# Explaining Evaluation Use

## A Qualitative Comparative Analysis of Factors Influencing Instrumental Use of Evaluations


**Research Master Political Science and Public Administration, Leiden University**

**Track: Political Science**

**Master thesis**

**Student: Marjolein Bouterse**

**Student number: 0900532**

**Supervisor: V.E. Pattyn**

**Second reader: N.J.G. van Willigen**


**Leiden, November 2016**

# Table of Contents

# Foreword

This thesis was written as the final project of the Research Master of Political Science and Public Administration at Leiden University. It marks the end of journey that lasted more than seven years. On this journey I have gained knowledge, skills, experience, and confidence through all the courses I took and activities I partook in. I am indebted for this to far more people than I could name on this page. For now, I will keep my thanks limited to those who supported and inspired me in this last project: the thesis. First and for most, two people need to be mentioned: Valérie Pattyn, my supervisor at Leiden University, and Wendy Asbeek Brusse, my supervisor at the Ministry of Foreign Affairs. They both encouraged me to work hard, be critical, and immerse myself in a subject I had little knowledge of before I started. And of course, they shared with me their valuable knowledge on and insights into evaluations and research designs. I am grateful to Professor Sandra Groeneveld, who first put me on the track of the IOB, for her effort in helping me obtain such a wonderful place to conduct my research. Furthermore, I am grateful to all working at the IOB for the time they gave to answer all my questions and for their useful comments on my research design. Lastly, I wish to thank all the respondents, who took the time to answer my questions and were so open and frank about their use of the evaluations.

Leiden, November 2016

Marjolein Bouterse

# Abstract

*This research contributes to the literature on the use of evaluation. Through the use of QCA it shows that the factors timing of the evaluation and interest shown by the policymaker(s) are necessary factors for instrumental use of evaluations. Three factors together are sufficient for instrumental use to occur: the timing of the evaluation, interest shown by the policymaker(s), and the inclusion of novel knowledge to the policymaker(s) in the evaluation. The research adds to the existing literature in two ways: First, it confirms that interest and timing are important. Moreover, in contrast to earlier studies it also shows when the timing is right: when the process of policy formulation and the process of evaluation run parallel. Second, the analysis accounts for causal complexity and equifinality, by allowing for multiple causal combinations of factors leading to the outcome and by showing in what way factors interact to affect the use of evaluations.*

# 1. Introduction[1]

When evaluators conduct policy evaluations, their goal often is to facilitate learning about policy. It should come as no surprise that the study of the usages of evaluations is a major field within the scientific study on evaluations (Henry and Mark, 2003: 293; Kirkhart, 2000: 5). Studying under what conditions learning can be promoted is not only valuable for scholarly purposes, but also for practitioners of monitoring and evaluation. Research into evaluations can help them improve the impact of their work. Studies on factors influencing the use of evaluations have been helpful in guiding evaluators and policymakers to adjust their processes to promote the use of the evaluations.

Although a lot of studies focussing on influencing factors have been published, the research field still struggles with two structural problems. First, evidence is often of anecdotal nature and it is unclear which factors influence what kind of use. Moreover, there is little variation in the methods used to study this. Most studies are qualitative case studies based on interviews and a document analysis. This limits the transferability of the findings from one context to another. The study of evaluation use needs to find more rigorous and structured methods to go beyond anecdotal evidence.

Second, although many case studies have been conducted, there is yet no clear perception of which factors are more important than others and how factors found in one study can be translated to other contexts. So far, many different factors influencing use have been found in the literature. These can be grouped together in four broad categories: first, personal characteristics of both the evaluator as well as (intended) user (Caplan, 1977; Patton et al., 1977); second, the political and administrative context (Balthasar, 2006; Ledermann, 2012; Patton et al. 1977); third, user involvement in the evaluation process (Leviton and Hughes, 1981; Shulha and Cousins, 1997); and fourth, the process of the evaluation and the (perceived) quality of the evaluation itself (Balthasar, 2006; Caplan, 1977; Ledermann, 2012). Although many factors have been identified, there is little understanding of the relation between the factors and how they interact in their effect on the use of evaluations (Johnson, 2009: 388). Also, it is often unclear which factors influence which type of use. The literature on evaluation use has identified many different types of use, but most studies are not very articulate about the specific type they study or how exactly they operationalize the type they study.

Thus, the main research question this thesis will focus on is: *What conditions influence the use of evaluations for policy improvement in the bureaucracy at the national level?*

[1] In order to comply with standards of good and transparent research, a replication document has been made, which can be requested from the author (marjoleinbouterse@gmail.com).

The study was conducted at the Policies and Operations Evaluation Department (IOB) of the Dutch Ministry of Foreign Affairs. Eighteen studies completed between 2013 and 2016 will be used in a qualitative comparative analysis (QCA). This method is appropriate to study complex causality and estimate what the effect on use is of combinations of factors.[2] Furthermore, it presents to the researcher a very systematic way to study conditions and forces her to be transparent about the choices and operationalizations made in the design.

Before turning to the theoretical framework of this study, some clarification is necessary on two points: the definition of an evaluation and who is considered its user. First, the definition of an evaluation that will be used is derived from the DAC (OESO).[3] According to this definition an evaluation is 'an assessment, as systematic and objective as possible, of an on-going or completed project, programme or policy, its design, implementation and results. The aim is to determine the relevance and fulfilment of objectives, developmental efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors.'[4]

Second, users can be many different people, for instance, policymakers, politicians, lobbyist, media, scientists, NGOs, etc. However, in this thesis only those who influence the formulation of national government policy are meant with the term 'users' or 'potential users'. This is a useful group to study because of the context they work in. Formulating national policy is a highly political endeavour which makes the use of evaluations difficult. This is especially the case for instrumental use (direct policy improvement), as in the end it are the politicians who determine what the policy will be. If there is no political will, or if policymakers perceive an absence of political will, recommendations and lessons from evaluations could easily be ignored. This study focuses specifically on this context and aims to show how evaluations can still be used for policy improvement.

[2] The notion of causal complexity entails that not one specific factor leads to an outcome, but multiple factors or even multiple combinations of factors do (Ragin, 1999).

[3] The definition is originally meant for organizations operating in the field of development; however it is easily transferred to other fields of policy. Although the terms 'recipients' and 'donors' refer specifically to actors in the development field, policies usually have both a provider of the money (e.g. Parliament) and a recipient of the money (e.g. the Minister or policy department).

[4] Development Assistance Committee (1991) *Principles for the evaluation of development assistance* Paris.

# 2. Theoretical framework

This section further highlights the literature on the use of evaluations and explains the choices that are made in this research.

## 2.1 Types of use

Evaluations have several goals; among other things, they can be meant for learning, policy improvement, legitimizing, accountability, and empowerment. It makes sense that, if evaluations have different goals, they will also have different uses. Table 1 shows the different types of uses named within the literature, of which instrumental, conceptual, and symbolic use are the most commonly distinguished. All three capture a different kind of goal for which an evaluation is used: to improve policy, to understand policy (effects), or to legitimize choices. This research concerns instrumental use, which will be explained into more depth in the following section. The section thereafter will show how use is related to learning; and the third will discuss the level on which use takes place. The last section will discuss other dimensions that have been employed to categorize uses, like timing, intent, and source.

### Instrumental use

Instrumental use has policy improvement as its goal; however policy improvement cannot be the definition of instrumental use. If policymakers make decisions on the basis of the evaluation, with the aim to improve, the evaluation is also used; even if, in the end, the policy might not actually have been improved. Therefore, instrumental use is defined as the use of the evaluation in order to inform policy decision making (Alkin and Taut, 2003; Ledermann, 2012). This includes substantive changes made to the policy, termination of a project or program, and changes in funding. Some authors specifically include that a change needs to be made to the policy (e.g. Ledermann, 2012: 160), while others only name that decisions need to be influenced (e.g. Alkin and Taut, 2003: 5). In the latter case, the decision to continue something would also count as instrumental use. This thesis adheres to the latter approach: there does not have to be an actual change to the policy, but the knowledge gained from the evaluation should inform decision making about the policy. This knowledge does not have to be novel for the policymaker. It is possible that the knowledge was already known, but that the evaluation causes the policymaker to take action upon the knowledge.

**Table 1: Types of evaluation use**

| Use | Definition | Based on | Learning | Intention | Timing | Source |
|---|---|---|---|---|---|---|
| **Instrumental use** | When an evaluation directly informs policy decision making | Ledermann (2012), Alkin and Taut (2003) | Learning might take place (mostly changes within knowledge structures) | Yes | Most likely during and immediately after the evaluation, but possible also much later (especially when there is a written report that becomes public). | Both findings and process use |
| **Conceptual use** | When an evaluation changes the understanding of concepts, underlying assumptions of policies, priorities, goals, etc. (in short, influences a policymaker's thinking about an issue) without directly informing policy decisions | Ledermann (2012), Alkin and Taut (2003) | learning takes place (mostly changes of knowledge structures) | Possible, but not necessarily | Possible during, soon after, but also much longer after the evaluation. | Both findings and process use |
| **Symbolic use** | When an evaluation is used to gain support or defend/legitimize an already-held opinion, or in order to display that the policymaker(s) were prepared to have their policy evaluated | Ledermann (2012), Alkin and Taut (2003) | No learning takes place | Yes | Most likely soon after the evaluation, when it has the most saying power. Less likely but possible during or long after the evaluation. | Findings use |
| **Tactical use** | When an evalution is used to gain time of avoid responsibility | Vedung in Widmer and Neuenschwander (2004) (Often included in symbolic use) | No learning takes place | Yes | Most likely soon after the evaluation, when it has the most saying power. Less likely but possible during or long after the evaluation. | Findings use |
| **Legitimizing use** | When an evaluation is used to legitimize decisions and resolutions | Vedung in Widmer and Neuenschwander (2004) (Often included in symbolic use) | No learning takes place | Yes | Most likely soon after the evaluation, when it has the most saying power. Less likely but possible during or long after the evaluation. | Findings use |
| **Accountability** | When an evaluation is used to account for how money was spend and whether goals were (not) attained | Azzam and Levine (2015), Sanderson (2002) | No learning takes place | Yes | Most likely immediately after the evaluation has been conducted. | Findings use |
| **Enlightment use** | When people and institutions, that were not included as the primary target group of the evaluation are influenced by the evaluation | Weiss (1998) (Sometimes used interchangably with conceptual use (e.g. Marra, 2004: 264; Widmer and Neuenschwander, 2004: 392)) | Learning might take place (changes possible both within and of knowledge structures) | Possible, but not necessarily | Possible soon after, but also much longer after the evaluation. | Findings use |
| **Empowerment use** | When the evaluation helps people to change their work, lives or ideas; or when the evaluation helps them to address social problems they are facing. | Fetterman (1994) | Learning might take place (changes possible both within and of knowledge structures) | Possible, but not necessarily | Most likely during and soon after the evaluation, but also possible longer after the evaluation | Process use mostly, but findings use is also possible |

It is not always easy to distinguish between instrumental and symbolic use. Symbolic use[5] occurs when an evaluation is used to gain support for or legitimize an already-held opinion, or in order to display that the policymaker(s) were prepared to have their policy evaluated (Alkin and Taut, 2003; Ledermann, 2012). When a policymaker explains that he was strengthened in a decision and they will continue on the same path, it might be either symbolic or instrumental use. The main difference between the two types lies in the sincerity of the usage. A policymaker that uses an evaluation in a symbolic way is not interested in the evaluation as such, but only to the extent that it can legitimize or defend his or her viewpoints and actions (Widmer and Neuenschwander, 2004: 392).

The types of uses are not mutually exclusive; more than one type of use can originate from one evaluation. One type of use can even be the basis of another type of use (Leviton and Hughes, 1981: 527). For instance, a policymaker can become aware of an underlying assumption of the policy he or she had not realized before (conceptual use) and then decide to change the policy to address that assumption (instrumental use).

### Learning

Often, it is supposed that evaluations contribute to learning. However, whether evaluations do actually contribute, is not always explicit. Some authors hold the view that policy improvement (instrumental use) always implies learning (e.g. Sanderson, 2002), and others that learning is the same as conceptual use (e.g. Henry and Mark, 2003). Furthermore, learning is also equalled to process use, a concept that will be discussed shortly (e.g. Alkin and Taut, 2003; Preskill and Torres, 2000).

In order to see the position of learning in evaluation use more clearly, it will be helpful to define the concept of learning itself. Learning assumes a change in or of knowledge structures on the basis of new information (Forss, Cracknell, and Samset, 1994: 574). That there is such a change in or of knowledge structured is clear for conceptual use. Per definition conceptual use includes learning, because it changes the *understanding* about an issue of the user. Instrumental use can, but does not have to, involve learning. If instrumental use happens, people start acting on the basis of information they receive from the evaluation. Often, something will have been changed in the knowledge structures to make them act. Even if the information itself is not new, then their interpretation of the information might have changed.[6] For example, reading information again has made the user realize the urgency of a

[5] As can be seen in appendix 1, symbolic use usually includes both legitimizing use and tactical use.

[6] Knowledge and information are different concepts. Information can be defined as an ordered collection of data and knowledge as 'interpreted information' (Alkin and Taut, 2003: 2). Thus, information does not have to be

problem or gave him or her a clue towards a solution. If there was no such change at all, the use might be of a symbolic instead of an instrumental nature. However, one can think of instances of instrumental use where no learning has taken place. For instance when the evaluation results helped a policymaker to defend his case before the management. He might not have learned anything new, but he could still use the evaluation to improve the policy.

### Level of use

So far, use has been defined on the individual level, yet it makes sense to view evaluation use as an organizational process. First, national policy is almost always made by more than one policymaker, sometimes several departments or ministries are involved in the process (Weiss, 1998). Thus, an individual policymaker cannot decide to use an evaluation in an instrumental way by him or herself, as it will necessarily include policy choices that must be approved by others. For instrumental use to happen the knowledge needs to be accepted within a broader segment of the organization. Second, knowledge can become institutionalized among policymakers. According to Radealli (1995: 178) institutionalization gives "…stability to shared causal beliefs, they set up structures of meaning, they create networks of actors, they constrain the perception of interests and of socio-economic change". The institutionalization of knowledge on the organizational level can be compared to the knowledge structures of Forss, Cracknell, and Samset (1994). In order for change to happen, not only the knowledge structures of individuals need to change, but also those of the organization. In this sense, conceptual use can also happen at the organizational level. The third reason why it is it important to look at levels of use, is that a focus on the organizational level highlights the complex interdependence and difficulty of learning (Freeman, 2009: 8) and focusses on the context in which learning is supposed to take place.

### Categorizing use

The types of use as discussed above can be categorized on several dimensions. Kirkhart (2000) names intention, source, and time as the key dimensions for studying use. East dimension is shortly explained.

The dimension timing is the most straightforward: it simply indicates when the use has taken place. This can happen immediately, even during the process of the evaluation; (shortly) after the evaluation is finished; and a longer period of time after the evaluation. 'Long-term' can be half a year after the evaluation, but could possibly also be years after.

---

new, but if is it is presented differently or in combination with other information, the knowledge it brings might be new.

**Table 2. Categorization of use**

| *Timing of use* | *Intention of use* | *Sources of use* |
|---|---|---|
| Immediate | Intended | Process |
| End-of-cycle | Unintended & aware[7] | Findings |
| Long-term | Unintended & unaware | |

Intention refers to the purpose users and the evaluators had in mind when conducting the evaluation (although they might not have made this explicit). Intended use happens if the evaluation is used in the same manner and by the same actor that the evaluators had in mind. Unintended use occurs if the evaluation was used in a manner and/or by users they did not have in mind. Symbolic use will usually not be intended use, and enlightment use is per definition unintended use. However, if the evaluators realized the possibility of this type of use happening, both might fall in the category unintended and aware.

The last dimension, source of use, distinguishes two different sources: process and findings. This dimension highlights that the findings are not the only (possible) influence of an evaluation on policymakers. They can also base their use on their involvement in the process of evaluation. Simply being involved in the process, might influence their understanding of the situation. They might engage in increased reflection of the policy due to conversations about the evaluation or the policy. The use of evaluation based on the process, is called process use. This term was first introduced by Patton (1997) and is often included among the types of uses, just like instrumental and conceptual use (e.g. Balthasar, 2006; Ledermann, 2012). The definition of process use resembles conceptual use save from explicit reference to the influence of involvement (Taut, 2007). However, in line with Kirkhart (2000) and Alkin and Taut (2003) it is argued here that process use is not a type of use such as instrumental or conceptual, but a source of use that can lead to, among others, instrumental and conceptual use. The definition still includes the influence of involvement, but it abandons the notion that it must involve learning, thereby allowing for other types of use.

These three dimensions can grant a researcher a clearer focus on the operationalization of the type of use studied. For example, a researcher can choose to focus only on immediate and intended use based on results or long-term and unintended use of the process.

[7] Kirkhart (2000) only distinguishes between intended and unintended, but Alkin and Taut (2003) show that unintended can even be separated further in 'unintended and aware' and 'unintended and not aware'.

**Table 3. A list of factors influencing use, as found in the literature**

| Main factor influencing use | Subfactors influencing use | Literature where factor is found/ suggested | Does the article include empirical research? | Kind of use for which the factor is found/ suggested | Relationship found with use |
|---|---|---|---|---|---|
| Involvement of policymaker | high level of involvement general | Marra (2004) | Yes | Instrumental and conceptual use | + |
| | High frequency of (face-to-face) contact | Preskill, Zuckerman, and Matthews (2003) | Yes | Undefined (learning) | + |
| | Presence of discussion of findings with policymakers (before report is final) | Marsh and Glassick (1988) | Yes | Instrumental and conceptual use | + |
| | | Greene (1988) | Yes | Intrumental /conceptual/symbolic use | + |
| | respectful facilitator of discussion (evaluator), makes sure all voices are heard | Preskill, Zuckerman, and Matthews (2003) | Yes | Undefined (learning) | + |
| Political context | High level of political conflict over | Ledermann (2012) | Yes | Instrumental use | + |
| | High level of political pressure for | Ledermann (2012) | Yes | Instrumental use | + |
| Timing | New policy formulated (before, during, or after evaluation) | Johnson et al. (2009) | Review of literature | Undefined | + |
| | | Leviton and Hughes (1981) | Review of literature | Instrumental use | before or |
| Evaluation/ Report characteristics | High quality of evaluation | Leviton and Hughes (1981) | Review of literature | Undefined | + |
| | | Johnson et al. (2009) | Review of literature | Undefined | + |
| | Theory-driven evaluation | Sanderson (2002) | No | Undefined (learning) | + |
| | | Marra (2004) | Yes | Instrumental and conceptual use | + |
| | | Coryn et al. (2011) | No | Undefined | + |
| | Explicit learning goal (not shared with accountability) | Widmer and Neuenschwander (2004) | Yes | Instrumental, conceptual, and interactive use | + |
| | High level of technical knowledge | Forss, Cracknell, and Samset (1994) | Yes | Instrumental and conceptual use | + |
| | Presence of qualitative information in report | Leviton and Hughes (1981) | Review of literature | Undefined | + |
| | Feasible recommendation | Leviton and Hughes (1981) | Review of literature | Undefined | + |
| | | Marsh and Glassick (1988) | Yes | Instrumental and conceptual use | + |
| | | Johnson et al. (2009) | Review of literature | Mostly instrumental, some conceptual use | + |
| Evaluator characteristics | Extent that promoting use is task of evaluator | Alkin et al. (1985) | No | Undefined | + |
| | | Patton et al. (1977) | Yes | Mixed instrumental/ conceptual | + |
| | Political sensitivity of evaluator | Alkin et al. (1985) | No | Undefined | + |
| | Credibility of evaluator perceived by policymaker | Alkin et al. (1985) | No | Undefined | + |
| | | Johnson et al. (2009) | Review of literature | Mixed instrumental/ conceptual | +/- |
| | | Leviton and Hughes (1981) | Review of literature | Undefined | + |
| Policymaker characteristics | Commitment to evaluations/ general idea that evaluations are useful | Leviton and Hughes (1981) | Review of literature | Undefined | + |
| | | Shulha and Cousins (1997) | Review of literature | Undefined | + |
| | | Patton et al. (1977) | Yes | Mixed instrumental/ conceptual use | + |
| | | Johnson et al. (2009) | Review of literature | Undefined | + |
| | Interest in the evaluation process | Preskill, Zuckerman, and Matthews (2003) | Yes | Undefined (learning) | + |
| | High position in organization | Johnson et al. (2009) | Yes | Undefined | ? |
| | Distance to evaluator | Balthasar (2006) | Yes | Instrumental use (findings) | +/- |
| | High level of novelty of knowledge | Johnson et al. (2009) | Review of literature | Undefined | - |
| | | Ledermann (2012) | Yes | Instrumental use | + |
| | Involvement in programming | Johnson et al. (2009) | Review of literature | Instrumental and conceptual use | + |
| | | Shulha and Cousins (1997) | Review of literature | Undefined | + |
| Organizational characteristics | Open communication about evaluation (results) | Hodges and Hernandez (1999) | Yes | Instumental use | + |
| | High level participation climate (cooperative goals and constructive controversy) | Turnbull (1999) | Yes | Instrumental and symbolic use | symbolic: + instrumental: - |
| | Managers as advocates of learning | Preskill, Zuckerman, and Matthews (2003) | Yes | Undefined (learning) | + |
| | High frequency of staff turnover | Hodges and Hernandez (1999) | Yes | Instrumental | - |
| | | Leviton and Hughes (1981) | Review of literature | Undefined | - |
| | Basing decisions on evaluation (results) is stimulated | Hodges and Hernandez (1999) | Yes | Instrumental use | + |

## 2.2 Factors

Apart from studying what type of use has occurred, the factors influencing use have also been a major strain in the research of evaluation use. Table 3 contains an oversight of the factors found in the literature. It also reports for which type of use a factor has found to be important; and whether the article was theoretical or contained empirical evidence of the factor's influence (or whether it was a review of the literature, in which case it is indirectly based on empirical evidence). Lastly, is shows whether the influence was found to have (or in case of theoretical articles: expected to be) a negative or a positive effect on use. That is, whether it diminished use, or whether it enhanced use.

Due to a great variety of factors, the evidence for individual factors is rather thin. Some factors have been studied rarely, while others, even though studied more extensively, have been defined differently by the various studies. This is supported by the review article of Johnson et al. (2009), in which it becomes clear that most factors are slightly differently measured or defined, even though they are labelled the same. Therefore additional research is necessary to improve the explanatory value of the influence that factors have.

# 3. Research design

## 3.1 Methodology

The literature dealing with the factors leading to or hindering use of evaluations is based on too little systematic empirical evidence. The empirical literature is mainly based on case-studies that take a very limited amount of factors (or even only one factor) into consideration. It is therefore difficult to generalize from the cases. The evidence is scattered and not systematic. Apart from the case-studies, there is some literature that suggests a broad array of influencing factors, but this is often not empirically studied at all. These factors are mostly expectations without evidence, although they are often presented as recommendations (e.g. Alkin et al. 1985; Feinstein, 2002; Morabito, 2002). This means that there is little more than anecdotal evidence about what factors actually make a difference in evaluation use (Ledermann, 2012: 159). In order to ensure that this will not be another anecdotal case-study, this research utilizes a systematic way to analyse the relation between factors and use.

A second problem that follows from the frequent use of case-studies with a limited amount of factors taken into account is that there is hardly an idea how factors might work together to produce use (Johnson et al., 2009: 388). Factors are almost always studied individually, but in reality it is very possible that factors have a combined effect.

In order to address these issues qualitative comparative analysis (QCA) is used. This is a qualitative method which aims to find causal relations between factors and outcomes, based on detailed case knowledge. One of its main strengths is its systematic way to compare between cases, which enhances generalizability (Rihoux, 2006: 680). In QCA the researcher selects, through a literature study possibly combined with an initial empirical study factors (named conditions in QCA terminology) that might influence the outcome of interest. It then scores all cases separately on the presence or absence of these conditions and the outcome. This way, paths leading to the outcome are created, which can indicate which (combination of) conditions are sufficient and/or necessary for the outcome to occur (Ragin, 1999: 1228; Rihoux, 2006: 682). A major benefit of QCA is its inclusion of the notions of equifinality and conjunctural causality in its quest for causal relations (Schneider and Wagemann, 2012: 5-6). Equifinality assumes that in reality there are usually multiple causal combinations of factors leading to one outcome. Multiple conditions might be sufficient for an outcome, of which none is actually necessary. Conjunctural causality refers to the idea that a combination of factors, instead of a single factor, might cause an outcome to occur. This makes QCA a good method to go beyond the effect of individual factors, and see in what way factors interact to produce an outcome.

The enormous diversity of factors influencing use found in previous studies, points out that there can be no one-way-fits-all approach to evaluation use, but that the context in which the evaluations are conducted and used is highly important. Factors are likely to vary in the effect they have (or whether they have one at all) over different contexts, resulting in causal complexity. In this regard, this study needs to be modest: in such a complex environment, no definite answers can be found (Ledermann, 2012: 160). The factors that might influence one case will not be important across all cases and situations.

However, this does not mean that it is useless to search for causality or that there are no patterns to be found. Instead, well-conducted research into evaluation use can help evaluators and policymakers to find ways to improve their process and enhance the use of evaluations. This requires a rigorous and structured research design that allows for a systematic comparison between cases.

## 3.2 The steps of the QCA

### Data collection

For all cases information on the selected explanatory conditions and on the outcome (evaluation use) has been gathered. Three main sources of information were used: First, the documents relating to the evaluations (Terms of Reference, partial studies, the final document, etc.) were studied. Second, per case a minimum of two interviews were held, one with the IOB-evaluator that was responsible for the evaluation and one with the most relevant policymaker. Third, a questionnaire was sent to other policymakers who were involved in the evaluation, for example policymakers who had partaken in the reference group, had been a respondent, or were responsible for writing the policy response on the evaluation. The information from the sources was compared to triangulate the findings.

In the appendix the number of interviews and returned questionnaires are listed per case.

### Selection of conditions

The study started with a list of more than twenty possible conditions, which have been presented in chapter 2. In order to execute a QCA, the list of conditions needed to be limited to four to six conditions; otherwise the analysis would have resulted in too many configurations which would have lowered the explanatory value of the findings. It is important that this is done carefully, for if the evaluation omits to include relevant conditions, it can lead to contradictory rows which are problematic for the analysis (Schneider and Wagemann, 2012: 120). The included conditions have been selected according to the criteria stated below. First, conditions should be deemed relevant in the existing literature. Second,

during the data collection, there should be indications that a condition is relevant. Third, the comparative nature of the research requires that there is enough variation on the factors between cases. Fourth, there needs to be a reliable and transparent way to measure and code the condition. These criteria brought the list down to these four conditions: political, timing, containing novel knowledge, and interest shown by the main policymaker(s).

### The analysis

On the basis of the information gathered, the four explanatory conditions and the outcome were coded and recorded in a raw data table. The coding process, or calibration as it is called in QCA terms, entails first, setting the standards for membership and non-membership and subsequently deciding for every condition per evaluation whether it receives membership or not. The calibration method prescribed by Basurto and Speer (2012) has, in an adapted form, been used to conduct the process.[8] This research has used crisp-set coding; meaning that only full membership (1) or full non-membership (0) has been awarded. The conditions, see chapter 4, are of such a nature that fuzzy-set coding, where gradations in membership can be awarded, is likely to become problematic. While it was possible to estimate from the interviews and the questionnaire whether the evaluation did or did not, for example, contain new knowledge, it was much more difficult, and less accurate, to estimate whether it should receive a 0.75 or a 1 for the newness of its knowledge.

After all the conditions, including the outcome, had been coded, the analysis was run in the software program R.[9] First, tables of the conditions were made to establish whether any of the conditions were necessary conditions for the outcome. Also the measures of fit were obtained and interpreted. Then, a truth table was made representing the possible pathways. From this the sufficient (combinations of) conditions that lead to the outcome were identified.

[8] Instead of using data analysis software, the entire process has done manually. First all the information from the interviews and survey was collected in an excel-document and assigned to a condition per evaluation. Then all conditions were calibrated. This indicated which conditions could be used in the final analysis. For these conditions a second round of calibration has been done. Per condition a table was made containing the reasons for awarding either membership or non-membership per case. In order to fulfil the demand for transparency (Basurto and Speer, 2012: 157) the raw data table, the codebook and the arguments per condition are included in the replication document.

[9] In this research the R-packages for QCA have been used to obtain the truth table and configurations. The 'How-to-guide' from Schneider and Wagemann (2012) has been very useful in the process.

## 3.3 The case

Several requirements of a good case for this research were maintained. The evaluations should be conducted by the same organization, be addressed to one clear target group, and be conducted within a reasonably short time span. This has two benefits: First, it makes clear in what context the study was conducted, and thus, the extent to which it might be generalized. Second, it limits the number of factors that are likely to influence use. Organizations vary in many ways; the possible factors influencing use will be fewer if only one organization is included (Ledermann, 2012: 164). An example is the institutional distance between the evaluator and the policymaker, which is found to influence use (Balthasar, 2006): This is much more likely to vary between organizations than within an organization. The same is true for target groups and time span. As mentioned, in QCA the researcher can include only a limited number of conditions in the analysis, but should not omit relevant factors. Consequently, it is sensible to try to limit the list of *possible* conditions.

Furthermore, the organization whose evaluations are analysed should have a reasonable experience in conducting evaluations, have a good reputation, and a clear target group of potential users. Also, the evaluated issues must be clearly policy-related, preferably at the national level. This will result in 'most likely' cases: It is expected that evaluations that meet these requirements have most potential to be used.

In order to satisfy the requirements stipulated above, the research will take place at the Policies and Operations Evaluation Department (IOB) of the Ministry of Foreign Affairs in the Netherlands.[10] The organization is part of the Ministry, but operates independently from the policy departments and has its own budget. The IOB was established in the seventies of the last century to study the effects of the Dutch governmental development aid. Since 1996, the organization evaluates all Dutch foreign policy. They have an established reputation and adequate experience with evaluation research. The IOB has a clear target group: the policymakers in the Ministry's policy departments. The main task of the IOB is to conduct high-quality evaluations for learning and accountability purposes.[11] Learning and policy improvement are major goals of the evaluations and the IOB wants to contribute to it actively.[12]

---

[10] A list of the evaluations that were used a case is included in the appendix.

[11] Ministry of Foreign Affairs (2009) *Evaluatiebeleid en richtlijnen voor evaluaties* 11. https://www.rijksoverheid.nl/documenten/brochures/2009/10/01/evaluatiebeleid-en-richtlijnen-voor-evaluaties. Accessed on April, 6th 2016.

[12] Apart from forming the basis of the analysis for the thesis that lies before the reader, the research conducted, was also used to draw up a report for the IOB and provide recommendations to enhance the use of their evaluations. This second purpose of the research is important, because the development sector operates in an

# 4. Operationalization of the conditions

QCA requires all conditions and the outcome to be operationalized in detail. It needs to be clear how the conditions and the outcome are measured, and when they receive membership or non-membership. In the following section, the outcome (instrumental use) will be operationalized. The next section will first present the conditions that were included in the analysis and then show how these were operationalized.

## 4.1 Outcome: Instrumental use

In the chapter on the literature, it became clear that many choices can be made regarding the focus of the use that is studied. In this study instrumental use is the type of use that was studied. Instrumental use was defined as the use of the evaluation for direct information of policy decision making. For the analysis, the outcome condition was operationalized as follows: Instrumental use is present if at least one major policy decision was influenced significantly by the evaluation. Policy decisions might entail: the termination or continuance of the policy; an important strategic change in the policy with consequences at the operational level; or a major change in funding.[13]

Instrumental use can, but does not have to be, written down (Leviton and Hughes, 1981: 530). Thus, for the measurement, self-reporting of users had to be relied upon to account for unrecorded uses. The thesis focused on intended as well as unintended use. In the interviews it was specifically asked whether and in what way recommendations were implemented; thus addressing intended use. However, if the policymakers stated that a certain decision was taken on the basis of the evaluation, it was considered use, regardless of the intention of the policymakers. Furthermore, this thesis focused on both immediate and end-of-cycle use, but not long-term use. It is expected that long-term use requires a different approach; one that lies outside the scope of this research, because in all likelihood long-term use will be less directly linked to the evaluation. Lastly, the study did not make a distinction between process use and findings use, as the two are difficult to separate. Conclusions and recommendations were often already known to the policymakers through their involvement, but also were also included in

---

increasingly complex environment, which strengthens the need to base policy decisions on evidence (Thomas and Tominaga, 2013: 58). This requires that the process of evaluations and especially the use of evaluations are embedded in the institutional context of the departments concerned with development aid.
In this way, this study did not only contribute to the scientific knowledge, but also served a societal purpose.

[13] This operationalization is an adaption of the operationalization Ledermann (2012).

the final report. Yet, in the end, most decisions that were made could be traced back to the final report, so it seems that for instrumental use findings were an important source.[14]

## 4.2 Explanatory conditions

### Political (POLC)[15]

The literature does not quite agree on the effect of political issues on evaluation usage. Ledermann (2012) found that an evaluation with a politically conflicted topic was used more, because, so runs her argument, these issues needed more substantiating. For the condition in this research I deviate from the idea that the topic needs to be politically *conflicted*. The main rational for this is that even if there is no conflict about an issue, as long as the politicians are somehow interested in or concerned about it, they are likely to influence it. This can either, as Ledermann found, give an extra opportunity for use; or it might hinder instrumental use. An example of the latter is when politicians use an evaluation only to support their pre-existing standpoints (symbolic use). Barrios (1986: 111) argues that evaluations on less politicized and less controversial issues are more likely to be used. If there are no political guidelines for an issue, policymakers will search for other sources to help them decide on policy decisions. Evaluations supply solutions for these problems.

So, if it is not necessarily political *conflict*, how is the condition 'political' defined? The issue needs to be high on the political agenda (and thus receive a lot of attention from politicians) or be politically sensitive. It is anticipated that a sensitive issue would be very quickly on the political agenda if something happens (for example, when an evaluation is published), even if the issue is not high on the agenda before.

The operationalization of political is:

*An evaluation is considered political if the topic of the evaluation is politically sensitive or high on the political agenda*.

### Timing (TIM)

Several studies have reported that the timing of an evaluation is important (Bober and Bartlett, 2004: 377; Boyer and Langbein, 1991: 527; Rockwell, 1990: 392; Shea, 1991: 107). All of them measure timing by asking the policymaker whether the evaluation was 'on time' or 'too late', sometimes also including

[14] Process use was mentioned separately from findings use, but only as conceptual use. Several respondents said that talking over the evaluation, either with policymakers or among each other, helped them to reflect on the policy and their own role in it.

[15] The capitalized word between the brackets is the abbreviation used during the analysis.

'too early'. Not surprisingly, for evaluations that were considered 'on time', use was more likely than for evaluations that were considered 'too late' or 'too early'. However, this does not tell us, what 'on time' actually means to policymaker(s).

In contrast to these studies, this study takes a more objective measure of timing: whether the process of the evaluation took place at the same time as the process of policy formulation. The rationale here is that when an evaluation process runs parallel to the process of policy formulation, there are many moments in which the evaluation can influence the policymakers. During the evaluation process there always is some contact between the evaluators and the policymakers. This allows the evaluators to pass on information and knowledge, and policymakers to ask specific questions on issues that they are unsure of.

The operationalization of timing is:

*An evaluation is considered to be timely when the policy department was working on new policy or major policy changes during the data collection and/or writing of the report phase. This includes a complete change of the existing policy, as well as major changes in focus, scope, and goals.*

### Containing novel knowledge (KNOW)

Whether the inclusion of novel knowledge in an evaluation promotes use is contested in the literature. Ledermann (2012: 173) shows that a high level of novelty of knowledge is a necessary condition for changed policy when an issue is conflict-laden. If the knowledge is not new, the result is more likely that the evaluation will be used to endorse, rather than revise. Johnson et al. (2009: 385), in their literature review, conclude that knowledge from an evaluation needs to be confirmed by observations from outside the evaluation. These two findings seem somewhat at odds with each other, as the conclusion of Johnson et al. seems to imply that evaluations with new knowledge might be easier put aside and not lead to changes.

Novel knowledge will thus be tested here again; the operationalization is:

*The evaluation is considered to have contained new knowledge if the main policymaker(s)[16] profess that the evaluation contained knowledge about the policy that was novel to him/her (or them).*

[16] The main policymaker(s) is defined as the policymaker(s) that has the direct responsibility for the formulation (and sometimes implementation) of the policy. This is not the same as relevant policymakers, which include are all policymakers that are directly or indirectly involved in the policy.

### Interest shown by the main policymaker(s) (INT)

Many studies have already shown interest, defined in just as many different ways, has a positive effect on the use of evaluations (e.g. Johnson et al., 2009; Leviton and Hughes, 1981; Patton et al., 1977, Preskill, Zuckerman, and Matthews, 2003). Due to the importance the literature attributes to interest, it is included in this research.

Measuring interest is difficult. One can choose either to plainly ask whether the department was interested and run the risk of receiving societally acceptable answers, or find some proxy measure to capture it. Although it was explicitly asked in the interviews, the answers did not seem reliable enough to use for the calibration process. Therefore, the proxy-method has been chosen. Interest can be shown in three ways. First, the policymakers can ask the evaluators, before the data collection starts, to include specific questions that they would like an answer to. Or the policymakers can discuss the findings within their department after the evaluation is finished. The latter can again be split in to two: they can discuss the results themselves with all the relevant policymakers or they can ask the evaluators to present the results to all relevant policymakers.

Thus, the operationalization of interest is:

*The evaluation is considered to have been interesting for the policy department if the main policymaker(s): communicated the findings via a presentation from the evaluators to the relevant policymaker(s); or communicated the findings by discussing them in a staff meeting with the relevant policymaker(s); or brought forward at least one question they hoped the evaluation would answer.*

# 5. Analysis

## 5. 1 Set-theoretic method

QCA is a set-theoretic method, which means that a group of cases is considered part of a larger group of cases. This way, relationships can be defined in terms of necessary and sufficient relations (figure 1). If a condition is a necessary condition for the outcome, the outcome variable is a set within the larger group of the necessary condition. If a condition is a sufficient condition for the outcome, the condition is a smaller set within the larger group of used evaluations. QCA specifically searches for these two kinds of relations. In the next section, the analysis of necessary conditions will be presented; the following contains the analysis of sufficient conditions.

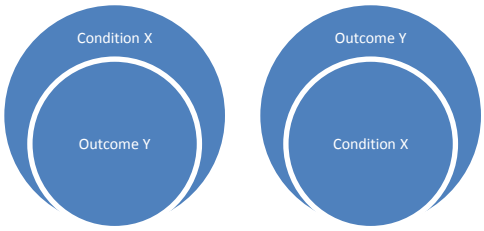


**Figure 1. Set-relations**

## 5.2 Necessary conditions

If a condition is necessary for an outcome to happen, there can be no instance of the outcome without the condition being present also.[17] From table 4 the extent to which the conditions can be considered necessary for the outcome can be deduced. Two of the conditions can be considered necessary: interest and timing. All cases which belong to the set of used evaluations, also belong to the sets 'timely evaluations' and 'evaluations for which the department showed interest'. The two other conditions, political and knowledge, are not necessary.

**Table 4. Necessity tables for POL, TIM, KNOW, and INT**

| | OUT | | | OUT | | | OUT | | | OUT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POLC | 0 | 1 | TIM | 0 | 1 | KNOW | 0 | 1 | INT | 0 | 1 |
| 0 | 5 | 1 | 0 | 9 | 0 | 0 | 11 | 1 | 0 | 6 | 0 |
| 1 | 8 | 4 | 1 | 4 | 5 | 1 | 2 | 4 | 1 | 7 | 5 |

In order to correctly establish whether conditions are necessary the measures of fit, the consistency and coverage scores, need to be interpreted. The consistency score highlights the extent to which the 'condition is consistent with the statement of necessity' (Schneider and Wagemann, 2012: 139). As can

[17] Of course, it might be the case that the absence of a condition is necessary for an outcome to occur; in that case, there can be no case where the outcome and condition are both present.

be seen in table 5, the conditions timing and interest are completely consistent, scoring a 1. The conditions political and novel knowledge are somewhat less consistent, and score a 0.8. Schneider and Wagemann (2012: 143) advise that consistency scores for necessary conditions should be at least as high as 0.9 to establish a condition as necessary. This means that timing and interest can be considered necessary conditions for the presence of instrumental use; but political and novel knowledge cannot.

**Table 5. Consistency and coverage scores for necessity**

| | Consistency for OUT | Coverage for OUT | Consistency for not-OUT | Coverage for not-OUT |
|---|---|---|---|---|
| POLC | 0.8 | 0.33 | 0.62 | 0.67 |
| TIM | 1 | 0.56 | 0.39 | 0.44 |
| KNOW | 0.8 | 0.67 | 0.15 | 0.33 |
| INT | 1 | 0.42 | 0.58 | 0.58 |

However, two other considerations are of importance here: First, the proportion between cases falling into the set of used evaluations and the cases falling in the set of not-used evaluations needs to be taken into account. There are only five of the eighteen cases part of the set of used cases. In comparison, thirteen cases did not receive membership of this set. This affects the consistency score, because with only a few cases in the outcome set, a high consistency score is easier obtained. A consistency score of 1 suggests that both timing and interest are important factors in the instrumental use of evaluations; however, due to the low number of cases, one must be careful in generalizing this conclusion.[18]

The second factor is the consistency score of the necessity for the outcome of non-use (those cases that scored a 0 in the outcome). As QCA assumes asymmetry, the scores for consistency of non-use are not simply the reserve of the consistency of use (Schneider and Wagemann, 2012: 81). In theory, it is possible that a condition is necessary for both use and non-use. In that case, obviously, the necessity of the condition would be rather trivial. As can be seen in table 5, none of the conditions come even close to the 0.9 threshold. This strengthens the finding that timing and interest are both important for use.

Apart from the consistency score the coverage score should also receive attention when interpreting the necessity of the conditions. The term coverage is somewhat misleading and using 'trivialness'

[18] There is little written about the exact consequences of this type of skewed data in QCA (Schneider and Wagemann, 2012: 248). This makes interpreting the data especially hard. Most importantly, it needs to be kept in mind that one must be careful in generalizing the findings. The results would become must stronger if they still hold in other, future, studies. Therefore, it is essential that the study is replicable. All documents of importance of replication can be requested at the researcher.

denotes better what is meant (Schneider and Wagemann, 2012: 144). If the coverage or trivialness score is low, it means that the number of cases in which the outcome occurs and has membership of explanatory condition X is small in comparison to the whole set of cases belonging in the set of condition X. In that case, a condition might be a necessary condition, albeit it is also a trivial one. The set of political evaluations consists of 12 cases, and the set of used cases within this set has only four cases; thus, giving it a low coverage score (0.33). Although, this is the lowest score among the four conditions, none of them has a score that could be called high. To some extend all conditions are trivial. Timing and interest are especially interesting, as they scored a 1 on consistency; and can be considered necessary conditions. Timing scores a 0.5: in five out of ten cases the outcome occurs; and interest scores even lower (0.42), since in five out of twelve cases the outcome occurs.

In conclusion, the conditions timing and interest are, according to these findings, necessary conditions for instrumental use of evaluations. All used evaluations were conducted during the drafting process of new policy. Also, for each of the used evaluation the policymakers showed an interest, either by proposing research questions or by communicating the findings to all relevant policymakers. However, both these conditions are to some extent trivial, meaning that there were many cases in which the conditions were present, but the outcome was not.

## 5.3 Sufficient conditions

Apart from necessary, a condition can also be a sufficient condition for an outcome to occur. If a condition is completely sufficient, there are no instances of the condition occurring, without the outcome occurring likewise. In this section the results of the sufficiency analysis for use and not-use will be presented. The interpretation of these results is the subject of the next section.

As can be seen in table 4, none of the conditions on itself can be considered sufficient. However, as has been mentioned above, QCA takes causal complexity into consideration, meaning that even if a single condition on its own is not sufficient, a combination of conditions might be. In order to find the paths leading to use, a so-called truth table has been composed.

The truth table shows all possible combinations of conditions, and whether a combination occurred in reality. Five of the combinations are not represented in any empirical case; these are the logical remainders. No assumptions on these remainders will be made, and these lines will be left out the analysis.

**Table 6. Truth table**

| Row nr. | POLC | TIM | KNOW | INT | OUT | PRI | Nr. of cases | Cases represented[19] |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2, 6, 9 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 16 |
| 3 | 0 | 0 | 1 | 0 | ? | - | 0 | |
| 4 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 12 |
| 5 | 0 | 1 | 0 | 0 | ? | - | 0 | |
| 6 | 0 | 1 | 0 | 1 | ? | - | 0 | |
| 7 | 0 | 1 | 1 | 0 | ? | - | 0 | |
| 8 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 13 |
| 9 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 11, 18 |
| 10 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 17 |
| 11 | 1 | 0 | 1 | 0 | ? | - | 0 | |
| 12 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 7 |
| 13 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 |
| 14 | 1 | 1 | 0 | 1 | 0 | 0.25 | 4 | 1, 3, 8, 15 |
| 15 | 1 | 1 | 1 | 0 | ? | - | 0 | |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 4, 10, 14 |

The other eleven combinations do represent at least one case. Of these eleven rows, there is one that contains a logical contradiction: the same row includes three cases with no membership in used evaluation and one case with membership (row 14). It is possible that this is due to a missing explanatory condition. Some attention to missing conditions will be given at the end of this chapter. In the analysis the contradictory row was considered a row with a membership score of 0 on the outcome, as the consistency score was set at 0.7.

**Table 7. Results from the sufficiency analysis for the presence of use (OUT)**

| | Political | Timing | Novel Knowledge | Interest | Consistency | Raw Coverage |
|---|---|---|---|---|---|---|
| 1: | | ● | ● | ● | 1 | 0.8 |
| Solution Consistency: 1 / Solution Coverage: 0.8 | | | | | | |

Configuration leading to the presence of the outcome (OUT): instrumental evaluation use. ● denotes presence of a condition in the solution; ɵ denotes the absence of a condition. Note: unique coverage is not displayed, as it is identical to the raw coverage.

[19] The numbers of the cases align with the list of cases presented in appendix 1.

From the truth table, a solution term has been created through logical minimization. As mentioned, no assumptions about the logical remainders are made, thus table 7 shows the complex solution. In formal terms the solution is:

TIM * KNOW * INT → OUT

A combination of the presence of three factors is sufficient to bring about the outcome: timing, novel knowledge and interest. This pathway has a consistency of 1, which means that there are no cases that have all three factors present, while not having the outcome present also. The coverage of the path is 0.8, as four out of the five used cases are represented by this path. The solution consistency and coverage are equal to the consistency and coverage of the path, because only one path came out of the analysis as sufficient.

Just as interesting as the path towards use is the path leads to the absence of use. Table 8 shows the findings for this analysis.[20] This analysis yields three different paths that lead to the absence of use. All have a consistency of 1. In formal terms this solution is:

tim * know + tim * INT + POLC * know * int → out

**Table 8. Results from the sufficiency analysis for the absence of use (out).**

| | Political | Timing | Novel Knowledge | Interest | Consistency | Raw Coverage | Unique coverage |
|---|---|---|---|---|---|---|---|
| 1: | | ɵ | ɵ | | 1 | 0.54 | 0.23 |
| 2: | | ɵ | | ● | 1 | 0.31 | 0.15 |
| 2: | ● | | ɵ | ɵ | 1 | 0.23 | 0.08 |
| Solution Consistency: 1 / Solution Coverage: 0.77 | | | | | | | |

Configurations leading to no evaluation use. ● denotes presence of a condition in the solution; ɵ denotes the absence of a condition.

The first path, with the largest coverage, shows that the absence of knowledge and the absence of timing combined are sufficient for the absence of use. The second path leading to the absence of use consists of the absences of timing and the presence of interest. The third path, with the lowest coverage (and only one unique case) consists of the presence of political and the absence of both interest and

[20] The inclusion score for the consistency level has again been set at 0.7. Consequently, the contradictory row has been included as not used.

knowledge. The paths together have a solution coverage of 0.77, meaning that three cases are not covered by one of these paths.

The next section will discuss the findings per condition, and also discuss how these paths fit into the literature and what the implications for policymakers and evaluators might be.

## 5.4 Interpretation

### Timing

The importance of timing should not come as a surprise: If changes to a policy are made anyway, it is much easier for the policymakers to incorporate the lessons of the evaluation. It might also be easier to gain support for the changes. As mentioned in chapter 4, this study implements a more objective approach to measure timing than earlier studies. In doing so, it does not just confirm earlier research saying that timing is important, but also shows *when* the timing is right. The results show that the use of evaluations is heightened when the processes of policymaking and evaluating run parallel. [21]

During the interviews multiple respondents said that they liked the processes to run parallel, as it gave them the opportunity to incorporate lessons learned immediately. For almost all evaluations there was frequent contact between the evaluators and the policymakers, in which early lessons could already be passed on. The respondents declared that they used the conversations with IOB not merely to give the evaluators information, but also to reflect on their policy and draw lessons.

### Political

In the QCA no convincing link between the politicization of an issue and the use of evaluations was found. The analysis of necessity does not convincingly establish that an issue needs to be political in order for policymakers to use the evaluation, as the consistency score was 0.8. In fact, during the interviews it became clear that in several cases use could happen in spite of it being a political issue, and not because of it. In these cases, decisions regarding the more politicized topics within the policy were untouched, but other parts of the policy were still changed. For the analysis of conditions leading to no use, the condition political scored the highest of all four conditions, although still well below the 0.9 threshold.

[21] There were no instances of an evaluation being immediately followed by a drafting process of new policy. Therefore, the effect of that specific timing could not be studied, and it is possible that it, likewise, will have a positive effect on use.

### Novel knowledge

The present research shows that novel knowledge in combination with an interested policy department and good timing strongly supports the use of evaluations. It cannot be called, with consistency scores of respectively 0.8 and 0.15, a necessary condition for use or for the absence of use.[22] It is not surprising that novel knowledge is not a necessary condition. During the interviews respondents mentioned that they sometimes could use the evaluation to convince others of the need of a decision, which they already knew was needed.[23] Or alternatively, sometimes the evaluation brought no new knowledge, but gave a good overview of all the problems in the policies. Such an overview could also help policymakers to find solutions.

### Interest

This study confirms the earlier findings: interest is a necessary factor of evaluation use. Sometimes the policymakers were simply not interested, because they felt themselves too busy to give it much attention; or because they did not think it could teach them anything worth knowing. Considering that there is no formal obligation to use the evaluation and nobody monitors whether policy departments act upon an evaluation, policymakers that are unwilling to use the evaluation can easily put it aside.

### The path leading to use

The sufficiency analysis showed one configuration, or path, with a consistency score of 1 and a coverage score of 0.8. The combination of good timing, novel knowledge and interest led in all four empirical cases to instrumental use. It is not difficult to see why this combination leads to use: timing provides the opportunity; interest shows that there is an incentive for use; and novel knowledge shows what might be changed.

Moreover, the three conditions can strengthen each other. First, it is argued that one way to show interest is to ask specific questions to be included in the evaluation. In this manner, the chances that new knowledge will be provided by the evaluation are increased. In one case, for example, a policy department asked the evaluators to estimate whether outsourcing had been beneficial; simply because they were it had been the case. Also, if new knowledge is provided, it might be more readily accepted by

---

[22] The reason that the consistency score for novel knowledge and the absence of use is so low (0.15) is, at least in part, because both the absence of knowledge and the absence of use are skewed in the data set (12 cases of ~know and 13 cases of ~use out of a total of 18 cases). If that is the case, the chances that ~Y will be a set of ~X become larger; and thus the changes of ~Y as a set of X smaller. See, Schneider and Wagemann, 2012: 232-237.

[23] Of course, the knowledge might still have been novel to whomever the policymakers needed to convince. However, here knowledge is operationalized as 'new to the main policymaker'.

the policy department than unrequested knowledge. Second, timing provides a strong incentive for the policy department to be interested and ask questions. Third, policymakers might also be more susceptible to new knowledge if they are actively thinking of ways to improve the policy. In the example of the policy department who reconsidered outsourcing, the department was at that point deciding on the new policy framework for the next couple of years. The timing thus offered them the perfect moment to ask such a question.

When interest, timing, and novel knowledge come together, it does not seem to matter whether the issue is political. As has been mentioned above, this might in part be the case because political interests are usually narrower than the complete policy issue; and considerable parts of political issues could still be changed.

### The path leading to the absence of use

The analysis for the sufficiency of conditions leading to the absence of use showed three possible paths. The first included the absence of novel knowledge and the absence of timing. It is not difficult to see why these two together are sufficient. When policymakers are not working on new policy or major policy changes, there is less opportunity for them to let the evaluation influence major decisions. If this is also combined with a lack of knowledge, incentives are lower as well: there is nothing that incites the policymakers to action. Besides, novel knowledge, if present, might show policymakers solutions they had not thought of before. This, too, is less likely without novel knowledge. Thus, when there is no novel knowledge and no new policy, the opportunity and the incentive for use are both low. In one case, the respondent explained that it was very difficult to change anything, because all the important choices had already been made, and there was no room to manoeuvre for the policy department. In such a case, novel knowledge might have helped to convince people of the need of some changes.

The second path, tim*INT → out, is not as easy to interpret. Four cases share the combination of the absence of timing and the presence of interest and all lead to the absence of use. That the absence of timing can contribute to the absence of use is not surprising. Yet, the combined effect of its absence with the presence of interest is not so clear. From the literature it becomes clear that interest is a strong factor is enhancing use, and also from the interviews it cannot concluded that interest would contribute in some way towards the absence of use. It should be noted though, that two of the four cases are also covered by the path 'tim*know → out'. For both, this does seem to be a more likely path to explain the non-use.

Reviewing the cases covered by the path 'tim*INT → out', it seems that the non-use might be due a third, missing, factor. It is not certain which factor this would be, although for at least one of the two cases only covered by this path, is becomes clear from the data that it might be due to the higher management's attitude towards the evaluation. In this case the policymakers were interested, but their managers were not willing to act upon the evaluation. In fact, it is not the only case where this was noticeable. In the interview became apparent, that more than the policymakers the managers focused on the implications of the evaluation for the Minister. All evaluations served the dual purpose of learning and accountability, and a clear tension between the two purposes was noticeable. If the focus of the managers was mainly on accountability, there was less space for learning. In those cases, the policy department tried to keep the evaluation low key. The best solution for the analysis would be to incorporate the attitudes of the management as a fifth condition. However, there is no (reliable) data of this for all the cases, which makes that an unviable option. Therefore, it is an important factor to investigate further in any potential follow-up research. The next section will discuss the problem of missing factors in more depth.

But before missing factors are discussed, the last path, POLC*know*int→out, still awaits explanation. In this path the presence of politics, combined with the absence of novel knowledge and interest creates a sufficient path for the absence of use. When there is no novel knowledge and no interest, there is little reason for policymakers to act upon the evaluation. And apparently, politicians are not likely to pressure the policymakers to use the evaluation in such a case. Ledermann (2012) concluded that politically contested issues are more likely to be used, because politicians need arguments. It seems however, that that is not the case here. In contrast, it is possible that politicians, when faced with an evaluation, are more likely to use it symbolically rather than instrumentally. One of the three cases that are covered with this path was called 'a compromise between left and right'. The respondents explained that both the parties at the left and at the right side of the political spectrum could create a politically expedient narrative around this policy: left saw it as development aid, and the right emphasized how it benefitted Dutch companies. When left and right both agree on this policy, there is little opportunity for the policymakers to decide on major changes. Furthermore, if this is then in combined with little interest from the policy department and no incentive in the form of novel knowledge, it is not surprising, that the evaluation ends up on a shelf.

### Missing conditions?

The study started out with a long list of possible conditions, of which only four ended up being used in the analysis. What does this mean for those that were left out? And, as mentioned above, what could the consequences for the analysis be that these were omitted?

It certainly does not mean that the omitted factors are unimportant. For instance, the factor involvement of the policymakers has many times been proved to be highly effective way to increase use. Johnson et al. (2009: 398) suggest that 'engagement, interaction, and communication between evaluation clients and evaluators is key to maximizing the use of the evaluation in the long run.' In the interviews held concerning the eighteen studied cases, this idea was confirmed. Many respondents mentioned that their contact with the evaluators helped them to understand their policies better and increased their interest in the evaluation process. Involvement was, however, not included in the design, because there was little variation between the eighteen cases. In seventeen cases, respondents said the quality of the contact was good. Also, although there were some differences in the frequency of the contact (ranging from once a week to once every few months; with on average about once a month), most respondents professed to be satisfied with the frequency.

The same accounts for some of the other excluded factors: respondents were almost all convinced of the credibility of the evaluators; almost all evaluators saw promotion of usage as part of their jobs; and most policymakers were satisfied with the quality of the evaluation itself. Seeing that these conditions all have been shown before to have an influence on use, it is likely that they have had an influence in these cases as well. Besides the conditions named in the literature, there were also factors the respondents mentioned in the interviews that might have an effect, but suffer from the same problem: lack of variation. Two possible factors that returned frequently were the lack of supervision on use and the lack of a standard reporting format. Both factors made it easier for departments to put the evaluation aside. However, without variation it is not possible to establish whether they are necessary or (in combination) sufficient.

Some conditions had to be left out for another reason: there was no reliable way to either measure or code them. Both the conditions 'learning atmosphere in the department' and 'feasible recommendations' were left out for this reason. This is more problematic than leaving conditions out because of the lack of variation; variables left out because of measurement issues might actually have made a difference between use and non-use in the cases present.

That there are probably factors missing in this analysis is noticeable in two ways. First, there were three uncovered cases in the analysis of none use and one uncovered case in the analysis of the

presence of use. This points out the likelihood that other factors than the four present have had an effect on the use. Second, one of the paths leading to non-use (tim*INT → out) suggests that there might be other conditions that influence use, for example the attitude of the management towards the evaluation. In future research this factor should be included in the analysis, if possible, to estimate whether it also has an effect.

# 6. Conclusion

The question this study sought to answer was: 'w*hat conditions influence the use of evaluations for policy improvement in the bureaucracy at the national level?'* From a list of more than 20 factors, four were selected for the qualitative comparative analysis. Eighteen cases were coded on either the presence of these four conditions or its absence, as was the outcome 'instrumental use'. The analysis showed that two factors are necessary factors to attain instrumental use, namely timing and interest. These two, together with knowledge, are a sufficient path to use.

The findings contribute to the literature in several ways. First, it confirms the importance of interest and timing for use, and moreover, it not only shows that timing matters, but also what good timing entails. Second, it shows that the use of evaluations is not hindered, but also not helped by an issue being political. Third, by its use of QCA, the research focusses not only on the individual factors, but also on the combined effect of conditions. It is shown that interest, novel knowledge and timing *together* are enough to cause instrumental use. Most studies before looked at the effect of individual factors, neglecting that factors can interact. And as this analysis shows, the factors do interact: good timing, novel knowledge and interest strengthen each other, giving the best possible position to an evaluation for use to take place. Fourth, this study has tried to conduct the research in as structured and transparent a way as possible. Many studies in the field of evaluation use lack transparency about the definitions they use and how they chose their cases. This hinders the application of the research as building blocks for further research. By being structured and transparent, the researcher has aimed to make replication and further research possible.

Two critical comments on the generalizability need to be made. First, this study is conducted in a specific context: at the national level, close to the politics, by evaluators that were internal to the organization. Many factors that might be important in influencing use were constant among these cases; but in other contexts, they can be still relevant. Therefore, the findings cannot simply be transferred to all other evaluation settings without further research. In addition, one cannot draw the conclusion that the factors that were not in this analysis are not relevant. Second, the second path in explaining the absence of use is difficult to interpret, suggesting that in this study one or more relevant factors are not included. Due to limitation of time and resources, it was impossible to collect more empirical data on all eighteen cases in order to construct additional conditions. Yet, future research should take into account the differences between the lower level and higher level interest in evaluation. The interviews suggest that higher level management and the political leadership have a different attitude towards evaluation:

They are more focussed on the political consequences of an evaluation, and less on the learning aspect of it.

These critical comments notwithstanding, the findings are valuable not only for theorists, but also for practitioners. If one wants to promote the use of one's evaluation, one should try to evaluate a policy when the policymakers are in the process of revising it and focus on including knowledge that is novel to the policymakers.

# Bibliography


Alkin, M.C., P. Jacobson, J. Burry, J. Ruskus, P. White, and L. Kent (1985) *A guide for evaluation decision makers* (Beverly Hills: Sage Publications).

Alkin, Marvin C. and Sandy M. Taut (2003) 'Unbundling evaluation use' *Studies in Educational Evaluation* 29 1-12.

Azzam, Tarek and Bret Levine (2015) 'Politics in evaluation: Politically responsive evaluation in high stakes environments' *Evaluation and Program Planning* 53 44-56.

Balthasar, Andreas (2006) 'The effects of institutional design on the utilization of evaluation: Evidenced using qualitative comparative analysis (QCA)' *Evaluation* 12 (3) 354-372.

Barrios, Nina Brown (1986) *Utilization of evaluation information: A case study approach investigating factors related to evaluation utilization in a large state agency* (Doctoral dissertation, Florida State University, ProQuest Dissertations Publishing).

Basurto, Xavier and Johanna Speer (2012) 'Structuring the calibration of qualitative data as sets for qualitative comparative analysis (QCA)' *Field Methods* 24 (2) 155-174.

Bober, Christopher F. and Kenneth R. Bartlett (2004) 'The utilization of training program evaluation in corporate universities' *Human Resource Development Quarterly* 15 (4) 363-383.

Boyer, John F. and Laura I. Langbein (1991) 'Factors influencing the use of health evaluation research in congress' *Evaluation Review* 15 (5) 507-532.

Caplan, Nathan (1977) 'A minimal set of conditions necessary for the utilization of social science knowledge in policy formulation at the national level' in: C. Weiss (ed.) *Using social research in public policy making* (Lexington: Lexington Books) 183- 197.

Coryn, Chris L.S., Lindsay A. Noakes, Carl D. Westine, and Daniela C. Schöter (2011) 'A systematic review of theory-driven evaluation practice from 1990 to 2009' *American Journal of Evaluation* 32 (2) 199-226.

Development Assistance Committee (1991) *Principles for the evaluation of development assistance* Paris.

Feinstein, Osvaldo N. (2002) 'Use of evaluations and the evaluation of their use' *Evaluation* 8 (4) 433-439.

Fetterman, David M. (1994) 'Empowerment evaluation' *American Journal of Evaluation* 15 (1) 1-15.

Forss, Kim, Basil Cracknell, and Knut Samset (1994) 'Can evaluation help an organization to learn?' *Evaluation Review* 18 (5) 574-591.

Freeman, Richard (2009) 'Learning in public policy' in: Robert E. Goodin, Micheal Moran, and Martin Rein (eds.) *The Oxford handbook of public policy* Oxford Handbooks Online. DOI: 10.1093/oxfordhb/9780199548453.003.0017.

Greene, Jennifer C. (1988) 'Communication of results and utilization in participatory program evaluation' *Evaluation and Program Planning* 11 341-351.

Henry, Gary T. and Melvin M. Mark (2003) 'Beyond use: Understanding evaluation's influence on attitudes and actions' *American Journal of Evaluations* 24 (3) 293-314.

Hodges, Sharon P. and Mario Hernandez (1999) 'How organization culture influences outcome information utilization' *Evaluation and Program Planning* 22 183-197.

Johnson, Kelli, Lija O. Greenseid, Stacie A. Toal, Jean A. King, Frances Lawrenz, and Boris Volkov (2009) 'Research on evaluation use: A review of the empirical literature from 1986 to 2005' *American Journal of Evaluation* 30 (3) 377-410.

Kirkhart, Karen E. (2000) 'Reconceptualization evaluation use: An integrated theory of influence' *New directions for evaluation* 88 5-22.

Ledermann, Simone (2012) 'Exploring the necessary conditions for evaluation use in program change' *American Journal of Evaluation* 33 (2) 159-178.

Leviton, Laura C. and Edward F.X. Hughes (1981) 'Research on the utilization of evaluations: A review and synthesis' *Evaluation Review* 5 (4) 525-548.

Marra, Mita (2004) 'the contribution of evaluation to socialization and externalization of tacit knowledge: The case of the World Bank' *Evaluation* 10 (3) 263-283.

Marsh, David D. and Judith M. Glassick (1988) 'Knowledge utilization in evaluation efforts: The role of recommendations' *Knowledge: Creation, Diffusion, Utilization 9 (3) 323-341.*

Ministry of Foreign Affairs (2009) *Evaluatiebeleid en richtlijnen voor evaluaties* 11. https://www.rijksoverheid.nl/documenten/brochures/2009/10/01/evaluatiebeleid-en-richtlijnen-voor-evaluaties Accessed on April, 6th 2016.

Morabito, Stephen M. (2002) 'Evaluator roles and strategies for expanding evaluation process influence' *American Journal of Evaluation* 23 (3) 321-330.

Patton, M. Q., P.S. Grimes, K.M. Guthrie, N.J. Brennan, B.D. French, D.A. Blyth (1977) 'In search of impact: An analysis of the utilization of federal health evaluation research' in: C. Weiss (ed.) *Using social research in public policy making* (Lexington: Lexington Books) 141-163.

Preskill, Hallie and Rosalie T. Torres (2000) 'The learning dimension of evaluation use' *New directions for evaluation* 89 25-37.

Preskill, Hallie, Barbra Zuckerman, and Bonya Matthews (2003) 'An exploration study of process use: Findings and implications for future research' *American Journal of Evaluation* 24 (4) 423-442.

Radealli, Claudio M. (1995) 'The role of knowledge in the policy process' *Journal of European Public Policy* 2 (2) 159-183.

Ragin, Charles C. (1999) 'Using qualitative comparative analysis to study causal complexity' *HSR: Health Services Research* 35 (5 II) 1225 - 1239.

Rihoux, Benoît (2006) 'Qualitative comparative analysis (QCA) and related systematic comparative methods; Recent advances and remaining challenges for social science research' *International Sociology* 21 (5) 679-706.

Rockwell, S. Kay, Elbert C. Dickey, and Paul J. Jasa (1990) 'The personal factor in evaluation use: A case study of steering committee's use of a conservation tillage survey' *Evaluation and Program Planning* 13 389-394.

Sanderson, Ian (2002) 'Evaluation, policy learning and evidence-based policy making' *Public Administration* 80 (1) 1-22.

Schneider, Carsten Q. and Claudius Wagemann (2012) *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis* (Cambridge, UK: Cambridge University Press).

Shea, Micheal Patrick (1991) *Program evaluation utilization in Canada and its relationship to evaluation process, evaluator and decision context variables* (Doctoral dissertation, University of Windsor, Canada, ProQuest Dissertations Publishing).

Shulha, Lyn M. and J. Bradley Cousins (1997) 'Evaluation use: Theory, research, and practice since 1986' *Evaluation Practice* 18 (3) 195-208.

Taut, Sandy (2007) 'Methodological and conceptual challenges in studying evaluation process use' *Canadian Journal of Program Evaluation* 22 (2) 1-19.

Thomas, Vinod and Jiro Tominga (2013) 'Development evaluation in an age of turbulence' in: Jan-Eric Furubo, Ray C. Rist, and Sandra Speer (eds.) *Evaluation and turbulent times: Reflections on a discipline in disarray* Comparative Policy Evaluation, vol. 20 (New Brunswick/London: Transaction Publishers) 57-70.

Turnbull, B. (1999) 'The mediating effect of participation efficacy on evaluation use' *Evaluation and Program Planning* 22 131-140.

Weis, Carol H. (1998) 'Have we learned anything new about the use of evaluation?' *American Journal of Evaluation* 19 (1) 21-33.

Widner, Thomas and Peter Neuenschwander (2004) 'Embedding evaluation in the Swiss Federal Administration; Purpose, institutional design and utilization' *Evaluation* 10 (4) 388-409.

# Appendix 1

| List of IOB evaluations included in the study | | | Interviews IOB (N) | Interviews policymakers (N) | Survey (N) |
|---|---|---|---|---|---|
| 1 | 2016 | Cultuur als kans. Beleidsdoorlichting Internationaal Cultuurbeleid (2009-2014) | 3 | 2 | 5 |
| 2 | 2015 | Vreedzame geschillenbeslechting en het tegengaan van straffeloosheid – Beleidsdoorlichting internationale rechtsorde | 1 | 2 | 4 |
| 3 | 2015 | Gender, peace and security: Evaluation of the Netherlands and UN Security Council resolution 1325 | 2 | 2 | 2 |
| 4 | 2015 | The Only Constant is Change: Evaluation of the Dutch contribution to transition in the Arab region (2009-2013) | 2 | 1 | 1 |
| 5 | 2015 | Work in Progress – Evaluation of the ORET Programme: Investing in Public Infrastructure in Developing Countries | 2 | 1 | 2 |
| 6 | 2015 | Met hernieuwde energie – Beleidsdoorlichting van de Nederlandse bijdrage aan hernieuwbare energie en ontwikkeling (2004-2014) | 1 | 1 | 2 |
| 7 | 2015 | Beleidsdoorlichting van de Nederlandse humanitaire hulp 2009-2014 | 2 | 1 | 3 |
| 8 | 2015 | Opening doors and unlocking potential: Key lessons from an evaluation of support for Policy Influencing, Lobbying and Advocacy (PILA) | 1 | 1 | 1 |
| 9 | 2015 | Aided trade – An evaluation of the Centre for the Promotion of Imports from Developing Countries (2005-2012) | 2 | 2 | 1 |
| 10 | 2015 | Evaluation of the Matra programme in the Eastern Partnership countries 2008-2014 | 2 | 1 | 3 |
| 11 | 2014 | Navigating a sea of interests: Policy evaluation of Dutch foreign human rights policy 2008-2013 | 1 | 3 | 0 |
| 12 | 2014 | Strategie bij benadering – Nederlandse coalitievorming en de multi-bi benadering in het kader van de EU-besluitvorming (2008-2012) | 2 | 1 | 1 |
| 13 | 2014 | Balanceren tussen koopmanschap en diplomatie – Evaluatie van de Netherlands Business Support Offices 2008-2013 | 2 | 2 | 1 |
| 14 | 2014 | Investeren in wereldburgerschap – Evaluatie van de Nationale Commissie voor Internationale Samenwerking en Duurzame Ontwikkeling (NCDO) | 2 | 1 | 0 |
| 15 | 2014 | Op zoek naar focus en effectiviteit – Beleidsdoorlichting van de Nederlandse inzet voor private sector ontwikkeling 2005-2012 | 2 | 1 | 1 |
| 16 | 2013 | Op zoek naar nieuwe verhoudingen - Evaluatie van het Nederlandse buitenlandbeleid in Latijns-Amerika | 1 | 1 | 1 |
| 17 | 2013 | Balancing ideals with practice: Policy evaluation of Dutch involvement in sexual and reproductive health and rights 2007-2012 | 1 | 1 (& 2 via e-mail) | 1 |
| 18 | 2013 | Investeren in stabiliteit: Het Nederlandse fragiele statenbeleid doorgelicht | 2 | 1 | 0 |