Exploring the Necessary Conditions for Evaluation Use in Program Change

Simone Ledermann

American Journal of Evaluation 2012 33: 159 originally published online 14 June 2011
DOI: 10.1177/1098214011411573

American Journal of Evaluation
33(2) 159-178

DOI: 10.1177/1098214011411573
http://aje.sagepub.com

Exploring the Necessary Conditions for Evaluation Use in Program Change

Simone Ledermann[1]

Abstract

Research has identified a wide range of factors that affect evaluation use but continues to be inconclusive as to their relative importance. This article addresses the complex phenomenon of evaluation use in three ways: first, it draws on recent conceptual developments to delimitate the examined form of use; second, it aims at identifying conditions that are necessary but not necessarily sufficient for evaluation use; third, it combines mechanisms of evaluation use, context conditions, and actor perceptions. The study reported here examines the use of 11 program and project evaluations by the Swiss Agency for Development and Cooperation (SDC). The article makes use of qualitative comparative analysis (QCA), a method that is well suited to the study of context-bound necessity. It is concluded that the analysis of conditions that are necessary to trigger mechanisms of evaluation use in certain contexts is challenging, but promising to face the complexity of the phenomenon.

Keywords

evaluation use, necessary conditions, qualitative comparative analysis (QCA), mechanisms, context

The questions of whether and how evaluations are used are nearly as old as the evaluation business itself. A first round of empirical research about evaluation use took place in the mid-70s, when evaluators realized that their results were not implemented as they had expected (Alkin, Daillak, & White, 1979; Weiss, 1972). However, empirical research on evaluation use has remained disappointingly inconclusive (Frey & Ledermann, 2010). In the 1980s, a rich body of mostly quantitative empirical studies identified many determinants of evaluation use, but their relative importance remained contested (Alkin, 1985; Cousins & Leithwood, 1986). Recent reviews of the empirical literature on evaluation use also identify many potentially relevant factors, but find it impossible to state which ones are most related to increasing evaluation use (Johnson et al., 2009). What has become evident through all the studies is that the phenomenon of "use" is multifaceted, and there have been numerous attempts to conceptualize it more clearly.

[1]University of Berne, Switzerland

Corresponding Author:
Simone Ledermann, Parliamentary Control of the Administration, Parliamentary Services, 3003 Bern, Switzerland
Email: simone.ledermann@parl.admin.ch

The study reported in this article follows the calls in recent years that it is time to abandon the ambition of finding "the important" characteristic for use and to adopt a focus on context-bound mechanisms of use instead (Henry & Mark, 2003; Mark & Henry, 2004). It draws on these contributions in order to conceptualize and explain use. This article remains modest as to the possible scope of explanation. Given the complexity of the phenomenon, we will never be able to fully account for any form of evaluation use. The "spirit of humility" advocated by Weiss, Murphy-Graham, and Birkeland (2005) is thus shared. Yet, this article aims to contribute to a better understanding of the mechanisms at work and to identify necessary—though not sufficient—conditions to trigger these mechanisms in specific contexts. I argue that in the face of complex phenomena like evaluation use, research on necessary conditions constitutes a good answer. The article presents a method for the identification of necessary conditions, which draws on qualitative comparative analysis (QCA; cf. De Meur & Rihoux, 2002; Ragin, 1987, 2000, 2008; Rihoux & Ragin, 2009; Schneider & Wagemann, 2007).

This method will be applied to evaluations commissioned by the Swiss Agency for Development and Cooperation (SDC). The SDC has a long evaluation tradition and belongs to the offices with the most evaluations in the Swiss Government (Balthasar, 2007; Spinatsch, 2002). There are different types of evaluations within SDC. The present study examines so-called *external* evaluations of programs and large-scale projects funded by SDC. These evaluations are commissioned by the desk manager who is responsible for strategic management of the program or project at SDC head office in Switzerland. External evaluations are carried out by external consultants and are usually scheduled toward the end of a program or project cycle in order to inform decision making by the desk manager about whether and how to continue.

The article begins with recent conceptual developments that will be drawn upon to clarify the focus of this study on decisions to substantially change the evaluated program or project. Then, hypotheses will be developed consisting of mechanisms through which an evaluation can lead to a decision of change and of conditions that are necessary to trigger these mechanisms.

These hypotheses will subsequently be tested in a qualitative comparison of the use of 11 external evaluations by SDC. It will be shown that depending on context, different mechanisms bring about a change decision and different actor perceptions are necessary to trigger these change mechanisms. The article concludes that a context-bound focus on mechanisms and on necessary conditions is expedient in the face of the complexity of evaluation use.

Conceptualization of Evaluation Use

The interest in use arises from the fact that evaluation draws its legitimacy in part from practical use, that is, its ability to help improve policy, programs, or projects (Patton, 1997). There is a normative expectation for evaluations to be used and SDC, for instance, has established principles to ensure successful use (SDC, 2004). From the 1970s onwards, three categories of effects of evaluations were distinguished (Sager & Ledermann, 2008): (a) *Instrumental use:* Evaluation recommendations and findings inform decision making and lead to change in the object of evaluation. (b) *Conceptual use:* Evaluation results lead to a better understanding or a change in the conception of the object of evaluation. (c) *Symbolic use:* Evaluation results are used to justify or legitimize a preexisting position, without actually changing it. Later, Patton (1997) extended the typology to include *process use*: Participation in an evaluation bringing about change, regardless of the evaluation results.

These categories of evaluation use have been widely applied in the relevant literature. In the last decade, however, the lack of a coherent definition has been criticized and several alternatives have been proposed. Mark and Henry (2004) distinguish between effects of evaluations at the cognitive level (thoughts and feelings) and at the behavioral (action) level. Conceptual use pertains to the first kind and instrumental use to the second, while process and symbolic use can be

both. Kirkhart (2000) proposes three dimensions according to which the effects of evaluations can be described: source for change, intention, and time. The *source for change* can be either the evaluation process or the evaluation results. This is what distinguishes process use from the other three types of use, which refer to the results of evaluations (cf. also Mark & Henry, 2004; Weiss et al., 2005). *Intention* is what characterizes symbolic use. In contrast, conceptual and instrumental use can be both intended or not. As to *time*, Kirkhart distinguishes between immediate, end-of-cycle and long-term effects of evaluations. Furthermore, Kirkhart (2000) maintains that "utilization" and "use" are inappropriate expressions, because they suggest purposeful action, whereas in reality, the effects of evaluations are more diffuse. As an alternative, she prefers the term "evaluation influence," denoting "the capacity or power of persons or things to produce effects on others by intangible or indirect means" (Kirkhart, 2000).

In spite of these conceptual developments, most recent studies work with the conventional types of evaluation use, even though many acknowledge their weaknesses and have welcomed conceptual contributions (e.g., Balthasar, 2006; Fleischer & Christie, 2009; Johnson et al., 2009; Weiss et al., 2005). As Weiss and her colleagues (2005) put it: "the three constructs of instrumental, conceptual, and political [here: symbolic] use appear to capture much of the experience in the empirical literature and practical experience."

This article combines old and new. It focuses on evaluation-based decisions to change the object of evaluation as a specific type of instrumental use and draws on the conceptual developments in order to delineate this type of use. Instrumental use is the most straightforward effect of evaluations (cf. Weiss, 1998; Weiss et al., 2005). In Mark and Henry's terms (2004), it is the behavioral outcome of "program continuation, cessation, or change" that is of interest. Change in existing structures—and cessation is just a form of radical change—is likely to lead to opposition. Under what conditions are SDC desk managers prepared to take a decision to change a program or project based on an external evaluation even though they might have to face opposition? This is the central question of the present article. The study concentrates on decisions with some bearing on the program or project (e.g., change of location or partner organization, modification of strategic orientation, termination of the program or of part of it, etc.), because it is assumed that in these cases, the preconditions for evaluation-based decision making emerge more clearly than in the case of decisions to continue with the status quo. Furthermore, the focus on evaluation-based change decisions corresponds to the purpose of SDC's external evaluations. The SDC's principles on external evaluation underline the importance of preparing for change throughout the evaluation process (SDC, 2004). The SDC has showed great interest in increasing this kind of use, also with the help of the present study.

With respect to the time dimension, I examine the end-of-cycle effects of evaluations that are looked at, while the more diffuse long-term effects are ignored. The study has covered the period of roughly half a year after the end of the evaluations. The change decisions had to have been formally taken, and there had to be a written "proof" for a decision to be counted as such. No attempt has been made to find out whether the decisions were actually carried out in the end. Neither did I distinguish between the process and the results as possible sources of change. The information from the evaluation certainly had to have some influence, or in other words "leverage" (Cronbach, 1982), on the desk manager's decision for it to be considered an evaluation-based decision. But I did not care whether the decision had been taken because of the process of evaluation or because of the findings. Given the situation that SDC desk managers are usually at a great geographical distance from the program or project site, the results of the evaluation are likely to be more important than in other contexts.

A Context-Mechanism-Actor Model of Evaluation Use

In Weiss' (1998) words: "Use is about change. Any theory of evaluation use has to be a theory of change." I propose a theoretical model of evaluation-based change constructed according to a

		Level of conflict			
		Low		*High*	
Pressure for change	*Low*	(a)	Awakener	(c)	Conciliator
	High	(b)	Trigger	(d)	Referee

Figure 1. Mechanisms of evaluations use in different contexts. Source: Valovirta (2002; Figure 2) with own adaptations.

"realistic" scheme of theory-building (Pawson & Tilley, 1997) consisting of different mechanisms of evaluation use and different contexts.

Mark and Henry propose a "theory of evaluation influence" with an abundance of interdependent mechanisms (Henry & Mark, 2003; Mark & Henry, 2004). While their list is comprehensive, the high number of mechanisms and possible interdependencies is likely to lead to pathways for evaluation use that look different for each case (Weiss et al., 2005). In order to prevent an "individualization" of each case, where explanatory patterns are not visible (Berg-Schlosser & De Meur, 2009), I only investigated a small number of mechanisms with a sound empirical foundation.

In a well-known review of research on evaluation use, Cousins and Leithwood (1986) propose a conceptual framework with 12 factors that are assigned to two dimensions—evaluation implementation and decision/policy setting. I selected two factors from each dimension. But instead of examining their separate influence on evaluation use, as it is usually done, I adopted a conjunctural approach (Amenta & Poulsen, 1994; Yamasaki & Rihoux, 2009), where it is combinations of factors that only together produce an outcome. This conjunctural approach is applied both in theory-building as well as in the empirical analysis (see next section).

In the dimension of evaluation implementation, the "truth test" described by Weiss and Bucuvalas (1980) singles out evaluation quality and the nature of the findings as two factors that are by now widely accepted to be important. They found that a study is perceived as useful, either if it confirms users' preconceptions or if it is considered good-quality research. The factors are interdependent: If a study challenges users' preexisting beliefs and reveals something new to them, it must be judged high quality to be considered useful, whereas quality is less important if the study confirms users' expectations. This means that the perceived quality of a study must be higher, the higher its "novelty value."[1] The novelty value and quality of evaluations are assumed to affect users' judgments about the usefulness of an evaluation and, as a consequence, are likely to be relevant to whether or not an evaluation is actually used as a basis for decisions about change in the program or project.

The added value of this article is that the truth test is put into context and further differentiated. This is done based on an empirically derived model by Valovirta (2002) that illustrates the significance of two specific characteristics of the decision and policy-setting dimension. According to Valovirta, evaluation use is an argumentative process in which evaluations provide arguments that correspond more or less to users' beliefs and expectations and which they can draw on or reject. The use of the arguments provided by an evaluation will depend on how actors perceive the study. This perception, according to Valovirta, depends on context. Based on empirical research about the use of evaluations in the Finnish Government, he singles out two major context factors: the *level of conflict* among the stakeholders and the amount of *pressure for change*. These two context conditions together result in four mechanisms in which an evaluation can bring about change as shown in Figure 1.

The cells of the figure can be explained as follows: (a) In a situation of low pressure for change and a low level of conflict, an evaluation can reveal unknown problems and act as an *awakener*.

Table 1. Hypothesized Necessary Actor Perceptions for Use in Different Contexts

Context Conditions			Assumed Necessary Actor Conditions	
Pressure for Change	Level of Conflict	Mechanism	Novelty Value	Evaluation Quality
Low	Low	Awakener	High	High
High	Low	Trigger	Irrelevant	High
High	High	Referee	Irrelevant	Irrelevant
Low	High	Conciliator	High	High

(b) In an environment of high pressure and low conflict, an evaluation can be a *trigger* for changes that are broadly accepted as necessary. (c) Under conditions of high pressure and high conflict, the evaluation is likely to function as a *referee*, deciding what ought to be done, even though the solution might not satisfy all the stakeholders. (d) In a situation of low pressure and high conflict, the evaluation is likely to be exploited to defend one's positions and criticize others, without necessarily being used as a basis for change. Exceptionally, however, an evaluation in such a situation can act as a *conciliator* between the conflicting parties, thereby enabling change to happen.

Will the actors perform a sober truth test on the evaluation in all of these situations before making use of it? According to the underlying hypothesis of this article, this is not the case. Depending on the context, actors' perceptions of the novelty value and evaluation quality are likely to be more or less relevant for whether an evaluation is taken as a basis for change decisions.

Table 1 summarizes the necessary actor perceptions for an evaluation to bring about change in the four specified context and through the four specified mechanisms.

It can be hypothesized that:

- To function as an *awakener*, the evaluation must first of all reveal something new to the users. Moreover, given the neutral context, actors are likely to perform a truth test, so the quality of an evaluation also tends to be relevant for the evaluation to be considered trustworthy.
- To act as a *trigger* for change in a consensual situation, where people are aware of problems, an evaluation is likely to be used even without revealing much new. The absence of conflict allows for a sober consideration of the evaluation, so its quality is assumed to be important.
- In a conflict-laden situation with a strong pressure for change, an evaluation is likely to be used as a *referee*, no matter its quality or novelty value, because something has to be done. Different groups are likely to draw on different parts of the evidence and interpret it differently (Jewell & Bero, 2007).
- Where there is a lack of problem awareness among stakeholders that are in conflict with one another, the chances for an evaluation to be used for change decisions are small. To act as a *conciliator*, it is to be assumed that an evaluation must be of high quality and must show some new ways out of the situation.

To sum up, this article examines the interplay between the policy and decision setting on the one hand and specific characteristics of evaluation implementation on the other hand. I investigate how the truth test, which is based on the actor perceptions about the novelty value and quality of an evaluation, works under varying levels of conflict and pressure for change. The question is whether these factors contribute to an explanation of why some evaluations are used as a basis for decisions to change the evaluated program or project while others are not.

It is not contended here that the context and actor-related factors fully account for the occurrence of any evaluation-based change decision. I will not try to identify all the ingredients needed for an evaluation to cause change. As past research has shown, evaluation use is a complex

phenomenon in which so many aspects are potentially relevant, that this would appear to be an excessive claim.

The assertion is a more modest one: To find conditions that are *necessary* for evaluations to cause change *in certain contexts*. In other words, I am interested in a specific set of change mechanisms, each operating in a specific context. I attempted to find some preconditions for an evaluation to be able to act as an awakener, trigger, referee, or conciliator under the above-mentioned contexts characterized by the presence or absence of conflict or pressure for change.

Method

The hypotheses in Table 1 are tested with a comparative case study design. This section describes the research context and case selection, data collection, and data processing. In particular, it introduces the method of QCA.

Research Context and Case Selection

The theoretical framework described above has been applied to a set of external evaluations of the SDC. Because organization matters for evaluation use (e.g., Weiss, 1998) and because organizational factors are not part of the research model, the focus on just one type of evaluation within only one organization is a crucial control. The SDC has internal guidelines about how to carry out external evaluations that guarantee a certain degree of similarity of the process. The research design thus corresponds to a "comparative-cases strategy" (Lijphart, 1975), in which cases are selected according to the "most similar" design (Przeworski & Teune, 1970) that maximizes the variance of the explanatory factors and minimizes the variance of the control conditions. There are about 30 planned external evaluations per year (cf. SDC, 2002 and following years) that provide enough variance. As it was not possible to ensure the "intimacy" with each of the 30 cases—something that is crucial for the application of QCA (Berg-Schlosser & De Meur, 2009; Ragin, 1994; Rihoux & Lobe, 2009)—11 cases were selected from among the population.

External evaluations are scheduled toward the end of a funding cycle in order to inform decision making about whether and how to continue with a specific program or project. They are carried out by one to three external consultants. The terms of reference with the evaluation questions are set by the desk manager at the head office in Switzerland responsible for the program or project, usually in collaboration with the local SDC field office. Field office staff frequently assist evaluators in organizing visits to program and project sites. At the end of the visit, a debriefing between the evaluators and the local SDC staff is common, sometimes together with program and project staff. However, the main intended users of the external evaluations are the desk managers. They receive the evaluation report and in most cases, a second debriefing with the evaluators takes place at SDC headquarters in Switzerland to discuss the report. Finally, the desk managers have to take the decisions about what to do with the programs and projects based on the evaluation.

In order to assure a certain breath, the cases were selected according to the structure of the organization (cf. Merkens, 2003). The basis of selection was the SDC evaluation program (SDC, 2002). Out of 30 external evaluations listed in the program, 24 had actually taken place and form the universe from which the cases were selected. As a first step, one external evaluation was drawn at random from each of the five organizational units (called departments) within SDC that had carried out at least one external evaluation. As a second step, seven more cases were selected at random to achieve a proportional representation of each department with respect to the total number of external evaluations. The case from the so-called Thematic Department had to be dropped in the course of data analysis due to a lack of information on its use by the desk manager. The final sample

Table 2. Cases by Department

SDC Department	Cases
Eastern Europe and Commonwealth of Independent States	A, B, C, D, E
Bilateral cooperation	F, G, H
Humanitarian aid	I, J
Multilateral cooperation	K

Note: SDC = Swiss Agency for Development and Cooperation.

listed in Table 2 consists of 11 cases and covers 4 departments. The cases cover evaluations of SDC funded program and projects all around the world. At SDC's request for anonymity, the cases are just referred to by random uppercase characters.

Data Collection

Given the desk managers' central role as commissioners and main users of external evaluations, I used interviews with the desk managers responsible for the programs and projects as the main source of information for the case studies. Due to SDC internal staff rotation, half of the desk managers had already left their position at the time of the interview and were often abroad. In some cases, it was necessary to interview more than one desk manager on the same project or program evaluation. In total, 14 desk managers were interviewed, half of them face-to-face, four on the phone, and three in written form by e-mail. The interview guides consisted of questions on the purpose and context of the evaluation, the selection of the evaluators and the process of evaluation, the quality of the evaluation, and its use for the desk managers and the organization. For each topic, there was a mix of closed questions where desk managers were asked to give their assessment on a 4-point scale and open questions where they were asked to elaborate on their view. Face-to-face interviews took about one and a half hours; telephone interviews were a bit shorter. In the case of the written interviews, there were at least two rounds in which desk managers were asked to give more detailed information on certain aspects. In addition to the interviews with desk managers, six telephone and four e-mail interviews were conducted with the external evaluators as a complementary source of information. The guideline consisted of open questions, especially on the context of the evaluation and the follow-up process. The telephone interviews took about 40 min.

Document analysis was used first to prepare the interview guides and second to cross-check the interview information. The documents included the evaluation reports, the terms of reference and the evaluation contracts, as well as program and project proposals for the following funding cycle. Empirical work was carried out mainly in 2004 roughly half a year after the end of the evaluations. At this point, the interviewees were still able to recall the evaluation process and the desk managers had already taken their decisions about how to continue.

QCA and the Concept of Necessity

Earlier it was noted that "intimacy" with each case is important for an accurate comparative case analysis. Therefore as a first step, each case has been regarded as a unique narrative of evaluation-based decision making and has been analyzed on its own. The interplay between actor perceptions, mechanisms, and context has been traced in the form of a "thick description" (Geertz, 1973) of how the evaluation has been conducted, in what context, how it has been perceived and used for decision making by the desk manager. The case narrative corresponds to the above-mentioned conjunctural approach (Amenta & Poulsen, 1994; Yamasaki & Rihoux, 2009) in that it traces the interplay between actor perceptions and context that in their combination lead to a change decision (or the lack of it).

As a second step, the information from the case studies was systematized in order to allow for the subsequent comparison of the cases by the method of QCA (Ragin, 1987). QCA uses set relations and formal logic to find commonalities between various cases with the same outcome. Contrary to statistical methods, which attempt to measure the "net effect" of single, independent variables on an outcome, QCA tries to explain outcomes through combinations of interdependent conditions or in other words through "configurations" (Ragin, 2008). This configurational thinking about causally relevant conditions that only together lead to an outcome is common in qualitative research in general. What is special about QCA is the degree of systematization that allows for the comparison of a greater number of cases.

Each case was systematically classified as to how well it fulfills each of the conditions (pressure for change, level of conflict, novelty value of evaluation, evaluation quality) and whether the outcome (evaluation-based change decision) is present or not. Once all the cases were classified, they were compared with each other by formal logic in order to find the configurations that are necessary or sufficient for the outcome. Given the interest of the present study in the interplay between contexts and actor perceptions, QCA as a configurational method was well-suited for the endeavor (Befani, Ledermann, & Sager, 2007).

There have recently been important methodological advances (e.g., Brady & Collier, 2004; Mahoney & Goertz, 2006). The concepts of necessity and sufficiency have been proven to be helpful to deal with the causal complexity we are confronted with in the real-world (Goertz, 2006a, 2006b; Goertz & Starr, 2003; Ragin, 2000). As the present article is interested in finding the conditions that are *necessary* for evaluation-based change decisions, it concentrates on the concept of necessity, leaving aside sufficiency.

The claim that a condition is necessary for an outcome is a strong one, because it means that this condition always has to be present if an outcome is to occur. Often, however, there are different causal paths to the same outcome, and it is only very trivial conditions that must always be present. Contrary to sufficiency, however, necessity does not imply that the condition will really always lead to the outcome. It is just a precondition for the outcome. In addition, this study will investigate *context-bound claims of necessity*, so the claim is a weaker one. It will try to identify necessary, but probably insufficient conditions for an evaluation to inform program/project change decisions *under particular circumstances*. The result is what George and Bennett (2005) call "contingent generalizations".

The analysis of necessary conditions is usually considered to be just the first step of a QCA analysis, before turning to the analysis of sufficiency (Schneider & Wagemann, 2007). But with regard to complex social phenomena, such as evaluation use, where so many factors might be relevant, context-bound necessity claims seem more adequate than sufficiency claims. This is why necessity actually deserves an analysis in its own right.

Data Dichotomization

Several QCA techniques have recently been developed (cf. Rihoux, 2006; Rihoux & Ragin, 2009; Schneider & Wagemann, 2007). The present article applies the basic technique of crisp-set QCA (cf. Grofman & Schneider, 2009, for a short general introduction). Crisp-set QCA requires a dichotomization. For every case, it has been decided whether the outcome (evaluation-based change decision) has occurred or not and whether each of the four conditions listed is high or low.

Mayring's (2003) method of qualitative content analysis was used to produce the dichotomized data matrix. For this systematic, rule-guided qualitative text coding, the categories are deduced from the theoretical context. In this case, it was the four conditions and the outcome of evaluation-based change decision that served as categories. After coding a subset of the interview transcripts and documents, the categories have been refined and then applied to the rest of the corpus. The codings

Table 3. Specification of Conditions and Outcome

Element	Fulfilled (=1), if:
Outcome: evaluation-based change decision	Desk manager mentions that based on the evaluation, she or he has decided to make significant changes in the program or project, such as: • Unplanned termination of (part of) the program or project • Change in partner organization • Important strategic change in the program/project with major consequences at the operational level The change decision has to be documented in some way (e.g., project proposal for next funding cycle, letter to partner organization, etc.). It is not necessary that the change decision has been implemented
Context condition 1: Pressure for change	Desk manager mentions problems with the program or project that required a change and that he or she was already aware of before receiving any evaluation information Interview information was cross-checked with evaluation purpose statement in the terms of reference
Context condition 2: Level of conflict	Desk manager or evaluator mentions specific conflicts in the project or program that affected the desk manager in his or her function
Actor condition 1: Novelty value	Desk manager mentions specific points she or he learnt from the evaluation or was surprised to hear from the evaluator
Actor condition 2: Evaluation quality	Desk manager gave positive ratings for the evaluation on the majority of the closed questions on the following accuracy standards: • Precise description of procedures • Appropriate application of research methods • Trustworthy sources of information • Substantiated conclusions • Neutral reporting

for each category have been summarized for each case, and based on these summaries, the cases have been compared and the categories have been rated high or low. It is crucial that this dichotomization is performed on the basis of the intimate knowledge of the cases. Table 3 specifies the categories for the outcome and the four conditions as applied.

In the rare instances of contradicting information on the same condition (contradicting interview statements or document information), the condition was only coded "high" if the majority of the sources pointed in this direction, otherwise it was assessed to be low. The dichotomization led to the data matrix in Table 4, which lists all the cases and codings.

In the following paragraphs, I have selected particular cases to illustrate the dichotomization of the conditions and the outcome. More detail on each case can be found below in the "findings" section of the article.

In all 6 out of the 11 external evaluations informed a decision to substantially change the evaluated program or project. In case C, for instance, the substantial change consisted in an additional component for the existing emergency medicine program (delivery of consumable material to local hospitals). In case E, the desk manager decided to terminate the project based on the evaluation, although this decision has finally not been implemented for political reasons. Given the focus of this study on decision making by the desk manager, the outcome is nonetheless rated positive. Five external evaluations did not lead to any important change decisions. For example, evaluation J, which concerned a building project, did not inform the desk manager's decision because he could not find

Table 4. Data Matrix

	Context Conditions		Actor Conditions		Outcome
Case	Pressure for Change	Level of Conflict	Novelty Value	Evaluation Quality	Change Decision
A	0	1	1	1	0
B	1	1	1	1	1
C	0	0	1	1	1
D	1	1	0	0	1
E	0	0	1	1	1
F	1	0	0	1	1
G	0	0	0	1	0
H	1	1	0	1	1
I	0	1	0	0	0
J	0	0	0	0	0
K	1	0	0	0	0

Note: 0 = condition low/outcome absent; 1 = condition high/outcome present.

an answer to his strategic questions in the evaluation, so the next funding cycle was started without any major changes. In case A, the decision to terminate the program had already been taken when the evaluation delivered its results. The lessons learnt from this evaluation had not informed any decisions about other programs by the time of the present study.

As to the context conditions, the pressure for change was graded high in case D, for example, because the desk manager had encountered major problems with an important partner all along. In three out of five cases, where the level of conflict was high, this was due to the fact that the program or project was very controversial within SDC or between SDC and a second funding body in the federal administration (cases A, B, and H). In the other two cases, there were severe conflicts between SDC and the local implementing agencies. In case F, there were conflicts between local stakeholders the desk manager did not know about and which did not affect him; the level of conflict is, therefore, considered low.

Concerning the desk managers' perceptions of the evaluations, the novelty value was rated high in four cases. For instance, the lack of consumable material in local hospitals, which evaluation C revealed, came as a great surprise to the desk manager. In all 7 out of the 11 evaluations were considered of high quality by the desk managers. The cases B, F, and H, for example, were rated positively on all five accuracy standards, whereas evaluation I was assessed negatively throughout. In case K, the description of the procedure, neutrality, and the soundness of the conclusions were negative.

Steps of a QCA Analysis

The data matrix in Table 4 served as a basis for the QCA analysis. The analysis of necessity focuses on the cases where the outcome has occurred, that is, the evaluations that have actually led to a change decision (cf. Schneider & Wagemann, 2007, for the steps of a necessity analysis). It is analyzed whether these cases show a common condition or a common absence of a certain condition. If they do, this is an indication that this condition or its absence, respectively, might be necessary for the outcome to occur. The claim that a condition is necessary must be substantiated in qualitative terms based on the knowledge of the cases (cf. Rihoux & Lobe, 2009). It must be argued why the presence of the necessary condition was in fact causally relevant for the occurrence of the outcome.

To further consolidate the necessity claim, the analysis can turn to the cases where the asserted necessary condition is absent. It follows logically that the outcome in these cases did

		Level of conflict	
		Low	*High*
Pressure for change	*Low*	(a) Awakener *C*, *E*, G, J	(c) Conciliator A, I
	High	(b) Trigger *F*, K	(d) Referee *B*, *D*, *H*

Figure 2. Distribution of cases by context. Characters in italics denote cases where the outcome is present.

not occur. Based on the case knowledge, it can be checked whether the absence of the necessary condition was causally relevant for the absence of the outcome. If it is, the necessity claim is substantiated further. However, there might be other reasons why the outcome did not occur. Even though the necessary condition is present, the outcome might be absent, but this does not challenge the necessity claim. For instance, an evaluation might not have been used for decision making because the information was not available in time. This, however, would not contradict the claim that good evaluation quality is necessary for evaluation-based decision making.

In the present study, the analysis of necessity is context-bound. This means that for each combination of the two context conditions "pressure for change" and "level of conflict," the analysis of necessity is carried out separately. First, we look at the cases with low pressure for change and a low level of conflict and examine whether the evaluations that were used for decision making under these context conditions show any common actor conditions that can be considered necessary. Then, we turn to the cases with low pressure for change, but high conflict, and check for necessity and so on and so forth.

In brief, data analysis with QCA involves a three-step procedure: First, each case is analyzed separately in order to trace the mechanisms that have led to a change decision or, on the contrary, have failed to do so. Second, based on the dichotomous data matrix, the cases in each context constellation are systematically compared with each other. Third, the results of this comparison are interpreted based on the intimate knowledge of each case. The findings from this analysis will be presented in the next section.

Findings of the Comparative Case Analysis

The data matrix in Table 4 shows that the cases with a positive outcome (desk manager has taken an evaluation-based change decision) do not share any common condition, nor do the cases with a negative outcome. This is to say that no single condition alone is necessary for the occurrence or nonoccurrence of an evaluation-based change decision. This complex structure in the data hints at the complexity of the underlying causal links and suggests that a context-bound view is more adequate. Analogous to Figure 1, Figure 2 shows the distribution of the cases along the two context dimensions and indicates the outcome, that is, whether in the respective cases an evaluation-based change decision has been taken or not.

The 11 cases are scattered among the four contexts. In the "awakener" and "trigger" contexts, there are both cases with a positive and a negative outcome. In the "referee" context, all the cases show a positive outcome, which means that all three external evaluations were used as a basis for a change decision. In contrast, in the "conciliator" context, none of the evaluations has been used in that way.

For each of the four contexts in Figure 2, I examine whether the claimed change mechanisms have been present and whether the necessity claims are substantiated or not.

Evaluation as an Awakener (Low Pressure, Low Conflict)

In the context of low pressure and low conflict, it had been claimed that an evaluation can cause change by awakening people, provided that it reveals something new and that it is of good quality. There are two cases, C and E, where the desk managers have actually taken a change decision, and two cases, G and J, where they have not.

Case C fulfills the two actor conditions that have been claimed necessary: The desk manager was convinced of the good quality of the evaluation, which had been carried out by two highly competent logistics specialists. Furthermore, the evaluation revealed something new, as it disclosed a severe shortcoming in the program (lack of consumables in emergency medicine) in one of the towns that had been considered the model project site. As a consequence, discussions were held with the state health ministry and the desk manager decided to integrate the missing component (consumable supply) in the program. The evaluation did clearly work as an awakener.

Case E also fulfills the two supposedly necessary conditions of novelty and good quality. The evaluation was assessed as good by the desk manager and gave her a lot of new information, given that she did not know much about the project before. In fact, the evaluation detected several deficiencies and suggested to terminate the project, which the desk manager decided to do. However, for political reasons, the SDC hierarchy intervened and decided that despite of the clearly negative evidence, the project was to continue. So the desk manager had taken a change decision based on the evaluation (positive outcome), but in the end it has not been implemented.

To further substantiate the claim that the two conditions of novelty and high quality are necessary for an evaluation to trigger change in a consensual situation with a low pressure for change, it is useful to consider the negative cases G and J, where one or both of the necessary conditions were absent. Evaluation G was considered good quality but did not disclose anything unknown. Instead, it confirmed the project strategy, which had been adopted before the evaluation was made. The lack of novelty was a crucial reason why the evaluation did not result in a change decision.

Evaluation J was estimated poor quality by the desk manager. The report consisted of only seven pages of unstructured text, describing the project site, which the desk manager knew from his own visits to the place. Many evaluation questions remained unanswered and the evaluation did not provide any new information. In this evaluation, bad quality and a lack of novelty go together.

Evaluation as a Trigger (High Pressure, Low Conflict)

In a consensual environment where stakeholders are aware of problems that must be solved, evaluations are assumed to trigger change only if they are of good quality. They do not need to show something new. It is sufficient if they confirm a strategy of action the stakeholders have thought of already.

Case F corroborates the account of how an evaluation can act as a trigger for change. There were several signs that the program needed modification (cancellation of the cooperation contract by the partner university, high staff fluctuation, lack of impact on government). The evaluation was carried out in a top-down manner, without much participation of the local stakeholders. The SDC's general strategy for the state in question was to foster decentralization to allow for a more balanced development of the country as a whole. So, when the evaluation recommended decentralizing the activities, the suggestion was taken up and the program was adapted accordingly. The evaluation triggered strategic program modifications SDC had already considered.

In case K, there were also signs for a need for change, such as high overhead costs of the organization in charge, but the evaluation was not used as a basis for a change decision in the end.

One of SDC's main evaluation questions was whether to continue the program or not and with what organization. However, the evaluation failed to address these questions. Rather, it restricted itself to how the program could be improved, without challenging the existing arrangement in a more fundamental way. The desk manager considered the evaluation as biased and of bad quality, which is why it failed to act as a trigger for a change.

On the whole, the evidence supports the assumption that an evaluation must be conceived good quality in order to trigger change in a consensual context with high pressures for change and that novelty is less an issue.

Evaluation as a Referee (High Pressure, High Conflict)

In a conflict-laden, high-pressure environment, it has been assumed that neither novelty nor quality is necessary for evaluations to be used as a referee to decide what to change. However, the conclusions of the evaluations are likely to be accepted only by one part of the stakeholders.

The three cases B, D, and H, which fall in this context, all led to decisions for substantial change in the evaluated programs. They differ with respect to their quality and novelty value, so the assumption that these actor perceptions are less important than in other contexts is confirmed. At a closer look, however, it appears that the quality is important if the evaluations challenge existing beliefs.

In case H, the evaluation showed that the development fund, which was the object of evaluation, had to change its strategy, because there were too few requests for financial assistance. The evaluation suggested that the fund start to develop its own projects instead of waiting for demands for financial assistance from other organizations. The evaluation confirmed the desk manager's opinion, but challenged the position of another involved Swiss federal agency, with which the SDC desk manager had been in conflict for a long time. The evaluation results also contradicted some of the local stakeholders, all of whom were part of the decision committee of the fund, which had to agree to the strategic change. So, even though the results were not new to the desk manager who took them as a basis to advocate a strategic change (positive outcome), they were new to some of the other decision makers. The latter were finally convinced by the evaluation not least thanks to its very good quality. The evaluation provided a good description of the procedure and the criteria.

Together with case H, evaluation B is the most sophisticated in the sample. The desk manager was totally convinced of the program, but there had been severe internal conflicts about its appropriateness for a development agency like SDC. The evaluation showed that the program was effective, and given good evaluation quality, internal criticisms toward the program ebbed down. The desk manager was supported in her decision to expand the program. In one respect, however, the desk manager's position was challenged as the evaluation showed that one of the project organizations was very expensive compared to others. The desk manager considered the evidence for this point to be sound and, as a consequence, drove down cooperation with this organization. Here again, good quality was necessary for the desk manager to take the evaluation as a basis for the change decision.

Evaluation D confirmed the desk manager's impression that the local project organization failed to implement an important part of the planned activities, which had been a cause of conflict throughout program implementation. As a consequence of the evaluation, the desk manager replaced the organization. The desk manager assessed evaluation quality rather negatively in the closed questions, but emphasized several times that she did not much care about it. It seems that evaluation quality did not matter to the desk manager, because the results were as she had expected.

The discussion of these three cases shows that in the context of high conflict and high pressure for change, evaluation users decided according to the truth test (Weiss & Bucuvalas, 1980): High evaluation quality is only necessary if an evaluation challenges preexisting beliefs of decision makers. Users were only willing to revise their opinion if they were convinced by the quality of the results.

Conversely, if evaluation results confirm decision makers opinions, quality is less an issue. None of the two conditions seems, however, always necessary.

Evaluation as a Conciliator (Low Pressure, High Conflict)

According to the hypothesis, in a situation of conflict, where stakeholders are not much aware of a need for change, substantial change decisions are only taken if an evaluation is regarded good quality and shows new ways out of disagreement. Situations where an evaluation acts as a conciliator are deemed to be rare.

Both evaluations in the sample, undertaken in a conflict-laden environment without pressure for change (cases A and I), were not used for a change decision, so strictly speaking it is not possible to assess the necessity claims.

In case I, there was a conflict between two sections within SDC about the aims of the evaluation. The role of the evaluators was not clear from the start. After several attempts to rewrite the intermediate evaluation report in order to suit both sections, the evaluation was stopped midway. Conflict was all-pervasive and determined desk managers' perceptions of evaluation quality and its novelty value; the involved desk managers were unable to give a differentiated judgment, so that the measurement of these two conditions is questionable. The desk managers were totally unwilling to take decisions upon the results of the evaluation in the intermediate report. Overall, the evaluation acted as a trigger for open conflict rather than as a conciliator.

In case A, there was clear mistrust between SDC and the project organization. Against the desk manager's expectations, SDC rather than the project organization was criticized by the evaluation. The desk manager considered the evaluation of good quality and accepted the critique, but did not act on it, because the decision to terminate the project had already been taken. The aim of the evaluation was merely to draw the lessons learnt from the project, but this did not have an impact on the decisions by the desk manager or SDC in general (at least in the period of study). The conditions that have been claimed necessary for an evaluation to act as a conciliator were fulfilled, but the evaluation was not used to for decision making because the decision had already been taken. Otherwise, it might have been used.

In the absence of cases with a positive outcome, it is not really possible to test the assumption that an evaluation must show something new and be considered of good quality to motivate a change decision in a conflict-laden environment, where the pressure for change is low. But case A points into this direction.

Discussion

The inherent complexity of the phenomenon of use is an important reason why research results on evaluation use have largely remained inconclusive so far. In order to address this complexity, I adopted three strategies: first, I built on recent conceptual contributions to confine the analyzed type of "use," namely evaluation-informed decisions to substantially change the evaluated program or project as one specific type of instrumental use. Change decisions imply that much is at stake, so it can be assumed that the ingredients that are necessary for such an important evaluation-based decision appear more clearly than in an analysis of minor decisions.

Second, I did not try to predict such change decisions. Rather, I examined how the decision makers (in this case, the desk managers in charge of the program or project) must perceive the evaluation so that they might take such a decision. In other words, I analyzed necessary conditions. Third, I presumed that the necessary conditions depend on the context, because depending on the context it is other mechanisms that bring about a change decision (cf. Astbury & Leeuw, 2010). Because of context dependency, this study has adopted a conjunctural approach to both hypothesis formulation and testing. According to this approach, it is combinations of factors instead of single conditions that cause an outcome.

Table 5. Results About Necessary Actor Perceptions for Evaluation Use in Different Contexts

Context Conditions			Necessary Actor Conditions	
Pressure for Change	Level of Conflict	Mechanism	Novelty Value	Evaluation Quality
Low	Low	Awakener	High	High
High	Low	Trigger	Irrelevant	High
High	High	*Endorser*	*Low*	*Irrelevant*
		Reviser	*High*	*High*
Low	High	Conciliator	?	?

Note: Italics indicates refinement of hypothesis; ? indicates necessity test not possible.

The empirical part of this article is exploratory in character. A total of 11 external evaluations on projects or programs funded by SDC have been used for a comparative case analysis. Decision making by the desk managers who commissioned the evaluations has largely followed the assumed patterns. Depending on the level of conflict and on existing pressures for change, evaluation information has been used differently, and it is different conditions that have been necessary to trigger the different mechanisms of use.

Table 5 summarizes the detailed results of the hypothesis test for the four context constellations that have been examined. For the first two context constellations, the cases confirm the assumptions that have been presented in Table 1: In a consensual environment with low pressure for change, evaluations can act as awakeners, if they disclose something unknown and are considered good quality. In high-pressure consensual situations, evaluations can function as a trigger for change decisions that are accepted as indispensable, but only provided that evaluation quality is deemed to be good. In contrast, novelty is not a precondition.

For the third context of high conflict and high pressure for change, the hypotheses can be refined based on the case studies. The evaluation does not simply act as a referee between different positions; rather the mechanism appears to depend on the novelty value. If the evaluation confirms the decision maker's expectations and does not provide much new information, it can be used to endorse one's decision as to how to change the evaluated program or project. There was one case in the sample which suggests that the quality of the evaluation does not matter much in this situation. In contrast, if the evaluation does not confirm the decision maker's expectations, it might still be used as a basis for a change decision, if the decision maker is persuaded that she or he had been wrong. In order for decision makers to revise their opinion, it is necessary that they consider the evaluation of high quality. This means that in this situation of high conflict and high pressure for change, decision makers actually performed a "truth test" in the way as described by Weiss and Bucuvalas (1980), where novelty and evaluation quality are interdependent, and none of the two alone is necessary.

Finally, the extent to which novelty and evaluation quality are necessary for an evaluation to act as a conciliator in a situation of conflict and low pressure could not be tested because the two cases in the sample that fall into this context were not used as a basis for a change decision. Given that, according to Valovirta (2002), it is rare that evaluations are actually used in such contentious situation where change is not pressing, this lack of an empirical case does not come as a surprise.

Based on the results of this article, certain recent findings about evaluation use can be somewhat differentiated. In a recent survey among American evaluators (Fleischer & Christie, 2009), beliefs and values of key stakeholders have been considered one of the main barriers to evaluation use. According to the above results, this is mostly true: if beliefs and values of decision makers are challenged by new evaluation information, the hurdle for its use for decision making is higher, as the evaluation needs to be considered of good quality, too. However, there are contexts where the

novelty value of an evaluation does not seem to be a relevant barrier to change decisions, namely where the pressure for change is high and stakeholders get along well with each other.

According to this study, evaluation quality appears to matter in most contexts, whereas in a recent review of empirical studies about use (Johnson et al., 2009) and in the above-cited survey among American evaluators (Fleischer & Christie, 2009), high standards of methodological rigor figured among the less important factors for use. Evaluation quality in the present article has been measured as perceived by the decision makers (the desk managers). The same external evaluations have been part of a meta-analysis that was made in parallel to the present study and where evaluation quality was assessed by professional evaluators according to evaluation standards similar to those of the Joint Committee (1994). In several cases, the desk managers' quality ratings were much higher than those of the meta-analysis, where many of the evaluations were considered of rather poor quality. The evaluations suffered from several shortcomings that according to Thomas (2010) are common in development evaluation. In fact, quality as rated by the meta-analysis is not a necessary condition for evaluation-based change decisions in any of the contexts. This underlines that, as Patton (1997) puts it, use depends on evaluation quality as "matters of subjective user judgment" rather than on preset standards.

Limitations of the Study

At a conceptual level, a first limitation follows from the focus on evaluation-based change decisions, which does not cover all forms of instrumental use. Conditions that appear necessary for the examined outcome of evaluation-based decisions to substantially change the program or project might be less so for minor change decisions or decisions to continue the program or project, because the hurdle is lower, but there is no reason to believe that results would be completely different. As an additional conceptual limitation, this contribution examines only four conditions, leaving aside many others. It is well possible that in certain contexts, other conditions that have been neglected due to a focus on a set of specific hypotheses are more relevant to evaluation-based change decisions than the ones examined. Even so, the selected factors are theoretically grounded and certainly not trivial necessary conditions that can always be taken for granted; the cases provide evidence that the perceived novelty value and quality of evaluations as well as the decision setting vary considerably. In tracing the mechanisms behind the change decisions, I am quite confident that the examined conditions are in fact relevant necessary preconditions for use, but they are indeed unlikely to be sufficient. So other conditions must be fulfilled too. One aspect that has repeatedly been claimed to be (most) central is stakeholder involvement (cf. recent studies like Fleischer & Christie, 2009; Johnson et al., 2009; Skolits, Morrow, & Burr, 2009). The key stakeholders in the present study are the desk managers. Their participation in the evaluation has been measured indirectly: it can be observed that the more they have been involved in the evaluations, the better their assessment of evaluation quality thanks to a better understanding of the evaluation process. Local stakeholder participation in the evaluation at the program or project site has not been included in the analysis, because it did not seem to affect decision making by the desk manager at the head office in Switzerland.

This study is based on theoretical claims that are context-bound, so that it can only lead to "contingent generalizations" (George & Bennett, 2005) that will remain restricted in scope. Given the explorative nature of the empirical part of the present study, there are further important limitations, in particular with respect to external validity. First, SDC has a particular evaluation tradition and a particular decision-making setting which differs from others. For internal validity, the fact that all cases have been chosen from the same organizational setting was important, given that organizational conditions were not included in the analysis. Furthermore, the SDC decision-making setting where responsibility for decision making relies on one person, the desk manager, has facilitated data collection. At the same time, this means that results cannot be simply generalized to other contexts.

Second, the study is based on 11 cases. The author was lucky that there were cases for each of the four context constellations of interest, even though cases with a positive outcome were missing for one of the constellations, so that the respective hypotheses could not really be tested. In the other three contexts, the number of positive cases lies between just one and three. The causal relationships that have been claimed are based on an analysis of the mechanisms underlying the processes that led to evaluation-based change decisions. So it is not just correlations based on a small number of cases. However, even in a qualitative case comparison, a higher number of cases is desirable because it helps to refine our understanding of the mechanisms that produce the outcome and of the conditions that are necessary to trigger these mechanisms. At the same time, it is indispensible to have an intimate knowledge of each of the cases to reveal the mechanisms in the specific context; mechanisms are hard to identify because they are usually hidden and context dependent (Astbury & Leeuw, 2010). So there is always a trade-off between a higher number of cases and a better knowledge of each case. With 11 cases, the present article lies well beyond the mean of 3.4 cases that are usually analyzed in studies about evaluation use based on case study method (Brandon & Singh, 2009).

As to internal validity, the study relies mainly on interview information from the desk managers that commissioned the analyzed external evaluations and were supposed to make use of them. There is a risk that desk managers overstated the quality of the evaluations they commissioned. I tried to deal with this in checking different dimensions of evaluation quality and in asking for the reasons of their quality ratings. The desk managers were interviewed about half a year after the end of the evaluation. Certain measures were taken to reduce the risk that they could not remember the evaluation. Before the interview, desk managers were prompted to provide certain documents, in order to get them to deal with the evaluation once again. Furthermore, interview information, namely information about change decisions, was cross-checked with document information. Interviews with evaluators provided supplementary information. In most cases, the data were rich, but in two or three cases where there had been a change in the desk manager during the evaluation or just after its completion, there remained certain gaps in the process reconstruction. However, these gaps are likely to exist not just with respect to this analysis but also in reality in the evaluation process and in the process of evaluation use.

Conclusions

Awareness for the complexity of the phenomenon of evaluation use has been growing in parallel with the body of research. This article has applied multiple strategies to handle this complexity: the form of use examined is clearly delineated; theoretical claims and empirical analysis are context-bound; a conjunctural approach is adopted in hypothesis formulation and in the QCA analysis; the study merely wants to find the necessary preconditions for change decisions in specific contexts and does not try to fully explain their occurrence or nonoccurrence. The explorative empirical analysis suggests that this combination of strategies is likely to pay off. The results suggest that even the understanding of well-known mechanisms can be refined: the "truth test," for instance, could only be observed under one specific circumstance.

Depending on the context, the perceived novelty value and quality of an evaluation seem to matter more or less. This is likely to be true for most of the factors that are related to evaluation use and has implications for research on this subject. It is high time that we do not just control for contextual factors but make context explicit. We need adequate theoretical models that depict interdependencies between causal mechanisms (or factors) and contexts. The realist methodology by Pawson and Tilley (1997) gives some guidance. The disadvantage is that even if we only examine a small number of factors (for example, four conditions as in the present study), theoretical models can get quite complicated. However, we might have to accept this higher degree of theoretical complexity if we strive for a better understanding of our complex reality.

Furthermore, we need adequate methods to test the complex theoretical models. QCA is one possibility. Recent developments of the method, namely "fuzzy-set QCA" (Ragin, 2000, 2005, 2008; Schneider & Wagemann, 2007), avoid dichotomization and accommodate more easily a higher number of cases, even though this goes to some extent at the cost of losing the direct link between the data and the cases, which is one of the strengths of the crisp-set application of QCA presented in this article. Another possibility would be a mixed-method design (e.g., Bergman, 2008), for instance, combining a statistical analysis of survey data with in-depth case studies. It is, however, important to pay attention to the "compatibility" of theoretical claims and method; in contrast to QCA, statistical methods do not allow for a test of necessity claims. Necessity claims are, however, very valuable to advance research on phenomena like evaluation use that depend on multiple and often unpredictable elements, which can never be fully accounted for. A possible way forward for research on evaluation use is a focus on mechanisms that lead to well-circumscribed forms of use and on the necessary ingredients to trigger these mechanisms in different contexts. Such knowledge about context-specific necessary conditions for different kinds of evaluation use is also valuable for evaluation practitioners seeking advice about how to promote the utilization of their findings.

Author's Note

An earlier version of this article has been presented at the ECPR Joint Sessions in Rennes, France, April 11–16, 2008.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Weiss and Bucuvalas were surprised to find that in addition to a study's consistency with users' knowledge and values (truth test), consistency with institutional norms (utility test) emerged as a separate dimension in factor analysis. In the present study, the two dimensions are taken together, because they could not empirically be isolated. The resulting "novelty value" factor shows whether the evaluation revealed something new or unexpected to the users that challenged either their personal beliefs or their institutional norms or both.

References

Alkin, M. (1985). *A guide for evaluation decision makers.* Beverly Hills, CA: Sage.

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations. Does evaluation make a difference?* Beverly Hills, CA: Sage.

Astbury, B., & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*, 363-381.

Amenta, E., & Poulsen, J. D. (1994). Where to begin: A survey of five approaches to selecting independent variables for Qualitative Comparative Analysis. *Sociological Methods & Research, 23*, 22-53.

Balthasar, A. (2006). The effects of the institutional design on the utilization of evaluation. *Evaluation*, *12*, 353-371.

Balthasar, A. (2007). *Institutionelle Verankerung und Verwendung von Evaluationen [Institutional anchorage and utilization of evaluations].* Chur/Zürich: Rüegger.

Befani, B., Ledermann, S., & Sager, F. (2007). Realistic evaluation and QCA: Conceptual parallels and an empirical application. *Evaluation*, *13*, 171-192.

Berg-Schlosser, D., & De Meur, G. (2009). Comparative research design. Case and variable selection. In B. Rihoux, & C. C. Ragin (Eds.), *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques* (pp. 19-32). Los Angeles, CA, London, New Delhi, and Singapore: Sage.

Bergman, M. M. (Ed.). (2008). *Advances in mixed methods research.* London, Thousand Oaks, CA, New Delhi, and Singapore: Sage.

Brady, H. E., & Collier, D. (Eds.). (2004). *Rethinking social inquiry.* Lanham, MD: Rowman & Littlefield.

Brandon, P. R., & Singh, J. M. (2009). The strength of the methodological warrants for the findings of research on program evaluation use. *American Journal of Evaluation*, *30*, 123-157.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research*, *56*, 331-364.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass.

De Meur, G., & Rihoux, B. (2002). *L'analyse quali-quantitative comparée (AQQC-QCA): Approche, techniques et applications en sciences humaines [Qualitative Comparative Analysis (QCA): Approach, techniques and applications in the humanities].* Louvain-la-Neuve: Bruylant-Academia.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American evaluation association members. *American Journal of Evaluation*, *30*, 158-175.

Frey, K., & Ledermann, S. (2010). Evidence-based policy: A concept in geographical and substantive expansion – Introduction to a special issue. *German Policy Studies*, *6*, 1-15.

Geertz, C. (1973). *The interpretation of cultures: Selected essays.* New York, NY: Basic Books.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences.* Cambridge: MIT Press.

Goertz, G. (2006a). Assessing the trivialness, relevance, and relative importance of necessary or sufficient conditions in social science. *Studies in Comparative International Development*, *41*, 88-109.

Goertz, G. (2006b). *Social science concepts: A user's guide.* Princeton, NJ and Oxford: Princeton University Press.

Goertz, G., & Starr, H. (Eds.). (2003). *Necessary conditions: Theory, methodology, and applications.* Lanham, MD: Rowman & Littlefield.

Grofman, B., & Schneider, C. Q. (2009). An introduction to crisp set QCA, with a comparison to binary logistic regression. *Political Research Quarterly*, *62*, 662-672.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation*, *24*, 293-314.

Jewell, C., & Bero, L. (2007). Public participation and claimsmaking: Evidence utilization and divergent policy frames in California's ergonomics rulemaking. *Journal of Public Administration Research and Theory*, *17*, 625-650.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*, 377-410.

Joint Committee. (1994). *The program evaluation standards.* Thousand Oaks, CA: Sage.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation, 2000*, 5-23.

Lijphart, A. (1975). The comparable-cases strategy in comparative research. *Comparative Political Studies*, *8*, 158-177.

Mahoney, J., & Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis*, *14*, 227-249.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation*, *10*, 35-57.

Mayring, P. (2003). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (8th ed.). Weinheim and Basel: Beltz.

Merkens, H. (2003). Auswahlverfahren, Sampling, Fallkonstruktion [Selection strategies, sampling, case construction]. In U. Flick, E. von Kardorff, & I. Steinke (Eds.), *Qualitative Forschung: Ein Handbuch* [Qualitative research: A handbook] (2nd ed., pp. 286299). Reinbeck bei Hamburg: Rowohlt.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA, London, and New Delhi: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London, Thousand Oaks, CA, and New Delhi: Sage.

Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry.* New York, NY: Wiley-Interscience.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies.* Berkeley, CA, Los Angeles, CA, and London: University of California Press.

Ragin, C. C. (1994). *Constructing social research: The unity and diversity of method.* Thousand Oaks, CA: Pine Forge Press.

Ragin, C. C. (2000). *Fuzzy-set social science.* Chicago, IL and London: The University of Chicago Press.

Ragin, C. C. (2005). From fuzzy sets to crisp truth tables. *Compasss Working Paper* No. 28.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond.* Chicago, IL and London: Chicago University Press.

Rihoux, B. (2006). Qualitative comparative analysis (QCA) and related systematic comparative methods: Recent advances and remaining challenges for social science research. *International Sociology*, *21*, 679-706.

Rihoux, B., & Lobe, B. (2009). The case for qualitative comparative analysis (QCA): Adding leverage for thick cross-case comparison. In D. Byrne, & C. C. Ragin (Eds.), *The SAGE handbook of case-based methods* (pp. 222-242). London, Thousand Oaks, CA, New Delhi, and Singapore: Sage.

Rihoux, B., & Ragin, C. C. (2009). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques.* Los Angeles, CA, London, New Delhi, and Singapore: Sage.

Sager, F., & Ledermann, S. (2008). Valorisierung von Politikberatung [Valorizing Policy Advice]. In S. Bröchler, & R. Schützeichel (Eds.), *Politikberatung [Policy Advice]* (pp. 310-325). Stuttgart: UTB.

Schneider, C. Q., & Wagemann, C. (2007). *Qualitative Comparative Analysis (QCA) und Fuzzy Sets: Ein Lehrbuch für Anwender und jene, die es werden wollen [Qualitative comparative analysis (QCA) and fuzzy sets: An course book for current and aspiring users].* Opladen and Farmington Hills, MI: Barbara Budrich.

SDC. (2002). *Ongoing evaluation programme for 2002-2003 of SDC.* SDC: Berne.

SDC. (2004). *10 principles to ensure successful use of evaluations*. SDC: Berne. Retrieved May 1, 2011, from http://www.deza.admin.ch

Skolits, G. J., Morrow, J. A., & Burr, E. M. (2009). Reconceptualizing evaluator roles. *American Journal of Evaluation*, *30*, 275-295.

Spinatsch, M. (2002). Evaluation in Switzerland: Moving toward a decentralized system. In J. E. Furubo, R. C. Rist,, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 375-391). New Brunswick, NJ and London: Transaction.

Thomas, V. (2010). Evaluation systems, ethics, and development evaluation. *American Journal of Evaluation*, *31*, 540-548.

Valovirta, V. (2002). Evaluation utilization as argumentation. *Evaluation*, *8*, 60-80.

Weiss, C. H. (1972). Utilization of evaluation: Toward comparative study. In C. H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education.* Boston, MA: Allyn & Bacon.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation*, *19*, 21-33.

Weiss, C. H., & Bucuvalas, M. J. (1980). Truth tests and utility tests: Decision-makers' frames of reference for social science research. *American Sociological Review*, *45*, 302-313.

Weiss, C. H., Murphy-Graham, E., & Birkeland, S. (2005). An alternate route to policy influence: How evaluations affect D.A.R.E. *American Journal of Evaluation*, *26*, 12-30.

Yamasaki, S., & Rihoux, B. (2009). A commented review of applications. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods: Qualitative Comparative Analysis (QCA) and related techniques* (pp. 123-145). Los Angeles, London, New Delhi, Singapore: Sage.