Critical Friend Approach: Policy Evaluation Between Methodological Soundness, Practical Relevance, and Transparency of the Evaluation Process

Andreas Balthasar
University of Lucerne (Switzerland)

Abstract
Methodological soundness, practical relevance, and a transparent evaluation process can be viewed as the three main elements of today's common understanding of evaluation. Based on these elements, a great diversity of evaluation approaches has been developed in evaluation theory and practice in the past years. In this paper, the "critical friend" approach is presented as a promising approach. This approach follows the common understanding of evaluation and takes into account empirical findings that showed that the use of evaluations depends decisively on the interest of the program managers and decision makers in the results of the study. The approach is discussed based on the example of the evaluation of the child care vouchers project in the city of Lucerne.

Zusammenfassung
Methodische Stringenz, praktische Relevanz und Transparenz des Bewertungsprozesses können als die drei Grundpfeiler des aktuellen Verständnisses von Politikevaluation verstanden werden. Darauf aufbauend sind in den letzten Jahren zahlreiche und sehr unterschiedliche Evaluationsansätze entwickelt worden. In diesem Artikel wird der „critical friend approach" vorgestellt. Dieser knüpft am aktuellen Evaluationsverständnis an und berücksichtigt aber zusätzlich empirische Ergebnisse der evaluationswissenschaftlichen Verwendungsforschung. Diese weisen darauf hin, dass die Nutzung von Evaluationsergebnissen stark vom Interesse der Programmverantwortlichen an der Evaluation abhängt. Der „critical friend approach" wird diskutiert am Beispiel der Evaluation des Pilotprojekts Betreuungsgutscheine für die familienergänzende Kinderbetreuung in der Stadt Luzern.

1 Introduction

Evaluation is marked by a great diversity of approaches and methods. In *Evaluation Theory, Models, and Applications,*

German Policy Studies
Vol. 7, No. 3, 2011
pp. 187-231

Stufflebeam and Shinkfield (2007) distinguished no fewer than 26 often-used evaluation models (see also House, 1987). In this paper, the “critical friend” approach is presented as a new and promising approach. Despite the great variety of evaluation approaches in practice, there are elements that characterize the current use of the term “evaluation”: First, evaluation is regularly understood as assessment (Stockmann, 2000, p. 12; Clarke & Dawson, 1999; Scriven, 1991). Rossi and Freeman (1993, p. 5) therefore defined evaluation as “the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs.” A second, central element of evaluation underlines its basis in social sciences methodology. This is apparent in Rossi and Freeman’s definition of the term as well. The third constitutive element is its practical relevance: Vedung (1999, p. 12) stated that the general practical orientation is so central that it must be a part of the definition of evaluation.

In this paper, we will first present the core aspects of three typical approaches representing methodological soundness, practical relevance, or transparency of the assessment process. We will then illustrate our “critical friend” approach based on the example of the evaluation of the child care voucher project in the city of Lucerne. The critical friend approach tries to find an optimal balance between methodological soundness, practical relevance, and transparency of the assessment process. In addition, it accounts for empirical findings showing that the use of evaluations in Switzerland depends decisively on the interest of the decision makers in the results of the study (Balthasar, 2007). At the end of the paper we will discuss possibilities and limitations of the critical friend approach in policy evaluation.

2 Diversity of evaluation approaches

As described above methodological soundness, practical relevance, and transparency of the assessment process can be considered as the three corners of the triangle in which most evaluation approaches can be positioned. The following sections sketch out core aspects of each of the three elements by presenting one typical approach for each element.

2.1 Methodological soundness: Donald T. Campbell

In the 1960s, the criticism that common evaluation practice was not scientific led Donald T. Campbell to demand scientific soundness when planning the evaluation design. Without a doubt, Campbell and Stanley’s (1963) *Experimental and Quasi-Experimental Designs for Research* and Campbell’s (1969) paper, “Reforms as Experiments,” were important foundations for the development of evaluation methodology. Campbell’s ideal of experimentally designed program evaluation continues to shape evaluation science in part up to today (Stufflebeam & Shinkfield, 2007). Similar to Campbell, also Chen and Rossi (1980; Chen, 1990) worked towards scientifically based evaluation research and drew up the central ideas of theory-driven evaluation. As stated in Rossi, Freeman, and Lipsey (1999):

> Every program embodies a conception of the structure, functions, and procedures appropriate to attain its goals. This conception constitutes the ‘logic’ or plan of the program, which we have called program theory. The program theory explains why the program does what it does and provides the rationale for expecting that doing things that way will achieve the desired results. (p. 156)

Methodologically oriented evaluation theories view evaluators foremost as independent specialists in methodological questions. Using clever and scientifically sound design, they provide firm answers to the question as to whether certain effects can be casually attributed to evaluated measures. In this tradition, evaluators are meant to mainly show the effects of programs. But they are supposed to abstain from assessing the results, so as not to jeopardize their independence.

2.2 Practical relevance: Michael Patton

There are excellent grounds for viewing Michael Patton as the father of evaluation approaches which focus on the practical relevance of evaluation results (Stockmann & Meyer, 2010, p. 120). Patton's "utilization-focused evaluation approach" aims to conduct evaluations in such a way that their results are used for policy practice: "The focus in utilization-focused evaluation is on intended use by intended users" (Patton, 1997, p. 20; see also Patton, 2000). For this reason, Patton places importance on having the users of an evaluation participate in important decision making from the very start of the evaluation process. The actual conducting of an evaluation is the responsibility of the evaluators, however, whose credibility and integrity depend on the quality of their handling of this communication task. In this regard, Patton assigns evaluators five functions (Patton, 1988, p. 15 ff.): First, they must help those involved overcome their fears concerning the evaluation. Second, they must make sure that the right questions are asked. Third, their task is to further develop the evaluation process taking into account and adapted to situational factors. Fourth, they must constantly keep in mind what has been done previously and draw conclusions as to how the evaluated program can be improved. Finally, evaluators must be advocates of the evaluation even when the process is at a critical point.

2.3 Transparent evaluation process: Michael Scriven

Evaluations always represent evaluative judgments (Stockmann & Meyer, 2010, p. 17). The assessments made in the context of an evaluation become comprehensible only when the criteria being used are disclosed. Transparent evaluation criteria and methods are therefore an essential determining feature of evaluations. It was especially Michael Scriven who studied this topic: "Evaluation is the process of determining the merit, worth and value of things, and evaluations are the products of that process" (Scriven 1991, p. 1). Scriven was unequivocal in his position that society requires valuing and that it is the role of the evaluator to do that job: "Bad is bad and good is good and it is the job of evaluators to decide which is which" (Scriven, 1986, p. 19). Scriven maintained that there is a science of valuing, and that it is evaluation. He acknowledged that "no single scientist ever observes 'reality' completely and undistorted, but rather observes through biased perceptual filters Using multiple perspectives helps construct a more complete picture of reality" (Shadish, Cook, & Leviton, 1991, p. 76). Using the perspectivism, or multiple accounts, that Scriven recommends for program evaluation corresponds to what in methodology is called triangulation (see section 3.4.1 below).[1]

3 The "critical friend" approach: Theory and practice

Methodological soundness, practical relevance, and a transparent evaluation process can thus be viewed as the three main elements of today's common understanding of evaluation. The problem is to find the optimal balance between these three challenges. Our own empirical findings showed that when scientific soundness is not questioned, the use of evaluations in Switzerland depends decisively on the interest of the program managers and decision makers in the results of the study (Balthasar, 2007). Taking into consideration this empirical evidence, we developed the "critical friend" approach, which tries to find an optimal balance between the challenges of evaluation theory and practice and the fact that the utilization of evaluations depends decisively on the decision makers' interest in the results.

The evaluation process is designed differently depending on the evaluation approach. This means that every evaluation approach yields different responses to the central questions that arise during the conducting of an evaluation. This is also true for the critical friend approach. For this reason, in this section the evaluation process is described along these questions. In the following, the critical friend approach will be outlined in general and then illustrated by taking the example of the evaluation of the child care voucher pilot project in the city of Lucerne. This evaluation was undertaken in the fall of 2008 and is ongoing until 2012. It began shortly prior to the launching of the pilot project

[1] In this paper, I use the term "political programs" to refer to programs, measures, and interventions of public institutions. On terms and definitions, see Scriven (1991).

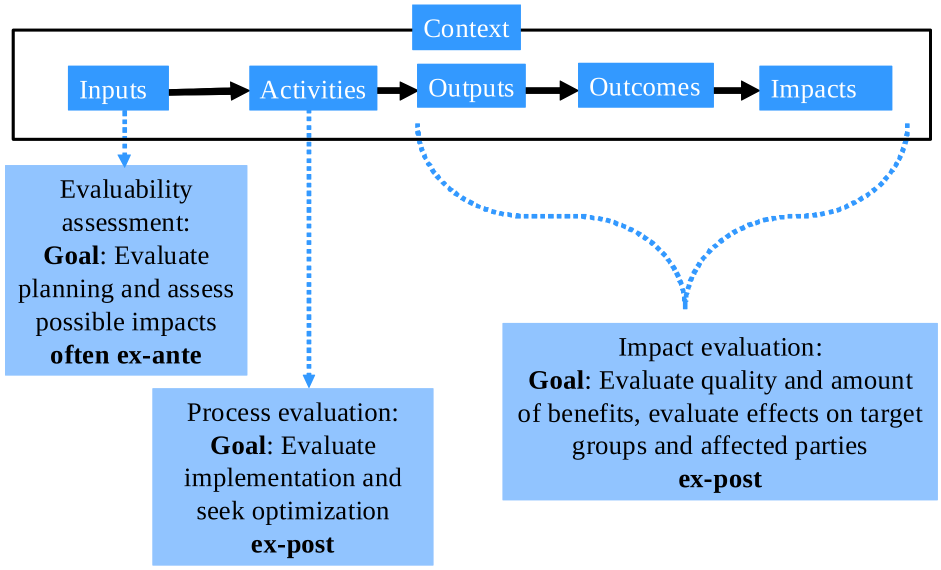
in Lucerne and will be completed only after the pilot project ends.

3.1 First question: What questions should the evaluation answer?

3.1.1 The idea ….

Every evaluation starts with politically relevant questions. In a first step, these questions must be stated more precisely as scientifically addressable evaluation questions. To do that, it is helpful to develop a "logic model" as proposed by Chen and Rossi (Chen & Rossi, 1987; Rossi et al., 1999). A logic model links impacts with program inputs and processes in a more or less linear way. The logic model facilitates thinking, planning, and communications about the objectives and actual accomplishments of a program. The term logic model is frequently used interchangeably with the term "program theory" in the evaluation field. Logic models can alternatively be referred to as theory, because they describe how a program works and to what end. Basically, a logic model is "a systematic and visual way to present and share your understanding of the relationships among the resources you have to operate your program, the activities you plan to do, and the changes or results you hope to achieve" (W.K. Kellogg Foundation, 2004, p. 1). A logic model is used to state the evaluation questions more precisely, to define the evaluation objects, and to identify the factors that may foster or hinder implementation and the effects. The logic model may be developed by those involved, by target groups, or by the evaluators of a program, depending on the purpose of the evaluation.

Figure 1: Logic model and evaluation objects



The basic components of the logic model shown in Figure 1 are defined as follows:

Inputs: The inputs of a program are its goals, the conceptual considerations concerning the postulated relationship between input and outputs, outcomes and impacts, and the financial and personnel resources a program has available. Wholey (1994) developed the evaluability assessment as an independent concept for evaluation of the input. An evaluability assessment examines in particular five elements that are considered essential for a successful program. First, clear objectives are a necessary condition for the evaluability of a program. In a political decision-making process, the ambiguity of objectives is often the premise to find a compromise supported by the majority. However, to evaluate program success, agreement on the benchmarks for the assessment is needed. Second, an impact-oriented program concept must have indicators based on which the success of the program can assessed. Third, there must be program activities to achieve the desired objectives. Fourth, the program must have resources to achieve the objectives. Fifth, a clear statement about the assignment of tasks and authority and about the role of relevant actors is required. A sound evaluability assessment is valuable in and of itself. But it also forms a good basis for later process and impact evaluations.

Activities: Program activities are what the program does with the resources. The focus of attention here are the structures and processes that are set up internally to provide the intended output of the program. Program activities themselves are frequently the object of process evaluations (Vedung, 1997; Patton, 1997). According to Vedung (1997), process evaluations are basically conducted to examine whether the government authorities, intermediary actors, and program managers have the understanding, the ability, and the willingness to implement the program in line with the intended objectives.

Outputs are the products and services delivered directly by a program and directed toward its target groups. These may consist in prohibitions and requirements, in monetary payments or claims, in checks, education benefits, care benefits, advice, network building, or many other forms of public incentives (Frechtling, 2007).

Outcomes are “the specific changes in program participants’ behavior, knowledge, skills, status and level of functioning” (W.K. Kellogg Foundation, 2004, p. 2). Outcomes have to do with the behavior of the target groups whose actions are intended to be directly affected by the program. Target groups include any persons or institutions that come into direct contact with the output and from whom a response is usually expected.

Impact is “the fundamental intended or unintended change occurring in organizations, communities, or systems” (University of Arizona Rural Health Office et al., 2000, p. 40) as a result of program activities. Those affected are persons or institutions that will have to expect advantages or disadvantages due to the changed behavior of the target groups.

Context: Finally, when developing a logic model it is important to also keep in mind the factors that have to do with the context of the program. If evaluators really wish to understand how programs affect target groups and society as a whole, they cannot simply limit the analysis to the program and its logic. Environmental factors that have effects during the whole process of policy decision-making and implementation contribute decisively to understanding the success or failure of a program (see Vedung, 1999, p. 184).

Based on the logic model, the central evaluation questions can be derived.

Table 1: Typical evaluation questions

Component	Typical questions
Input	- Are there objectives or at least a common understanding of the benchmarks for the assessment? - Is it clear when the objectives have been met (evaluative elements)? - What measures are planned for implementation of the objectives? - What resources will be available? - What is the plan for assigning/handling responsibilities?
Activities	- Are the program managers in a position to carry out implementation (resources, know-how)? - Are the program measures implemented in the expected way? - Are the responsibilities with third parties arranged clearly and practically?
Output	- Are there outputs, and do they accord with the plan? - Were they produced on time and efficiently?
Outcome	- Were target groups reached? - What is there response? Is the target group reacting in the intended way (effectiveness)? - Are there desired or undesired side effects?
Impact	- Can the intended change on affected parties be observed? - What other effects influence the impact (reinforcing or neutralizing)? - Has the targeted problem been solved? - Is the chosen way to solve the problem efficient?
Context	- Was the program approved with strong political support, under political disputes, or in consensus? - What is the development of the relevant policy area in the international, national, and regional context? - Are there relevant political changes in the environment of the program?

Like all model-type representations, the logic model is a rough simplification of reality, and it is not a detailed causal model postulating scientific hypotheses about the relationship between intervention and reaction. Causal models may be necessary for a scientifically sound analysis of specific aspects of the logic model, as proposed by Pawson and Tilly (1997).

A logic model does not have to satisfy the scientific requirements of a theory. In most cases it will contain an implicit theory of the actors. Although the theory is not stated and there is no logical rationale in the form of hypotheses and chains of causes and effects, it serves as practical guidelines for the planning of the evaluation.

3.1.2 … and its practical implementation

Child care outside the home not only provides children with opportunities for important learning experiences but also makes it easier for parents to be employed. These are goals that are highly accepted in terms of education, integration, social, and economic policy (OECD, 2004). Unfortunately, in most European countries the demand for child care services exceeds supply (Enste & Stettes, 2005). Up to now the state has usually responded to the problem by subsidizing private day care providers or establishing state-run day care facilities (BMFSFJ, 2008). Thanks to these supply-side payments to child care services providers, in many places the number of children served has increased considerably (OECD, 2007).

However, for many parents the situation continues to be unsatisfactory. In many regions sufficient child care services are still not available. In addition, the prevailing practice has the disadvantage that only those parents or guardians profit from the subsidies that have the fortune to obtain a place in a subsidized child care facility. Also, in the current financing model, parents can only rarely select child care according to their personal preference.

For these reasons, mainly economists have proposed the idea of redirecting funding for child care from providers to parents (Blau, 2001; Stutzer & Dürsteler, 2005). Rather than subsidize institutions that offer child care, the payments are made directly to parents. The hope is that this will result in more child care services, which will better meet the demand and possibly also become less expensive. In addition, the aim is to create equality among families that depend on child care outside the home.

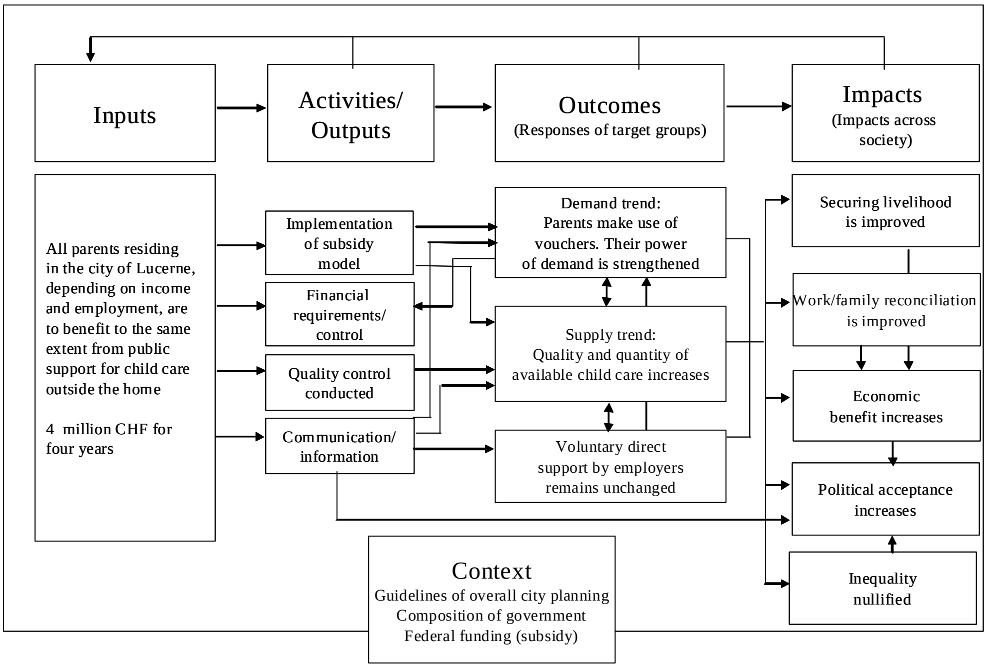
Various European countries (United Kingdom, Germany, Finland, Austria) and North American states have tried out child care “vouchers,” or certificates, for preschool care and for care before and after school. An international comparison showed that voucher systems have not been one hundred percent successful or ineffective (Gerlach, Damhus, & Anschütz, 2007). In German-speaking regions, in addition, the full-coverage voucher system for child care introduced in 2003 in Hamburg has received some attention in the media due to critical evaluation results (Hilgers Kastner, & Strehmel, 2007).

Following a recommendation of the Organization for Economic Cooperation and Development (OECD) (2004, p. 21), the Swiss Federal Council decided in 2007 to fund pilot projects on child care vouchers (Schweizerischer Bundesrat, 2007). Based on that, the city of Lucerne launched a child care voucher pilot project in April 2009; to date, it is the most important child care voucher pilot project in existence in Switzerland.[2] The project is funded in part by the federal government. With the goal to learn more about the practical feasibility of the voucher approach, the federal government required that an evaluation be conducted. This is the classic starting point for an evaluation of a political program!

The city of Lucerne did not launch the child care voucher pilot project on the basis of a logic model that it developed itself. The logic model presented in the following was developed by the evaluation team in the course of preparing for the evaluation and was then adjusted in dialogue with the people responsible for the program. It has since become a central element of the program that is used almost every time the pilot project is presented, whether the main topic of the presentation is evaluation or not.

[2] For more information on the pilot project and its design, see: http://www.stadtluzern.ch/de/themenbereiche/?action=showthema&themenbereich_id=7&thema_id=401 (retrieved June 27, 2011).

Figure 2: Logic model of child care vouchers model



The logic model of the child care voucher pilot program shows the connections between the intervention by the city of Lucerne in the area of child care and the family policy objectives of the city in a simplified way. For clarity, it is linear in design and does not show more feedback loops than necessary. In addition, activities and outputs were grouped together for simplification.

The model shows that the financial and personnel resources of the authorities and the conceptual foundations are viewed as the starting point of the intervention “child care vouchers.” The city implements this concept, pays out the subsidies, takes quality control measures, and informs the public about the project. The four activity areas implementation of subsidy model, financial requirements/control, quality control, and communication/information, together with their outputs, are not only a part of the logic model. They also form the administrative organizing principle of the responsible social department for childhood, adolescence, and families in the administration of the city of Lucerne. From the political point of view, the monitoring of the pilot project’s budget conformity is of special importance. The government and also the parliament of the city place great importance on the project staying within the available budget. If it did not, the project would probably be discontinued relatively quickly. The authorities therefore insist on having the option to adjust the subsidy model, if it should happen that its financing gets out of hand.

Further, the logic model in Figure 2 shows that the voucher system aims to influence demand and supply in the area of child care services (outcomes). Finally, the program aims for effects at the societal level that accord with the policy goal of promoting Lucerne as a family friendly city (impact). Evaluation is planned for all of the objects mentioned. For this reason, the central questions were formulated in accordance with the logic model.

Table 2: Evaluation questions for the child care voucher pilot project in Lucerne

Evaluation object	**Evaluation questions** (selected examples)
Inputs	- Does the subsidy model prove itself? Does it need adjustment? - Does the implementation model prove itself? Does it need adjustment?
Process/ Outputs	- How do the city, the child care facilities, and the parents rate the administrative processing? Is there a need to optimize? - During the pilot phase, how did the costs develop for the city of Lucerne and for the federal government? Did costs stay within the budget?
Outcomes	*Parents* - What consequences does the introduction of the new model have for parents who up to now had a place in a subsidized child care provider? What is the development of demand for child care services? *Child care facilities* - How is the development of supply of child care services? Does supply meet demand? What is the development of costs and prices of child care outside the home? *Employers* - How do employers who have up to now financed child care services for their employees respond to the child care vouchers?
Impact	- Do what extent does the child care voucher model contribute towards strengthening livelihood security? Specifically, the evaluation should discover whether and to what extent parents of young children claim no (or less) welfare assistance due to the child care outside the home. - What is the benefit of child care vouchers for the economy (for example, more tax revenue)?
Context	- How, and with what influence, is federal and cantonal policy on child care outside the home developing? - What is the development of the financial situation of the city of Lucerne? What effects does this have on the pilot project? - Have there been changes in city policy due to elections?

The evaluation does not seek to answer all possible (and relevant) questions. It focuses on those questions that emerged as relevant from discussions between the evaluators and the responsible departments in the administration of the city of Lucerne, the Canton of Lucerne, and the Federal administration, which is partly financing the evaluation.

3.2 Second question: Why is the evaluation being conducted?

3.2.1 The idea ….

Basically, two types of evaluation can be distinguished. The second question of the evaluation process intends to clarify which of these two types is the focus:

Summative evaluations assess the worth of a program at the end of the program activities (Vedung, 1999, p. 9). They are usually designed such that they contain recommendations on whether a program or project should be continued, expanded or reduced, or discontinued. The purpose of summative evaluations is to make possible basic decisions on the evaluation object.

In contrast, *formative* evaluations serve to provide early feedback to the relevant actors, so that it is possible to optimize implementation of the program (Wollmann, 2000, p. 198). The purpose of a formative evaluation is to help those responsible and those involved to improve the evaluation object and to use the available resources as efficiently as possible. Formative evaluations are primarily prospective, whereas summative evaluations are primarily retrospective.

The distinction between summative and formative evaluations goes back to Scriven (1972). With clear scientific working methods in mind, the distinction aims to clarify the intention and goal of an evaluation. With summative evaluations, special attention must be paid to the independence of the examination: The distance between evaluators and those being evaluated has great importance in this type of evaluation. With formative evaluations, in contrast, it is especially important that the people involved in implementation of the evaluation object participate actively in the evaluation process, so that the findings can quickly inform practice. Thus, in a formative evaluation the valorization of evaluation results is supported by a certain degree of closeness between evaluators and those being evaluated (see Balthasar, 2007). However, formative evaluations should not be confused with self-evaluation. A formative evaluation can be carried out in the form of a self-evaluation. In this case, the assessment is performed by those managing the program. In other instances, a formative evaluation is conducted as an external evaluation by independent evaluators (see also section 3.3.1).

Table 3: Differing types of evaluation

	Evaluation for accountability (summative evaluation)	**Evaluation as a learning process (formative Evaluation)**
Who will use the evaluation results?	More likely the sponsors and funders	More likely directly involved implementation managers and stakeholders
What type of questions is especially appropriate?	Basic questions about the effectiveness and necessity for a project (retrospective)	Specific questions about possible ways to optimize and fine-tune the activities (prospective)
Can the evaluation be conducted by those involved in program management?	Mostly not; in demand are independence and neutrality, which external experts are more likely to guarantee	Mostly yes; in demand are good knowledge of specific measures and short information paths
Should all of the causal relationships be examined or only a part of them?	All relevant causal relationships (breadth)	A part of the relevant causal relationships (depth)

3.2.2 … and its practical implementation

The evaluation of the child care voucher pilot project in the city of Lucerne pursues both formative and summative goals. The first goal of the evaluation is *continuous improvement of the pilot project*: The evaluation is supposed to continuously inform the responsible department of the city administration and the executive authority about the project. It has to ensure that, if required, necessary adjustments to the pilot project can be instituted immediately. Second, the evaluation serves *quality assurance and development* of available child care services. It should help to identify undesired changes in the quality of child care, so that – if necessary – immediate countermeasures can be taken. It also aims to identify incentives that improve the quality of child care services in Lucerne. Third, the evaluation should provide the city of Lucerne *bases* for the future form of subsidization of child care outside the home. It should provide the federal government and interested cantons a basis for the development of comparable projects. And finally, the evaluation follows the clearly *summative goal* of examining whether the expectations and objectives of the pilot project can be met. This includes accountability to the parliament and federal government concerning the use of the financial means.

The previous explanations clearly show that the evaluation follows primarily *formative goals*. Aspects of "learning" stand in the foreground. For this reason, relevant actors are systematically included in the evaluation process. Events are held with responsible persons of the administration and with day care personnel as well as with persons who work at the childminder exchange or work as childminders. The parents themselves are invited to group meetings. The responsible parliamentary committee and the city council are regularly informed about the findings of the evaluation; any new questions that arise are taken up and – as far as possible – followed. The evaluation team serves as the moderator in a number of these forums. As external partners (not a part of the administration) and in the sense of a critical friend, it brings another way of looking at the program, its context, and its effects.

Experiences made with the evaluation of the child care voucher pilot project as well as the findings of other evaluations show that it is possible in an evaluation to address formative and summative concerns in parallel. From a methodological perspective, this is not a strictly "clean" approach, for summative and formative goals suggest distinctive methods. In everyday evaluation practice, however, both independent assessments and advice on improvement possibilities are usually expected. This is also the case with the evaluation of child care voucher pilot project, as the questions listed in Table 2 make clear. For this reason, the critical friend approach combines summative and formative matters. However, it attaches importance to always clarifying the roles of the evaluators. In addition, it tries to extensively substantiate the summative aspects with independently collected facts such as official statistics, surveys by third parties, and so on. Finally, it attaches importance to careful documentation of the databases, to make the conducting of a possible later, and more strongly independent, summative evaluation easier.

3.3 *Third question: Who is responsible for conducting the evaluation?*

3.3.1 The idea…

Responsibility for the conducting of evaluations can be variously arranged. The distinction is often made between self-evaluation and external evaluation.

With *self-evaluations*, those responsible for the program are identical with those evaluating the program; this means that the actors evaluate their own activity (Kirkhart, Morgan, & Sincavage, 1991, p. 482; Müller-Kohlenberg & Beywl, 2003, p. 65). *External evaluations* are performed by people outside of the program being evaluated. Here, experts obtain information on the evaluation object and provide feedback on their assessment of it to those responsible. Mayne (1994) emphasized that self-evaluations are more suitable for eliciting learning processes not only on the part of program managers but also on the part of decision makers.

External evaluations are often performed by specialists from universities or other institutes. These external evaluators are considered to be relatively independent. Davies, Nutley, and Smith (2000) put forward the argument that this independence increases acceptance of the findings. Ritz (2003) attributed this to the fact that external evaluators are in less danger of perceiving the problems one-sidedly (“professional blinkers”), have no “history” in the organization concerned, and are subject to no employment relationship. Undesirable evaluation results have no direct effect on their jobs or careers (Ritz, 2003, p. 52). However, Wollmann (2003, p. 344) pointed out that the intellectual independence and the scientific objectivity especially of external evaluators working in the private sector may be challenged by their need for follow-up contracts.

Starting out from some dissatisfaction with the findings of self-evaluations on the one hand, and the use of the results of external evaluations on the other hand, in recent years in-between forms have been increasingly practiced. One of these is the “hybrid evaluation” proposed by Bitel and Läubli-Loud (2010), which “combines the best of both the internal (self) and the external evaluation worlds” (p. 1). Bitel and Läubli-Loud (2010) stressed that hybrid evaluation:

> requires a role change for the external evaluator – from doing to supporting the evaluation process. Data collection is planned together, is principally collected internally, but validated by the external evaluator. (…) The hybrid evaluation process encourages organisations to develop the use of evaluation logic and skills to foster and enhance practice. (p. 1)

Similar to that approach, the critical friend approach proposed in this article aims to support the program managers and decision makers in a direct manner. However, in contrast to the hybrid evaluation approach, the critical friend approach is not in essence planned as an internal process. Rather, it conceptualizes evaluation as a partnership between the people responsible for program management and the people responsible for the evaluation. Through basing the program and the evaluation along a common logic model, planning and conducting the evaluation around common central questions that concern all elements of the program and are relevant for practice, and holding regular information and “reflection” meetings for evaluators and those being evaluated, it is assured that the different parts of the evaluation are well-coordinated with the program and its needs for information.

The critical friend approach is based on the following principles:

The role of the evaluator as a critical friend means that a (more or less classic) external evaluation is conducted. But the evaluators are at any time available in an advisory capacity to the program managers. They are open to new questions for evaluation. And they see themselves as active external observers, who pose critical questions or point out a need for action.

The critical friend approach starts out from the conviction that *evaluations must not only be methodologically sound but also useful* for program managers and decision makers. The evaluation must make effect-oriented, strategic, and operational management of the program easier by delivering methodologically sound assessment at the right time.

The evaluation concept of the critical friend approach places importance on *coherence* between and within the individual parts of the evaluation. In particular, it is critical that the evaluation concept takes into account the program logic and that it contains questions and indicators concerning all evaluation objects of the program.

The critical friend approach sees evaluation as a part of program management, with its limited financial and time resources. For this reason, it has to be theoretically coherent and methodologically stringent but *pragmatically practicable* in practical application. The evaluation concept must stay within the given financial and time guidelines and nevertheless respect the quality standards of evaluations.[3]

The approach also has an *empowering* component. In the sense of the empowerment evaluation approach proposed by Fetterman (2001), the people involved in a program should be enabled to judge its elements independently and on this basis to improve them systematically. This is supported in particular by logic model developed at the start of the project and by the fact that evaluation again and again makes the causal connections and the intended effects of the program the subject for discussion.

Finally, the critical friend approach attaches importance to the *participation* of the program managers and the decision makers in the evaluation process. To this purpose, regular information and discussion meetings are held between the evaluators, program managers, decision makers, and further relevant actors.

---

[3] On evaluation standards, see section 4.

3.3.1 … and the practical application

The evaluation of the child care vouchers pilot program in the city of Lucerne has elements of both external evaluation and self-evaluation. First, the evaluators supported the city of Lucerne already during preparation for the pilot project, in that they developed subsidy models and financing options (Balthasar et al., 2008). In addition, the evaluators had already conducted a study for the responsible federal department in 2005 on the bases of and options for child care vouchers (Balthasar, Binder, & Götsch Neukom, 2005). The people responsible for the program in Lucerne contracted them to conduct the evaluation because they did not want to do without their expertise. Second, the city council of Lucerne wanted the pilot project to be seen as a learning program from the very start. For this reason, they expected the evaluators to deliver assessments and results as rapidly as possible, so that they could be used to optimize the program. In other words, the decision makers of the child care vouchers pilot program in the city of Lucerne wanted to set up intensive feedback loops between evaluation and program.

Third, the evaluators had extensive experience in conducting external evaluations as well as in supporting self-evaluation processes. One of the important lessons they had learned from the numerous previously conducted evaluations was that evaluation is most useful for program development and program success when it is intensively integrated into the conceptualization of the program. This condition was fulfilled with the child care vouchers pilot project, so that from the point of view of the evaluation team, the prospects for a useful evaluation were good.

To guarantee a balanced relation between closeness and distance between evaluators and the program managers, the evaluation is based on three types of empirical data. The first type of information comes from system data that documents the program implementation process. This data is collected by the local administration according to instructions by the evaluators and analyzed regularly by the evaluators. Next, the evaluators conduct standardized surveys of parents, heads of child care facilities, and other actors to document the development of the child care services landscape. As in a classic external evaluation, this second type of data is collected independently from the program managers. The third database is called “reflections.” Reflections include a self-evaluation conducted by the program managers, which the evaluators moderate and document. Another element of the reflections is discussions between target groups, such as heads of child care facilities, and responsible politicians. This dialogue is a kind of political consultation procedure organized by the city of Lucerne. The evaluators contribute to the discussions by presenting results of their work, and they make use of the results of the exchange as additional empirical material.

The empirical bases of the evaluation were chosen with the intention of closely coordinating the evaluation process with the needs of the program managers and still retaining the distanced view of the evaluators.

The evaluation team has been consistently included in the information of the city parliament, the city council and the public. The evaluators were and are included also in difficult discussions with target groups. In addition, the evaluators participate in the monthly strategy meetings of the responsible department of the city administration. For one, the tight integration of the evaluation in the pilot project gives the evaluation team insight into central challenges of the pilot project without any extra effort required on the part of the program managers. The evaluators are always up-to-date and informed about pending decisions, options under discussion, and the solutions chosen. In this way, no factor can be forgotten that promotes or hinders the impact of the program. For another, the evaluators are in this way competent discussion partners for the program managers. Despite their closeness to the program, they can provide advisory support from an external perspective. The role of a critical friend is also strengthened by the fact that the evaluators also support other towns and cities in the planning and implementation of similar projects and are versed in general in family policy.

Along with the strengths of the critical friend approach mentioned above, there are also weaknesses. For one, the role of the critical friend – between bearing some responsibility for the program and serving as an independent, external expert – is not always clear cut. For another, the political acceptance of the results produced by this kind of evaluation may be rather low. If in the end the evaluation yields positive results on the program, opponents of child care vouchers will likely cast doubts on the independence of the evaluation findings. At the present time, however, these problems have not yet appeared. It is hoped that the results of the pilot project will be so clear that opponents and supporters will be convinced by the facts.

3.4 Fourth question: What evaluation design should be the basis of the evaluation?

3.4.1 The idea …

In the literature, the term evaluation design is not used uniformly. For some evaluators, deciding on the design of the examination starts already when stating the problem and ends with conducting "feedback workshops to assist the client group in reviewing and discussing draft evaluation reports" (Stufflebeam & Shinkfield, 2007, p. 493 ff.).

Here we argue for using the term evaluation design in a narrower sense. The determination of the benchmark for assessing the examined object is considered the heart of the evaluation design. This is because, following Scriven's postulated science of valuing (see section 2.3), we conceive of evaluations as an act of transparent assessment of the merit, worth, and value of programs. This is not possible without making comparisons. The ideal of the controlled experiment as described by Campbell and Stanley (1963) can be seen as a starting point for the development of an evaluation design. This does not mean that we view the experimental design as the gold standard of evaluation research. But it does mean that we find it decisive for the evaluation design to make clear to what benchmark the assessment of the program being evaluated will be compared.

It has been shown that it is in most cases neither possible nor adequate to implement the classic experimental design in everyday evaluation practice. This has several reasons: First, it is usually not at all possible to undertake random assignment to intervention group and control group. Second, in many cases ethical and political reasons demand equal treatment of all citizens, so that it is not possible to form a control group with the same starting conditions. And third, there are justified methodological reservations concerning the meaningfulness of evaluations that are based on experimental designs (Pawson & Tilley, 1997; Farrington, 2003). For these reasons, method-oriented evaluation theoreticians have developed numerous alternative designs to achieve isolation of the effects of the programs (Fitzpatrick, Sanders, & Worthen, 2004). The point is always to establish what data should be collected when (before and/or after starting the political program) and on what objects (test group or some type of control group).

In the evaluation literature the distinction is often made between comparative and noncomparative evaluation designs (Stufflebeam & Shinkfield, 2007, p. 22). Whereas comparative designs often stand in the tradition of Campbell, especially single case studies are held to be the prototype of noncomparative designs (Klöti & Widmer, 1997). As mentioned, we are convinced that evaluations always contain comparative aspects. Here, three types of comparisons can be distinguished: *Target/actual comparisons* compare the objectives pursued with the actual results of a program. In evaluation practice single case studies are often based on (implicit) target/actual comparisons. Single case studies are highly accepted in evaluation practice, because they make possible richly faceted insights into the connections between intervention and program. With this approach, causal relationships connecting structures, processes, and results can be identified (Bude, 2003; Blatter, Janning, & Wagemann, 2007). *Longitudinal comparisons* base on observation at different time points, or on the gathering of retrospectively assessed situations. They are used when the temporal development of a project is to be studied. Longitudinal comparisons can be planned as a simple comparison of the situation before and after the introduction of a program (Kirchgässner & Wolters, 2006). But it is more correct methodologically to carry out measurements over a longer time period. Finally, in *cross-sectional comparisons* different groups are compared with respect to the factors of interest (Klöti & Widmer, 1997, p. 189).

Many times, the assessment of a program takes place on the basis of a comparison and synthesis of the results of various comparisons. The classic experimental design, for example, can be seen as a combination of a longitudinal comparison and a cross-sectional comparison. In the case of combination of different comparisons, evaluation science speaks of triangulation (Flick, 2008). Triangulation, which is commonly used particularly in the social sciences, serves to provide higher plausibility and credibility to research findings, because a series of various measurements or analyses produce a more valid analysis in the aggregation. The idea is that we can be more confident about a result, if different methods or perspectives lead to the same finding. Weaknesses of the one method are balanced out through the strengths of the other, because in this way, errors, problems, or measurement inaccuracies can be revealed. Denzin (1970) identified four basic types of triangulation: data triangulation, investigator triangulation, theory triangulation, and methodological triangulation. In 1994 we developed the idea of conceptional triangulation, meaning that different sorts of comparisons can be combined in one single evaluation (Balthasar & Knöpfel, 1994).

3.4.2 … and the practical application

The evaluation of the child care vouchers pilot project uses target/actual comparisons, longitudinal comparisons, and cross-sectional comparisons.

Target/actual comparisons are carried out in particular with regard to development of child care supply and the financial requirements. The target values are the values listed in the city council's (*Stadtrat*) report to the legislature (*Grosser Stadtrat*), with figures on supply development and the associated costs of the pilot project (Stadt Luzern, 2009). Further target values were established with the project management before the start of the project.

Longitudinal comparisons are used for almost all questions. For example, it is examined whether the introduction of child care vouchers improves the supply situation in the area of child care services.

C*ross-sectional comparisons* are used, for one, in the evaluation of the child care vouchers where different types of child care facilities are compared (such as day care facility with previous subsidies versus day care facility with no subsidies up to now; day care facilities versus childminders; company day care versus other day care facilities). For another, the developments of the child care supply in Lucerne are compared with the development in cities without a voucher system (comparison with/without measure).

Table 4: Comparison levels of the evaluation

Levels of comparison	**Data collection**
Target/actual comparison	Survey of parents Survey of day care facilities/childminders Self-evaluation by project management Discussions with actor groups Interviews with experts
Longitudinal comparison	Monitoring of system data Documentation of discussions at project events
Cross-sectional comparison	Interviews with experts “Observation”

In the “reflections” part of the evaluation the results of the different comparisons carried out during the evaluation are compared and cross-checked, in the sense of triangulation. This procedure accords with the perspectivism, or multiple accounts, propagated by Scriven (Shadish et al., 1991, p. 76). Unanimous triangulation findings are not interpreted as a more exact but rather as a more stable picture of reality. Conflicting findings between the different methods prompt the carrying out of more in-depth analyses.

3.5 *Fifth Question: How is the information gathered and analyzed?*

3.5.1 The idea …

Methodologically speaking, evaluation is empirical social research. This means that it is based on empirical material and that it analyzes the material adequately. For this reason, in the context of the evaluation it must be clarified how the information that is essential for assessment of a program should be gathered and analyzed. The literature on information gathering and information analysis in the context of evaluation is broad (see, for example, Meyer, 2010; World Bank, 2010), and it can be extensively complemented by the literature on the techniques of social research (for example, Flick, 2005; Diekmann, 2007).

It is not the place here to discuss even some of these methods. However, it is important to make sure that all evaluation questions can be answered with the methods being applied. The critical friend approach attaches great importance to program managers' and decision makers' understanding how the answers to the questions were reached. To reach that goal, quantitative and qualitative methods can be used alongside each other with equal value. In addition to information from surveys and interviews, also official or unofficial statistics, counts, and observations are often used to answer the evaluation questions. If surveys are to be conducted, it is important to first identify the persons or groups whose feedback can best be used to answer the evaluation questions. Depending on the questions, these can be the target groups, those commissioning the evaluation, the program managers, or external experts. In special cases it can also make sense to survey people who are completely uninvolved – for example, when the question is about the image or perception of a program by the public.

After collection of the information it needs to be analyzed by processing the information into data. In this process, a systematic comparison of parallelism and differences in the data and a critical examination of the results are considered to be a standard in qualified social science (Meyer, 2010, p. 216; see also Roth, 1999).

3.5.2 … and the practical application

The evaluation of the child care vouchers pilot project uses a combination of quantitative and qualitative methods. It is based in total on three information bases.

Information Basis I: Monitoring of system data

In continuous monitoring, information that is produced in connection with the carrying out of the project by the city is recorded electronically and analyzed by the evaluation team.

Table 5: Monitoring data

Source of information	**Information (examples)**
Application form for parents (including confirmation by the city)	Number of children who receive child care vouchers Number and composition of households with subsidized children (one-parent or two-parent household; number of subsidized children in the household, etc.) Extent and type of employment outside the home of parents Taxable income and assets
Confirmation form of day care facility /childminders exchange	Name of the child care provider Mutually agreed number of hours of care per child and day Day care facility: parents’ contribution per day
Accounting data	Total value of vouchers Administrative costs

In the starting phase, the most important data on development of supply and demand and the accompanying financial requirements were analyzed at relatively short intervals (monthly) by the evaluation team. Since 2010 these data have been analyzed quarterly.

Information Basis II: Standardized surveys on development of the child care services landscape

The second information base comprises quantitative standardized survey data that make possible systematic assessment of the development of the childcare services landscape in the city of Lucerne.

For this, there was first a written questionnaire for parents who have children in child care outside the home. The questionnaire focused on assessing satisfaction with the child care services provider (criteria for selecting the child care provider, rating of education and care, etc.). The parents were surveyed for the first time (baseline measurement) in 2009. The second parent survey was conducted in the spring of 2011. A written questionnaire was chosen, because this is the best way to reach the parents. The questionnaires are sent by postal service in cooperation with the child care services providers. For their part, they gain the possibility to use the survey results (in anonymized form) for their internal quality development processes.

Second, there is a survey of the providers of child care services (on form of organization, financial bases, number of child care slots, etc.). This survey was also conducted at two time points (baseline measurement in January 2009; second survey in the spring of 2011).

The analysis of these data is done by the evaluators using descriptive and analytical methods.

Information Basis III: Reflections

It is the explicit aim of the evaluation to have the project management and various relevant actor groups (child care services providers, parents, employers, representatives of government, possibly politicians) comment on the status of implementation, topic-specific emphases, eventual problems, and possible ways to optimize. The various assessments are gathered qualitatively via a total of four information sources, and they form the Information Basis III of the evaluation:

The first instrument is a systematic process of *self-evaluation* by the project management, supported by the evaluation team. The content of the self-evaluation focuses on the implementation process, the outputs, and the outcome with the target groups. In this way the strengths and weaknesses of the program are elaborated from the viewpoint of the program management, and the optimization need for the following year is determined. The assessment is made, among other things, according to previously defined goals and indicators. The program management carried out the self-evaluation in May 2009, April 2010, and April 2011. The results, along with further information and evaluation results, are discussed at a meeting with the advisory group of the program whose members are representatives of the city, canton, the federal government, and the evaluation team. This synthesis serves optimization of the project and forms the basis of the definition of program management's objectives for the following year.

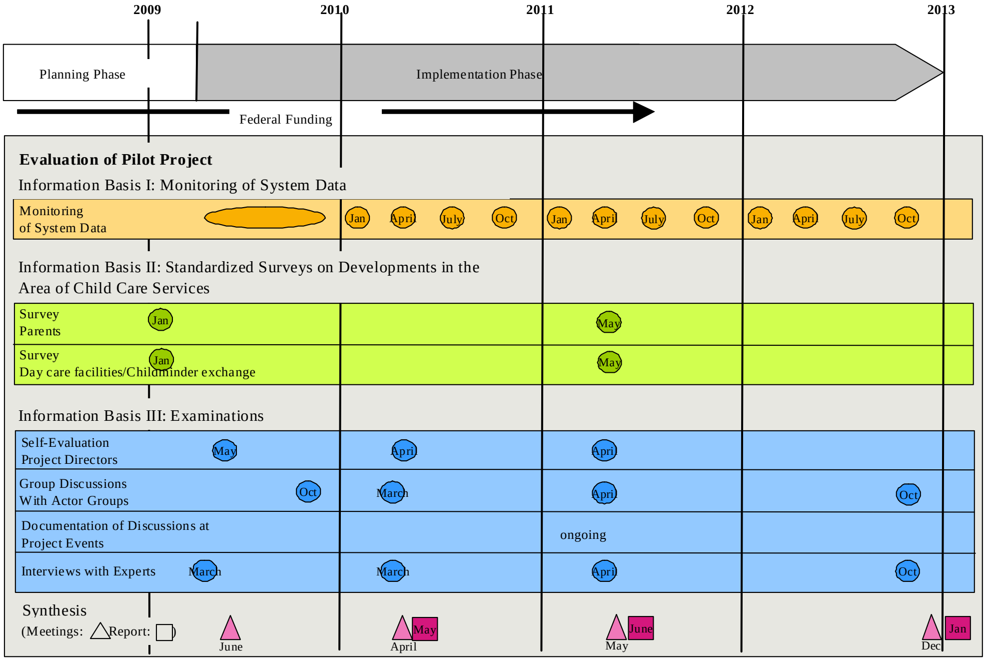
Second, the evaluation includes *group discussions* with actors involved in the pilot project on various topics. The discussions are conducted based on the methodology of the focus group discussion (Morgan & Krueger, 1998). In groups of maximal 10 persons, perceptions, opinions, and ideas are exchanged, and they are documented by the evaluators. The positions of fellow discussion participants work as a stimulus, so that this method clearly produces more information than interviews with single persons conducted in succession do. Group discussions are held with representatives of the child care services providers (day care facilities, childminders), parents, and employers.

Third, during the course of the project the city holds various *events* for the actor groups involved (such as information events, echo groups with child care facilities, workshops on quality development, etc.). The evaluators document the discussions at these events and include them in the ongoing evaluation.

To include also the assessments of further relevant actors, there is a plan for additional *interviews with experts*. These experts will be, for example, representatives of the administration (such as the welfare department), the social services director of the city of Lucerne, representatives of municipalities in the agglomeration involved in the project, and so on).

The following figure shows data collection and submission of reports in the planned time sequence.

Figure 3: Empirical bases and chronology of the evaluation



3.6 *Sixth question: What is planned to promote and support use of the evaluation?*

3.6.1 The idea …

"In the end, the worth of evaluations must be judged by their utility" (Rossi et al., 1999, p. 431). The fundamental practical orientation of evaluations is hardly disputed by evaluation specialists (see section 2.2). However, the extent to which evaluations are of use in practice is disputed. Mainly older studies found that evaluation findings are not utilized sufficiently (Stockmann & Meyer, 2010). However, in those studies usefulness was understood exclusively as direct practical utility, for example in the form of the implementing evaluations' recommendations. In addition to this instrumental utilization, more recent studies identified also conceptual, process-related, symbolic, and other use dimensions (Balthasar, 2007). The conceptual utilization of evaluations has to do with indirect effects through "gradual sedimentation of insights, theories, concepts and ways of looking at the world" (Weiss, 1977, p. 535). Process-related utilization takes place when actors gain insights into the program that affect their behavior due to their participation in the evaluation process alone – and not due to the evaluation results (Vedung, 1999, p. 242). Symbolic utilization of evaluation findings occurs when decision-makers utilize evaluations to confirm their views and to obtain legitimation (Henry & Rog, 1998, p. 90). If the analysis is extended to the various dimensions of utilization, the judgment on the utilization of evaluations is more optimistic. A study that we conducted in 2007 of all of the about 300 evaluations commissioned by the federal government of Switzerland between 1999 and 2002 (Balthasar, 2008) revealed that two-thirds of the evaluators rated the utilization of the evaluations that they had conducted as high or rather high.

Just as important as insight into the diversity of the dimensions of utilization is the fact that the utilization of evaluations and their findings is determined by very different factors (Balthasar, 2007; see also Cousins & Leithwood, 1993). Empirical analyses showed that the utilization of evaluations in Switzerland depended decisively on interest in the results on the part of the program managers (Balthasar, 2007). The critical friend approach takes these findings into account. It attaches importance to ensuring that the evaluation answers the questions that are important to program managers and decision makers, that the results are communicated in their language, and that care is taken to deliver the information to potential users at the right point in time. This is in line generally with Patton's concept of utilization-focused evaluation (Patton, 1997).

3.6.2 … and the practical application

The step named “Synthesis” in Figure 3 shows only parts of the actual contributions of the evaluation that inform the process of further developing the pilot project. As explained above in section 3.3.1, the program managers and the evaluators of the child care vouchers pilot project in the city of Lucerne are in steady communication with each other. No week passes without the mutual exchange of information. In this way there is steady utilization of the knowledge that accumulates through the evaluation.

To discuss the results of the three information bases I, II, and III, semi-annual meetings are held with the project management, the decision makers, and the advisory group of the project. After the first pilot project year, a synthesis report had to be written and presented to the city council in the summer of 2010 (Müller & Bürgi, 2010). The report was presented at a workshop to an interested audience in September 2010. A second, more extensive synthesis report was written in mid 2011. That report will be the basis for the planning and design of the child care system in the city of Lucerne after conclusion of the pilot project at the end of 2012. After the conclusion of the pilot project a short, summarizing synthesis report will be written. For one, this report has legitimizing character, as it documents the results of the pilot project in Lucerne over the entire time period. For another, it can also be useful to cantons and municipalities that also intend to introduce child care vouchers.

4 Conclusion

The critical friend approach was described above along six central questions. The explanations made it clear that the approach has various theoretical roots and that it cannot be assigned completely to any of the three branches of evaluation theory outlined in section 2. In the tradition of the methods-based evaluation theories, the approach attaches importance to methodological rigor. This explains the importance of the logic model as the starting point for the critical friend approach and the importance of a clearly recognizable design for methodologically traceable isolation of the program effects. In the tradition of the values-based evaluation theories, the critical friend approach sees evaluation as assessment on the basis of comprehensible criteria. The evaluator is expected to deliver clear statements on the strengths and the weaknesses of the project. This assessment is usually carried out on the basis of multiple perspectives, as proposed especially by Scriven and as practiced in qualitative research under the keyword “triangulation”: “Perspectivism accommodates the need for multiple accounts of reality as perspectives from which we build up a true picture, not as a set of true pictures of different and inconsistent realities” (Scriven, as cited in Shadish et al., 1991, p. 76). Finally, the critical friend approach also stands in the tradition of Patton’s utilization-focused evaluation, for evaluations aim to contribute to the optimization of policy making. In contrast to Patton, however, the critical friend approach stresses the importance of a logic model as a common ground for all parties involved in the evaluation process. In addition, the critical friend approach focuses – in contrast to Patton – on data that does not need further interpretation by the evaluator or by the program management. There is an affinity with the CIPP approach by Stufflebeam (2000). Stufflebeam attaches importance to an all-embracing evaluation of context (C), input (I), process (P), and product (P) of a project. Similar to the critical friend approach, the CIPP model is used to assess both formative and summative assignments, and evaluation is seen as an “ongoing monitoring” of the implementation process. Stufflebeam’s approach can help to plan activities and to improve the project. The CIPP model is different from the critical friend approach in that it is not based on a common logical model, and it does not stress the importance of unambiguous indicators. In Stufflebeam’s model the evaluator more or less gives up his critical distance and becomes part of the project management.

In the following, we examine the possibilities and limitations of our approach. We do so upon the background of quality standards of evaluation that have become largely accepted in the United States and also in Europe. Finally, we broaden our view beyond the critical friend approach and discuss the problems of evaluation as a part of policy analysis.

4.1 Possibilities and limitations of the critical friend approach

The acceptance of evaluations depends strongly on their quality. Whereas broadly accepted quality standards have existed for a longer time in classic social science research, at least for quantitative research, corresponding considerations for evaluation research are more recent. In the course of the professionalization of evaluation, various initiatives were taken in the United States in the late 1970s to capture the diverse dimensions of evaluation using a set of criteria (Widmer, 2000). Particularly important were the Standards for Evaluation of Educational Programs, Projects, and Material published in 1981 by the Joint Committee on Standards for Educational Evaluation (Joint Committee & Sanders, 2000). With regard to their methodological orientation, these standards follow a relatively open approach. They form the basis for the evaluation standards developed in recent years in several European countries. The evaluation standards of the Swiss Evaluation Society, for example, distinguish four quality dimensions, each having between three and four individual standards (Widmer, Landert, & Bachmann, 2000):

A total of eight *utility standards* “guarantee that an evaluation is oriented to the information needs of the intended users of the evaluation” (Widmer et al., 2000, p. 4). Required here are, for example: explicit identification of persons participating in, and affected by, an evaluation, clarification of the objectives of an evaluation, and transparency of the results for the intended users of the evaluation.

Three *feasibility standards* ensure that an evaluation is conducted in a realistic and cost-conscious manner. Evaluation procedures should be designed such that there is minimal disruption of the evaluation object; the cost-benefit relationship should be as optimal as possible; and “the various positions of the different interests involved are taken into account in planning” (Widmer et al., 2000, p. 7).

Six *propriety standards* “ensure that an evaluation is carried out in a legal and ethical manner” (Widmer et al., 2000, p. 7). This means, among other things, that written agreements specify the duties of the parties who agree to conduct an evaluation, the personal rights of individuals are ensured, and the strengths and weaknesses that exist in the object being evaluated are reported completely. One especially challenging standard for the critical friend approach is the requirement to declare conflicts of interests. Evaluators should be especially explicit about their own interests and values concerning the conduct and outcomes of an evaluation.

Finally, ten *accuracy standards* ensure that an evaluation is conducted with scientific rigor, so that "the evaluation produces and disseminates valid and usable information" (Widmer et al. 2000, p. 10). It must be guaranteed that the object of the evaluation, context, methods, and sources of information are documented, that the reliability of the analyses is assured, and that "the conclusions reached in an evaluation are explicitly substantiated" (Widmer, 2000, p. 12).

This list sounds clear and comprehensible, but the individual standards are difficult to meet in everyday evaluation practice. In particular, the accuracy standards often collide with the standards for timely reporting and the standards on data protection, so that frequently, the four groups of criteria can be fulfilled only in part. The evaluation standards are therefore to be understood as the ideal – not as minimal standards but more as maximal standards (Widmer 2000, p. 92). A closer look also shows that cutting across the four dimensions and the diverse individual standards, there are mainly three groups of principles for all evaluations that should be applied at all levels: These are methodological soundness, practical relevance, and transparency of the evaluation process.

These three principles are also the guidelines of the evaluation of the child care vouchers pilot project in the city of Lucerne. Experiences show that the critical friend approach is proving its worth for ensuring the utility standards. The evaluation has already now been able to establish itself as an important element of a learning organization in the administration of the city of Lucerne. Members of the evaluation team are consulted when important strategic decisions are pending by the city council; evaluation team members participate at internal meetings held for the purpose of clarifying fundamental questions; and they are important contacts for public relations work. At the same time, their closeness to the program makes it possible to take up topics relevant to the evaluation in a timely manner and – as far as possible – to integrate them into the evaluation or have them handled by the administration.

Whether or not the approach sufficiently satisfies the requirements of methodological soundness and transparency cannot be assessed conclusively by us. However, it is our opinion that the approach will largely fulfill the demands for methodological soundness, because it means that scientific methods are applied correctly. The publication of the intermediate and the final report gives a third party the opportunity to conduct a meta-evaluation. With this tool, the adequate application of the techniques used for data collection and analysis can be examined independently. In addition, to enable a meta-evaluation, attention is drawn to an accurate documentation of the databases. Documentation is also an important requirement to secure the necessary transparency of the evaluation process. Another critical element of transparency is the importance given to unambiguous indicators. This measurement helps to prevent the evaluators from making biased judgments due to their limited distance to the project. In addition, the different roles of the evaluation team – between consulting and external observation – must be stressed at all times.

However, in the disclosure of interest conflicts the limitations of the approach become discernible. But this is a general problem of evaluation. Most evaluators are aware that the line between scientific objectivity and political conviction is often very fine. But the problem emerges particularly clearly with the critical friend approach. Whereas it can already be stated at the present time that the evaluation, through its work and its contributions to project development, has earned the respect of the program managers and decision makers, it remains open whether the evaluation results will be accepted also by outside persons. This may be doubtful in the case of those stakeholders who have been critical of the child care vouchers pilot project from the very start. It is therefore conceivable that when the evaluation following the critical friend approach has been completed, an independent evaluation may have to be conducted. The evaluation ongoing at present can do no more than to present and to archive the data collected in such a way that independent summative evaluation and a meta-evaluation will be possible.

4.2 Challenges to evaluation as part of policy analysis

Evaluations have become an important steering instrument of government authorities. As a consequence, the evaluation field of activity is strongly influenced by politics. Evaluation is predominantly contract research. Although it seldom happens that those commissioning an evaluation want to directly influence the evaluation results, they do decisively determine the evaluation object, in that they select the questions that the evaluation should answer. For this reason, in many places evaluation activity is under criticism (Matthies & Simon, 2008). For instance, the added value of evaluations is disputed, and the conducting of evaluations is seen as a new disease called "evaluitis" (Frey, 2008, p. 125). More differentiated opinions criticize that as a advisory intervention evaluation neglects the reference to central political science, especially democracy theory, that evaluation is too strongly instrumentally oriented, and serves even as an elitist instrument (see, for example, Héritier, 1993, p. 11).

Naturally, evaluations have limitations. Besides the points mentioned above, these limitations are due to the fact that it is often only partly possible to reliably trace policy effects back to specific measures. Moreover, due to the time and cost pressures on administrations, increasingly methodologically simple if not, to put it crudely, "quick and dirty" evaluations are requested. Fortunately, this development is being countered by a promising professionalization of the field. To this belongs the development of the discussion on quality standards mentioned above (see also Rieder, 2009).

The conducting of evaluations is without doubt not a sufficient condition for rational policy, but it is at least a necessary condition, and it contributes towards democratic governance. Without transparency of the results achieved by the actions of government, democratic opinion formation is not possible. By contributing to transparency, well-conducted evaluations support democratic opinion formation. In addition, they help to enlarge the empirical basis for policy research. To understand and explain public polities adequate knowledge of the policies is needed. One of the tasks of evaluation is to provide that knowledge.

References

Balthasar, A. (2007). Institutionelle Verankerung und Verwendung von Evaluationen: Praxis und Verwendung von Evaluationen in der Schweizerischen Bundesverwaltung. Chur: Rüegger.

Balthasar, A. (2008). Der Einfluss unterschiedlicher Institutionalisierungsmodelle auf die Verwendung von Evaluationen: Erfahrungen aus der Schweiz. Zeitschrift für Evaluation, 7(2), 243–271.

Balthasar, A., Binder, H.-M., Götsch Neukom, R. (2005). Kinderbetreuungsgutscheine: Diskussionspapier zuhanden der Zentralstelle für Familienfragen des Bundes. Lucerne: Interface Institut für Politikstudien.

Balthasar, A., Knöpfel, C. (1994). Umweltvorschriften und technische Entwicklung. Eine politikwissenschaftliche Evaluation am Beispiel der Hausheizungen. Ökologie & Gesellschaft 8. Basel: Helbing & Lichtenhahn.

Balthasar, A., Feller-Länzlinger, R., Müller, F. (2008). Betreuungsgutscheine in der Stadt Luzern: Konzeptionelle Grundlagen. Bericht zuhanden des Luzerner Stadtrates. Lucerne: Interface Politikstudien Forschung Beratung.

Bitel, M., Läubli-Loud, M. (2010, October). Learning to Learn: A “Capacity Building Approach” Integrating Evaluative Practice & Programme Intervention. Paper to be presented at the 9th European Evaluation Society EES International Conference, Prague, Czech Republic.

Blatter, J., Janning, F., Wagemann, C. (2007). Qualitative Politikanalyse: eine Einführung in Forschungsansätze und Methoden. Grundwissen Politik Vol. 44. Wiesbaden: VS Verlag für Sozialwissenschaften.

Blau, D. M. (2001). The Child Care Problem: An Economic Analysis. New York, NY: Russell Sage Foundation.

BMFSFJ (Bundesministerium für Familie, Senioren, Frauen und Jugend) (2008). Gutscheine: Gezielte Förderung für Familien. Monitor Familienforschung 12. Berlin: BMFSFJ.

Bude, H. (2003). Fallrekonstruktion. In R. Bohnsack, W. Marotzki, et al. (Eds.), Hauptbegriffe Qualitative Sozialforschung: ein Wörterbuch. UTB für Wissenschaft: Uni-Taschenbücher, 8226. Opladen: Leske + Budrich.

Campbell, D. T. (1969). Reforms as Experiments. American Psychologist, 24: 409–429.

Campbell, D. T., Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), Handbook of Research on Teaching (pp, 171-246). Boston, MA: Houghton Mifflin.

Chen, H. (1990). Theory-Driven Evaluation. Newbury Park, NJ: Sage Publications.

Chen, H., Rossi, P. (1980). The Multi-Goal, Theory-Driven Approach to Evaluation: A model for Linking Basic and Applied Social Science. Social Forces, 59, 106–122.

Chen, H., Rossi, P. (1987). The Theory-Driven Approach to Validity. Evaluation and Program Planning, 10(1), 95–103.

Clarke, A., Dawson, R. (1999). Evaluation Research: An Introduction to Principles, Methods and Practice. London, UK: Sage Publications.

Cousins, J. B., Leithwood, K. A. (1993). Enhancing Knowledge Utilization as a Strategy for School Improvement. Knowledge: Creation, Diffusion, Utilization, 14(3): 305–333.

Davies, H. T. O., Nutley, S. M., Smith, P. C. (Eds.) (2000). What Works? Evidence-Based Policy and Practice in Public Services. Bristol, UK: Policy Press.

Denzin, N. K. (1970). The Research Act in Sociology. Chicago, IL: Aldine.

Diekmann, A. (2007). Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen. Hamburg: Rowohlt.

Enste, D., Stettes O. (2005). Bildungs- und Sozialpolitik mit Gutscheinen, zur Ökonomik von Vouchers. IW-Analysen:

Forschungsberichte aus dem Institut der deutschen Wirtschaft Köln 14. Cologne: Deutscher Instituts-Verlag.

Farrington, D. P. (2003). Methodological Quality Standards for Evaluation Research. The Annals of the American Academy of Political and Social Science, 587, 49–68.

Fetterman, D. M. (2001). Foundations of Empowerment Evaluation. Thousand Oaks, CA: Sage Publications.

Fitzpatrick, J. L., Sanders, J. R., Worthen, B. R. (2004). Program Evalu–ation: Alternative Approaches and Practical Guidelines (3rd ed.). Boston, MA: Allyn and Bacon.

Flick, U. (2005). Qualitative Sozialforschung: eine Einführung (3. Aufl. der vollst. überarb. und erw. Neuausg.). Reinbek bei Hamburg: Rowohlt.

Flick, U. (2008). Triangulation: eine Einführung (2. Ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Frechtling, J. A. (2007). Logic Modeling Methods in Program Evaluation. San Francisco, CA: Jossey-Bass.

Frey, B. S. (2008). Evaluitis – eine neue Krankheit. In H. Matthies & D. Simon (Eds.), Wissenschaft unter Beobachtung: Effekte und Defekte von Evaluationen. Leviathan Sonderheft 24 (pp. 125–140). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gerlach, I., Damhus, C., Anschütz, C. (2007). Politische Rahmenbedingungen bei der Einführung von Betreuungsgutscheinen: Expertise für das Kompetenzzentrum für familenbezogene Leistungen im Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ). Berlin: BMFSFJ.

Héritier, A. (Ed.) (1993). Policy-Analyse: Kritik und Neuorientierung. Politische Vierteljahresschrift Sonderheft 24. Opladen: Westdeutscher Verlag.

Hilgers, A., Kastner, J., Strehmel, P. (2007). Die Lage in den Kitas im Jahr 2006: Vol. 1. Endbericht für die Max-Traeger-Stiftung. Hamburg: Gewerkschaft Erziehung Wissenschaft. Retrieved May 1, 2010, from http://www.gew-hamburg.de/Binaries/Binary1392/Kita_Endbericht_Bd_1.pdf

House, E. R. (1987). Assumptions Underlying Evaluation Models. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), Evaluation Models: Viewpoints on Educational and Human Services Evaluation (pp. 45–64). Boston, MA: Kluwer-Nijhoff Publishing.

Joint Committee on Standards for Educational Evaluation, Sanders, J. R. (Eds.) (2000). Handbuch der Evaluationsstandards: die Standards des 'Joint Committee on Standards for Educational Evaluation'. Opladen: Leske + Budrich.

Kirchgässner, G., Wolters, J. (2006). Einführung in die moderne Zeitreihenanalyse. Munich: Vahlen.

Kirkhart, K. E., Morgan, R. O., Sincavage, J. (1991). Assessing Evaluation Performance and Use. Test-Retest Reliability of Standardized Instruments. Evaluation Review, 15(4), 482–502.

Klöti, U., Widmer, T. (1997). Untersuchungsdesigns. In W. Bussmann, U. Klöti, & P. Knoepfel (Eds.), Einführung in die Politikevaluation (pp. 185–213). Basel: Helbing & Lichtenhahn.

Matthies, H., Simon, D. (Eds.) (2008). Wissenschaft unter Beobachtung: Effekte und Defekte von Evaluationen. Leviathan Sonderheft 24 (pp. 125–140). Wiesbaden: VS Verlag für Sozialwissenschaften.

Mayne, J. (1994). Utilizing Evaluation in Organisations: The Balancing Act. In F. L. Leeuw, R. C. Rist, & R. C. Sonnichsen (Eds.), Can Governments Learn? Comparative Perspectives on Evaluation and Organizational Learning (pp. 17–43). New Brunswick, NJ: Transaction Publishers.

Meyer, W. (2010). Informationssammlung und -bewertung. In R. Stockmann & W. Meyer (2010), Evaluation: eine Einführung (pp. 191–234). Opladen: Barbara Budrich.

Morgan, D. L., Krueger, R. A. (1998). The Focus Group Kit. Thousand Oaks, CA: Sage Publications.

Müller, F., Bürgi, M. (2010). Evaluation des Pilotprojekts Betreuungsgutscheine für die familienergänzende Kinderbetreuung in der Stadt Luzern. Zwischenbericht 2010. Lucerne: Interface Politikstudien Forschung Beratung. Retrieved July 21, 2011, from http://www.interface-politikstudien.ch/downloads/deutsch /Be_Betreuungsgutscheine_LU_Evaluation_2010.pdf

Müller-Kohlenberg, H., Beywl, W. (2003). Standards der Selbstevaluation – Begründung und aktueller Diskussionsstand. Zeitschrift für Evaluation, 2(1), 65–75.

OECD. (Ed.) (2004). Babies and Bosses: Reconciling Work and Family Life: Vol. 3. New Zealand, Portugal and Switzerland. Paris: OECD.

OECD. (Ed.) (2007). Babies and Bosses: Reconciling Work and Family Life: Vol. 5. A synthesis of findings for OECD countries. Paris: OECD.

Patton, M. Q. (1988), The Evaluator's Responsibility for Utilization. Evaluation Practice, 9(2), 5–24.

Patton, M. Q. (1997). Utilization-Focused Evaluation: The New Century Text (3rd ed.). Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (2000). Utilization-Focused Evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), Evaluation Models: Viewpoints on Educational and Human Services Evaluation (2nd ed.) (pp. 425–438). Boston, MA: Kluwer Academic Publishers.

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

Rieder, S. (2009). Anwendung der SEVAL-Standards: Ergebnisse aus vier Fallstudien. LeGes – Gesetzgebung & Evaluation, 20(3): 387–397.

Ritz, A. (2003). Evaluation von New Public Management: Grundlagen und empirische Bewertung von Verwaltungsreformen in der schweizerischen Bundesverwaltung. Bern: Haupt.

Rossi, P. H., Freeman, H. (1993). Evaluation: A Systematic Approach (5th ed.). London, UK: Sage Publications.

Rossi, P. H., Freeman, H., Lipsey, M. W. (1999). Evaluation: A Systematic Approach (6th ed.). London, UK: Sage Publications.

Roth, E. (Eds.) (1999). Sozialwissenschaftliche Methoden: Lehr- und Handbuch für Forschung und Praxis (5th reviewed ed.). Munich: Oldenbourg.

Schweizerischer Bundesrat. (2007). Verordnung über Finanzhilfen für familienergänzende Kinderbetreuung vom 9. Dezember 2002 (Stand am 1. Oktober 2007) [SR 861.1]. Retrieved June 29, 2010, from http://www.admin.ch/ch/d/sr/861_1/index.html

Scriven, M. (1986). New Frontiers of Evaluation. Evaluation Practice, 7(1), 7–44.

Scriven, M. (1972). The Methodology of Evaluation. In C. H. Weiss & A. Etzioni (Eds.), Evaluating Action Programs: Reading in Social Action and Education (pp. 123–136). Boston, MA: Allyn & Bacon.

Scriven, M. (1991). Evaluation Thesaurus (4th ed.). Newbury Park, CA: Sage Publications.

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

Stadt Luzern. (2009). 1. Statusbericht per 30.6.2008: Pilotprojekt Betreuungsgutscheine. Retrieved June 30, 2010, from http://www.stadtluzern.ch/de/dokumente/publikationen/welcome.php?action=download&amt_id=&id=27230

Stockmann, R. (2000). Evaluation in Deutschland. In R. Stockmann (Ed.), Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder (pp. 11–40). Opladen: Leske + Budrich.

Stockmann, R. (2007). Handbuch zur Evaluation: eine praktische Handlungsanleitung. Sozialwissenschaftliche Evaluationsforschung 6. Münster: Waxmann.

Stockmann, R., Meyer W. (2010). Evaluation: eine Einführung. Opladen: Barbara Budrich.

Stufflebeam, D. L. (2000). The CIPP Model for Evaluation. In D. L Stufflebeam, Madaus G. F., & Kellaghan, T. (Eds.), Evaluation Models: Viewpoints on Educational and Human Services Evaluation (2nd ed.) (Evaluation in Education and Human Services Volume 49) (pp. 279-317). Boston, MA: Kluwer Academic Publishers.

Stufflebeam, D. L., Shinkfield, A. J. (2007). Evaluation Theory, Models, and Applications. San Francisco, CA: Jossey-Bass.

Stutzer, A., Dürsteler, R. (2005). Versagen in der staatlichen Krippenförderung – Betreuungsgutscheine als Alternative (CREMA Working Paper Series, 26). Basel: Center for Research in Economics, Management and the Arts (CREMA).

University of Arizona Rural Health Office, University of Arizona College of Public Health, Arizona Board of Regents. (2000). Community Health Worker Evaluation Tool Kit. Tucson, AZ: Arizona Board of Regents. Retrieved June, 6, 2010, from http://www.publichealth.arizona.edu/chwtoolkit/toolkit.htm

Vedung, E. (1997). Public Policy Program Evaluation. New Brunswick, NJ: Transaction Publishers.

Vedung, E. (1999). Evaluation im öffentlichen Sektor. Vienna: Böhlau Verlag.

Weiss, C. H. (1977). Research for Policy's Sake: The Enlightenment Function of Social Research. Policy Analysis, 3(4), 531–545.

Wholey, J. S. (1994). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), Handbook of Practical Program Evaluation (pp. 15–39). San Francisco, CA: Jossey-Bass.

Widmer, T. (2000). Qualität der Evaluation – Wenn Wissenschaft zur praktischen Kunst wird. In R. Stockmann (Ed.), Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder (pp. 77–102). Opladen: Leske + Budrich.

Widmer, T., Landert, C., Bachmann, N. (2000). Evaluation Standards of SEVAL, the Swiss Evaluation Society (SEVAL Standards). Bern, Geneva: SEVAL. Retrieved June 27, 2010, from http://www.seval.ch/en/documents/SEVAL_Standards_2000_en.pdf

W. K. Kellogg Foundation. (Ed.) (2004). Using Logic Models to Bring Together Planning, Evaluation, and Action: Logic Model Development Guide. Retrieved June 6, 2010, from http://www.ncga.state.nc.us/PED/Resources/documents/LogicModelGuide.pdf

Wollmann, H. (2000). Evaluierung und Evaluationsforschung von Verwal-tungspolitik und -modernisierung – zwischen Analysepotential und -defizit. In R. Stockmann (Ed.), Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder (pp. 195–230). Opladen: Leske + Budrich.

Wollmann, H. (2003). Kontrolle in Politik und Verwaltung: Evaluation, Controlling und Wissensnutzung. In K. Schubert & N. C. Bandelow (Eds.), Lehrbuch der Politikfeldanalyse (pp. 335–360). Munich: Oldenbourg.

World Bank. (2010). International Program for Development Evaluation Training (IPDET). Retrieved June 6, 2010, from http://www.worldbank.org/ieg/ipdet/modules.html