute more importance to process use than evaluators

> *HYPOTHESIS #2*
> *Perceptions on participatory approach and process use differ depending on one's role (evaluator versus non-evaluator) and level of experience with evaluation.*

Proposition 1: Evaluators and stakeholders with a higher experience in evaluations show a higher level of awareness of both participatory approach to evaluation and process use than non-evaluators and stakeholder's with a lower level of experience with evaluation

Proposition 2: Evaluators and stakeholders with a higher experience in evaluations attribute more relevance to both participatory approach to evaluation and process use, than non-evaluators and stakeholder's with a lower level of experience with evaluation

> *HYPOTHESIS #3*
> *The relevance attributed to both participatory approach and process use can be explained by the perceived experience of this events controlling for background variables*

Proposition 1: One's (awareness of) direct experience of process use depends on his/her (awareness of) direct experience of a participatory approach to evaluation controlling for his/her role, academic background and general experience with evaluation

Proposition 2: The relevance attributed to participatory approach and/or process use can be predicted by one's (awareness of) previous experience of participatory approach and evaluation process use

4. Study Context

As discussed by Forss et al (2004), the field of development cooperation is particularly interesting for organizational research, and I would add, in this particular moment, for

evaluation theory. Foreign aid is a vast and complex industry, including multilateral agencies such as WHO; bilateral agencies (the largest of which is USAID); global non-governmental agencies, regional, national, and local NGOs; national and local governments, and many other layers. Ideally, all of the learning produced by the projects these organizations fund would be shared and used by this entire complex (Levine, 2006). Among other, the need for organizational learning is particularly high for the following reasons: 1. The activity is international by definition, i.e. the program is prepared and implemented in two different countries. 2. The environment is highly turbulent in terms of technology, politics, economics, and social structures. 3. The objectives are contradictory and confused (Forss et. al, 1994). Unfortunately, evaluations have traditionally been designed to measure whether the money was spent as planned — as an accountability instrument, not to measure whether the spending was effective at achieving the stated outcomes. It looks like aid organizations have so far failed to learn from experience. To see why learning is not integrated, it is helpful to examine The World Bank (hereafter WB), the world's largest funder of development projects. In 1996, Bank President James Wolfensohn announced his vision for a "Knowledge Bank", a great repository of information about development to be gathered and disseminated by the WB. Over the next few years the Bank took many important steps to advance this goal (see, e.g., Wood and Hamel 2002). But despite some progress, the Bank has not yet institutionalized learning in its major activity: making loans to facilitate major development projects. Let's look at the process. A loan begins with a program proposal worked out by a team of Bank officials and national officials. The Bank officials are rewarded for success in getting loans out the door (Levine, 2006). Thus, time spent designing a rigorous evaluation impedes the Bank officials' achievement of their lending goals. Government officials face similar incentives to reach a deal. Bank and government officials typically devote a great deal of time to preparing the documents describing the various projects that will be funded by a loan, and provide little attention to designing a logical framework and an M&E plan, much less a rigorous evaluation. Yet, things are changing. The stimulus for this study came from the author's own experience working on impact evaluations at the WB, while the international organization was (and still is) experiencing a particular momentum as far as impact evaluations and producing evidence of what works is concerned. In fact, in the past 4 years the WB interest and investments in evaluations increased substantially, as it is proved by the growing number of impact evaluations and funds devoted to this type of evaluations[10]. The WB is implementing innovative approaches for mainstreaming impact evaluation in the early stages of project preparation.

Proposals are still written to get loans and not much attention nor budget is devoted to the M&E of the programs funded by the loan. Yet, the growing demand for generating evidence on which to base decisions about future investments has created easier access to

[10] Add data on number of impact evaluations and web page with Trust Funds specific to IEs

experts in evaluation, and funding for Bank and local government staff that wanted to design an evaluation. Also, while until 4 years ago the large majority of evaluations was designed long after loans are given and long after the funded projects have been designed and put in place, more recently there has been a strong push for prospective impact evaluations based on the estimation of a counterfactual (i.e. what would have happened without the program). The evaluation is designed ex-ante and built-in project operations. The evaluation team is composed by evaluation specialists, the program manager, the field implementation team, the government counterpart and other specialists, such as sampling experts or specialists of the area of program implementation. The collaboration focuses in the design of the evaluation by defining the identification strategy (i.e. selecting treatment and control groups), the analytical framework, program outcomes, indicators and measurement instruments. Also, during program implementation, the collaboration continues in order to assure that the program and the evaluation strategy are implemented as planned (e.g. control and treatment are not confounded etc...). Data analysis, interpretation and dissemination are also made in consultation in order to assure meaningfulness of results and provide policy feedback.

Ultimately, as suggested in the toolkit "Impact evaluation and the project cycle" (World Bank, 2006) the evaluation specialist should become intimately familiar with the project, the country and institutional context, the design options that are being considered, and the details of the roll-out and execution. On the other side, the program manager and client need to buy into the logic of the evaluation to understand what project design and implementation elements are critical for carrying out an evaluation that will contribute to improving the success of the project. Ideally, these two sides should come together as the manager, evaluator and counterparts work through the design choices of the project to identify which of these choices need to be tested. An interesting quote from the concluding remarks of a presentation at the World Bank: "…making evaluation work requires a change in the culture of project design and implementation, one that maximizes the use of learning to change course when necessary and improve the chances for success…Impact evaluation is more than a tool is an organizing analytical framework for doing this and can add value to the project under evaluation" (Legovini, 2006).

In this phase of particular attention to impact evaluations, the WB is also providing learning events to its staff and government counterparts. The involvement of different stakeholders (i.e. participatory approach) underlies the approach taught at the learning events. In fact, a fundamental factor of success of most experimental and quasi-experimental impact evaluation techniques is the buy in of program management and government counterparts which allows for the alignment of the evaluation design with project operations. These events will be described in more detail in the sampling strategy paragraph.

5. Research Methodology

This Section describes the instrumentation, the sample and the procedures utilized to investigate the hypothesis.

5.1 Data collection

5.1.1 The instrument: questionnaire development and pilot testing

The instrument was developed and pilot tested by the researcher. The content of the questionnaire was initially derived from a survey of the relevant literature, looking at the instruments previously used for similar studies (Preskill and Caracelli, 1997; Cousins et. al, 1996; Taut, 2007), plus the researcher's own knowledge and experience working on evaluations of WB programs. The validity of the instrument was addressed in two ways. Firstly, by comparing the content of the questionnaire with those used in the past by other researchers, and secondly, by conducting semi-structured interviews with six key informants within the WB (mainly senior staff involved in evaluation related activities). Each interview lasted between 40 and 60 minutes, and the objective was to test the relevance of the research question and validity of the instrument. Each interview was tape-recorded and transcribed and informed the development and finalization of the questionnaire. The instrument was examined by native English speakers who worked at the WB and had different levels of experience with impact evaluations (i.e. not necessarily familiar with evaluation jargon and able to assess understandability for non-specialists as well). The questionnaire was then translated into Spanish, with the effectiveness of the translation being checked by native speakers. The researcher personally tested the translation by back-translating the instrument from Spanish to English.

The instrument was pilot tested for reliability, readability and understandability using a sample of 53 participants to WB events on impact evaluations. 21 questionnaires were filled. Data from the pilot was used to ensure that the wording of the instructions made sense and expressed similar meaning to all participants removing sources of possible confusion. The phrasing, order and number of items of the questionnaire were revised. Ambiguity was reduced as far as possible. The length was also reduced by eliminating questions that presented no variation or were left blank by almost all respondents. The pilot was also used to test the sampling strategy. The expected variation among respondent in terms of backgrounds, role in the evaluation team, and experience with evaluation, was achieved.

With respect to the content, the questionnaire surveyed perceptions and experiences of evaluation team members on i) different purposes of evaluation, ii) process use/use of findings iii) participatory approach of evaluations. It is structured in 2 main Sections, plus an open-ended comments Section at the end. The first Section contains background questions to be used as control variables, such as respondent's academic background (economics, management, health/social protection specialists, etc.),

most frequently played role in the evaluation team (project manager, part of implementation team, evaluation specialist), and overall experience with evaluations (high, medium, low, none). The second Section contains three main questions that collect *Likert scale* responses on a total of 35 items intended to capture respondent's opinions and perceptions on their experiences with program evaluations.

Finally respondents had an open-ended Section for comments at the end where they were able to add their thoughts on evaluations. A vast majority of respondents used this opportunity for further communication. The data generated was analyzed and used as a source of additional useful information instrumental in enhancing the interpretation of the *Likert* scale responses.

5.1.2. The sampling strategy: participant recruitment

Given the nature of the study, the sample was not random, but purposefully selected to possess certain predefined characteristics, such as a minimum level of experience with evaluations. Yet the aim of the study was to collect experiences and perceptions of different stakeholders, not only evaluators. Also, since process use is an outcome which may vary quite significantly, the author decided to collect perception of respondents exposed to a specific type of context and approach to evaluations, which as discussed above, generally requires high participation by different stakeholders. Finally, the perceptions were studied in "real time", which is uncommon in research on evaluation use (Ciarlo, 1981), but is especially helpful to examine the use of the evaluation process as it is indicated by cognitive, affective and behavioral changes (Preskill et al., 2003).

The main study sample was recruited from: i) the database of program managers and evaluation principal investigators of ongoing WB impact evaluations and ii) the database of participamts of three international WB workshops "Impact Evaluations: Turning Promises into Evidence" organized in Colombia, Mexico and Argentina in 2006/2007. The country based workshop approach aims at mainstreaming impact evaluation in the early stages of project preparation. It brings together project managers with government counterparts and carefully selected local and international researchers. The participants are exposed to international practice and methodological training on impact evaluation, and are required to apply this knowledge to the design of impact evaluation for specific projects. Project teams and government counterparts define priority questions for impact evaluation and design the framework to be adopted under the guidance of impact evaluation experts. Research teams are formed and project-specific collaboration agreed in principle. Given the compositions and approach of the learning activities, participants were considered a particularly approriated sample for investigating the perspectives of different stakeholders involved in the evaluation process, their preceptions and the correlation with their experiences. To ensure a reasonable sample size, another sample was also recruited from participants of 'evaluation clinics', i.e. informal sessions that support program managers in the design of

impact evaluations for their projects which are based on the same approach of the workshops but are organized in the WB headquarters and usually connect government counterparts via videoconference.

Participants were contacted 10-12 months after the event in order to collect information on the status and possible challenges encountered by the evaluations discussed at the workshop/clinic. The manager of the unit organizing these activities was asked permission to combine the standard folow-up activities with the survey designed for the current research. An email (prepared by the researcher) with an explanatory cover letter soliciting voluntary participation was sent by the manager. The e-mail explained the purpose of the study and provided an *htlm* link to the online study instrument. The instrument was hosted on a popular web-based survey service. Participants did not receive compensation for participation.

All the participants to the workshops and clinics described previously were asked to participate in the study. The participants were of course free to contact the researcher if there was any difficulty involving completion of the questionnaire. Data were collected over a 4 month period May-August 08, and a total of three reminders were sent. Ultimately a total of 581 individuals were contacted to participate in this study, and 197 (34%) completed the questionnaire. Some 16 returns were discarded because they had not been filled out correctly.

Descriptive statistics of the sample is presented in Table 1. More than half (54%) of the respondent were government officials, 33% were World Bank staff, while the other category which accounted for 13% of the sample was working at universities, centers of reasearch and foundations. 13% where coming of the In terms of academic background, almost half of the sample has a degree in economics, while only 4% has a degree in evaluation. With respect to the self-reported level of experience with evaluation, 48% declare to have a medium level of experience, and only 16% overall have a low level of experience with evaluations (all have been part of an evaluation team at least once because of how the sample was selected, i.e. from a list of participants to evaluation projects). 40% define themselves as evaluation specialist (I will consider them as evaluators) while the rest represents the other stakeholders involved in the evaluation process: program managers (25%) and more generally members of the project implementation team (35%). To be noted that crosstabulations of the data show that evaluation specialists do not necessarily declare to have a high level of experience with evaluations. This will be taken into consideration in the analysis.

5.2 Analysis

5.2.1 Construct Measurement

The methodological challenges when investigating perceptions is related to how respondents interpret complex and ambiguous concepts such as use and learning

especially given that we argue that they might interpret things with different mental models. For this reason, in the process of *operationalization*, the author's main goal was that of specifying the dimensions of the constructs for which the lexicon might be ambiguous in order to limit the number of possible alternative interpretations of the results, by knowing exactly how respondents interpreted the concepts and whether their perceptions are based on similar assumptions about participatory approach, process use, use of findings and learning from evaluations (see Preskill et al, 2003 for a discussion on the methodological challenges of measuring concepts such as process use and learning).

Especially given the decision to survey a sample which does not include only evaluation experts, the author made the assumption that the lexicon associated with process use might not be as familiar to the respondents as the better known category of use of findings. The survey provided an opportunity to operationalize process use and to determine whether practitioners recognized and attached importance to the specific dimensions selected to define the construct. Also, to overcome the possible limitation concerning respondents' perception and understanding of concepts such as "organizational learning" (see Preskill and Caracelli, 1997) it has been defined as change of organizational knowledge, which is not a unanimous definition, nor the author believes it is exhaustive, but it focuses the responses on one possible aspect of a complex phenomenon as organizational learning.

The main concepts of i) participatory approach to evaluation, ii) process use of evaluation iv) use of findings, where operationalized by questionnaire items as described in the Annex. Synthetic measures were elaborated to account for the level of awareness and relevance attributed to each of the concepts. For the calculation of measures and indexes, the *Likert* scale responses were coded in the following manner: Strongly disagree/Not at all important = 1, Disagree/Little important = 2, Somewhat agree/ Somewhat important = 3, Agree/ Important = 4, Strongly Agree/Very important = 5, I don't know = 0 or Missing Value (depending on the question). Each of the measures was constructed summing the value of the responses given to each of the items, i.e. given to each of the dimensions underlying the constructs.

5.2.2 Empirical specification and results

Testing differences in perceptions

To assess the first hypotheses about how perceptions change according to one's role and experience with evaluations, I used *two-sample t tests with equal variances* when comparing evaluators versus non evaluators, and ANOVA (Analysis of Variance) when comparing different levels of experiences. The t statistic, confidence interval and p-value are used to assess whether the means of two groups are *statistically* different from each other. The statistical hypothesis for the "t" test is stated as the null hypothesis concerning differences. If you can reject the null hypothesis, than you can conclude that the

difference between the means for the two groups is statistically different (considering the variability of the responses). When analyzing the differences according to level of experience with evaluation (which in our cases create 4 groups) I performed ANOVAs, to be able to set one alpha level and test if any of the groups differ from one another.

Differences in perceptions where compared with respect to:

1. the single items of the survey, with particular attention to the questions relating to the purposes of evaluation,
2. the following latent variables: awareness of having experienced a participatory approach to evaluation (aw_part), awareness of having experienced process use (aw_pu), relevance attributed to a participatory approach to evaluation (rel_part), relevance attributed to process use (rel_pu).

Descriptive statistics

Before proceeding with the analysis based on the constructed latent variables described below I will discuss the results of the analysis of response distribution reported in Table 3 and 4. Respondents were asked to i) state their level of agreement with respect to having experienced certain situations which generally characterize a participatory approach to evaluation, process use and use of findings (Table 2) and ii) to assess the importance of similar situations in determining a successful evaluation (Table 3).

As background information, respondents were asked if sufficient time and resources were available for staff to be involved in evaluation activities. This item received the highest level of disagreement, which is important to keep in mind as a general constraint to fully apply a participatory approach to evaluation.

Looking at the tables, the questions which received answers that appeared statistically different depending on the role one covers in the evaluation (column before the last), and on the level of experience with evaluations (last column) show stars which represent ho statistically different were the responses. Generally speaking, we observe more variability in the responses to the experience question and less in the ones asking about assessing the relevance of different factors for which there seems to be more agreement among respondents.

Looking at the single items, the strongest agreement is found with respect to the importance of stakeholders' ownership of the evaluation in determining a successful evaluation both in theory and in practice (meaning in both the questions on experience and the ones on relevance). Also, the value of the item relating to attention to contextual factor (the third from last in Table 3 and the third in Table 4) increases monotonically with the level of experience with evaluation. A possible interpretation is that more experienced members are able to recognize and consequently value this patterns more than less experienced ones.

Focusing on Table 3, the lowest mean values were associated with having experienced evaluations that have limited program flexibility, which I interpret as an indirect indicator of coordination between evaluation and program implementation.

Also "changes in the organization operations encouraged by evaluation findings" received a very low rate, but was also the item with the highest rate of non response, possibly meaning that individuals do not think they have enough information on the organizational level to respond to this item.

The highest level of agreement was associated to the following statements: "… rigorous evaluation findings enhanced the quality of decision making", and "…rigorous evaluations helped improving program designs as they were implemented". In both cases, evaluators seem to have experienced these situations more often than non-evaluators (the difference is statistically significant at the 5% level). Also the mean values increase as the level of experience increases, which can be explain with an increased capability to recognize these patterns by having observed them more often.

With respect to the relevance attributed to different factors in making an evaluation successful, after the buy-in of stakeholders, project team attitudes toward the evaluation is the most important factor. While the least relevant is encouraging innovative ideas for project operations, and developing evaluative skills (which are effects that we call process use).

Which purpose of evaluation is more important and is there any significant difference in perceptions?

I would like to discuss these results (see Table 2) in light of two other studies which empirically analyzed how evaluators characterized the purposes of evaluation. The traditional purpose of evaluation, determining merit or worth of a program, was the primary purposes in most of the case studies analyzed by Fitzpatrick (2004), while knowledge development, which would place it in the realm of research, was always mentioned, but not as the primary purpose of evaluation. According to Preskill and Caracelli's (1997), survey results, nearly all respondents agreed that the purpose of evaluation is to provide information for decision-making and improving programs.

According to the current survey results, *providing information for decision making* (4.63) remains the most important of all purposes, followed *by determining the merit or worth of a program* (4.38) and *facilitating organizational knowledge* (4.36). To be noted that facilitating organizational learning was the item with the highest number of missing values. Even if the concept was explicitly defined as "changes in organizational learning", more than 3% of the respondents did not respond.

Generally, there is substantial agreement on the purpose of evaluation between evaluators and non. The only significant difference can be found with respect to the value associated to *investigating the merit or worth of a program* which is believed to be very important as compared to other purposes by both groups, yet it has a higher mean value for evaluators

(4.53 versus 4.29 for non-evaluators, statistically different at the 5% level). The different purposes of evaluation were chosen for the questionnaire to reflect the distinction between use of findings and process use. In particular, the purposes of *rigorously question all program assumptions* and/or *develop team's systematic inquiry skills and/or facilitating individual learning* are dimensions of process use, while *providing information for decision-making* and/or *generating knowledge* and/or *assessing the merit or worth of a program* are more closely related to the use of findings. Given this interpretation, we can say that, generally speaking both evaluators and non-evaluators attribute more relevance to the use of findings (mean values above 4.4) than to process use (mean values below 3.9). Yet, even if minimal and not statistically significant, the trend seems to reverse: while evaluators seem to attribute slightly more relevance to the purposes associated to the use of findings, the sign of the difference in means changes for *developing inquiry skills* and *questioning program assumptions*, which seem to be more the important for non-evaluators. This hypothesis is not confirmed given the non-statistical difference between the values, but it is further investigated below.

Do perceptions on participatory approach and process use differ depending on one's role and level of experience with evaluation?

Evidence of process use and on stakeholder's involvement in evaluation processes has been primarily based on evaluators' perceptions (cite article Preskill and Caracelli, 1997; Cousins et al., 1996; Harnar and Preskill, 2007; Chen, 2006 is one more rare example of study that surveys non-evaluators' perceptions). The results of the present study (see Table 6 and 7) show statistically significant differences in terms of the awareness of participatory approach and process use, both with respect to role and experience levels. Evaluators seem to be more aware of both. Coherently, the awareness increases with the level of experience. On the other hand, the level of relevance attributed to these factors is not significantly different nor show a defined pattern according to role or experience. In particular, evaluators seem to have experienced participatory approaches to evaluations and process use more than non evaluators and the difference is statistically significant (the mean values are higher, see Table 6). Assuming, on the basis of how the sample is constructed, that the two groups have, on average, experienced the same approach to evaluation, and given that we have limited the discretionarily and ambiguity in the interpretation of "participatory" approach to evaluation, I would argue that the significant differences may be explained by the different mental models which lead to a different interpretation of the specific events. Looking at Table 7 we can see that the level of awareness increases monotonically with the level of experience, which would confirm the assumption that the ability to interpret certain phenomena increases as one repeatedly experiences it leading to increase sharedness of mental models. Yet the relevance attributed to these aspects does not differ depending on role and experience. Both evaluators and non value as important both the participatory approach to evaluation and

process use, and a lower level of experience does not correspond to a lower relevance attributed to them.

The factors underlying perceptions

A logistic regression was estimated to determine the predictors of the four main variables of interest: awareness of being exposed to a participatory approach to evaluation (aw_part), awareness of being exposed to process use (aw_pu), relevance attributed to a participatory approach to evaluation (rel_part), relevance attributed to process use (rel_pu). The logistic regression is specified as follows:

$$\text{Odds}\ (Y_i = 1 | x_i \ldots x_n) = \exp\ (b_o + b_1 x_1 + \ldots + b_n x_n + b_i x_i * x_n)$$

where Y is a dummy variables assigned a value of 1 if the measure exceeds the sample average and a value of 0 if the measure is below or equal to the sample average. The independent variables vary for each regression as described below. The role played in the evaluation (role), level of experience with evaluations (exp), and academic background (acad) are introduced in the equation as control variables. Moreover, the following interaction effects are considered in the full model: role*experience, experience*academic background, role*academic background[11]. Following are the full regression equations estimated.

Regression 1

$$\text{Odds}\ (\text{aw\_part01} = 1 | x_1 \ldots x_3) =$$
$$= \exp\ (b_o + b_1\ \text{role} + b_2\ \text{exp} + b_3\ \text{acad} + b_{12}\ \text{exp*role} + b_{23}\ \text{exp*acad} +$$
$$+ b_{13}\ \text{role*acad})$$

where aw_part01 is the dummy variable assuming value 1/0 if awareness of participatory evaluation is above/below the sample average.

Regression 2

$$\text{Odds}\ (\text{aw\_pu01} = 1 | x_1 \ldots x_3) =$$
$$= \exp\ (b_o + b_1\ \text{aw\_part} + b_2\ \text{role} + b_3\ \text{exp} + b_4\ \text{acad} + b_{23}\ \text{exp*role} + b_{34}\ \text{exp*acad}$$
$$+ b_{24}\ \text{role*acad})$$

where aw_pu01 is the dummy variable assuming value 1/0 if awareness of process use is above/below the sample average.

Regression 3

[11] As noted by Peng and colleagues (2002), "examining the possibility of interaction between predictors is an essential step in model building strategies"

Odds (rel_pu01 = 1| $x_1 \ldots x_3$) =
= exp (b_o + b_1 aw_part + b_2 aw_pu + b_3 role + b_4 exp + b_5 acad + b_{34} exp*role +
+ b_{45} exp*acad + b_{35} role*acad)

where rel_pu01 is the dummy variable assuming value 1/0 if relevance of process use is above/below the sample average.

Regression 4
Odds (rel_part01 = 1| $x_1 \ldots x_3$) =
= exp (b_o + b_1 aw_part + b_2 aw_pu + b_3 rel_pu + b_4 role + b_5 exp + b_6 acad +
+ b_{45} exp*role + b_{56} exp*acad + b_{46} role*acad)

where rel_part01 is the dummy variable assuming value 1/0 if relevance of participatory evaluation is above/below the sample average.

In order to examine the usefulness of the selected set of predictors in explaining the response, a backward elimination procedure has been selected and performed with STATA 9.0 with the *stepwise logit* command. Starting with a model that contains all the predictors included in the regression equation specified above, the procedure systematically removes the largest p-values terms until only the subset that consists of entirely statistically significant terms is left. In order to test the appropriateness of the selected model, I performed two goodness-of-fit tests, the Hosmer-Lemeshow and the Pearson Tests. In both cases the hypothesis is that the models fit. The results of the analysis are presented in Table 8 and discussed below.

The odds-ratio represents the change in the odds of a particular variable (above average awareness/relevance) relative to the reference category (below average awareness/relevance) that is associated with a one-unit change in a particular independent variable holding constant all other variables (Peng, So, Stage, & St. John, 2002).

Which factors explain above average awareness and relevance of participatory approach?

Role and experience and their interaction explain awareness of participatory approach. Being evaluators (vs non) increases the odds of recognizing the approach as participatory of 8.6 times (significant at 5% level). Also a higher level of experience alone has almost the same effect (8.0 – significant at 1% level).

With respect to the relevance attributed to participatory approach, the strongest predictors are the awareness of the participatory approach (1.3 significant at the 1% level), i.e. the fact of having experienced it, and the relevance attributed to process use (1.213 significant at the 1% level). One possible interpretation of the latter result is hypothesized

in the analytical framework, i.e. that a participatory approach is seen as valuable also because of its effects in terms of process use.

Which factors explain above average awareness and relevance of process use?

Again, even if less significant from a statistical point of view, experience counts as the strongest predictor of awareness (2.332 significant at the 10%), and as hypothesized, being aware of participatory approach increases the odds (1.354 significant at the 1% level) of being aware of process use. This is a very relevant result because it sheds light on the assumption that process use occurs through the engagement in evaluation processes.

With respect to the relevance attributed to process use, having experienced a participatory approach increases the odds by 1.642 (significant at the 5% level), while having experienced process use doesn't seem to have an impact per se.

To sum up, role and experience seem to have the stronger explanatory power in the awareness of participatory approach. While experience alone explains awareness of process use. Other control variables have smaller (yet statistically significant) effects on awareness and no explanatory power for relevance variables. Having experienced participatory approaches explains the relevance attributed to both concepts and having experienced and attributing relevance to process use influences the level of relevance attributed to participatory approaches.

A structural equation model to assess causal link

Finally, to further analyze the hypothesized relationship of this research and assess Hypothesis 3, I employed a structural modeling approach. I evaluated the structural model (henceforth referred to as path model) represented in Figure 2-3.

The path diagrams represent the theorized relationships among the research variables and explain the numerical values associated with each path arrow, consequently indicating the causal influence of each. In the *a priori-determined recursive* path model, as customary, the one-way straight arrows represent regression paths for presumed causal relationships. The arrows indicate the primary causal direction flowing from the observed variable on the left to the one on the right of the arrowhead. While the curved double-headed arrows represent assumed covariances among the *exogenous variables* (such as role and experience). The endogenous variables are depicted with associated error terms. Endogenous variables, conversely, are causally dependent on the other variables in the model, and their causal source is internal (awareness/relevance of participatory evaluation/process use). Recursive path models have no feedback loops; consequently, all causal effects are unidirectional. Standardized path coefficients (a measure of to what extent a change on the variable at the tail of the arrow is transmitted to the variable at the

head of the arrow, holding all other variables constant; Loehlin, 1992) were calculated by STATA 9.0 with the *pathreg* command.

The first model tested here (depicted in Figure 2) is the saturated model where the three exogenous variables (role, academic background, and experience) are modeled as being correlated and as having both direct and indirect effects (through the awareness variables) on the relevance variables (the dependent or 'endogenous' variables). As in most real models, the endogenous are also affected by factors outside the model (including measurement error). The effects of such extraneous variables are depicted by the 'e' or 'error' terms in the model (indicated by the unidirectional arrow pointing only to the endogenous variables).

The saturated 'a-priori' model is specified by the following path equations:

Equation 1. $rel\_part = b_{11}\, role + b_{12}\, experience + b_{14}\, aw\_part + b_{17}\, relev\_pu + e_1$
Equation 2. $rel\_pu = b_{21} role + b_{22} experience + b_{25} aw\_pu + e_2$
Equation 3. $aw\_pu = b_{31} role + b_{32} experience + b_{33} acad\_back + b_{34} aw\_part + e_3$
Equation 4. $aw\_part = b_{41} role + b_{42} experience + b_{43} acad\_back + e_4$

where the b's are the regression coefficients and their subscripts are the equation number and variable number.

Other two reduced models are tested and compared on the basis of fit tests. They are based on same theory *a priori*, and on the elimination of paths that show non significant coefficients at least at the 5% level. The model that best fits the data is depicted in Figure 3.

What appears when interpreting this model, is that role and academic background have no effect, direct or indirect on the endogenous variables, while experience has a strong (0.39 significant at the 1% level) direct effect on the awareness of participatory approach, but no effect on awareness of process use. Awareness of participatory approach has a direct effect (0.17 significant at the 5% level) on relevance of participatory approach and a very strong link (0.62 significant at 1% level) with awareness of process use. Awareness of process use has a direct effect (0.28 significant at the 1% level) on the relevance of process use which has an effect (0.37 significant at the 1% level) on the relevance of participatory approach. The highest and most significant coefficient links awareness of participatory approach and of process use, which confirms the result of the logit model of a correlation (which the structural model confirms to be causal) between experiencing a participatory approach and experiencing process use.

6. Conclusion and Discussion

To sum up with a general conclusion, it appears that experience and role don't really influence the perception of evaluators versus non evaluators, nor more experienced versus less. The level of relevance attributed to the concepts of process use and participatory approaches are generally high. Yet the level of awareness, which is a proxy of direct experience, changes according to role and even more as the experience with evaluation grows. It appears that the level of awareness increases monotonically with experience and coherently it is higher in evaluators rather than non evaluators. The analysis also confirms the relationship between process use and participatory approaches, and the fact that the relevance attributed to participatory approaches could a result of the relevance attributed to process use.

Opportunities abound for managers to learn from the evaluation profession and to employ evaluators to support performance based management. There is ample evidence of renewed interest in what evaluation of programmatic performance can offer management. Yet the challenge is to identify ways and means for program managers to develop evaluation skills and to create an organizational culture wherein systematic thinking about how and why programs work becomes the norm (Newcomer, 2001). I suggest that a participatory approach to evaluation accompanied by awareness creation and reflection upon the learning effects of the evaluation process could contribute to this end, by accelerating the creation of shared mental models among the heterogeneous range of professionals involved in the evaluation process. More attention should be given to the multidisciplinary composition of evaluation teams and the fact that different perceptions need to be taken into consideration, especially when trying to leverage the opportunity of process use by stimulating reflection on the evaluation process, which is characterized by tacit knowledge (see, e.g. Marra, 2000; Marra, 2007, my conceptual framework described in paper 1). In fact, the problems with mental models doesn't lie in whether they are right or wrong - by definition, all models are simplifications. The problems with mental models arise when the models are tacit - when they exist below the level of awareness (Mathieu et. al, 2000). Mathieu et al.'s (2000) findings suggest that researchers and practitioners should conduct thorough team task analyses to identify the most critical knowledge requirements for a given situation and which part of that knowledge must be shared to enable effective team work. Institutionalizing the process of reflecting on and surfacing mental models requires the development of mechanisms that make these practices unavoidable (Senge, 1992).

References

Amo C. and Cousins J.B. (2007), "Going through the process: an examination of the operationalization of Process Use in Empirical Research on Evaluation", *New Directions for Evaluation*, 116: 5-26

Chen K-N (2006), "Library Evaluation and organizational learning. A questionnaire study", *Journal of Librarianship and Information Science*, 38(2): 93

Cousins J.B. and Earl L.M. (1992), "The Case for Participatory Evaluation", *Educational Evaluation and Policy Analysis*, 14(4): 397

Cousins J.B., Donohue J.J., Bloom G.A (1996), "Collaborative Evaluation in North America: Evaluators' Self-reported Opinions, Practices and Consequences", *American Journal of Evaluation*, 17(3): 207

Cousins J.B. and Leithwood K.A. (1986), "Current Empirical Research on Evaluation Utilization", *Review of Educational Research*, 56(3):331

Cousins J. B., Goh S., Clark S., and Lee L. (2004), "Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge". *Canadian Journal of Program Evaluation*, 19(2), 99-141.

Dyer J. L. (1984), "Team research and team training: A state-of-the-art review", in F. A. Muckler (Ed.), *Human factors review,* Santa Monica, CA: Human Factors Society.

Forss K., Cracknell B., Samset K. (1994), "Can Evaluation Help an Organization to Learn?" *Evaluation Review*, 18(5): 574

Forss K., Rebien C.C., and Carlsson J. (2002), "Process Use of Evaluations: Types of Use that Precede Lessons Learned and Feedback", *Evaluation*, 8(1): 29

Greene J.G. (1988), "Stakeholder participation and utilization in program evaluation". *Evaluation Review*, 12(2): 91

Harnar and Preskill (2007), "Evaluator's descriptions of process use: an exploratory study", *New Directions for Evaluation*, 116: 27-44

Henry G.T. and Mark M.M. (2003), "Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions", *American Journal of Evaluation*, 24: 293

Johnson-Laird P. (1983), "*Mental models".* Cambridge, MA: Harvard University Press.

Kirkhart K. (2000), "Re-conceptualizing evaluation use: An integrated theory of use" *New Directions for Evaluation*, 88: 5

Legovini A. (2006), "Impact Evaluation and the project cycle" presentation at the World Bank PREM (Poverty Reduction and Economic Management Network) Week, May 2, 2006. Washington DC: World Bank

Loehlin J.C. (1992), *Latent variable models: An* introduction to factor, path and structural analysis (2nd ed.). Hillsdale. NJ: Erlbaum

Levine D. I. (2006), "Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry", Berkeley: Haas School of Business, University of California.

Lincoln Y.S. and Guba E.G. (2004), "The roots of fourth generation evaluation: theoretical and methodological origins" in in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Marra M. (2000), "How Much Does Evaluation Matter? Some Examples of Utilization of the Evaluation of World Bank's Anticorruption Activities", *Evaluation* 6(1): 22

Marra M. (2004), "Knowledge: The Case of the World Bank The Contribution of Evaluation to Socialization and Externalization of Tacit", *Evaluation* 2004; 10: 263

Myers-Walls J.A. (2000), "An Odd Couple with Promise: Researchers and Practitioners in Evaluation Settings", *Family Relations*, 49(3): 341

Newcomer K. and Neill S. (2001), "Issue Introduction", *Evaluation and Program Planning*, 24: 61

O'Sullivan R.G. and D'Agostino A. (2002), "Promoting Evaluation through Collaboration Findings from Community-based Programs for Young Children and their Families", *Evaluation,* Vol 8(3): 372

Mathieu J.E., Goodwin G.F., Heffner T. S., Salas E., Cannon-Bowers J. A. (2000), "The Influence of Shared Mental Models on Team Process and Performance", 85(2): 273-283

Patton M.Q. (1997), *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Patton M.Q. (2004), "The roots of utilization-focused evaluation" in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Patton M.Q. (2007), "Process Use as Usefulism", New Directions for Evaluation, 116: 99-112

Patton M.Q. (2008), *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Peng C.-Y. J., So T.-S. H., Stage F.K. and John E.P. St. (2002), "The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999", *Research in Higher Education,* 43(3): 259

Preskill H. (1994), Evaluation's role in enhancing organizational learning: A model for practice. *Evaluation and Program Planning*, 14 (2), 291

Preskill H. and Caracelli V. (1997), "Current and Developing Conceptions of Use: Evaluation Use TIG Survey Results", *American Journal of Evaluation;* 18: 209

Preskill H., Zuckerman B. and Matthews B. (2003), "An Exploratory Study of Process Use: Findings and Implications for Future Research", *American Journal of Evaluation*; 24: 423

Preskill H. (2004), "The transformational power of evaluation: passion, purpose, and practice" in Alkin M.C. (2004), "Evaluation Roots: Tracing Theorists' Views and Influences", Thousands Oaks, CA: SAGE.

Preskill, H. and Torres T. R. (2000), 'The Learning Dimension of Evaluation Use', *New Directions for Evaluation*, 88: 25

Preskill H. and Torres R.T., 1999), "Building Capacity for Organizational Learning Through Evaluative Inquiry" *Evaluation*, 5(1): 42

Pritchett L. (2002) "It pays to be ignorant: a simple political economy of rigorous program evaluation" *Policy Reform,* 5(4): 251-269

Senge P.M. (1992), "Mental models. Putting strategic ideas into practice)", *Planning Review*, 20(2): 8

Shulha L.M. and Cousins J.B. (1997), "Evaluation Use: Theory, Research, and Practice Since 1986", *American Journal of Evaluation*, 18(1): 195

Rouse W. B., and Morris N. M. (1986), "On looking into the black box: Prospects and limits in the search for mental models", *Psychological Bulletin*, 100: 349

Rynes S. L., Bartunek J.M., Daft R.L. (2001), "Across the Great Divide: Knowledge Creation and Transfer between Practitioners and Academics", *The Academy of Management Journal*, 44(2): 340

Rentsch L R., Heffner T. S., and Dully L. T. (1994), "What you know is what you get from experience: Team experience related to teamwork schemas", *Group and Organization Management,* 19: 485

Shrivastava P. and Mitroff I.I. (1984), "Enhancing organizational research utilization: The role of decision makers' assumptions", *Academy of Management Review*, 9: 18

Torres R.T. and Preskill H. (2001), "Evaluation and Organizational Learning: Past, Present and Future", *American Journal of Evaluation*, 22: 387

Weiss (1998), "Have we learned anything new about the use of evaluation? American Journal of Evaluation", 19:21

Wilson J. R. and Rutherford A. (1989), "Mental models: Theory and application in human factors", *Human Factors,* 31, 617-634

Wood R.C. and Gary H. (2002). "The World Bank's Innovation Market." *Harvard Business Review*, November 1:2

World Bank (2006), Impact Evaluation and the project cycle, *Doing Impact Evaluation Series*, No.1 Washington DC: The World Bank

ANNEX

Variables Operationalization and Definition

"Level of awareness of experiencing a participatory approach" (aw_part) is composed of 4 survey questions (answers were coded 0-5 depending on how much the respondents would agree to have experienced the following dimensions of participation – "I don't know" is coded 0): 1. …the decisions about the evaluation activities (e.g. the design, the information to be collected, the timing of data collection) were made collaboratively by all team members 2. …the evaluation design did not limit the flexibility of the project 3. …frequent consultation with project management and local counterparts took place in order to design the evaluation 4. ...stakeholders' (e.g. government counterpart) felt ownership of the evaluation.

"Relevance attributed to participatory approach" (rel_part) is composed of 5 survey questions (answers were coded 1-5 depending on how much importance would the respondents attribute to the following dimensions of a participatory approach to evaluation - "I don't know" is coded as a missing value): 1. …attention to contextual factors through active involvement of different stakeholders in the evaluation (e.g. government counterparts, project management) 2. …evaluation is conducted in a participatory way by all relevant stakeholders (e.g. government counterparts, project management) 3. …sufficient buy-in for the evaluation from the relevant stakeholders (e.g. government counterparts, project management) 4. …evaluation's independency from project management to assure objective findings (reverse score) 5**.** …evaluation design alignment with project operations (i.e. the evaluation strategy does not limit the flexibility of the project implementation).

"Level of awareness of experiencing process use" (exp_pu) is composed of 7 survey questions (answers were coded 1-5 depending on how much the respondents would agree to have experienced the following dimensions of process use – "I don't know" is coded 0): 1. ...the evaluation helps question tacit assumptions about the program theory 2. ...the evaluation helps improving programs design as they are implemented 3. ...the evaluation stimulates team discussions on innovative ideas for project implementation 4. ...the evaluation increases team's shared understanding of the pjt 5. ...the evaluation increases the ability to solve problems during pjt implementation 6. ...the team has to pay more attention to changes in the external environment and contextual factors 7. ...team attitudes toward evaluation have become more positive after their involvement in the evaluation

"Relevance attributed to process use" (rel_pu) is composed of 7 survey questions (answers were coded 1-5 depending on how much importance would the respondents

attribute to the following dimensions of process use - “I don’t know” is coded as a missing value): 1. …the capacity to question program theory assumptions (i.e. the logic framework that links inputs-outputs and outcomes) 2. …encouragement of innovative ideas for project operations 3. …project team members’ attitude towards evaluation 4. …team members’ development of evaluative skills 5. ...facilitate individual learning 6. ...rigorously question all program assumptions 7. …develop team’s systematic inquiry skills.

“Role in the evaluation team” (role_dummy) a dummy variable coded 1 if the respondent is reports to be the evaluator specialist and 0 otherwise (i.e. project manager or implementation team member)

“Experience with evaluations” (exp_num) is a numerical variable assigned values 1-4 depending on the self-reported level of experience (being “None” = 1 to "High" = 4)

“Academic background” (acad_num) is a numerical variable assigned values 1-5 as follows: 5 if academic background in "Evaluation"; 4 if in "Economics"; 3 if in "Engineering"; 2 if in "Public Policy / Administration / Management"; 1 if in "Human Development“, "Sociology", or "Other”

Table 1 Characteristics of the Sample (n=181)

	№	%
Organization (n. %)		
Government	97	*54%*
World Bank	60	*33%*
Other	24	*13%*
Academic Background/Degree Program (n. %)		
Evaluation	8	*4%*
Economics	89	*49%*
Engineering	19	*10%*
Public Policy and Management	23	*13%*
Human development	21	*12%*
Sociology	12	*7%*
Other	9	*5%*
Level of experience with evaluations (n. %)		
None	7	*4%*
Low	29	*16%*
Medium	87	*48%*
High	58	*32%*
Role generally played in evaluation teams (n. %)		
Project Management	46	*25%*
Member of project implementation team	63	*35%*
Evaluation Specialist	72	*40%*

Table 2 Purposes of evaluation – response distribution and differences in perceptions according to role (evaluators vs non evaluators)

Please rate how IMPORTANT is each of the following purposes of evaluation	**Not at all/Little important**	**Somewhat Important**	**Important/Very**	**Un-decided/No answer**
Investigate the merit or worth of a project	2.2%	9%	86%	2%
Facilitate individual learning	10.5%	31%	57%	2%
Facilitate organizational learning (i.e. changes of organizational knowledge)	2.2%	10%	84%	3%
Provide information for decision-making	1.7%	4%	93%	1%
Generate new knowledge	2.8%	12%	85%	1%
Rigorously question all program assumptions	12.2%	16%	70%	2%
Develop team's systematic inquiry skills	7.2%	20%	70%	2%

			Role		*Diff*	
Purposes of evaluation	Mean	Std Dev	Evaluator	Non-Eval		
Provide information for decision-making	4.63	0.68	4.64	4.62	0.02	
Investigate the merit or worth of a project	4.38	0.78	4.53	4.29	0.24	**
Facilitate organizational learning (i.e. changes of organizational knowledge)	4.36	0.79	4.40	4.33	0.07	
Generate new knowledge	4.27	0.84	4.31	4.24	0.06	
Develop team's systematic inquiry skills	3.95	0.97	3.88	4.00	-0.13	
Rigorously question all program assumptions	3.85	1.05	3.82	3.88	-0.06	
Facilitate individual learning	3.59	0.89	3.61	3.58	0.03	

* significant at 10% level (p< .10)
** significant at 5% level (p< .05)
*** significant at 1% level (p< .01)

Table 3 Perceived experiences with evaluation (awareness)

Thinking back to your personal experience, to what extent do you agree or disagree with the following statements? IN MY EXPERIENCE….	**Strongly Dis/Disagree**	**Somewhat Agree**	**Agree/Strongly Agree**	**Un-decided/No answer**	**Mean**	**Diff wrt Role**	**Diff wrt Experience**
...rigorous evaluations helped question tacit assumptions about the program theory	6.6%	19%	65%	9%	3.5	**	***
…rigorous evaluations have partially limited the flexibility of project implementation	42.5%	24%	27%	6%	2.6		
…rigorous evaluations helped improving program designs as they were implemented	8.8%	19%	68%	4%	3.7	**	***
… rigorous evaluation findings enhanced the quality of decision making	6.1%	17%	71%	6%	3.8	**	***
…sufficient time and resources were available for staff to be involved in evaluation activities	48.1%	20%	24%	7%	2.5		
…findings of rigorous evaluations made it easier to convince external stakeholders of the merit or worth of a program	8.3%	17%	61%	14%	3.3	*	**
…the decisions about the evaluation activities (e.g. the design, the information to be collected, the timing of data collection) were made collaboratively by all team members	16.6%	23%	54%	7%	3.3	**	***
…evaluation stimulated team discussions on innovative ideas for project implementation	7.2%	20%	65%	8%	3.6		
…the evaluation increased team's shared understanding of the project	7.2%	19%	66%	8%	3.5	**	***
…changes in the organization's operations were encouraged by evaluation findings	14.9%	30%	31%	23%	2.5		
…frequent consultation with project management and local counterparts took place in order to design the evaluation	8.3%	23%	56%	13%	3.2	**	***
…the evaluation increased the ability to solve problems during project implementation	16.0%	24%	44%	17%	2.9		
…due to the evaluation, the team had to pay more attention to changes in the external environment and contextual factors	10.5%	22%	52%	16%	3.1		***
…team attitudes toward evaluation have become more positive after their involvement in the evaluation	8.3%	16%	62%	14%	3.3		
...stakeholders' (e.g. government counterpart) ownership of the evaluation is central to its success	7.2%	9%	74%	10%	3.8		*

* significant at 10% level (p< .10)
** significant at 5% level (p< .05)
*** significant at 1% level (p< .01)

Table 4 Perceived relevance of participatory evaluation and process use

According to you, how important are each of the following factors in determining a successful evaluation?	Not at all/Little	Somewhat Important	Important/Very	Un-decided/No answer	Mean	Diff wrt Role	Diff wrt Experience
The capacity to objectively assess the merit or worth of a project for knowledge generation	4.4%	13%	77%	6%	4.1	*	
The capacity to question program theory assumptions (i.e. the logic framework that links inputs-outputs and outcomes)	1.1%	14%	82%	3%	4.2		
Attention to contextual factors through active involvement of different stakeholders in the evaluation (e.g. government counterparts, project management)	0.6%	13%	85%	2%	4.3		***
Evaluators' familiarity with the project, country and institutional context	0.0%	10%	87%	3%	4.4		
Evaluation is conducted in a participatory way by all relevant stakeholders (e.g. government counterparts, project management)	3.3%	12%	83%	2%	4.2		
Encouragement of innovative ideas for project operations	5.5%	21%	71%	3%	3.9		
Project team members' attitude towards evaluation	1.7%	9%	87%	2%	4.4	*	
Sufficient buy-in for the evaluation from the relevant stakeholders (e.g. government counterparts, project management)	0.6%	5%	92%	3%	4.5		
Evaluation's independency from project management to assure objective findings	3.3%	12%	83%	2%	4.3		
Evaluation design alignment with project operations (i.e. the evaluation strategy does not limit the flexibility of the project implementation)	2.2%	11%	84%	3%	4.2		
Team members development of evaluative skills	3.3%	20%	74%	2%	4.0		
Planning the use of findings at the beginning of the evaluation	5.0%	12%	78%	4%	4.2		

* significant at 10% level (p< .10)
** significant at 5% level (p< .05)
*** significant at 1% level (p< .01)

Table 5 Latent variables

Level of awareness	Mean	Std. Dev	Min	Max
Partecipatory approach	13.4	3.7	0	20
Process use	23.6	7.3	0	35
Level of relevance				
Partecipatory approach	18.8	1.8	14	23
Process use	28.0	4.0	15	35

Table 6 Differences in Perception according to role

	Role				
	Evaluator	Non-Eval	Diff	*Pvalue*	
Level of awareness					
Partecipatory approach	14.4	12.7	1.7	***0.0026***	***
Process use	25.1	22.5	2.6	***0.0184***	**
Level of relevance					
Partecipatory approach	18.8	18.9	-0.1	*0.7566*	
Process use	28.1	28.0	0.1	*0.9012*	

Table 7 Differences in Perception according to experience

	Experience					
	None	Low	Medium	High	*Pvalue*	
Level of awareness						
Partecipatory approach	6.9	12.1	13.6	14.4	***0.0000***	***
Process use	15.9	20.4	23.9	25.6	***0.0004***	***
Level of relevance						
Partecipatory approach	17.9	18.5	18.8	19.2	*0.1211*	
Process use	28.7	26.7	28.0	28.6	*0.2348*	

** significant at 5% level (p< .05)
*** significant at 1% level (p< .01)

Table 8 Logit estimation -

	Awareness				*Relevance*			
	Part Approach		*Process Use*		*Part Approach*		*Process Use*	
Predictors	OR	Std. Err.	OR	Std. Err	OR	Std. Err	OR	Std. err.
Awareness participatory approach	-	-	1.354	0.092 ***	1.290	0.093 ***	1.642	0.386 **
Awareness process use	-	-	-	-	0.941	0.034 *	-	-
Relevance process use	-	-	-	-	1.213	0.063 ***	-	-
Role in team (evaluator/non)	8.582	8.049 **	-	-	-	-	0.659	0.142 *
Experience with evaluation	8.035	5.196 ***	2.332	1.045 *	-	-	-	-
Academic Background	-	-	0.399	0.145 **	-	-	-	-
Role*Experience	0.495	0.147 **	0.686	0.112 **	-	-	-	-
Experience*Academic Background	-	-	-	-	-	-	-	-
Role*Academic Background	-	-	0.686	0.220 **	-	-	-	-

* significant at 10% level (p< .10)
** significant at 5% level (p< .05)
*** significant at 1% level (p< .01)

Figure 2 Path model n.1

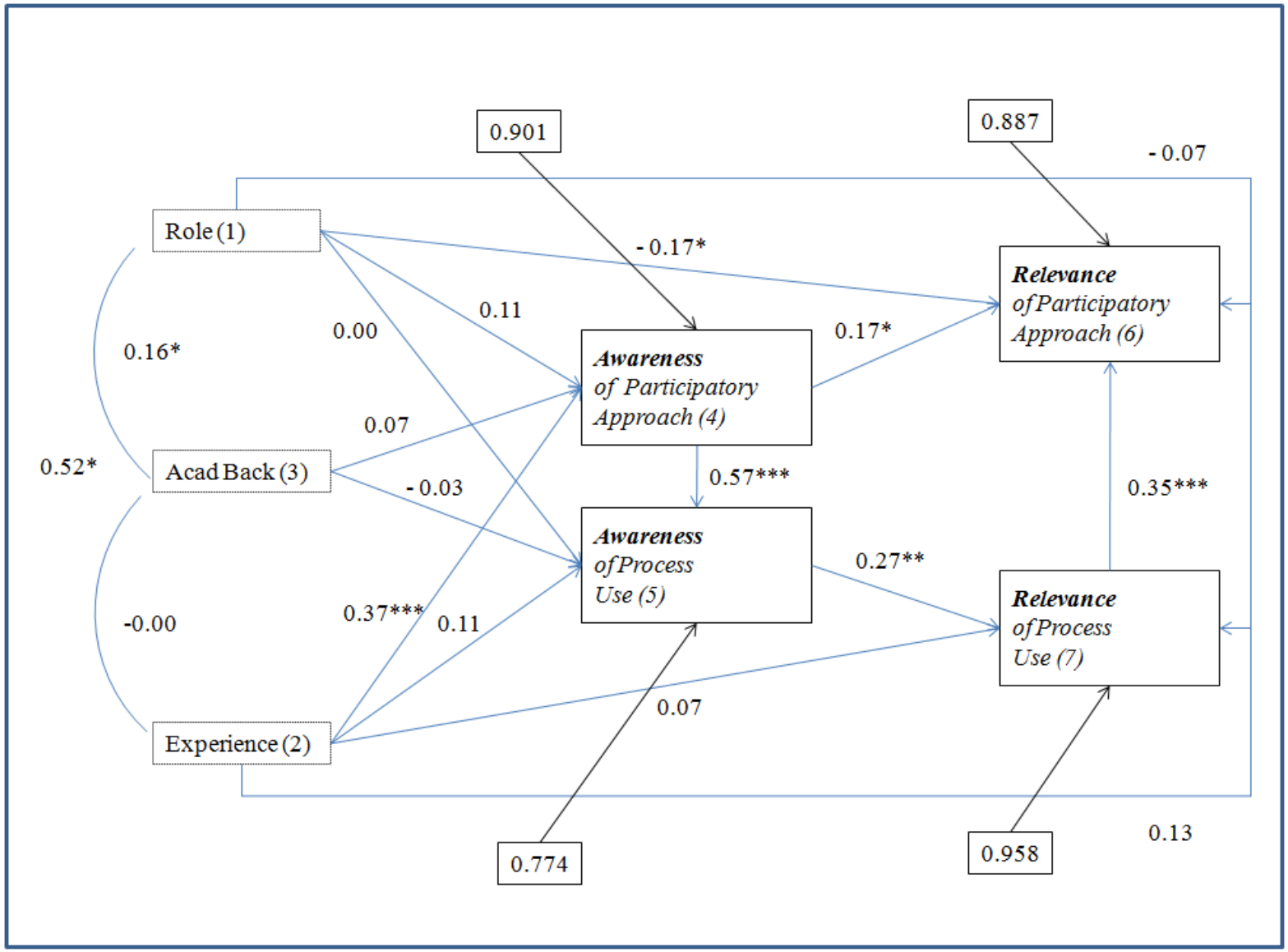
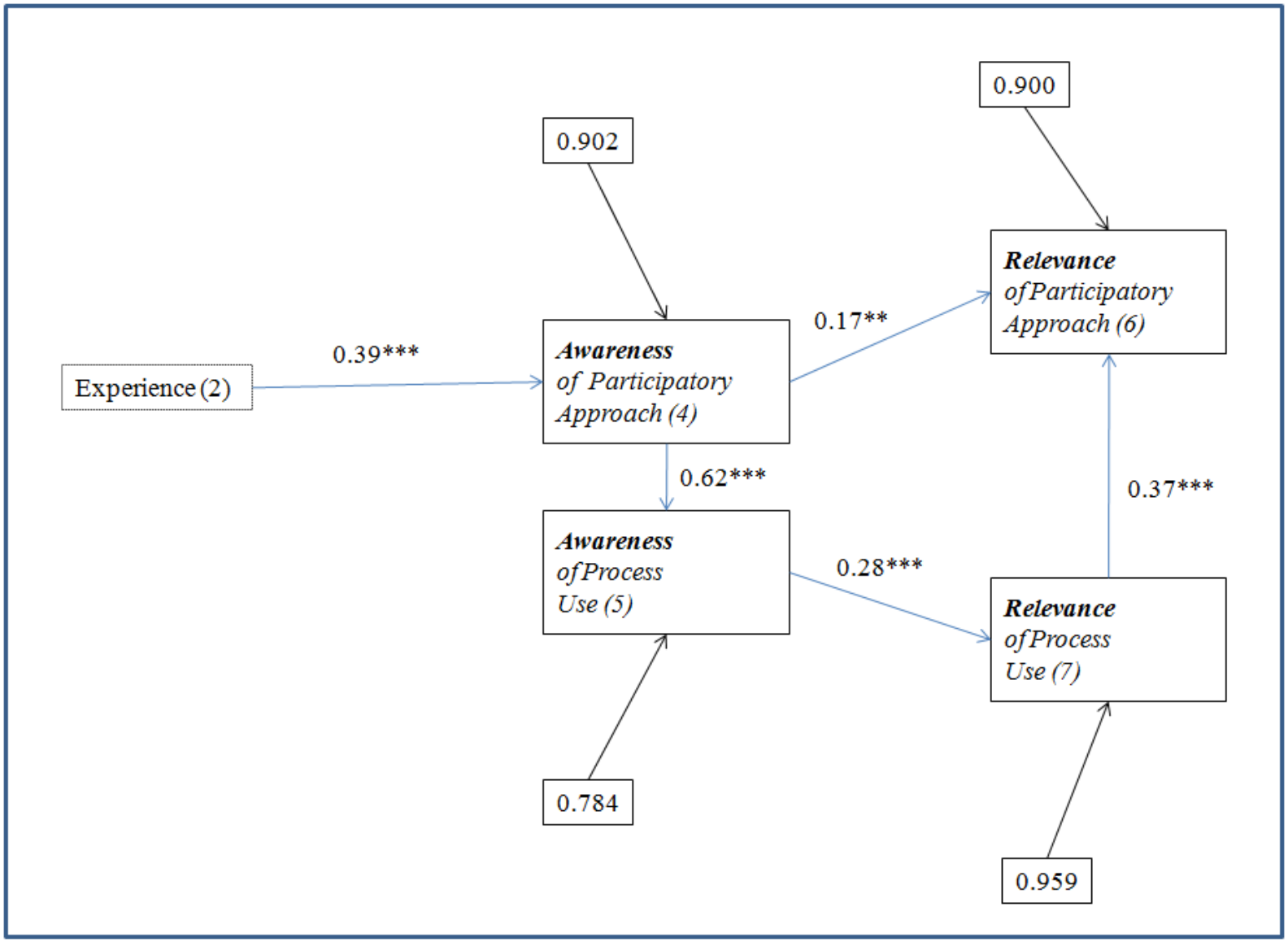


Figure 3 Path model n.2



PROCESS USE
AND
EVALUATION CAPACITY BUILDING
IN EXPERIMENTAL IMPACT EVALUATIONS:
THE VIRTUE OF NECESSITY

Candidate: Silvia Paruzzolo
PhD in Business Administration and Management
XX cycle - Track: Public Management
Paper 3/3

OUTLINE

1. **Introduction**
2. **Study Objectives and Research Questions**
3. **The Case Study**
    a. **Data and Methods**
    b. **The Evaluand**
    c. **The Impact Evaluation**
4. **The Case Study Findings**
5. **Reflections**
6. **Study limitations**
7. **Discussion and next steps**

Annex
References

1. Introduction

The concept of use is central to evaluation literature. Vedung (1997), who maintains that role of evaluation in political and administrative contexts is "many sided, subtle and complex" suggests that an evaluation can be used in many different ways: instrumentally, conceptually, legitimizing, interactively, tactically, ritually and as a process. Other scholars have debated the definition of evaluation use, and some common categories exist: instrumental, conceptual, and persuasive or symbolic (Peck and Gorzalski, 2008). As Peck and Gorzalski define these categories:

> *Instrumental use* refers to the direct use of an evaluation's findings in decision making or problem solving. *Conceptual use* refers to an evaluation's contribution to changes in thinking or to the general knowledge base. *Persuasive use* refers to an evaluation's use to convince others of a political position.

According to Patton (1997), the three primary *instrumental uses* of evaluation findings included (1) judging merit or worth, (2) improving the program, and (3) generating knowledge. The first of these uses can aid in "go/no-go decisions" and may be the result of a summative evaluation (Scriven, 1996). The second form of instrumental use presumably helps to develop the needed rationale for action and may be related to the notions of a formative evaluation (Scriven, 1996). The third form of use of evaluation findings - that of generating knowledge - tends to focus on the scholarly and policy-making aspects of evaluation. Furthermore Patton (1997, 2007, 2008) introduced the term *process use* to describe the uses that derive from engaging in an evaluation process in contrast to using evaluation findings. Process use, therefore, enables the evaluators and the evaluation stakeholders and audiences to make use of the logic and process incorporated into the evaluation itself. As elaborated by Russ-Eft et al. (2002):

> The processes of participation and collaboration have an impact on those who participate beyond whatever tasks they may accomplish by working together. In the *process* of participating in an evaluation, participants are exposed to and have the opportunity to learn the logic of evaluation and the discipline of evaluation reasoning. Skills are acquired in problem identification, criteria specification, and data collection, analysis, and interpretation. Acquisition of evaluation skills and ways of thinking can have a longer-term impact than the use of findings from a particular evaluation study.

This study answers to a call for better understanding of the factors that affect evaluation process use in the hope that we can design and implement evaluation in such a way to maximize process use and evaluation capacity building (Lawrence et al., 2007). In fact, most empirical accounts of process use focus on the effects of being involved in an evaluation, but much less attention has been given to understanding how the effects of process use are brought about. The existing literature points out several factors which appear to influence the likelihood that those involved in evaluation processes will learn from their participation (Preskill et al., 2003, 2005): i) how evaluations meetings are facilitated; ii) the extent to which, and the ways in which, management and leadership support participants' involvement in the evaluation process; iii) participants' personal characteristics and experiences with evaluation in the program being evaluated; iv) the frequency, methods, and quality of communications between and among stakeholder participants; v) organizational characteristics.

The present work is a case study of a large evaluation that underwent a complex process which involved many stakeholders over the course of a 4-year period. Large can be defined in several ways. If we can describe evaluations as large because they cost a great deal of money, include a large number of participants, include multiple sites or because they evaluate large programs (Bickman and Mulvaney, 2006), then the present case is definitely about a large evaluation according to all the criteria above. The case study is an account of how the World Bank project managers (hereafter TTLs – Task Team Leaders) initiated the process and requested evaluators within the WB to design a prospective and experimental evaluation of a multi-donor development program. In carrying out the analysis I was particularly concerned with how the stakeholders viewed this process and how and if their accounts differed depending on their role and level of involvement in the process.

The paper also contributes to the current debate on the distinction between Evaluation Capacity Building (ECB) and Process Use (PU) (Patton, 2007; Preskill and Boyle, 2008; King, 2007, Harnar and Preskill, 2007). I personally argue that PU and ECB are distinct effects of evaluation processes, and they are both worth being intentional and exploited dimensions. I agree with Patton in maintaining that deepening shared program understanding and strengthening the *evaluand* (examples of PU), are not types of ECB. On the other side, teaching how to implement one or more steps of the evaluation process, or how to become more educated consumers of evaluations are ECB activities. The findings of the present analysis will be categorized according to this distinction.

Table 1 Matrix of Intentionality ad Use/Influence

	Finding Use and Influence	Process Use and Influence
Intended	Intended use by intended users	Includes explicit, planned ECB, as well as other process uses
Gray Area	Intentionally focused on primary intended users, but planned dissemination hopes for broader influence	Evaluators facilitate the evaluation process ALSO to build capacity but this is implicit and those who are involved are motivated by, and focused on, findings use
Unintended	Unplanned influence of findings beyond primary intended users-and even beyond original dissemination	ECB implicit (an artifact of participation in the evaluation)

Source: Patton, 2007

The case study analyzed in this paper exemplifies what described in the gray area (left side) in the matrix above. As it will be discussed in depth, in the evaluation of the PBF in the country selected for the present analysis (hereafter xxx), capacity building was mentioned in the documents justifying the need for an evaluation. Yet very little was done intentionally to create capacity at the local level. Only towards the end of the process some investment were made due to necessity more than choice. I argue that this positions the type of process use found in the present case study in the gray area described by Patton (2007).

The paper is structured as follows. I begin by justifying the selection of the case study methodology and of the particular setting and by providing a brief overview of the program and the evaluation design. Next I describe the PU and ECB resulting from the evaluation process, and examine the factors influencing this type of use and the ways in which the context or the environment for the evaluation may have impacted both use (and non-use). I conclude by reflecting on the findings and identifying some limitation of my work, as well as some future areas of investigation.

2. Study Objectives and Research Questions

The present study provides empirical evidence of how process use occurs in practice, which type of cognitive and behavioral changes result from the engagement in the evaluation, which might be the factors that determine the types and levels of process use, according to the perspective of stakeholders highly involved in an ongoing evaluation, and relevant evaluation documentation. The analysis focuses on three main dimensions of the evaluation: 1. the process (i.e. the process of initiating, defining and implementing the evaluation design); 2. the approach, in terms of the level of participation of different stakeholders to the evaluation process; 3. *process use*, i.e. the changes in thinking, behavior and in the *evaluand* (i.e. the program under evaluation) and learning occurring as a consequence of the impact evaluation (hereafter IE).
The main research questions addressed in the paper are:

1. *Which are the steps that require collaboration to produce meaningful findings in a prospective experimental impact evaluation?*
2. *How does Process Use occur in practice? Which manifestations of Process Use (and/or ECB) are evident to those involved in the evaluation process?*

3. The case study

3.1 Data and Methods

The case study was selected as the appropriate method for this exploratory investigation (Yin, 2003). Given the variability of PU (Forss et al., 2002) and the fact that the definition is situational and context-dependent, focusing on one case is important to investigate variations in perceptions of the stakeholders exposed to the "same events".

The focus on an experimental prospective impact evaluation aims at providing some insight on which kind of process use is specific to this evaluation design since also the choice of an evaluation method will have an impact on process use (Forss et a., 2002). The choice is based on the assumptions elaborated during the author's experience working at the WB and also present in the literature (Roessler, 1980, Bickman, 1996): the definition of a counterfactual (e.g. choice of how to select treatment and control) ex-ante requires i) high coordination and/or participation of implementers in the evaluation process (e.g. evaluators need to know precisely how the program will be implemented to elaborate the evaluation plan ex-ante); ii) alignment of evaluations with project implementation (e.g. the evaluators need to be aware of possible contamination of the "experiment" and need to be aware of the status of the program to be prepared for the timely data-collection process – follow up data need to be collected right after the treatment window closes, especially if the control areas are receiving the program right after); point I) and ii) imply the necessity of having the iii) buy in and compliance to the evaluation plan of the actors implementing

the program. The xxx case was chosen after a preliminary analysis of available documentation which provided me with reasonable evidence and examples of integration and coordination of the evaluation and the program operations and possible influence of evaluation on program implementation. Also, given my position as a consultant within the World Bank I had easy access to documentation and key informants. Yet, the choice was not to write a case study on an evaluation I was directly involved in to strike a balance between being close enough to understand part of the process as an insider (which was appreciated by the interviewees, since they could skip explanation on standard WB processes) and being biased by my role as evaluator in the case analyzed. The case was also selected based on the timing of the evaluation. In fact, to overcome previous studies limitations the decision was to focus on an ongoing IE which is uncommon in research on evaluation use (Taut, 2007). According to Preskill et al. (2003), the timing of research in investigating the learning and changes that occur during the evaluation (process use). The case selected is a four year long evaluation which is still ongoing. Three key informants have been involved in the past three years, but are still involved in it. I assume that waiting to analyze the process after its conclusion would have jeopardized the richness of the account due to memory losses. On the contrary, by being still involved, the actors seem to have good memory of the process, which is close to conclusion, so their accounts should represent a complete view on (almost) all the important steps[12].

The data collection took place between July and December 2008. The main sources of data were interviews and written official documentation. Eight face to face interviews and three phone interviews (with people who are located in xxx) aimed at exploring the experiences, perspectives and perceptions of key informants highly involved in the process of the evaluation. Key informants were chosen purposefully, with a snowball technique. The initial list was drafted with a researcher's colleague directly involved in the impact evaluation, to include, at minimum, a representative from i) the government side, ii) the WB operational team and iii) the WB evaluation team. The objective of this initial selection was to be able to triangulate the information reported. Each person interviewed was asked to suggest additional knowledgeable people for interviewing. The process continued until no new names were suggested. Only two high-level officials contacted were not available for the interview. Open-ended questions and probes were crafted to encourage spontaneous responses. Few questions were included in the interview protocol to allow the interviewee enough time to reflect on and make sense of their experience with the evaluation. Each participant was provided, prior to the interview, with a short presentation of the topic to be discussed during the interview. The interview protocol enabled the interviewer to be adaptive to each individual. The interviews lasted between 30 and 90 minutes, depending on availability, and were recorded with the interviewees' permission. In order to avoid the possibility of interviewees providing socially desirable responses, we provided confidentiality of their responses. For the purpose of maintaining confidentiality and some privacy regarding the person's work and experiences, we do not identify key informants, but the general nature of the evaluation and project work, as necessary for the analysis and interpretation, remains intact. The interviews were transcribed and coded. The analysis consisted in categorization, interpretations, and revisions of the category systems. To increase the validity of the data, I corroborated as much as possible the

[12] I would expect the informants to have forgotten some parts of the process, but this doesn't necessarily bias the account, since it is what is not forgotten that actually represents process use.

information provided by the respondents with existing documentation. Public and international organizations are notable for their attention to documentation and reporting. There is careful preparation and retention of reports and extensive minutes of meetings are taken. When, as in this case, part of work is conducted by consultants and paid through public funding sources such as Government Trust Funds, relevant information can be found in TORs (Terms of References) for consultant work and missions to the field, and BTORs (Back to the Office Reports). In addition, the researcher had access to i) evolving versions of evaluation concept notes, ii) evaluation proposals for fundraising, and iii) relevant email exchanges.

Table 1. List Of Key Informants Interviewed
Local Government
Director of PBF system – Ministry of Health Member of TWG for PBF implementation - MSH
World Bank - Operational Side
Task Team Leader for General Health – World Bank Task Team Leader for HIV/AIDS – World Bank
World Bank – Evaluation side
Principal Investigator for HIV/AIDS – World Bank Principal Investigator for General Health – World Bank Joint Evaluation Coordinator – World Bank
Non World Bank - Evaluation side
Prinicipal Investigator – Berkeley University Economist - Evaluation Specialist – INSP
Local researchers
Field survey coordinator and member of TWG - SoPH Responsible for process evaluation - Consultant

3.2 The Evaluand

The program under evaluation is a Performance-Based Financing (hereafter PBF) scheme for healthcare service delivery in xxx. The PBF is one of the major strategies adopted by the Ministry of Health for improving access to quality healthcare for the population of xxx. PBF is defined as a method of healthcare services management which seeks to increase the volume and quality of healthcare services provided to the population. Performance based financing increases funds available at the operational level to increase health worker motivation through a system of complementary remuneration based on performance. Performance based financing operates through contracts between those providing the financing and the various local actors in the health system. The theory behind this scheme is that PBF facilitates efficiency and cost-effectiveness in the utilization of health resources, and is more effective in achieving results than input-based financing because it motivates workers to achieve better performance and it also ensures that funds arrive at the health facility levels instead of tickling down from higher levels in the system.

Context and Background In xxx, human capital is a key component for the delivery of health care, including HIV/AIDS services. Motivating health workers and keeping them in the public sector is challenging, especially since many of them work under difficult conditions and in remote areas. Several experiences of PBF schemes have been initiated in an effort to reinforce health providers' incentives. For example, in 2002 the NGO's Cordaid and HealthNet introduced performance based contracting for general health services in health centers in two provinces of xxx. In 2005, the Coopération Technique Belge (CTB) introduced PBF in other 3 provinces. In 2005, PBF was also introduced for HIV/AIDS Services one of these provinces. While the details of the different experiences vary, the current PBF schemes rely essentially on paying a premium to the health facilities depending on the number of services delivered for a set of previously agreed indicators. In other words, the contracting agency "buys" a package of services at a predetermined rate.

Over the course of several years, it has been decided by the government, in consultation with several other donors - including the WB - that the PBF approach had to be extended on a large scale in the health sector. Three types of expansion were planned.[13] First, the PBF was going to be extended to most frontline health centers in the country. Second, the package of services covered by PBF at the health centers was going to be extended to a large spectrum of HIV/AIDS services. Finally, PBF was going to be introduced for community based health activities.

The impact evaluation rationale The first PBF pilot initiatives had been documented and the available information pointed to increases in the quantity of reported services. While these were very encouraging results, it was argued that such a large scaling-up effort needed to be accompanied by a rigorous prospective evaluation of the impact of PBF on health services delivery, as well as on the health status of the population.

In particular, the study is designed to test the following hypotheses:

- PBF increases the quantity of health services delivered
- PBF improves the quality of the services provided
- PBF improves the health status of the population
- PBF improves the mix of health services provided
- PBF improves the motivation and behaviors of the health providers.

The study was designed to test the effectiveness hypothesis for the following interventions: 1. the planned expansion of performance based contracting in health centers; 2. the planned expansion of performance based contracting for HIV/AIDS services. For ethical reasons, it was not desirable to evaluate the impact of performance based contracts separately for HIV/AIDS services and for other services in health centers. Hence, it was proposed that these two interventions be implemented as a "package". In the impact evaluation, it is this package that will be evaluated rather than the two interventions separately.

Due to the limiting factors of financing and central administrative capacity to implement the schemes, it was proposed to the Ministry of Health to phase-in the schemes in such a way that the phase-in could be used for evaluation purposes.

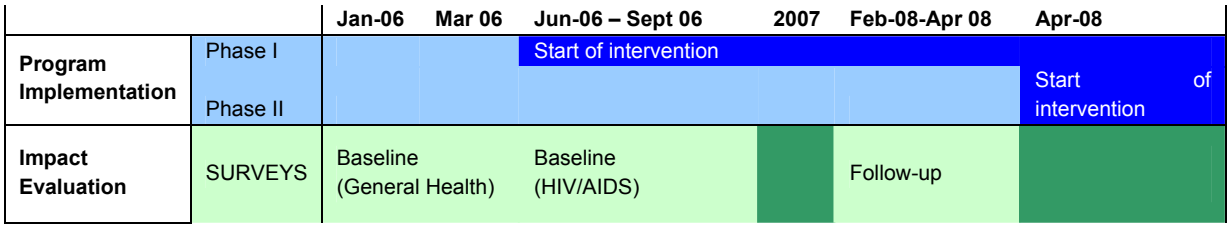
[13] The extensions are not mentioned in chronological order.

Evaluation stated objective The evaluation's objective was to create evidence of what works and what doesn't and help the government decide how to expand the PBF. Plus, the monitoring system and data collection set in place could have served as the basis for a long-term monitoring system of the impact of health interventions. Also, since **t**he program WB Task Team Leader (TTL) identified weak capacity for evaluation design and analysis, the proposed study aimed at building capacity at the national level for conducting evaluations of policy interventions.

3.3 The Evaluation Design

The evaluation took advantage of a prospective randomized experimental design. The identification strategy made use of the expansion of the PBF program over time. For the purposes of the study, Phase I districts were defined as those districts where PBF for health services was going to be implemented by June 2006. These districts were going to be considered the "treatment" areas. Phase II districts were defined as those districts where PBF was going to be implemented by April 2008. These districts were going to be considered the "control" areas. The impact evaluation strategy was to measure the health situation before the start of the package, in both Phase I and Phase II areas (the "baseline"), and to measure the health situation again after the exposure period and before the roll-out of the package in the Phase II areas (the "follow-up" survey) (see Figure 1 below). This gave an exposure window of an estimated 19-22 months in order to capture the impact of PBF on the main indicators.

Figure 1. The roll out plan



The evaluation strategy did not call for any withholding of financial resources to districts located in Phase II areas. In order to determine whether or not the incentives based approach has a more positive impact on health services than lump sum payments, health centers located in Phase II districts received "input based" funding relatively equal in size to the amount dispersed to Phase I health centers as "output based" funding. In this manner, no financial resources were withheld from any group. In effect, what was tested was the difference between receiving lump sum payments (input-based), and PBF payments (output-based).

The objective of the identification strategy was to produce a balanced sample between Phase I and Phase II districts. If the Phase I and Phase II districts were balanced at baseline, then differences at follow-up for health outcomes and other indicators could be attributed to the PBF program, rather than to some pre-existing difference between the two groups.

For administrative reasons, the Government required that the rollout of PBF take place at the District level. Therefore, all facilities in a district had to be in either the first or the second phase of the rollout. In order to identify a sample of districts which was relatively balanced at baseline, we followed several steps. First, the areas of the country without performance based contracting in health centers were paired into couples based on similar observed characteristics including rainfall,

population density and livelihoods. Once districts were grouped into similar couples, the districts within couples were randomly assigned to treatment and control groups.

4. The Case Study findings

The evaluation process As discussed one of the study objective was to reconstruct the steps that were undertaken to initiate, design and implement evaluation giving particular attention to those that required most collaboration between the stakeholders involved in the process. The main steps, according to the interviews and the documentation analyzed were the following (clearly the steps didn't necessarily occur one by one): i) the demand for an evaluation: the WB program TTLs saw the potential value of a prospective evaluation in validating the program theory that they strongly supported and decided to ask the evaluator opinion on whether one could be designed for their program; ii) the *evaluability* assessment: the TTLs presented the program's implementation design and theory to the evaluators who came up with an hypothetical design, the feasibility of which had to be verified through broad consultations with the government counterpart and the local stakeholders (e.g. the donors) involved in the program; iii) the design of the evaluation strategy: evaluators had to deal with many design issues: defining the identification strategy, specifying the appropriate unit of random assignment, setting a desired level of statistical power, dealing cross-over and attrition, confounding factors and design methods to preserve randomization; iv) the data collection: evaluators had to find and contract the local institutions which had to collect the baseline (and follow up data), and work with them to define the sampling strategy and do the actual data collection.

According to the evaluators the most difficult part of the process was that of getting the buy-in of the government and the donors involved in the PBF. Below are extracts from meeting minutes were concerns were raised with respect to the roll-out defined for the evaluation.

> *Extract From Minutes of Meeting with donors 1.9.2006*
> xy also stated that our rollout plan is really not a top priority, as the MoH is currently working on securing funding and actually convincing donors of the PBF approach. There are several actors who are not convinced the PBF approach is the most effective method for service delivery.
> Concerning the timeline, xy believes it is unlikely that the treatment roll out will occur anytime in the next quarter. It actually seems unlikely that the treatment will roll out any time soon – there is no way to determine a timeline until they have some idea of funding source/amount.
> There is funding secured by the WB to fund PBF expansion – however it is unclear whether or not the government has a timeline for when this expansion will take place, and whether or not they have fully committed to the rollout plan (i.e. To rollout in the districts designated as treatment, and remain out of the districts we have designated as control).These are issues that should be addressed by xy with the directors and SG, and by the WB team in a Steering Committee meeting.

> *Extract From Minutes of Meeting of the Evaluation coordinator & PBF Director 1.18.2006*
> Due to decentralization process, everything has been set back. They are still trying to organize who is in charge (hospital, health facility, district levels) and who will be responsible for arranging the contracts. The director believes it will be difficult to adhere to the roll out plan. They have marketed the PBF approach to a large extent, and he cannot see how they will be able to deny treatment to Phase II districts who request the PBF approach.

The evaluators had to do a lot of **e**xplaining of the purposes, advantages and disadvantages of experimental designs to the high-level government officials and the ones implementing the program that is being studied and dealing with and countering objections to randomization. A very intense process of consultation was required to find a situation in which randomization was going to be most easily implemented, the idea of comparing PBF with input-based contracting. When defined, the roll-out plan had to go through a process of socialization. Finally, the program roll-out had to be monitored closely to ensure the integrity of the randomization process and ensure exposure to PBF for enough time by avoiding contamination. This required: i) regular participation in the PBF technical committee meetings by the impact evaluation team members ii) monitoring the threat to internal validity of the sample from political pressure to expand PBF into Phase II districts before 2008 and from possible imitation of PBF in districts of Phase II (many facility directors and providers in Phase II districts heard of PBF through colleagues or media so attempted to imitate treatment). Also, additional data collection efforts were required to assess the fidelity of the intervention being tested. Finally, for this strategy to work the follow-up data collection activities had to be completed before the expansion of PBF into the Phase II districts. The evaluation team had to be always informed of the implementation status in order to ensure timely and coordinated implementation of program and evaluation. Also the definition of the measures of impact to be collected required some consultation with the local stakeholders. Even though everybody agreed that it would be useful to show the impact of the program on health indicators, the implementation team and donors voiced concern with the feasibility of seeing the impact in the timeframe selected. The design has been partially adapted to fit the perceived constraints (increased treatment window), but this was still not met with 100% approval.

The evaluation team composition When asked who was part of the team working on the evaluation, the answer was coherent across interviews: the operational TTLs at the WB, the evaluators, the government officials involved in the PBF and the local researchers involved in the data collection. The "catalyst or leaders" of the evaluation, as defined by the evaluation coordinator, were the two TTLs who initiated the process and the renowned evaluation specialist who was asked by them to assess the feasibility of an impact evaluation for the PBF system. The composition of the team was considered especially important by one of the TTLs who called it the "magical team" and described it as follows:

> ...the team included...the Minister of Health (MoH), the Director of Planning, the Director of the Program, and the Minister, who was particularly supportive, sometimes tough because he was very starightforward on his political constraints … So that was a key group. The second part was the School of Public Health (SoPH), a local research Institution, and in that we were incredibly lucky to have of course, xy who is a brilliant thinker…And then really having an academic that is extremely creative and extremely strong like xy and team…It is interesting because I think that without these three actors it would have not worked the same way. Each of them has been really critical. And I think there is a fourth actor, and I don't think I am in the best place to talk about it because it is at the Bank and it's me, the project manager.

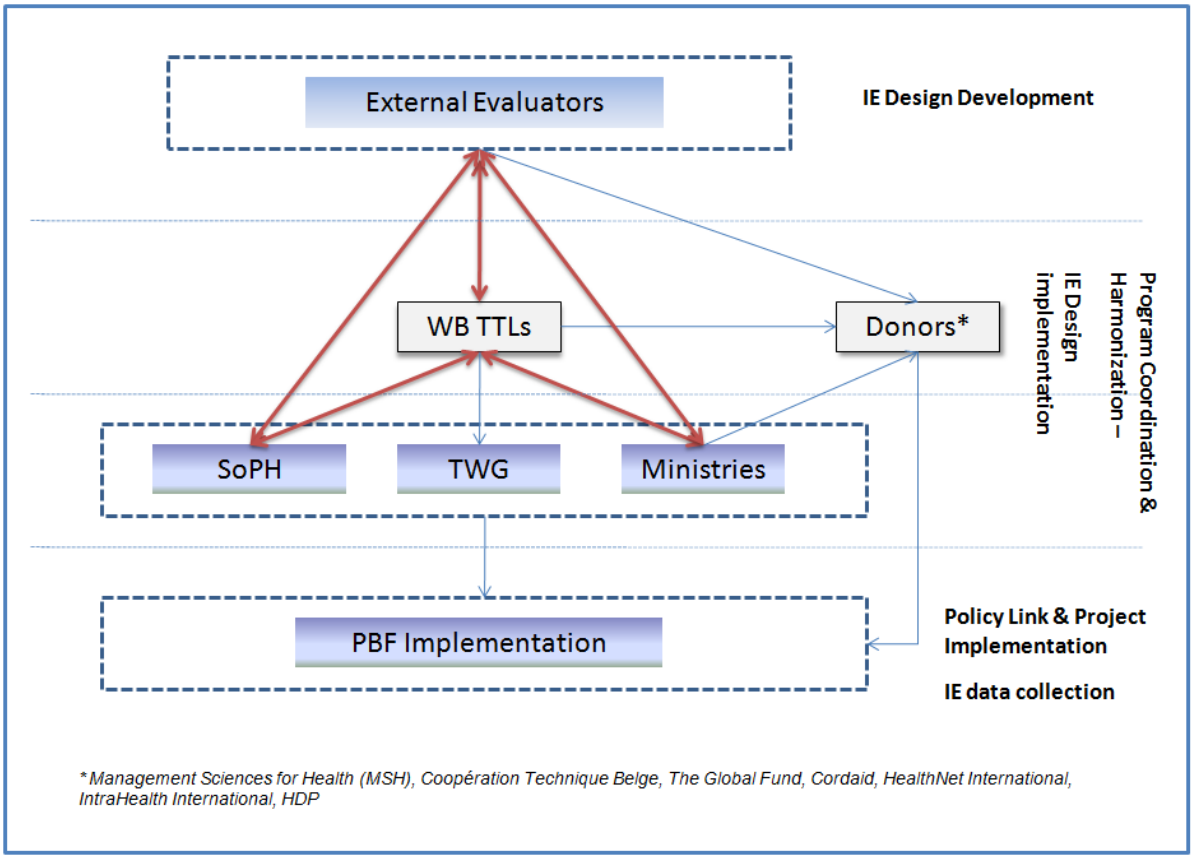
As reported by the local focal point, the TTLs' role was fundamental to first "set up a good environment for the evaluators and provide advice on how to integrate the evaluation into operations and on how to work in collaboration with the government. As commented by the TTL:

> ...this would also not have happened if I had not constantly integrated the problem of the IE into the broader dialogue and sometimes I have put my weight into discussion to make sure that this would not be derailed. And also you know help design the program in a way that would allow be politically and oprationally feasible… So that was really the contribution of this force group which consisted of myself and the team. We really constantly tried to think how you can find ways of implementing the program that would accommodate the IE yet without jeopardizing the program, kind of creative ways that would accommodate both goals, the goals of scaling up and the goal of actually measuring the impact…

Another relevant actor on the government side was a high-level official at the National AIDS Commission, who contributed substantially by guiding the process, making sure that the questions asked were relevant and asked in the right way. Also, a strong mandate in favor of the IE came from the Secretary General, who did a lot of work to get everyone to agree to the initial design.

Stakeholder participation/Interaction between stakeholders The main evaluation stakeholders and their relations are represented in an simplified way in the Figure below which replicates the governance structure of the evaluation according to the documentation and information collected through the interviews.

Figure 2. The Evaluation Governance Structure



In a seminal article, Cousins and Whitmore (2007) distinguish three separate dimensions of participation and collaboration: (1) who controls the evaluation process (a researcher-practitioner continuum); (2) stakeholder selection for participation (a continuum from all legitimate groups to

just primary intended users); and (3) depth of participation (a continuum from consultation to deep participation). According to this criteria I would categorize this evaluation as participatory: i) the researchers had definitely more control over the evaluation design, but the practitioner had the power to actually compromise its integrity, if it did not feel ownership over the evaluation; ii) the stakeholder selection for participation was limited to the decision-makers, since the interaction with treatment areas, program staff and clients were voluntarily kept at minimum, but the primary audiences were the people influencing the policy, high-level officials; iii) the stakeholders depth of participation varied according to the stakeholder; in a continuum from consultation to deep participation, I would choose an average, since everything was defined in consultation, but the implementation was left to the locals when the international evaluators would go back to the US, and the final analysis was left to the external evaluators to avoid compromising the validity of the results. The level of participation was dictated by necessity.
There were several types of actors involved with their own motivations, forms and levels of participation. As described above the actors mostly involved were the members of the core evaluation team, the members of the government team responsible for the implementation of the PBF and the local researchers initially responsible for the data collection process. Participation in the core evaluation processes has been both direct (attending meetings) and indirect. Given the distance many communication means were used: VC, telephone calls, emails. Focal points were in charge of briefing the core team members of possible relevant information. For instance, the focal point working at the SoPH became member of the PBF Technical Committee, the Technical Working Group (TWG) where issues regarding the implementation were raised so that he could act as liaison with the evaluators and keep them up to date on implementation issues of interest for the evaluation.

There was some criticism from one member of the TWG who maintained he never met the evaluators. He commented that the only person he had contacts with was one person from the SoPH, but he didn't even meet the other people at the SoPH (which, he says, is indicative). During the interview he recalled that "there was a lady presenting the results of the IE baseline at a workshop and I have never met her, in three years that I have been working here…" He was also critical of the choice of the measures of impact; he stated that "the measures they use are esoteric, far-fetched…" He was consulted once and tried to criticize something on the phone with one of the evaluators, but according to him his concern was not taken into consideration. Yet, the rest of the non-evaluators, believe that the collaboration with the evaluators has been, on the whole, very good.

The evaluation was once presented at the WB HQ, and the following is an extract from the description of the session:

> In the case of xxx, the Government has been in the driving seat, ensuring that the general health and the HIV/AIDS sector program implementation and evaluation were well synchronized by leading a strong coordination among the donors. The result was an impact evaluation design and a geographical and chronological roll-out plan of PBF in xxx that was carefully adhered to by all stakeholders.[14]"

[14] Seminar held in November 2008 at WB HQ.

According to the local researchers, the effort was really collaborative because when the evaluation team left, the local actors where the ones making this happen by sticking to the plan. On the other side, when the evaluators would come on mission to xxx, meetings were organized with the stakeholders to consult and define the work in collaboration.

Also the evaluator's responses indicated that the collaboration between them and the local counterparts had worked well. There was the right balance between being involved in the evaluation design built into operations but not in how they program was actually implemented. In fact the evaluation was designed and carried out as an external evaluation. In explaining what is meant by "external" evaluation, one of the researchers said:

> …you always need the government to be involved… if you are answering a question that is relevant for them, than you have higher chances to have a good evaluation…It is important to have them involved to feel ownership of the results. But it also key that the people who implement the evaluation design measuring and analysis cannot be part of the program…all has to be done in collaboration with the government. You can have more and less involvement of the people in the government…

With respect to the actual interaction, the coordinator felt that there were some information gaps:

> ...there will always be information gaps between the evaluation team and the program team…even if everyone is on the same team, not everyone knows what everyone should know and when they should know it and xy (a member of the TWG) kind of helped us to know what was going on in the implementation of the program, not only implementation design, but in the actual implementation of the program...

When asked about the level of participation, the coordinator said:

> It is hard to say, I think there could always be more…But at least the challenge we faced with this evaluation was that as time went on, more and more people were involved…and it was really hard to get everyone on the same page... There was always a lot of dialogue going on, but it was with different people at different points in time and about different things…I would say out of the whole group, xy (the SoPH focal point), more than myself, more than anyone, understood what was going on from any point of view at any point in time…because he had both stories…

Manifestations of process use According to the analysis, although the real impact of the evaluation will depend on the findings, the evaluation process contributed to some of the types of process use described by Patton (2008) and Forss et al., (2002). The complete typologies are described in the Table 2 below. Other effects attributed to the involvement in the evaluation process have been classified under capacity building. In particular, I have identified the following types of PU and ECB.

Infusing evaluative thinking Mainstreaming evaluation depends on the capability to internalize it as a value, integrating evaluation into the organizational culture (Sanders, 2003; Patton, 2008).
The government officials and the WB managers demonstrated to have mainstreamed evaluative thinking in their work by providing support for the adoption of new prospective evaluations,

notwithstanding the substantial efforts they had to put into the evaluation of the PBF described in the study. In fact, even if the on-going evaluation has not produced its findings yet, the Director of the PBF and the WB operational side involved in the PBF are already thinking about how to design the evaluation for the next round of the program. The evaluation coordinator observed this change in the stakeholder's attitude towards evaluation. The program director that is now involved in the consultations for the new round of evaluation, speaks about the roll-out plan as if he perfectly understood what it means and implies, while in the past evaluation, everything the process brought with it was a "continuous surprise".

Table 2. Typologies of Process Use
Patton (1997, 2008)
1. Infusing evaluative thinking into the organizational culture
2. Enhancing shared understandings
3. Supporting and reinforcing the program intervention
4. Instrumentation effects and reactivity
5. Increasing participants' engagement, self-determination, and sense of ownership
6. Program and organizational development
Forss et al. (2002)
1. Learning to learn
2. Developing professional networks
3. Creating shared understanding
4. Strengthening the project
5. Boosting moral

According to one of the evaluators, one of the biggest value-added of the evaluation process was represented by the "general thinking about what is an evaluation and how it can be helpful" in the client country and WB managers.

For instance, at the beginning it was very difficult for people to understand how and why an evaluation is designed before the start of program implementation …A local researcher recalls someone from the government stating "these Americans, they come on mission even before we start implementing the program…they just want consultancies". Yet now, "after a lot of explanation, a lot of presentations", the researcher notes that "the people involved really understand…they send me emails because they want to know the results of the analysis". "And also, they really want to understand if the program works and if something is not working, why it is not working…so there is no pressure from the government to come up with positive results", says the researcher from the SoPH.

It was recognized that the government was initially "not very favorable". Many times the TTLs received calls from the evaluator to request support thinking that the government had changed their mind and didn't want to stick to the evaluation plan. So the TTL had to go back and discuss with the government why to do the evaluation. And now, they are talking about the next

evaluation….and the government immediately thinks about testing different variations of the same program, as in the current evaluation. “

According to the two most experienced evaluation specialists:

> “...the government became more and more enthusiastic … it is an incremental process which accumulates over time. It starts with the involvement of the government in doing impact evaluations of its programs, and which “steamrolls” and reaches a “tipping point” after which evaluations becomes a “culture”. According to the specialists, the cultural change occurs when the government “starts seeing that the evaluation is about the program and is not about questioning their competence, but the intervention”. Also, at some point it becomes like a “social norm…if you do evaluations, you are modern and transparent…it’s a cultural change”. Yet, again, “it does not happen in one step…it is more than capacity at training, it entails changing the way politicians see social policy and how important it is to learn and improve, and how external evaluation gives them this opportunity”.

I maintain that also the following is a sign that the evaluation impacted on stakeholder’s views. During the interview, the Director of the PBF pointed out that without the IE they could have not known if there is any improvement produced by the program, and that this knowledge is important also for other countries.

According to one of the TTLs, the local counterparts are now more educated consumers of evaluations, “initially the WB was leading the process and they were collaborating with us, but now we need a paradigm shift: they are now leading their own research, there are keen on doing this themselves”.

Also other actors thinking and awareness on the evaluation logic changed. The researcher from SoPH feels that “now when they say IE, I know exactly what they mean.”

Another important change in attitude and thinking due to experience of the evaluation process is represented by the following quote from the WB TTL:

> ....from day 1 when we started thinking about testing contracting for results I was always haunted by the XX study which shows that contracting works better than non-contracting...and there was no point in testing that... We really wanted to control for the money affect which was interestingly where there was complete convergence of the interest of the impact evaluation and the interest of government because they also didn’t want to give money to one group and not to another…So in fact, for me, that experience showed that it’s matter of commitment because there are not such contradictions between the goals of the policy maker and those of the evaluator. If you really think about it and go through it and think carefully, actually they are quite consistent and convergent…

The TTL learned that it is a matter of how you present the evaluation to the client, since there is basically no contradiction. The conflict would exist if the evaluation would impose not implementing the program at all in the control groups, which according to the TTL “is not an option. In fact the idea is that one can implement a program introducing variations to answer

unresolved questions, does not produce any conflict… The TTL believes that this approach to evaluation became "probably more of a philosophical view".

Moreover, one of the evaluators recalled observing a big change when going back to the field after at least a year from her last visit. She was very involved in the design phase and was very surprised that everybody in the field was now talking very confidently about Phase 0 – Phase 1 – and Phase 2. Also in this case, the evaluator maintained that the way you present the design and the terminology used was key in increasing the receptivity of the actors involved. The technical terminology "Control and Treatment would have not stuck" maintains one of the evaluators.

To sum up, changes in thinking and the diffusion of evaluative thinking were represented by the TTLs new "philosophical view" of experimental impact evaluations, the government buy-in of the evaluation and willingness continue evaluating its own programs, and everyone's belief in the value of evaluation and the importance of learning and validating theories behind programs.

Enhancing shared understanding The very process of formulating goals so that they can be evaluated usually has an impact on how people think about what they're trying to accomplish, long before data are actually collected to measure results. The logic and principles of evaluation also can be useful in negotiations between parties with different perspectives (Patton, 2008).

According to the WB TTL, the first and most important step of the whole process was to get all the actors (different donors) that were working on the PBF independently and, under the leadership of the MoH, actually define what it is that was going to be scaled up. In particular, the TTL commented:

> "It was extremely difficult to actually get people to clearly define what is it that they are going to do" … "you first need to know what you are going to do and then think about what it is that you are going to test" …"it seems trivial but it is one of the major externalities of the impact evaluation… you know even without results, the whole process has been very useful! It pushed people to be much more rigorous about the way they were implementing the program"

Yet this added value might not be attributed to the IE by the other stakeholders involved, such as the government. In fact, the WB TTL quote continued as follows:

> I think they (*the government*) appreciated the added value of better defining the program...and so on…but they may not have perceived that the IE was a driver in that…because you know, thing are not always that clear-cut, but there was no question in mind where i was sitting that a lot of my thinking on the program was informed by the rigorou of the ie and what we are going to try to test

The TTL's view was shared by the evaluation coordinator who believes that the evaluation increased the necessity to agree upon what was going to be measured. The coordinator remembers being involved in a meeting were the donors where questioning the fact that the IE was going to measure the program impact in terms of health improvements at the HH level, while, according to their view, the program was supposed to increase services at the health facility. According to the evaluation coordinator, the dialogue around this point made them think that the results chain

"doesn't just stop there" at the output level, consequently increasing their shared understanding of the program objectives.

According to a local researcher, the understanding of the program was improved by the existence of treatment and control. Those implementing the system had to know very well which were the differences between the controls and treatments, since, a part from the PBF incentive system, everything else was supposed to be exactly the same in all districts. Also, due to the evaluation and the necessity to isolate possible confounding factors, the stakeholder's where paying more attention to the environment and contextual factors and it appears that a big concern in every discussion with relevant stakeholders was how to control for the external factors. The effect was that the evaluators had many occasions to explain the logic behind the randomized design and the fact that if the "other things going on" happened at random, that would have not compromised the evaluation, and that the existence of a comparable control group would have maintained the validity of the results accounting for unobservable factors, increasing the stakeholders' understanding of the evaluation. Also, this discussion would increase the shared understanding of the program by promoting a more broad and open reflection on all the factors possibly influencing it.

Another element, which is very context specific, but possibly happening in many developing countries where there is the presence of a large donor community, was that the evaluation was used and represented a tool to facilitate harmonization.

While the IE was designed, the government was really "trying to reassert the leadership role of the MoH in the donor coordination", says the evaluator who was present at a meeting where a very high-level official presented the idea of an IE executed under the leadership of the WB and that "he expected everyone to follow the same plan". The government's negotiations became more powerful when the necessity to stick to the national plan was further motivated by the evaluation plan. The evaluation team drafted a letter sent out by the Ministry asking to stick what their roll out plan (see letter attached). The evaluator commented that "they kind of asked the government to send the letter, but they were convinced enough to actually do it".

The relevance of the harmonization role of the evaluation was highlighted by the evaluation coordinator:

> ...you just realize that with the political landscape and everything and the donors…the money coming in and different objectives from different programs, different ministries, and people working in the government, there are just so many things gong on at once that you can see like, at least for me when you had an evaluation into place it helped to try and line up incentives among the different groups.... It makes sense to have something to line up objectives of all these different groups... and this (*the evaluation*) helped to get a system!

To sum up, the necessity to come up with an agreement on what exactly was going to be measured, to discuss the possible confounding factors, and the necessity to adhere to one common national plan, forced the stakeholders to discuss more in depth and broadly the program under evaluation, thus clearly increasing the shared understanding of those involved.

Strengthening (supporting and reinforcing) the program intervention The IE was not properly an intervention-oriented evaluation as defined by Patton (2008), yet the evaluation process improved program implementation from the very beginning by focusing staff implementation efforts. The program staff used the evaluation design (roll out plan) formulated by the external evaluators as their framework for their plan of work (Patton, 2008). Moreover, since this effect was not intentional, there is no doubt that the evaluation credibility was not compromised, nor the capacity to render an independent summative judgment.

According to one of the WB TTL, the idea itself of implementing one national program was triggered by the evaluation. In fact, she says, "the moment you want to go for an impact evaluation, you need a sizeable sample…and it is much more interesting at the political level to do an evaluation of a national program instead of a micro-project". Yet, the quote continues "You need to have a program" as to imply that that was not the case when the decision to have a prospective IE was first made. One of evaluators also described the choice to define a national model as part of the "*evaluability* process". He stated that to design an IE with statistically significant results they "needed a model of PBF that was not exactly the same but very similar, at least very comparable in all the treatment areas" and that "required a lot of convincing from the government to the donors actually implementing the plan".

On the other hand, according to one of the evaluators, the decision to have a national model and which model, was not made by the evaluators, since that would have violated the external evaluation logic, and it was too politically sensitive to get involved in it. Yet the evaluator remembers the local actors asking about the possibility of evaluating the relative impact of the different models. According to one evaluator, the fact that the evaluation of different models would have been much more difficult and less rigorous (since it would have been difficult to understand the choice of a model in each area) "might have set them thinking, but I can't say".

In the opinion of the evaluator renowned evaluation specialist, one of the changes brought by the evaluation in xxx ("it happens everywhere – he added"), was related to strengthening the intervention in response to the increased accountability caused by the existence of an evaluation. As an example, he reported the fact that the government decided to expand the exposure period of the treatment areas to two years before starting the PBF in the control areas "because they really wanted to make sure that things work". Also, "they became more precise on what they want to do ex-ante so the program got designed better and faster…and much more attention was paid to implementation".

Finally, the existence of an impact evaluation called for constant monitoring and attention to preserve the integrity of the roll-out plan, and which has strengthened the program. As reported by one of the TTLs:

> Once the roll out plan was out there, what became absolutely critical was to constantly be on top of things, monitoring the process, being in constant contact with both the technical and the political level so that the government would stick to the roll out plan, because of course they have all sorts of other priority would could have completely diverted the direction...

In fact, as the study progressed, concerns were raised regarding the control districts. The MoH and SoPH teams expressed the concern to the evaluators that some facilities in the Phase II districts were developing a method for imitating PBF. With this in mind, the IE team, in coordination with the SoPH and MoH instated a monitoring system. Also, the existence of possible confounding factors (e.g. the roll out of community insurance schemes) required the collection of additional data which are now used by the evaluators to produce additional analysis on the ongoing healthcare system and reforms in xxx, and not only on the PBF. In the same way, some of the field work and preliminary results triggered the necessity to collect more quantitative data, but also accompany the quantitative study with a qualitative analysis of the how the incentives were working at the health facility level. A process evaluation was financed by the WB and an additional case study is currently being designed by another research institution. This was supported by the local officials and researchers especially since it will provide a more immediate feedback mechanism to the program implementation level.

In the context of strengthening the program, the evaluators' impression was that the saying "what gets measured, gets done" was occurring with this evaluation: "When people know they are going to be studied, and they know that you are evaluating them…more effort goes into the program, because no one wants to show that the program has no impact". The effort was put into program implementation, even if the evaluators kept their role of external evaluators and never got involved into the actual implementation of the program. As an example, the researcher carrying out the process evaluation noticed that there had been some changes in how the money was spent in the health facilities during the period they had been using the PBF system. Inquiring about the reasons for this change, the consultant felt interpreted as a result of the evaluation. In fact, even if the staff at the health level facility was not involved in the evaluation (not to be influence by it), the district level team was "trying to influence the health centers to work better" since they knew about the evaluation and the objectives to be measured.

To sum up, the decision to evaluate the program obliged the implementation level to think about one national model that could work most efficiently. Even if the evaluators were not getting involved in the implementation, the fact that the program was under evaluation created this incentive to do things better, in line with the saying "what gets measured gets done".

Increasing participants' engagement, self-determination, and sense of ownership Actors' role and engagement in the evaluation changed in the course of the evaluation. An example is represented by the local researchers of the SoPH. During the baseline data collection, they were just implementing the research project as it was given to them by the WB evaluation team. They didn't feel ownership of the evaluation, nor the data collection, which according to the evaluation coordinator was the reason why they didn't provide too much contribution to the development of the instruments, and made some mistakes in the data collection. The SoPH focal reported that some of the errors in the data collection where due to the fact that the survey administrators didn't understand how the evaluation was working. On the other hand, by the time of the follow-up data collection process, the focal point at the SoPH became integral part of the evaluation team, and felt ownership of the process. His role was fundamental in many steps due to his understanding of the context, the problems and constraints. The SoPH was instrumental in crafting a more locally sensitive instrument which increased substantially the reliability of the data. The government also

incrementally felt stronger ownership of the evaluation, especially due to the strong belief in the program, and as a consequence, of the roll-out plan designed for evaluation purposes. The areas where the implementation of the PBF schemes was delayed were "their controls". Yet some of the government is still not convinced about the possibility of demonstrating the impact on the health outcomes, and I imagine that it will not take ownership of the results, especially if not showing the desired impact. Yet this remains to be seen. According to one of the TTLs, the one thing she would do differently is involving the countries and the people at the programmatic level since the first step, asking them to identify the research questions, also in order to increase their sense of ownership since the beginning. Ownership was definitely felt by the WB TTLs who believe to have played a key role in making this evaluation happen.

It appears that the ownership of the evaluation grew incrementally as the engagement of the different stakeholders increased and so did the understanding of the evaluation logic.

Developing professional networks Evaluation appears to be a way of rapidly interacting with many people, and hence an opportunity to build networks (Forss et al., 2002). This was a clear effect of the evaluation analyzed in the present study, and is particularly relevant in the context of development work. The evaluation team invited the PBF implementation officials, and the SoPH local researchers to various international workshops to discuss their experience. Especially the focal point at the SoPH was engaged in workshops in Berkeley and at the WB HQ, where he had an occasion to meet impact evaluation experts and academics that are currently also supporting his work for his PhD dissertation. Also for the WB TTLs and the government, having evaluated the program represented a reason to meet other country officials and discuss their PBF model, by becoming part of a thematic cluster on impact evaluations of PBF systems in different developing countries.

Capacity Building As described above, one of the stated objectives of the evaluation was to create capacity in the client country. Yet, according to the information collected, it did not appear like an intentional effort. Quoting one of the interviewee it appeared more like it was written because "you always have to write that to get the money". In any case, it appeared from all the interviews that capacity has been created both intentionally and un-intentionally, and everyone involved seemed to have learned a lot from this experience.

In particular, the interviewees all agreed that the evaluation process had a substantial impact on the SoPH. The local researchers of the school were actually involved in all the steps of the process. They were first involved only in the baseline data collection, but finally they kept collaborating with the evaluators, and are now considered part of the research team. Yet the collaboration and capacity building process came "little by little" and was not initially intentional. According to one of the evaluators, the initial choice to work with someone locally was mainly driven by the constraints of time and financial resources that impeded them to hire an international survey firm. The SoPH was already collaborating with the MoH and was a natural choice, yet, as stated by the evaluation coordinator, they couldn't know if the collaboration would have continued throughout the evaluation, especially given the initial scarce resources and limitations of the Bank procurement policies, which do not allow many small contracts with the same firm. The uncertainty in terms of

the length of collaboration with different actors seems to be a reason why, at least initially, the effort to create capacity was not intentional.

All the evaluators believe that the collaboration with the SoPH really created capacity at the local level. One of the evaluators believes that "it is one of the best things that happened". The evaluation coordinator, who worked most closely with the SoPH, commented as follows:

> I think initially it was a little slow, it was just, you know, we work with the local researchers, the people involved but we didn't know how long term it was going to be…but then because of the interaction with XX and XX (the researchers), it became obvious that if we invested a little more time with them working on the research, getting them more involved with the data, that was a really good opportunity for them to become better researchers and I think that became intentional capacity building
>
> ……
>
> …it is ridiculous what happened to some of the people from the SoPH through this project. They also got to work with a quality assurance firm financed by the evaluation (which in the meantime raised a lot of funding) that came in for months and worked for them not only on data collection, but also on field management, data entry skills, very technical skills…

Everybody agreed that also the government learned from the involvement in the process, at least in terms of what is the logic of a phased roll-out plan, and what it takes to do an impact evaluation, i.e. what are all the main steps required to conduct an IE. As reported before, the government is now involved in a new round of evaluation and knows what does an impact evaluation mean and as opposed to the first experience, everything that comes with the evaluation doesn't come as a total surprise". The TTL who had most contacts with the government believes that the "three top-managers that were really involved will be certainly left with a more in-depth understanding of evaluations". Yet, she believes that not enough has been done to achieve this ECB goal, and probably more work should be done.

With respect to the devaluation design, and the government's understanding of the evaluation logic, the evaluation coordinator feels that it was a form of unintentional capacity building but "it happened for sure":

> I think we took for granted that people would just understand if we presented the roll-out plan or the evaluation design….But I think it took the last 3 years of Phase I, Phase II, why Phase I and Phase II, what does that mean for the evaluation design, for them to really start understanding evaluation, at least one form of evaluation…and what is it that this is giving us rather than what we typically use which is before and after…and I saw lot more questioning on what was used in the past as standard…but i don't think it was us going in and thinking let's educate them on evaluation design…

In the opinion of the coordinator, the roll-out plan became a tool to try to educate and inform the government on the evaluation. It was also a useful mechanism to start a dialogue on how the actual implementation was going to be. The evaluators were drafting the roll-out plan on the basis of the information provided by the government and the donors and was going back to them to get their

confirmation that the plan reflected what they agreed. This created a back and forth which, according to the informants, enhanced the government and other actors evaluation capacity by improving their understanding of the logic behind having districts randomly assigned to treatment and control areas.

The evaluator involved in the initial design of the roll-out plan, also highlighted the importance of that step in "making the government realize they needed to do something about their data". In trying to design a map and collect the data necessary to define which districts were going to be part of which phase of the roll-out, the evaluator and the field coordinators needed to manually draw a map and color it according to the information personally collected from the donors on who was working with which model and where. Each donor had its own information and database, the system was very fragmented, and even if the government was aware of this situation, the necessity to design the roll-out plan provided them with the incentive to do something. The evaluator said that she was "impressed" when she went back after a while and realized how things had changed. According to the evaluator, the working on the evaluation design "brought out the weak spot they had, raised a little consciousness and allowed some discussions....it is when you start using the data that you strengthen the system." Yet the evaluators could not intervene directly to improve the system because there weren't enough resources.

In the same way, the idea to hire a local firm to collect the baseline was not driven by the willingness to create local capacity, but by the scarce resources that wouldn't have allowed them to hire an international firm. Yet the collaboration with the local institution, the SoPH, continued all along the process and capacity was created and the evaluators learned that investing in local capacity was the only way "if they wanted to do it right!"

With respect to the operational side of the WB, both TTLs believe they have learned a number of things with this evaluation, such as the importance of evaluating programs prospectively and of sustaining the impact evaluations in the long term, the feasibility of doing fundraising to avoid using the scarce resources of the government, and the importance of taking a holistic and integrated approach in the evaluation. Also, one of the lessons learned is the critical role of the composition of the team working on the IE. As pointed out by one of the TTLs:

> And I see that is something that could be the lessons learned, I mean having a team composed of these 4 groups: a government, a group of academics (a group of local academics, a group of external academics) who bring in the best in terms of knowledge and then the partner agency, a donor agency that is bringing the money and that is open both to academic discussion and a policy discussion.

Finally the evaluators themselves believe the evaluation was a learning experience. The interviewed provided more or less concrete examples of their learning. In general, their comments can be represented by one of the evaluator's statements: "I learned a lot about all the going from a nice research design on paper to implementing it in practice". Furthermore, the learning was evident in some changes occurring during the process. For instance, the decision to invest in creating capacity and building a sense of ownership in the local researchers seemed to be the result of the realization that to "do things right", as reported by the evaluation coordinator, they needed the contribution of

good researchers with a local understanding (letting alone the availability of more funding). Also, one of the evaluators had to face some problems with the authorization required for the baseline to go in the field, which made her realize that it was a step that the team didn't foresee "didn't anticipate it could take so much time". Clearly that was an important lesson learned.

Different actors, different PU and ECB The bulk of process use that occurred was represented mostly in changes in *evaluand* and the local stakeholders' capacity and thinking about evaluation. As previous case studies revealed (see also Lawrenze et al., 2007) process use decreases as the level of participation becomes less direct. The types of PU are quite different from each other, but the categories are not independent from each other: increasing shared understanding of a program has clearly an effect in terms of strengthening the program. And the same goes for increasing engagement and shared understanding. However the present study seems to confirm Forss et al. (2002) finding that different categories of PU will affect different stakeholders in different ways. As summarized below, each person was differentially affected by participating in the evaluation process, depending on which step they were mostly involved in, and by their personal level of knowledge and behavior (Lawrence et. al, 2007).

The evaluation side: the local researchers and evaluators The local researchers were the ones that appeared to have benefited the most from the evaluation process. They became the focal point for the evaluators abroad, participated to all the meeting and have been trained on evaluation design and on innovative data collection and management techniques. Plus they are involved in the analysis and will publish scientific article jointly with renowned evaluators who made them part of an international network of experts in evaluation and PBF systems.

The evaluators learnt what it means to translate a nice research design into a real operationally feasible impact evaluation. They also learnt how to facilitate discussion about evaluation in a policy challenging environment, and learnt about the benefits and how to make use of the capacity building of local counterparts. The evaluators were exposed to a broader range of ideas about what might be appropriate outcomes and measures for the program through the meetings and workshops organized with local researchers and the international community of practice involved in PBF.

The operational side: government & WB TTLs learned what it takes to do an evaluation and developed an interest in doing it. The discussion of different IE models resulted in a greater understanding and appreciation for the chosen design. They also have an increased understanding of the program logic and increased the precision of the implementation plan ex-ante.

Assessing the experience of the evaluation process as a whole All in all, the experience of the evaluation was defined as an unexpected success. Unexpected because the evaluation process encountered different non-cumbersome obstacles and yet the integrity of the design held for 2/3 years, and the follow up data has been collected successfully. Obstacles included the initial unfavorable environment, the concerns related to the feasibility of implementing a harmonized national plan given the presence of many donors with individual objectives and interests. The evaluators were not totally prepared for the poor quality of the data available from the centralized monitoring systems, and some parts of the work of evaluation process had proved to be considerably more demanding and time-consuming than they had predicted (e.g. receiving the authorization for the baseline survey). Notwithstanding all these obstacles, the evaluation was

considered a success and was described as having been a "rewarding and worthwhile personal learning experience". The majority of evaluators, the local researchers and one of the TTLs (the one that was more involved in the process), by and large, believe that the benefits obtained can commensurate with the resources invested. On the other hand, the government and one of the TTLs (who was more concerned with the level of financial investment in the evaluation) believe that if the costs commensurate with the benefits depends on what will the evaluation findings look like.

The success factors Everybody agrees that the main success factor has been the strong buy-in and commitment of government. The second most important variable was what Patton (1994) defined the "personal factor". Most of the informants attributed a key role to the renowned evaluation specialist, a "creative" and "brilliant" academic and his capacity to "sell" the impact evaluation to the government and get their buy-in, and to a very high-level official who was able to convince the donors to adhere to the plan. He made a difference by being "forceful" and "dynamic". The government showed a lot of confidence in the competence of the team of evaluators partially because of the credibility of the renowned specialist. One of the evaluator commented as follows:

> I had gone to xxx ..... I went back in September 2005 and that mission had a lot more traction because xy (the evaluator) was coming and he had a lot convening power…I go on my own and nobody pays attention….and when xy is coming…and...everybody pays attention to him!

Also the local researcher have been attributed a substantial part of the merit, and again because of their dynamic personalities and commitment to their work. Also the TTLs received credit in integrating the IE into operations and maintaining the dialogue with the government in order not to lose their support. Moreover, what contributed to the success was also the fact that the TTLs and evaluators that started the process were "risk-takers" and decided to start the effort with very limited funding, believing that they would have found a way to tap in with new sources along the way. As reported by one of the TTLs, this could have not been the case: not everyone would take the risk of investing in something that might not get the funding to move further. The experimental design also seemed the appropriate evaluation activity since the credibility of the results is of mayor importance and the setting is highly complex and political (see Taut, 2007). Using randomly assigned control groups, and finding a solution to be equitable by not denying funding to any district, by changing only the finance mechanism helped enormously in establishing credibility (the same result was reported in Riccio and Fitzpatrick, 1997). The government didn't really believe in the evaluation per se, but strongly believed in the program and believed that the experimental evaluation could have validated their model in a very rigorous way that could have attracted more funding. As reported by one of the TTLs:

> The government was on board with it (*the evaluation*), was committed to it, saw the importance of it…understood that at the end of the day…if you show good results and you have evidence based approaches that provide good results, then you are going to get more money from the donors…they are very keen on this kind of strong approach (*meaning experimental evaluation*)…

In addition to each person's specific role, the approach and method, what seems to have made a difference is the team factor. As described by one of the TTLs, each of the members of the team had

a strong intrinsic commitment to the evaluation. The evaluators have a commitment to evaluation and knowledge generation on what works and what doesn't as a public good; the researchers had a commitment in learning from the experience, so to be able to lead their own research efforts; the TTLs and the government had a commitment in demonstrating the impact of the PBF. Even if the motivations were the different, the commitment was aligned and made the evaluation happen.

Other specific anecdotes helped this evaluation achieve its completion, according to the informants. First the fact that the roll-out plan was broadly "socialized", i.e. it was "sufficiently well-explained and sufficiently well spread; everyone knew about the roll out". The documents on the roll-out contained all the relevant information and were immediately translated in the local languages. One anecdote reported by an evaluator described the first and "most important" meeting with all the donors and the government, as key in obtaining the buy-in of the key stakeholders. During the meeting, the renowned evaluator explained the how and the why of doing IE. The most interesting thing of that meeting was that since not all the audience spoke good English, everything was translated in French by another evaluator. The fact that the concepts were repeated twice for who understood at least a little of both language, was crucial and very powerful in conveying the message about the value of IE. The government asked the evaluators to help them design the evaluation for the PBF right after that meeting. Another element that appeared to have made the difference in obtaining the buy-in and socializing the randomized evaluation was the vocabulary. Calling the treatment and control areas, Phase I and Phase II seemed to have had a role in increasing the adoption of that terminology which had diffused much more than if the evaluators would have used only the technical terminology, and increased the familiarity with the roll-out plan.

PU and ECB I would argue that PU and ECB both occurred as a result of the involvement in the evaluation process, and were mostly an unintentional/unintended outcome. They were at best instrumental to the design and implementation of the evaluation. Evaluators actively involved government officials (who were needed in the planning of the evaluation and to preserve its integrity during program implementation) and local researchers (who were involved initially due to lack of resources to hire an international firm, and later on because of their unique understanding of the local context). Yet the emphasis was not on capacity building which was not the evaluation's advance organizer (as defined by Stufflebeam, 2001). In evaluations where ECB is considered an advance organizer, training of staff, for example, becomes a priority (Fitzpatrick, 2004), even if under a tight budget. In the PBF case, communication with the government and the local researchers was mostly done to design the evaluation according to program implementation and understand the local context to be aware of possible research constraints and frame the most meaningful questions. One of the comments of the renowned evaluator was very indicative of his thoughts of the ECB role of the study, noting that it was stated as an objective in the evaluation concept notes prepared by his team. When asked about it he said:

> I think the best strategy is to get them (*the local stakeholders*) involved in these IE trainings and I think that is a huge advantage in terms of making the evaluation go better, but also it creates receptivity in part of the government officials...(*and then referring to a team working on a IE who had attended an IE training)*...look at the xxx team, look at the stuff they do. They have now assimilated all of this knowledge …they now are meticulous in their evaluations…

Nonetheless, my understanding of what has happened in the course of this evaluation is that both PU and ECB occurred and benefited the stakeholders mostly involved; also, process use and ECB did not occur in a single direction, flowing from the evaluation specialist to the non-evaluation specialist (see also Lawrenz et al., 2007). The core evaluation team was also affected by participating in the evaluation, especially in terms of ECB, their capacity to do evaluation of government-led programs increased. With respect to the non-evaluators, it appears from the following quotes that given the technical nature of the IE, ECB would need to be intentional to be fully effective either through training (confirming what noted by the evaluator above) on the basics of IE or through active learning-by doing through extensive involvement.

> "I think the evaluation is difficult to understand for many people…if you really want to understand, you need more than a 20 min presentation, you need to know what the logic is and go into depth, look at it for some months…"

> "When I was actively involved in all the meeting and emails exchanges.., that's when I really started to understand (*the evaluation)*"

> "I was initially trained by my colleagues at the SoPH, but it is the fact that I was working in the field and I was involved in the different steps that helped me to understand more the process.

The long-distance nature of the collaboration of this evaluation had a both a positive and a negative effect on PU. On the one side, the quality of the communication is a fundamental factor in stimulating PU (Preskill et al., 2003) and in person meetings were not always possible, but on the other side, the fact that the evaluators would leave the daily supervision to the local researchers could have possibly compensated with a reverse positive effect on PU. Another factor that could have fact a positive and a negative effect on PU was the length of the evaluation. In a lengthy evaluation as this one considered, some PU should be achieved through many presentations, discussions, and briefings with the involved stakeholder groups. On the other side, the long duration of the process associated with the staff turnover reduced the abovementioned effect (see also Preskill and Boyle, 2008), since few people were able to be present to all the presentations, discussions and briefings from the start to the end. Finally the number of actors involved on the one side increased the number of people actually aware and possibly influenced by the evaluation, but on the other side when many actors are involved in such a long process it appeared difficult to always involve all the relevant actors.

No reflection upon the experience and the lessons learned has been undertaken so far. Yet it is not excluded that this will happen after the evaluation is concluded. Only two interviewees reported that they had an occasion to share their experience with this IE, when asked to do so during knowledge sharing events on PBF systems. No spontaneous knowledge sharing has yet occurred. The interviewed posit that there is no time to reflect upon their experience and share it with other colleagues. One of the TTLs participated to a workshop organized specifically to share experiences on PBF systems and said that she would have liked to share more of her lesson learned but did not want to monopolize the discussion.

Difference in perspectives The informants interviewed agreed on the process, main steps, who was involved, and on the fact that the IE was a "learning" experience. Yet some differences were pretty evident: many changes were not attributed to the IE by the government, while they were by the WB managers, and partially by the IE specialist who appeared not to have thought about it much. With respect to the latter, generally speaking, it felt like their learning about the process and the effects on the non-evaluators and the program were taken for granted. During the interviews, my impression was that while the TTLs didn't have an occasion to share their lessons learned yet, but will probably do it (depending on the findings), for the evaluators this part of the story (the learning during the process) will most likely remain tacit, because it is taken for granted.
Whether one is an evaluator versus a non evaluator, or a TTL versus a government official appears to be less relevant in determining the level of perceived process use and ECB[15]. What matters most is the level of involvement in the process. In fact, given the open-ended nature of the interview I was able to capture who was spontaneously attributing more relevance and value to the outcomes of the process (simply by the fact that I needed to use more or less probes in the interview to direct the discussion). A big difference was noted in the issues addressed by the two TTLs interviewed. The one that was actually more involved in the process of designing the evaluation and supervising the implementation was much more aware and positively affected by the process use and ECB, than the one who was less involved. Yet, formally their role was the same, being TTLs of the program under evaluation. The difference in perspective and real changes occur when people start understanding the logic of evaluation, as reported also by the TTL whose own attitude changed when she understood the logic of IEs and elaborated her "new philosophical view on evaluations" which is based on the belief that having a counterfactual does not necessarily mean comparing something to nothing, since variations of the same program theory can be applied to answer unresolved questions. Yet to arrive to this realization, a lot of discussions and thinking had to be done, with the evaluators and the government counterpart. Finally, I noted that one of TTLs repeated twice that what one should be interested in are the results on the ground and "not writing papers on the findings", as she implied, would be the academics interest.

6. Reflections

It may be argued that what have been described are examples of unplanned instrumental use of a formative nature. On the contrary I would like to argue that these exemplify process use and ECB, in that the evaluation process itself triggered this reflection during and on practice, leading to a more rigorous program implementation, enhanced evaluative thinking, engagement and ownership among the stakeholders, and both unintentional and intentional ECB.
Thinking about the use of this evaluation, the PBF study has a clearly stated summative objective. It addressed questions of impact, and was undertaken to render a summary judgement on certain critical aspects of the program's performance. No formative type of evaluation activity (i.e. activity that would guide program improvement) was intentionally undertaken by the evaluation team, who was actually very careful in maintaining an "external" role. Yet we cannot deny that even if rigor was adopted by the team as the standard and producing knowledge as the advance organizer - as opposed to immediate utility – (see Marra, 2000), the evaluation reached more than just a summative objective. Furthermore participation was not broad and was not used as an instrument to

[15] This confirms the results of Paruzzolo's paper 2 "Modelling perceptions.. " (2009)

create capacity. PU and ECB, are a result "making a virtue of necessity": the RBF evaluation could have not happened without the collaboration of the local stakeholders, and PU and ECB could not have happened without participation. As discussed, stakeholder participation can be a necessity (see above), and can also have the purpose of ensuring validity (Brandon, 1998). For example, the success and reliability of an evaluation rests heavily on the quality of the data used (Paruzzolo et al., 2007). Consequently data collection strategies need to be carefully considered, and especially when the context is not familiar to the evaluator (often the case in international development), the collaboration of local experts is utmost important. The scope and context of the evaluation influence the nature of the involvement. As previously discussed, in a randomized controlled experiment, the interaction with treatment areas, program staff and clients are voluntarily kept at minimum, while the primary audiences become the people influencing the policy, generally high-level officials. For example, the evaluators discussed with the implementation team and argued for measuring health outcomes, the ultimate program goal, and to maintain the integrity of the design by trying to limit imitation of PBF in non-PBF areas, but they didn't distance themselves from people implementing the program, and actually were constantly seeking to obtain government ownership of the IE. The government was very supportive because it trusted the evaluators competence, credibility, and integrity.

It also needs to be noted that what happens during evaluations of donor funded impact evaluations, could also be the result of power dynamics. Power[16] is central to management in any organization (Pfeffer, 1992) and especially in aid relationships (Ros Eyben, 2008). It was not the scope of the study to analyze the power relationships of the stakeholders involved in the evaluation, but from the interviews it appeared that what was very specific of this case was the "power" of the government. As confirmed by others, a local researcher stated that "In xxx, you don't just come and impose things…This evaluation was really a collaborative effort…" As reported in a seminar introduction on the PBF IE "*An impact evaluation at the national-level required the study to be strongly owned by the Government and local stakeholders. The Government has been in the driving seat."* Yet, this doesn't exclude that the evaluation was initiated just to signal compliance to the donors, given their power to decide whether or not resources will be invested in the program. Yet, the power relation was balanced by the fact that once the IE started the government possessed the power to sabotage the process and or to try to influence the results. Yet, as reported by the interviewees, the government collaborated with the evaluators in the common goal to obtain results that could inform on the performance of the program. What made the evaluation a successful experience seemed to be the alignment of interest and objectives to carry out the evaluation, even if the ultimate reasons behind the interest in the evaluation certainly differed.

7. Discussion and next steps

My work represents an innovative perspective in the field of evaluations applied to international development where the need to create evidence of what works is combined with the call for capacity building in client countries. In fact, in developing countries, participatory social research has been linked to community development and capacity-building (e.g., Salmen and Kane 2006). The case study highlights the great potential also (not only) for experimental impact evaluations to

[16] Power is vested in us by the dependence of others, and that dependence is a function of how much others need what we control, as well as how many alternative sources for that resource there are (Pfeffer, 2002).

strengthen program implementation, foster engagement and continuous learning, build relationships, and create evaluation capacity in a number of stakeholders, while maintaining the evaluation integrity. The RBF evaluation demonstrates that every good evaluation has the potential to be good research, have policy relevance, and build capacity by becoming a learning experience. I agree with Bickman (2002) that "it's a waste" not to see the research opportunities as part of an evaluation, as it is a waste, I would add, not to see the evaluation capacity building opportunity. Also I would like to use this example of a prospective randomized evaluation to clarify my position on the following issue. The existing categories of evaluation use (instrumental, conceptual and political) mirror somewhat the result-based (use of findings) and process-based (process use) tension that exists regarding evaluation use (Peck and Gorlanski, 2008). That is, the instrumental use of evaluation tends to expect that the findings of an evaluation are what will be used directly, and in the short-term after the evaluation is concluded; whereas, the conceptual use of evaluation is more likely to reflect longer-term process-based influences, again after the evaluation is concluded. Conceptual use relates strongly to learning, i.e. "changing a mental framework that may or may not influence *future* thinking and decision making" (Forss et al., 2002). I argue that all these types of uses can occur simultaneously and both during the process and after the evaluation is concluded. The case study reported provides an example of how process stimulates uses that can be instrumental, by entailing the action of creating a national roll-out plan that will allow an evaluation instead of an opportunistic roll-out plan based on single donor decisions; but it can also be conceptual, by changing stakeholder's perspective on evaluations and the program outcomes; and it can also be political, by providing the government with a tool to reassess leadership in providing a harmonized healthcare reform[17].

Moreover, as pointed out by Patrizi (2003) "evaluation is still often caught between the false choice of accountability or learning". This tension is discussed in the OECD-DAC (2005) report and very well explained in the following quotes:

> Evaluations that give clear answers versus evaluations that examine complexity without single cause-effect relationships (Patrizi, 2006; Wisely, 2002)
>
> Evaluations for learning and inquiry as opposed to evaluations that determine accountability to results (Patrizi, 2003)
>
> … having diminished expectations for evaluations to measure outcomes and determine causality, but increased expectations for its role as a management and learning tool (Kramer et al., 2007). In Behrens (2008)
>
> As the 21st century began, logic models and randomized controlled trials were two of the predominant design approaches for evaluation as funders sought to understand what works and what does not. During the last eight to ten years, a new paradigm began to gain currency: the "learning organization", popularized by Senge (1990). In Behrens (2007).

I argue that this tension should not exist. I see no inherent contradiction in pursuing the two objectives of accountability and learning, and I hope the dynamics described in the present investigation provide evidence for this argument. The PBF evaluation is an example of a

[17] On the different types of uses, see also Vedung (1997) and Rist (1999).

randomized experiment whose main objective is that of determining the merit or worth of a program, and producing new knowledge as a public good for other countries willing to introduce PBF systems. Yet it is also an example of an evaluation where the making virtue of necessity by maintaining continuous dialogue with the local stakeholders, produced process use and ECB. Finally, a very important clarification. I decided to study a randomized experiment, not at all because I consider them as the premier model for public program evaluation, but because given their predominance and increasing diffusion in organizations such as the World Bank, we should get the most out of them!

One major limitation of the present study is that it is a single case study, taking place within a particular context. Consequently the processes may be unique to this particular evaluation. Although other evaluations in other contexts will differ, the case study may illuminate the opportunities and challenges to enhance process use and capacity building through prospective experimental evaluations. Yet, similar evaluation efforts are becoming very common in the development arena and this case study can contribute to the literature in the field. Certainly the use of multiple case studies would also help to strengthen an examination of evaluation process use. Another extension of this work would involve a longitudinal follow-up after the evaluation findings have been produced. Such a follow-up could be undertaken with all of the previous participants. In a follow-up study, it would be interesting to examine the extent to which the perceptions of the evaluation and the evaluation process are affected by the evaluation findings and its use/non use.

ANNEX 1

LETTER FROM THE GOVERNMENT REQUESTING ADHERENCE TO ROLL OUT PLAN

Dear _____;

As you know, the Ministry of Health of xxx is dedicated to ensuring that funds in the health sector are allocated through the most effective and efficient methods. Due to the presence of many different donors contributing to achieving the targets set forth by Vision 2020, the Poverty Reduction Strategy Proposal and the Millennium Development Goals, the government is taking the lead in harmonizing and aligning the activities of donor organizations.

Performance Based Contracting of general health and HIV/AIDS services is an important strategy to strengthen the health sector, and the MOH is committed to implementing a rigorous evaluation of the impact of this approach. The results from this study will greatly contribute to our knowledge of what contracts work, and under what conditions, in order to continue improving our health care system.

To ensure the quality of the impact evaluation, the MOH has adopted a roll-out plan for Performance Based Contracting. Districts were allocated to Phase I and Phase II of the implementation of Performance Based Contracting on an equitable basis. Please note that the roll-out plan does not involve withholding financial resources from facilities, as facilities will either move to performance based contracting or receive input based funding. Donors were briefed on the roll-out plan and given an opportunity to comment. As the impact evaluation depends on the roll-out plan, it is very important that all donors adhere to it. More in particular, donors are asked to commit not to switch to performance contracts in the Phase II areas until the evaluation in Phase I has been completed.

Your organization has contributed greatly towards improving the quality and quantity of health services offered in xxx by means of financial resources, as well as valuable technical assistance. We would like to confirm your continued commitment to the government-led coordination of the expansion of the Performance Based Contracting scheme following the attached roll-out plan. Please take this opportunity to review the roll-out plan, and specifically the list of health care facilities in Phase I and Phase II districts. We also ask you to take this opportunity to confirm the donor information for the sample of health facilities, and contact us with any necessary additions or revisions.

We kindly request a written confirmation of your commitment to the roll-out plan by March 10, 2006. Please send via email to _____ or to the above address. We greatly value the partnership we have developed between our government and your organization and thank you for the opportunity to work together towards achieving our common goals.

Sincerely,

References

Behrens, T. R., & Kelly, T. (2008), "Paying the piper: Foundation evaluation capacity calls the tune", *New Directions for Evaluation*, Vol. 119

Bickman L. (1996), "A continuum of care: more is not always better", *American Psychologist*, Vol. 51

Bickman L. and Mulvaney S. (2006), *Large-Scale Evaluations of Children's Mental Health Services - The Ft. Bragg and Stark County Studies* in "Book Handbook of Mental Health Services for Children, Adolescents, and Families" Publisher Springer US

Brandon P.R. (1998), "Stakeholder Participation for the Purpose of Helping Ensure Evaluation Validity: Bridging the Gap Between Collaborative and Non-collaborative Evaluations" *American Journal of Evaluation*, Vol. 19(3)

Cousins J.B. and Whitmore E. (1998), "Framing participatory evaluation", *New Directions in Evaluation*, 80 (5)

Fitzpatrick J.L. (2004), "Exemplars as Case Studies: Reflections on the Links Between Theory, Practice, and Context", *American Journal of Evaluation,* Vol. 25

Forss K., Rebien C.C., and Carlsson J. (2002), "Process Use of Evaluations: Types of Use that Precede Lessons Learned and Feedback", *Evaluation*, 8(1)

Harnar and Preskill (2007), "Evaluator's descriptions of process use: an exploratory study", *New Directions for Evaluation*, 116: 27-44

King J.A. (2007), "Developing Evaluation Capacity Through Process Use", *New Directions for Evaluation*, 116: 45

Kramer M., Graves R., Hirschhorn J., and Fiske L. (2007), *From insight to action: New directions in foundation evaluation.* Boston: FSG Social Impact Advisors.

Lawrenz F., Huffman D., McGinnis J.R. (2007), "Multilevel Evaluation process Use in Large-Scale Multisite Evaluation", *New Direction for Evaluation,* 116

Marra M. (2000), "How Much Does Evaluation Matter? Some Examples of Utilization of the Evaluation of World Bank's Anticorruption Activities", *Evaluation* 6(1): 22

Paruzzolo S., Vivo Guzman L.S., Martinez S., McGinnis L., Gertler P., Lundberg M. (2007), "Evaluating Youth Projects", *World Bank publication,* Youth Development Note, Vol II (5)

Patton M.Q. (1997), *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Patton M.Q. (2007), "Process Use as Usefulism", New Directions for Evaluation, 116: 99-112

Patton M.Q. (2008), *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patrizi P. (2003), The inside perspective: An assessment of the current state of the evaluation field, Parts 1 and 2. Retrieved October 12, 2007, from http://www.geofunders.org

Patrizi P. (2006). The evaluation conversation: A path to impact for foundation boards and executives: Vol. 10. *Practice matters: The improving philanthropy project.* Retrieved October 12, 2007, from http://www.foundationcenter.org/gainkowledge/practicematters/

Pfeffer J. (2002), *Managing with power: Politics and influence in organizations*, Harvard Business School Press, Boston

Preskill H. (1991), The cultural lens: Bringing evaluation utilization into focus. In C. L. Larson & H. Preskill (Eds.), *Organizations in transition: Opportunities and challenges for evaluation*, New Directions for Program Evaluation, 49 (pp. 5–15). San Francisco, CA: Jossey-Bass.

Preskill H. and Boyle S. (2008), "A Multidisciplinary Model of Evaluation Capacity Building", *American Journal of Evaluation*, Vol. 29(4) 443-459

Preskill H., Zuckerman B. and Matthews B. (2003), "An Exploratory Study of Process Use: Findings and Implications for Future Research", *American Journal of Evaluation*; 24: 423

Preskill H. (2005), Appreciative inquiry, in S. Mathison (Ed.), Encyclopedia of evaluation. Thousand Oaks, CA: Sage.

Rist R. (1999), *Program Evaluation and the Management of Government*, Transaction Publishers

Roessler R.T. (1980), "Impact of Social Programs: Issues in Implementing Research", *Knowledge, Diffusion and Untilization*, Vol 1(4)

Rossi P.H., Lipsey M.W., Freeman H.E. (2004), "Evaluation – A systematic approach" (7th ed.), Thousand Oaks, Calif.: SAGE Publications.

Russ-Eft D., Atwood R. and Egherman T. (2002), "Use and Non-use of Evaluation Results: Case Study of Environmental Influences in the Private Sector" *American Journal of Evaluation*, Vol. 23(1)

Salmen L.F. and Kane E. (2006), *Bridging Diversity – Participatory Learning For Responsive Development*, The World Bank, Washingon DC

Sanders J.R. (2003), "Mainstreaming Evaluation", *New Directions for Evaluation*, Vol. 99

Scriven M. (1996), "Types of Evaluation and Types of Evaluator", *American Journal of Evaluation*, Vol. 17(2)

Senge P.M. (1990), *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday.

Stufflebeam D. L. (2001), "Evaluation models", *New directions for evaluation*, Vol. 89

Taut S. (2007), "Studying self-evaluation capacity building in a large international development organization", *American Journal of Evaluation*, 28(1)

Vedung E. (1997), Public Policy and Program Evaluation, Transaction Publisher

Wisely D. S. (2002), “Parting thoughts on foundation evaluation”, *American Journal of Evaluation, 23*(6)

Yin R.K. (2003), *Case Study Research: Design and Methods*, Third Edition, Applied Social Research Methods Series, Volume 5, SAGE Publications, Thousand Oaks London, New Delhi