



UMEÅ UNIVERSITY

National environmental evaluation systems – guiding towards sustainability?

Magnus Larsson

Department of Sociology
Umeå 2018

This work is protected by the Swedish Copyright Legislation (Act 1960:729)
Dissertation for PhD
ISBN: 978-91-7601-897-2
ISSN: 1104-2508
Electronic version available at: <http://umu.diva-portal.org/>
Printed by: Print & Media, Umeå, Sweden 2018

Table of Contents

- List of original papers in the thesis iii**
- Abstract..... iv**
- Acknowledgments v**
- Introduction.....1**
 - Conceptual note 4
- Previous research on evaluation and evaluation systems 6**
 - Development of evaluation and evaluation research 6
 - Defining evaluation and evaluation research 8
 - Research on evaluation use and evaluation systems..... 9
 - Evaluation use*10
 - Evaluation systems* 11
 - Environmental evaluation research.....14
 - Challenges to environmental evaluations*14
 - Research for environmental evaluation* 15
 - Research on environmental evaluation* 17
- Framework to critically explore environmental evaluation (systems)19**
 - Exploring evaluation from a governance perspective 20
 - Management by objectives*21
 - Network governance*.....21
 - Policy instruments*.....22
 - Critical program theory analysis*.....23
 - Exploring the validity and relevance of evaluations..... 23
 - Validity*.....24
 - Relevance*24
 - Exploring the usability, use, and influence of evaluations..... 26
 - Usability*.....26
 - Use*.....27
 - Influence*.....28
 - Exploring the contribution of evaluations to sustainable development..... 28
 - Notions of sustainable development* 29
 - Exploring the consequences of environmental evaluations.....31
- Institutional setting 33**
 - The Local Investment Program and subsequent evaluations 34
 - The National Environmental Quality Objectives system and evaluations 36
- Materials and Methods 39**
 - Narrative synthesis 40
 - Program theory analysis 43
 - Directed content analysis 44
- Results – summary of papers 46**

Paper I. Effects on sustainable development from large environmental programs: a review of 16 evaluations.....	46
Paper II. Evaluation in management by objectives: A critical analysis of Sweden's national environmental quality objectives system	47
Paper III. Translations of knowledge from an environmental evaluation system – do they contribute to usability of evaluation knowledge?.....	48
Paper IV. Stakeholders' use of environmental evaluations (system) and their contribution to sustainable development	50
Conclusions.....	52
Concluding discussion	57
Contributions of the thesis	59
References	62

List of original papers in the thesis

- I. Larsson, Magnus and Hanberger, Anders. 2015. "Effects on sustainable development from large environmental programs: a review of 16 evaluations." *Journal of Integrative Environmental Sciences* 12(2): 85–105.
- II. Larsson, Magnus and Hanberger, Anders. 2016. "Evaluation in management by objectives: A critical analysis of Sweden's national environmental quality objectives system." *Evaluation* 22 (2): 190-208.
- III. Larsson, Magnus. "Translations of knowledge from an environmental evaluation system – do they contribute to usability of evaluation knowledge?" Submitted.
- IV. Larsson, Magnus. "Stakeholders' use of environmental evaluations (system) and their contribution to sustainable development." Submitted.

Abstract

Background. Dealing with environmental threats is one of the largest, if not the largest, challenge contemporary societies face. One way to better deal with this challenge would be to produce knowledge that can be used to improve environmental work and environmental policy and thus ultimately contribute to sustainable development. National environmental evaluations, which this thesis explores, could potentially fill this function because they are supposed to generate applicable and useful knowledge for improving environmental policy and practice for a sustainable transition. However, what different environmental actors view as useful knowledge varies, and needs to be empirically investigated. Against this background, the aim of this thesis is to investigate whether, and how, national environmental evaluation systems contribute to key actors' environmental work and sustainable development. The thesis explores two national environmental evaluation systems in Sweden.

Method. A mixed methods approach is applied that combines three methods. Firstly, a narrative synthesis is developed and applied to compile a summary of sustainable development effects from national environmental evaluations. Secondly, a critical program theory is used to investigate the evaluation system's underlying logic and to assess the likelihood of it achieving its intended effects. Thirdly, a directed content analysis is used to explore the usability and use of environmental evaluations and evaluation systems. The methods are applied to various documents, interviews with key actors, and observations at two environmental seminars.

Results. The results show that, to contribute to sustainable development in the context of evaluation systems and network governance, environmental evaluations need to be of sufficient quality and meet different stakeholders' knowledge needs. However, only some evaluations meet this demand. The main value of national environmental evaluations and evaluation systems is that they reinforce the national objectives, provide a recurrent report on achievement of objectives, and push actors to take responsibility to improve their environmental work.

Acknowledgments

Writing a dissertation is a formidable task. It would have been an insurmountable task without the support of many people over the years. While I am responsible for all the potential errors, ambiguities, misconceptions, and faults, I could not have written the thesis without the support of all of you. Thank you!

First, I want to thank my supervisors Anders Hanberger, Katarina Eckerberg and Daniel Larsson I could not have wished for more guidance, knowledge and support for my project. Thank you Daniel for sharing your broad knowledge of sociology in general, and for asking the hard questions when needed. Discussing sociology with you is a delight. Katarina, thank you for your thorough reading and pragmatic suggestions for how to improve all my texts. I am grateful to you for taking me to seminars and conferences. You have introduced me to so many people and networks. Finally, thank you Anders for all the energy and time you have put into reading and commenting on my drafts. You have taught me so much about academic writing. Most importantly you have shown me what it takes to put together a coherent argument without compromising complexity. I am so glad that you introduced me to the field of evaluation research.

Dividing my time between the sociology department and the Umeå Centre for Evaluation Research (UCER) has been a challenge and made me feel, at times, disconnected. However, thanks to the wonderful people at the department of sociology and the department of applied educational science (where UCER is located) these feelings have been few and fleeting. Thank you to all my fellow doctoral students for the fun times (trench coats etc.) and the great camaraderie. I especially want to thank Michael Dahlberg-Grundberg and Magdalena Sjöberg. I am grateful that I had your moral support and informal meetings in the beginning when we were all just trying to figure it out. As a PhD student, one of the most rewarding activities for me has been participating in several networks including UCER, Critical Organization Studies (COS) and the environmental network at the department of political science. Thank you for all the interesting, gratifying, and fun seminars, travels, dinners, and discussions.

In my time as a PhD student, I have also had the opportunity to work with many great teachers. Britt-Inger Keisu, thanks for including me in your course with flipped classrooms. I learned so much about engaging students and learner-centered teaching. Thank you Åsa Gustafson for believing in me and generously sharing your pedagogical expertise. Your insights have made me a better teacher. Finally, thank you Ingrid Schild for all the courses we taught together. I so enjoyed working with and our discussions about teaching, evaluation, grading and other related (and unrelated) matters.

My dissertation has benefited greatly from comments and suggestions from several readers over the years. I am grateful for the useful and insightful feedback Åsa Persson provided me at the mid-seminar. And I appreciate the guidance and clear comments and suggestions from Göran Sundqvist at the end-seminar. Malcolm Fairbrother, thank you for useful comments at the final stage. Per Wisselgren, you gave me valuable comments on how to increase the sociological relevance of my thesis and, perhaps more importantly, for showing an interest in my PhD project. It meant a lot to me to have your helpful comments from the beginning of the dissertation writing process.

One challenge of working in two departments is the additional administrative work. Thankfully, both Sociology and UCER have competent, professional, and helpful administrative personnel. I owe a special thanks to Helene Risberg, Barbro Hedlund, Gunilla Renström, Sofia Wård and Liselott Engström for your patience with my confusion and swiftly answering my questions.

I am especially grateful to Lasse Reinikainen and Ulrika Schmauch for introducing me to academic life so many years ago in Sundsvall, and later in Umeå. Your kind and helpful comments were invaluable as I adjusted to my new role as instructor. I also want to thank Jonas Edlund, Ingemar Johansson Sevä, and especially Annica Brännlund for graciously involving me in the development new research projects.

I was fortunate to share my office with Malin Benerdal. I have benefited immensely from our discussions on evaluation research. I am particularly in debt to you for introducing me to viewpoints from political science - in the end these perspectives greatly influenced my thesis. I also want to thank Sara Carlbaum for interesting discussions and helpful comments on a late version on my kappa.

Thank you to all my friends for unforgettable vacations (swallows can be mean!), book clubs, football, brännboll, and everything else. A special thanks to “gänget” and my running club Velox Fortis. Finally, thank you Daniel Papacosta for late study nights and invigorating discussions on many subjects, big and small.

Maria, Mats, Hannah, and Lina thanks for laughing and arguing with me. You are the best siblings' one could ask for. Thank you Mom and Dad for always helping when I need it! Most importantly, I would like to thank my family. Lisa, thank you for putting up with my shifting moods and sharing the highs and lows of this journey with me. I could not have done this without you. Tim and Ellie, thank you for getting me to think about other, more important, things when I was getting too obsessed with my thesis.

Magnus Larsson, Umeå, May 2018

Introduction

Dealing with environmental threats is one of the largest, if not the largest, challenge contemporary societies face (Allen et al., 2014; World Bank, 2012). These threats include, but are not limited to, climate change, loss of biodiversity, ocean acidification, and chemical pollution (Rockström et al., 2009; Steffen et al., 2015). Many of these threats originate from central features in contemporary societies, such as consumption, transportation, agriculture, forestry, and so on (Allen et al., 2014). A radical transition of our societies towards sustainable development (SD) is therefore needed, according to the majority of experts in the environmental field (ibid.; Baker and Eckerberg, 2008; WCED, 1987).

There is, however, less consensus around the urgency of these challenges and how they should be dealt with. This is mirrored in the multiple understandings of SD, and in how environmental problems and solutions are discussed, regarding whether there is a need for new knowledge to manage the problems, for example. Sustainable development itself is a highly contested and ambiguous concept (Baker, 2006; Du Pisani, 2006; Persson et al., 2016). The concept's ambiguity and its positive connotations can explain why many people and governments advocate and support a transition towards SD (Persson et al., 2016), but when it comes to deciding on action it is obvious that they have vastly different understandings of what SD implies and how the problem should be dealt with (ibid.).

Hopwood et al. (2005) distinguish between three distinct understandings of transition towards SD: status quo, reform, and transformation. Those advocating *status quo* transition acknowledge that changes need to be made, but that these adjustments can be carried out without any fundamental changes to the structure of society, to how decisions are made, or to existing power relations. Supporting technological solutions should be the main focus in transitioning society towards SD. From this perspective, existing knowledge or new knowledge for single-loop learning (Argyris and Schön, 1978) is sufficient and there is no demand for novel insights. *Reformers*, on the other hand, recognize that “large shifts in policy and lifestyle, many very profound, will be needed at some point” (Hopwood et al., 2005: 43). They typically highlight technology, scientific knowledge, and modifications and restrictions of the market and reforming government as important reform areas in the transition towards SD. There is a knowledge need to establish what works to help improve the effectiveness of technical solutions to alleviate environmental problems. By contrast, *transformationists* see the essential problems as stemming from our current way of life, which is based on the exploitation of the majority population and the environment by a few. The transformation view of SD is the most heterogeneous of the three perspectives;

however, a central notion of this third view is introducing more direct democracy and equity. Organization of popular action outside parliamentary and state settings is viewed as the most effective approach to restraining environmental degradation and ultimately putting a stop to it. The transformationists' knowledge need is radically different from the status quo group's and reformists' need and includes knowledge to support problem solving in multi-actor settings (Schwandt, 2016), for example knowledge on how to develop and maintain institutional capital and local knowledge (Hajer, 2003). All three perspectives therefore agree that a transition is needed, but what this transition should entail and what knowledge is needed continues to be debated.

The point of departure of this thesis is in line with the reformist and transformationist notion of transition towards SD, however acknowledging that different actors can conceive and advocate the three perspectives and that their knowledge needs differ. I also recognize that some researchers claim that in environmental decision making and environmental policy, the use of scientific knowledge remains scarce (Kirchhoff et al., 2013; Owens et al., 2006). It has been suggested that one reason for this gap is that scientific knowledge is rarely actionable or applicable in a policy setting (Caplan, 1979). If this is correct, evaluations which this thesis sets out to explore could potentially bridge this gap. Environmental evaluations are initiated to generate applicable and useful knowledge for decision making and to improve practice (e.g., environmental work) (e.g., Fitzpatrick et al., 2012; Patton, 2008). However, I recognize that when evaluations provide "useful knowledge" they will reflect and promote either a status quo, or a reformist or transformational perspective towards SD. Moreover, that environmental evaluations can provide decision makers and other stakeholders with different knowledge about the effectiveness of environmental programs and policies, including their goal attainment, and how to improve environmental work towards SD. Therefore, this thesis pays attention to what conception of SD transition the evaluations reflect. In discussing the achievement of the Sustainable Development Goals (SDGs), Schwandt et al. (2016: 1) suggest that evaluation can "examine policy and programme implementation and effectiveness, and build well-reasoned and supported cases for claims of progress." They add, however, that the SDGs are interrelated in complex ways and "... cannot be neatly captured in a linear straightforward cause-effect relationship" (ibid., 3). Although environmental evaluations may be used and referred to as providing uncontested truths about effects of environmental policies and programs this thesis recognizes that evaluations are value-laden and each supports one or the other perspective on SD. This is not to say that they cannot be a valuable tool in the transition towards a more sustainable environment and society, but in what way they may be valuable is an empirical question.

Furthermore, for evaluations to be able to contribute to SD, they need to be reliable and valid and, moreover, need to be able to discern effects of an intervention from other effects. In addition, evaluations should meet the key stakeholders' knowledge needs as they are jointly responsible for addressing the problems and promoting the transition. These are all basic quality requirements for evaluations to be productive tools for supporting change towards SD. But as said I recognize that the actors' understanding of the challenges and their knowledge needs, as well as the evaluations' contributions to SD, depends on whether this is viewed from a status quo, reform, or transform perspective.

An assumption in this thesis is that the ways in which the evaluative knowledge is produced, organized, and used becomes critical to whether, and how, it contributes to the transition towards a more sustainable society. However, we do not know enough about whether environmental evaluations actually contribute to environmental work and SD in practice, and therefore are part of a certain solution, or whether they are a waste of time and resources – or, even worse, are used to justify inaction and are therefore part of the problem. This thesis seeks to shed light on this issue by investigating the extent to which evaluation may be a useful policy instrument. More specifically, the thesis will examine the contribution of national evaluations and evaluation systems to national environmental actors' environmental work and SD.

To reflect the value and impact evaluations may have, the evaluation needs to be understood as a social phenomenon situated in a political context (Schwandt, 2015) and this is how the thesis explores evaluations. The context in this case is two multi-actor settings in Swedish national environmental politics, namely, the institutions and actors involved in, or affected by, two national environmental evaluation systems and subsequent evaluations. The first is the largest environmental program in Swedish history so far, the Local Investment Programs (LIP). The second is the national policy framework titled “National Environmental Quality Objective (NEQO).” Systematic evaluation was (LIP) and is (NEQO) a major part of these initiatives. These evaluations are initiated by the state and are objectives-oriented, that is, aimed to support the achievement of the national environmental objectives. Hence, the evaluations provide knowledge for improving the presiding environmental politics. This thesis takes its point of departure in the state's environmental problem framing and objectives but employs a critical approach to its assumptions. This implies that the thesis problematizes the problem framing, the environmental objectives, and the adopted notion of sustainable development.

Various objective-oriented environmental evaluations exist on the global level (e.g., evaluations of the SDGs) and on national, regional and local levels in many countries (Persson et al., 2016). The Swedish case is a compelling as Sweden has

a longstanding tradition of being at the forefront of environmental work (Lafferty and Eckerberg, 1998; Liefferink and Skou Andersen, 1997). Sweden has also developed national environmental programs and policies that have been thoroughly evaluated, which means that it is possible to study the impact of environmental evaluations in depth.

The purpose of this thesis is to investigate whether, and how, national environmental evaluation systems contribute to key actors' environmental work and SD, and if they have other implications.

The thesis seeks to answer the following two research questions:

Under what conditions, in the context of evaluation systems, can environmental evaluations contribute to sustainable development?

Have environmental evaluations (and evaluation systems) contributed to national actors' environmental work and SD and, if so, how have they contributed?

Mainly, to answer the first question, the thesis will investigate the quality, relevance, and usability for key stakeholders, of evaluations and their contribution to SD. These factors are generally considered important in research on evaluation use (Cousins and Leithwood, 1986; Johnson et al., 2009; Shulha and Cousins, 1997). To answer the second research question, additional factors and conditions presumed to be important in terms of evaluation use in evaluation systems will be explored, such as actors' proximity to the policy cycle, their knowledge needs, forums for deliberation, and stakeholder involvement (Brandon and Fukunaga, 2014).

The subsequent sections are constructed as follows: first, previous research on evaluation and evaluation systems with special focus on environmental evaluation is put forward. Thereafter, the framework I have developed to critically explore environmental evaluation (systems) is described. The third section places the study's objects in their institutional setting to gain a better understanding of the two evaluation systems. The subsequent section presents the materials and methods applied and outlines the content of the four individual papers. Finally, conclusions are drawn from the four papers as a whole, guided by the aim and the two research questions, followed by a concluding discussion.

Conceptual note

Evaluation is in this thesis defined as the activity of systematically determining the merit, worth, or significance of something, usually a program, policy, or

practice (Scriven, 1991). In addition, evaluation is recognized as a social and political practice with societal outcomes, intended or not (Schwandt, 2015).

Environmental evaluation refers to an evaluation where the evaluand (i.e., the object evaluated) concerns some aspect of the environment; but environmental evaluations by necessity also involve social aspects (Rowe, 2012).

The term *evaluation system* refers to “the procedural, institutional and policy arrangements shaping the evaluation function and its relationship to its internal and external environment” (Liverani and Lundgren, 2007: 241).

Sustainable development (SD) is defined by the three pillars, or dimensions (economic, ecological, and social), commonly related to SD (WCED, 1987)¹. In addition, a distinction is made between weak and strong SD (Cabeza Gutiérrez, 1996; Nourry, 2008).

Environmental policy refers to a multitude of activities including public policies, programs, projects, specific measures, and work intended to improve the environment. In addition, “policy work,” “environmental work,” and “environmental interventions” are terms used to indicate public environmental policy.

¹It is recognized that SD is a highly contested term (Baker, 2006) with many different definitions and a long history (Du Pisani, 2006). In this thesis, no distinction is made between the concepts of sustainability and SD.

Previous research on evaluation and evaluation systems

The role of knowledge in public policy has been studied in numerous ways in different disciplines. When studying the interaction between knowledge and policy in the environmental field, focus has been on, among other things, how science advisors (Jasanoff, 1990) and scientific commissions can shape policy and influence policy making (Owens, 2015). The research question has primarily been whether and, if so, how, *scientific* knowledge informs and permeates policy. Even though this thesis is concerned with the relationship between knowledge and policy, its focus is on evaluative knowledge which largely is produced outside academia and inside, or closely connected to, the policy realm. The following review of previous research will therefore focus on the evaluation field, which this thesis seeks to contribute to. Specific attention will be given to topics that are central to the research question, namely: evaluation systems, evaluation use, and environmental evaluation. This section provides a brief rendition of the development of evaluation and evaluation research and definitions of evaluation, and specifies where, in the field of evaluation research, the thesis is situated.

Development of evaluation and evaluation research

Evaluation is something human beings have always conducted on an informal cognitive basis as a way to appraise the world. Formal evaluation, as a practice in itself, gained traction in relation to the emergence of advanced capitalistic societies (House, 1993) and is today described as a global phenomenon (Furubo et al., 2002).

The development of evaluation research has been described in a number of ways. The most recognized analogies used to describe the development of evaluation and evaluation research are roots (Alkin, 2013), generations (Guba and Lincoln, 1989), and waves (Vedung, 2010). These three authors with their different approaches provide insights into the development of evaluation and evaluation research in different ways. Alkin (2013) focuses on evaluation theorists and scholars and places them in different areas (use, methods, and valuing) of evaluation research to explain the development of evaluation theories and practice. Guba and Lincoln (1989) suggest that evaluation theories and practices can be divided into four generations where the fourth generation highlights stakeholder involvement. While both Alkin (2013) and Guba and Lincoln (1989) have an American or at least Anglo-Saxon context, Vedung (2010) has focused mainly on the Swedish context. Vedung states that the developments (waves) he has identified can also be found in North America and other places, but to a lesser extent. In addition, Vedung places evaluation in a policy setting and thereby

highlights the importance of the political context for the development of evaluation. This thesis underscores the importance of the policy context for understanding the value and consequences of evaluations. This makes Vedung's political framing and wave analogy the most suitable to apply in this thesis.

According to Vedung (2010), evaluation has gone through four waves of development, that in some way changed the manner evaluation has been practiced. While the analogy of the wave is meant to illustrate how one particular mode of evaluation, after another, became prevalent Vedung also points out that the previous waves do not disappear when a new wave emerges. The waves leave traces and therefore remain present although not as the most dominant mode of evaluation. Below follows a short description of each wave.

The first wave, the "science-driven wave," emerged in the early 1950s. This wave was based on an idea of radical rationalism. In terms of this, as long as evaluations were carefully planned and systematically implemented, as well as conducted by highly trained academics, it would be possible to achieve "the great society." Political decisions were to be based on knowledge and careful planning and implemented through large-scale programs. This was believed to be the way forward and it was believed that this largely was how society was governed. This wave was in line with the expansion of the welfare state in Western societies at this time.

By the early and mid-1970s, the assumptions of the society of piecemeal planning and of the science-driven wave were highly questioned and a new form of evaluation began to emerge. The second wave, acknowledged by Vedung (ibid.), is called the "dialogue-oriented wave." While the science-driven wave mainly focused on the front end of the policy cycle, such as planning and design, the dialogue-oriented wave focused on processes, stakeholder participation, deliberation, and democracy. As the oil crisis loomed large and the idea of the "great society" became more and more questioned, evaluators and commissioners too became more attuned to questioning the assumptions of the scientific wave. Instead, evaluators now focused on stakeholder opinion and deliberated on alternative ways of resolving issues and assessing programs. Experts were no longer recognized as better suited at providing a definition of what a good society was. Rather, these questions were to be answered by different stakeholders, assisted by evaluators, through deliberation. This wave was largely based on a social constructivist understating of knowledge. It was focused less on retrieving objective facts and more on how different artefacts were given meaning.

At the end of the 1970s, a new wave emerged, the neo-liberal wave, reflecting a change in public governance. This wave was right wing-oriented in contrast to the

first two waves. The most prominent set of ideas behind this wave was the so-called “New Public Management (NPM)” approach. This way of thinking suggests that the best way of governing the public is by exposing it to market logic. That is, treating the stakeholders and users as customers, exposing public institutions to competition, and shifting the control from bureaucrats to leaders and the leadership ideals that are prevalent in the open market. Evaluations developed during this wave reflect the knowledge needs of NPM, namely, to provide recurrent performance information to be able to govern by objectives and results.

In the mid-1990s, a fourth wave surfaced in the US, which reached the Nordic countries by the early 2000s. This wave is called the “evidence wave.” It was introduced through the tag line “what matters is what works.” Government programs and policies, in terms of this, should be based on best available evidence. The question of what constitutes best evidence, following assumptions of this wave, is determined by the so-called “evidence hierarchy” that grades methods according to their proclivity for measuring effects. This in turn means that experimental methods are considered to be a gold standard for gaining hard evidence. During this wave, evidence-based evaluations were requested and commissioned.

According to Vedung (*ibid.*), we are at the moment wrapped up in the evidence wave even though traces of all previous waves linger in the form of sediments. In addition to these waves, there is another trend which is only briefly mentioned by Vedung, which is the increasing number of evaluations conducted. Since the 1990s, basically all types of interventions and practices have been potential subjects for evaluation. It can be argued that this trend can be largely attributed to the neo-liberal wave and the practice of NPM since they occurred simultaneously; and, further, that the NPM paradigm focused on output rather than process (Hood, 1995), meaning that measuring and evaluating outcomes becomes a central practice in governing.

Contemporary evaluation practices will vary between regions, fields of inquiry, and levels of governing. Vedung’s wave metaphor (2010) provides an understanding of different trends in evaluation, showing that traces of “older” waves or trends may still be prevalent as sediments in contemporary policy settings. Therefore, we can expect to find all of these evaluations in the environmental field, but perhaps evaluations from one or two waves are more prevalent.

Defining evaluation and evaluation research

Evaluation has been defined in numerous ways (Fitzpatrick et al., 2012; Patton, 2008; Rossi et al., 2003; Scriven, 1991; Shaw et al., 2006; Vedung, 2010; Weiss,

1998a). No single definition has taken hold in the evaluation literature, which should not be surprising given the diverse features of evaluation across disciplines. Scriven's (1991) definition of evaluation, as the activity to systematically determine the merit, worth, or significance of something, usually a program, policy, or practice, is a broad and often cited definition. Shaw et al. (2006) and Schwandt (2015) suggest that evaluation also needs to be recognized as a social and political practice. That is, evaluation is always embedded in political and social systems. The main contribution of this rendering of evaluation is the focus on the context surrounding all social practices, including evaluation. Subsequently, evaluation activities may have different consequences and effects given the social and political circumstances. Evaluation is therefore not solely understood as a rational–logical activity, but also as a social and political activity. Evaluation research should by this standard include contextual factors and unintended effects to fully comprehend the societal consequences of evaluation activities. Therefore, as stated above, evaluation is in this thesis defined as the activity to systematically determine the merit, worth, or significance of something, usually a program, policy, or practice (Scriven, 1991). In addition, as already mentioned, evaluation is recognized as a social and political practice with societal outcomes, intended or not (Schwandt, 2015).

Evaluation research includes a wide variety of topics and theories (Alkin, 2013). A separation can be made between “research *for* evaluation” and “research *on* evaluation.” Research *for* evaluation focuses on developing approaches and methods to support evaluation practice. The approaches and methods typically address how to conduct evaluation to generate knowledge to meet someone's knowledge needs (Fitzpatrick et al., 2012; Rossi et al., 2003). Research *on* evaluation focuses on generating knowledge of evaluation as a political and social phenomenon. Examples of this are research on historical trends in evaluation (Vedung, 2010), evaluation use, and consequences of evaluation. Although many research topics transcend this binary construct it is a useful demarcation for understanding previous research in the evaluation field. This thesis explores topics related to research both “for” and “on” evaluation.

Research on evaluation use and evaluation systems

One central research area in contemporary research on evaluation discussed in this thesis is evaluation use. A second research area in the thesis is evaluation systems, which are scarcely researched. To be able to address the second research question, of whether the evaluations and evaluation systems contribute to key actors' environmental work and SD, a necessary step is to investigate whether evaluations are used in the context of evaluation systems. Hence, research on evaluation use and evaluation systems is briefly reviewed and discussed below.

Evaluation use

Use is one of the longest still ongoing research areas in evaluation research (Ledermann, 2012; Weiss, 1979, 1998b). Research on evaluation use has principally focused on different types of use (Vedung, 2015; Weiss, 1998a), ways of enhancing use (Patton, 2008), and factors affecting use (Cousins and Leithwood, 1986; Shulha and Cousins, 1997).

Evaluation researchers have developed and discussed different types of use to cover the many ways evaluations may be used and impact society (Højlund, 2014a, 2014b; Johnson et al., 2009; Patton, 2008; Vedung, 1997; Weiss, 1998b). The archetypal way of using evaluation refers to when an actor applies the knowledge from the evaluation to guide decision making, usually called “instrumental use.” Other ways of using evaluation include: learning from the evaluation but without that knowledge impacting decision making directly (“conceptual” or “enlightenment use”) or using evaluations to justify a predetermined position (“legitimizing use”).

This thesis applies Vedung’s conceptualization of evaluation use (2015), which is comprehensive and provides clear demarcations between the types of use that have been developed over time. It includes six types of use: instrumental, conceptual, legitimizing, tactical, ritual, and constitutive. In addition, Vedung separates these uses into product and process use (see Højlund, 2014a, and Kirkhart, 2000, for similar conceptualizations), meaning that the six types of use can be a case of either product or process use. “Product use” refers to the “employment of evaluative findings and recommendations from final reporting. Final reporting may include written, oral, and audio-visual final reporting as well as similar information efforts” (Vedung 2015: 205). “Process use” is defined as “the utilization of any evaluation activity taking place before final reporting. This means that the use of findings from draft final reports and interim reporting during the evaluation process, written as well as orally and audio-visually communicated, is assigned to process use” (ibid., 205–206). Use induced from the announcement of an upcoming evaluation and learning while working with an evaluation is also considered process use.

In recent years, the research on use has been criticized for being too actor-focused and not taking broader consequences of evaluations into account (Dahler-Larsen, 2011). Subsequently, use has been complemented with concepts such as functions (Hanberger, 2011), constitutive effects (Dahler-Larsen, 2014), and influence (Henry and Mark, 2003; Kirkhart, 2000). In order to examine how evaluations can contribute to SD, the larger impacts of evaluations including types of use, and broader influences and unintended consequences of evaluations should be taken into account. This thesis intends to contribute to this expanded understanding of evaluation use, in particular within the environmental policy field and in the

context of evaluation systems. How this is done is expanded on in the section “Framework to critically explore environmental evaluation (systems).”

In two reviews of research on factors enhancing, promoting, and explaining use of evaluation (Cousins and Leithwood, 1986; Shulha and Cousins, 1997), the factors and characteristics can be separated into two categories: (1) characteristics of evaluation implementation; and (2) characteristics of the decision or policy setting. Findings show that evaluator competence, evaluation quality, credibility, relevance, communication quality, findings, and timeliness are factors explaining use. The second category includes information needs, decision characteristics, political climate, competing information, personal characteristics, and commitment or receptiveness to evaluation as explanatory factors for use. Recently, research has added stakeholder involvement as an important explanatory factor for use (Johnson et al., 2009).

Ledermann (2012) points out that these reviews have been helpful in finding relevant and potentially determining factors to explain evaluation use, but that they do not to the same extent ascribe weight to the different factors, by showing which factors might be more, and which factors less, central. Saunders (2012) widens the notion of what could impact use by distinguishing between use and usability where “usability” refers to the communicative aspects of the evaluation. An evaluation may in this sense have high usability (i.e., be well written and widely disseminated), but may not be used (e.g., due to lack of interest from the potential users). Furthermore, the recent discussion on evaluation use pays more attention to the setting it is used in. A new setting for many evaluations is the evaluation system in which it is embedded (Leeuw and Furubo, 2008). This thesis explores evaluation use in this new setting and aims to contribute to research on evaluation use and influence within evaluation systems. Hence, in the next section, research on evaluation systems is discussed.

Evaluation systems

Evaluation researchers have described the development of evaluation systems in terms of how such systems have become institutionalized in the public and private sector in the last decade either as a complement to, or to replace, stand-alone evaluations. Rist and Stame (2006) suggest that a stream is a feasible metaphor for understanding development towards evaluation systems: the typical single-study evaluation gives way more and more to a continuous flow of information that resembles the structure of a stream. The reasons given for this qualitative change in the way of performing evaluation have been suggested to come from different directions. One explanation comes from a logic–rational perspective. Single studies can often be perceived as uninformative if they are not supported by other findings. Performing ad hoc studies can also increase the risk

of administration amnesia (ibid.). That is, the same questions and the same answers are repeated in a loop and no real progress is made. Another explanation for the increasing number of evaluation systems is demand. New information needs from governmental agencies call for a steady flow of information in order to produce timely performance information. As discussed above, this knowledge need is apparent in NPM-oriented governance reflected by the second evaluation wave (Vedung, 2010). Moreover, as demonstrated in paper II (Larsson and Hanberger, 2016), it is apparent in governance through management by objectives (MBO). The general trend of a “knowledge society” that infuses more knowledge into organizations also plays an important role for the development of evaluation systems (Dahler-Larsen, 2011).

Evaluation systems have been defined in different ways (Dahler-Larsen, 2011; Hanberger, 2011, 2016; Leeuw and Furubo, 2008; Liverani and Lundgren, 2007). Leeuw and Furubo (2008) propose four criteria for defining evaluation systems. An evaluation system should, according to them, consist of: a shared epistemological perspective, some level of institutional organization, recurring or permanent production of evaluation knowledge, and focus on intended users. Other definitions have widened the focus from the recurring production of knowledge (Dahler-Larsen, 2011) to also include the response aspect of evaluation systems (Hanberger, 2011) and how institutions and organizations establish routines for dealing with ad hoc evaluations and recurring streams of evaluation knowledge (Hanberger, 2016). Evaluation systems can vary in terms of how they are constructed, produce knowledge, and intend to be used. This thesis recognizes the different reasons for developing evaluation systems and applies the following definition of evaluation systems: “the procedural, institutional, and policy arrangements shaping the evaluation function and its relationship to its internal and external environment” (Liverani and Lundgren, 2007: 241).

Research *for* and *on* evaluation systems is still scarce, particularly in the environmental field, but has recently garnered increased attention. In the research *for* evaluation systems, focus has been placed on how evaluation systems should be constructed to best fill their function. A common way to organize evaluation systems is through so-called “monitoring and evaluation (M&E)” systems. Kusek and Rist (2004: 12) define M&E systems as the “feedback system on the outcomes and consequences of government actions.” They suggest ten steps to build an effective M&E system. One, stakeholders should be consulted and engaged in setting targets; two, targets that can be monitored and evaluated must be chosen. Three, key performance indicators should be set to monitor progress. Four, baseline performances indicating achievements should be decided. Five, interim, or milestone, targets are decided to chart the way towards long-term targets. The sixth step includes monitoring for results and

implementation by collecting quality performance data. The seventh step sets up the evaluation function of the M&E system. In step eight, the timing of the evaluation results is organized to accommodate intended users' needs. In step nine, focus is placed on how to report the findings to make sure that they come to use, including ways to ensure learning and knowledge sharing within the organization (or government). The final step, ten, deals with sustaining the M&E system by ensuring clear roles, responsibilities and trustworthy and credible information, as well as capacity and appropriate incentives.

The main critique towards organizing and understanding the role of evaluation systems in Kusek and Rist's (2004) framework is that it implies a rational policy and planning perspective. Following the definition of evaluation in this thesis, that evaluation activities by necessity are tied to their political and social context, with many stakeholders involved may affect how the evaluation systems are actually set up, implemented, and used.

The empirical findings from research *on* evaluation systems relate to how they operate and are perceived in different policy contexts. Dahler-Larsen (2006) focuses on the construction of evaluation systems by comparing two different evaluation systems in Danish medical policy. Even though both systems claim to be evidence-based they use two very different approaches to meet this claim. One approach is more context-based and has been criticized on that basis for not holding up universal objective truths. The other evaluation system is based on universal criteria taken from a US perspective. This in turn has caused larger resistance from the street-level stakeholders who feel that many of the universal criteria lack relevance in the Danish context. Dahler-Larsen concludes that the construction of evaluation systems must be viewed through the lens of their social and organizational context.

Hanberger (2011) explored a management response (MRE) system in foreign aid that structures the final part of the evaluation system. While the intended principal use of the MRE systems was policy improvement, very little evidence for this type of use was found. Instead, MRE was used in a legitimizing fashion. Strengthening organizational legitimacy and creating a feeling of assurance were reported to be key functions of MRE systems.

Dahler-Larsen (2011; 2012) highlights additional problematic issues with evaluation systems. He suggests that the era of reflexive modernity was craving more and more evaluation to be able to better make decisions in an increasingly more complex and risky world. However, the dream of making better, more rational decisions in some ways has become a nightmare. The increasing pressure to document and evaluate parts of working life for doctors and teachers has backlashed by becoming what can be described as an aversion for these practices.

While teachers and eldercare staff are subject to increased demands from the state to provide monitoring data for various M&E systems they also question the value of this, arguing that it takes time away from teaching and caregiving (Hanberger et al., 2016; Hanberger and Lindgren, 2018).

Not much research has been reported on evaluation systems in the environmental field. In the Introduction, objective-oriented environmental evaluation systems for the SDGs were mentioned. These efforts are still in an early stage and Persson et al. (2016) suggest that some aspects need to be resolved regarding what commitments exactly should be followed up, and why. Most research on environmental evaluation is not on systems, but on stand-alone evaluation. In the next section, this research will be briefly discussed.

Environmental evaluation research

So far, research *for* environmental evaluation (developing various approaches for conducting environmental evaluation) dominates, whereas research *on* environmental evaluation (i.e., research on use, effects, and consequences of environmental evaluations) is scarce.

Challenges to environmental evaluations

There are some fundamental challenges facing environmental evaluation that must be dealt with. Firstly, there is the issue of time, scale, and space. It is recognized that environmental problems are not usually confined by geographic borders and are not easily predicted because they span large time frames (Gysen et al., 2006). Scale is depicted as a central issue in dealing with environmental problems and a challenge to environmental evaluations. This is because environmental problems transcend social and political borders and therefore become hard to manage as the effects from an intervention may have consequences in remote places (Bruyninckx, 2009). The practical implication of this is that local or regional environmental problems quickly become national and transnational in character, and vice versa, further complicating the evaluation of effects from policies and programs.

Secondly, environmental effects over time are rarely linear. Instead, they often display discontinuities and threshold effects, which makes them hard to grasp and predict (Hildén, 2009). This issue is especially germane to environmental evaluation because, here, a balance must be struck between social and ecological time frames, as environmental evaluations must be timely in both policy and ecological cycles. Furthermore, as Rowe (2012) argues, social and ecological systems and time frames must be kept apart because they adhere to different logics. In some cases, environmental evaluations even need to address the fact that social and natural values may be in conflict (Scolobig et al., 2008). The

complex composition of environmental interrelations (social and ecological) can therefore be difficult to accommodate in evaluations.

Issues concerning measurement are a central challenge in any evaluation concerned with effects, but the problem could be said to be somewhat different and even more challenging for environmental evaluation, given how environmental effects extend over time and space. Stakeholders' value systems regarding ecological and social aspects also constitute a fundamental challenge to environmental evaluation. Different stakeholders may value natural resources differently, which can cause conflict (Scolobig et al., 2008). For example, a forest may be valued as the place for bird watching, walking, fishing, or driving a dirt bike. While all stakeholders may value the forest in general the particulars that are valued in the forest can invite conflict between the different stakeholders. In addition, an environmental evaluation needs to separate the effectiveness of one environmental policy instrument from others that are also in place (Hildén et al., 2014). This is, however, a challenge for policy evaluation in general.

Dealing with risk and innovation is another problem for environmental evaluation. Innovation is an important part of mediating and adapting to environmental concerns (Hildén et al., 2014). Therefore, environmental evaluations need to be able to evaluate innovative and new policy instruments in terms of their effectiveness. To be able to do this, environmental evaluations must be able to address ways of reducing risk when it comes to new and innovative policy instruments (Hildén et al., 2014). This is not an easy task for any evaluation, but it might be even more complex in the context of environmental evaluation because time and scale often display non-linear effects. A policy instrument could be effective in the short term but less effective in the long term, or vice versa.

Given these challenges to environmental evaluation, a number of methods and approaches have been developed, which will be discussed in the next section.

Research for environmental evaluation

Research *for* environmental evaluation includes developing methods for, and approaches to, conducting evaluation. *Modus narrandi* is an evaluation method/approach that focuses on closing the gap between measuring effects in a rigorous way and embracing the complexity involved in environmental issues (Gysen et al., 2006). Because of the non-linearity of environmental effects of interventions when it comes to scale and time, Gysen et al. (2006) suggest that effects should be narrated in a consistent effect line and effect chain. The method focuses especially on accounting for both the effects of a policy and its side effects.

Rowe (2012) suggests that evaluations that include both natural and social aspects should not inevitably require increased rigor in the evaluation process. Instead, priority should be placed on the perspective of decision makers and stakeholders. Rowe suggests that usability should be the main concern for environmental evaluation. Following this approach, the knowledge produced is geared towards users of environmental evaluations rather than to the internal validity of the evaluation.

Along similar lines, Lehtonen (2005) proposes combining evaluation for accountability with learning and argues that problems of performance measures can be eliminated this way. He underlines that peer review evaluation, combining evaluation for accountability and learning, could be an effective mechanism for promoting SD. This method attempts to deal with a common problem in evaluation between the two functions of evaluation: accountability and learning. Both functions are needed to promote SD, but evaluations often end up focusing on either accountability or learning, according to Lehtonen (2005).

Multi-criteria evaluation is yet another method/approach focused on dealing with complex environmental evaluation objects with different values at stake (Scolobig et al., 2008). The central tenet behind multi-criteria evaluation is to incorporate all conflicting values and criteria in the method to produce a broad picture of the context at hand. This differs from classic cost-benefit analysis as it does not render all the criteria in monetary terms, which in turn can provide a broader and more complex image of what is at stake (Crabb and Leroy, 2012).

Ekins and Medhurst's (2006) approach to environmental evaluation, which entails the three dimensions of SD, is of special interest for this thesis. It was used to inform the analysis in paper I. They suggest that four types of capital indicators can be used to measure the impact of programs for SD. "Manufactured capital indicators" (e.g., infrastructure for transport, telecommunications), "environmental capital indicators" (e.g., energy consumption, water pollution), "social capital indicators" (e.g., trust, income equality, participation in planning processes), and "human capital indicators" (e.g., wellbeing, education, and motivation) are used to measure changes in the four "stocks of capital" and the "flows associated with them." Largely, these indicators reflect the three dimensions of sustainability, where manufactured capital indicators represent economic sustainability, and social and human capital indicators depict social sustainability.

Taken together, environmental evaluations, informed by the cited research, should produce knowledge that accounts for complexity (including aspects of time and space), stakeholder perspectives, accountability and learning, and social, economic, and environmental dimensions of SD. Whether, and how, these

aspects are met in Swedish national environmental evaluation is an empirical question that paper I attempts to answer.

Research on environmental evaluation

A few studies have been conducted on environmental evaluation. A topical study on climate policy evaluation in Europe is one instance (Huitema et al., 2011). Drawing on a synthesis of 259 climate policy evaluations from six European countries and the EU, Huitema and colleagues (2011) analyzed the evaluations and their characteristics. They found that by far the most common criteria in the evaluations were goal achievement and effectiveness, followed by efficiency and cost-effectiveness. They also found a considerable gap between evaluation theory and practice (ibid.). The gap between evaluation theory and practice had been recognized previously in the evaluation literature (Khakee, 2003).

Evaluation theory highlights that the application of multiple criteria and the use of various evaluation methods are needed to recognize complexity. However, the vast majority of evaluations included in the referred study did not sufficiently observe complexity (Huitema et al., 2011). In addition, the majority of evaluations did not question formal policy goals and were therefore not regarded as reflective, an important aspect of environmental evaluation. The overwhelming majority of the evaluations did not include stakeholder participation in developing criteria or methodology. Furthermore, discussions and reflections on evaluation use were not included in any of the evaluations.

Sandin et al. (2017: 1813) reviewed 30 evaluations in Sweden that looked at policy instruments for energy efficiency in buildings, and concluded that the design, methods, and data used did not include aspects of “side effects, attribution, rebound effects, and triangulation.”

Environmental evaluation systems and their contribution to SD remain underresearched. This provides one of the principal rationales for this thesis. Firstly, the thesis aims to contribute to research for evaluation by developing a framework for analyzing environmental evaluations (presented in the next section) and by developing a narrative synthesis for synthesizing knowledge from evaluations and their contribution to SD (paper I). Secondly, it aims to contribute to research on environmental evaluation by generating knowledge about the use, usability, and influence of evaluation, specifically in the context of national environmental evaluation systems set up for the achievement of national environmental objectives and transition towards SD (all four papers). As research on environmental evaluation (and evaluation systems) is scarce, the thesis will largely refer to research on evaluation in general. Building on the cited research,

in the next section I develop a framework to critically explore environmental evaluations (systems) in their political context.

Framework to critically explore environmental evaluation (systems)

This section presents the analytical framework developed for the thesis. Evaluation is understood as a social and political phenomenon and consequently there is a need to analyze it as such. This calls for some kind of framework, theory, or approach. As there was no existing framework that I could apply, I have developed one for the purpose of this thesis. This section presents the analytical framework developed to frame and facilitate the analysis of national environmental evaluations.

The framework aids in exploring evaluations from a governance perspective by considering the validity and relevance of environmental evaluations in their governance context. Furthermore, the framework supports analysis of the usability and use of evaluations and their contribution to SD and of the consequences of evaluations and evaluation systems including any negative consequences. Overall, the framework helps to depict a many-sided representation of environmental evaluation systems and their implications.

The analytical framework as a whole was not developed at the outset of the research process, but was successively worked out along with the papers. The framework also serves to bind together the papers into a whole entity, thereby allowing for wider and deeper analysis by drawing together evidence and conclusions from each of the papers individually. Different aspects of this framework, as discussed below, were applied in the four papers, and they are reiterated and linked here. I recognize that if more aspects of the evaluation (and evaluation systems) are uncovered the validity of the analysis of environmental evaluations increases.

Firstly, the framework takes into consideration the governance context in which evaluations are embedded. This is in line with the understanding of evaluation as a political and social phenomenon. The governance context is described in terms of ways that government can govern society and the environment. Two different governance models, “management by objectives (MBO)” and “network governance,” as well as different policy instruments, are given attention. In addition, so-called “critical program theory (PT) analysis” is applied to empirically analyze evaluation systems from a governance perspective.

Secondly, the framework illuminates and helps explore the validity and relevance of environmental evaluations. How these concepts, reflecting two key quality criteria of evaluations, are conceived and treated is discussed below. If

evaluations are not valid and relevant to stakeholders sharing responsibilities to improve the environment, their presence can be questioned already in the first place.

Thirdly, the framework is also developed to explore usability, use, and influence of environmental evaluations. It is recognized that an evaluation's usability is of importance, although evaluations may still not be used even if they have high usability. Therefore, different factors known to enhance use and the wider influence of evaluations are discussed in this framework.

Fourthly, the framework helps to explore the contribution of evaluations to SD. Different ways of interpreting SD are taken into consideration in this analysis. The different interpretations of SD are further described, as is how they are applied in the framework.

Finally, the framework helps to explore the wider influence and potential consequences of evaluations. How actors can use evaluations in this respect, and the different functions and influences that the evaluation system could have on environmental policy and work and SD are accounted for. This also includes any negative consequences that might arise.

The framework can be used either in its entirety or in part in future studies. I welcome a critical discussion about its advantages and limitations.

Exploring evaluation from a governance perspective

As already mentioned, evaluations (and evaluation systems) are here conceived and analyzed in their governance context, as a form of policy instrument and in light of prevailing governance models. Partly, this is done by using a critical program theory analysis approach. This part of the framework is mainly applied in paper II when the NEQOs and its evaluation system is explored.

It has been suggested that a shift from government to governance has taken place in the last three decades (Pierre and Peters, 2000; Jordan et al., 2005; Pierre, 2000; Rhodes, 1997). In Sweden, this has been described as a shift from a “strong state” to a more decentralized governance approach that requires coordination at multiple levels (Lindvall and Rothstein, 2006; Premfors, 1998). However, this does not necessarily imply a reduced or non-essential role for the state (Reed and Bruyneel, 2010). The strength of the governance concept is “its capacity – unlike that of the narrower term ‘government’ – to cover the whole range of institutions and relationships involved in the process of governing” (Pierre and Peters, 2000: 1). Investigating the environmental policy sector, Jordan et al. (2004) suggest that governance complements, competes, and fusions with government rather

than eclipsing it. In line with this understanding, the framework departs from a polycentric approach of governance to denote the multiple levels, numerous policy instruments, and actors involved in making and implementing environmental policy, including traditional ways of governing, for example through laws and regulation.

Management by objectives

Setting up environmental objectives, such as the LIP and the NEQOs, and organizing monitoring and evaluation of the objectives, enables the state to steer and organize national environmental policy. This mode of governance is often referred to as “MBO” and is based on the assumption that an agency becomes more effective when managed by explicit government objectives (Pollitt, 2013). Agencies should be committed to achieving these objectives, but free to decide how to organize their work. They should, however, be requested to report regularly on their performance and, if the government is dissatisfied with the results, to react and take proper action. However, research on how MBO works in practice indicates that even if it does not work as intended in the public sector it is still applied and perceived as attractive to many decision makers (Brunsson, 2002; Wildavsky, 1979).

Moreover, the evaluation systems should be analyzed in view of a setting with many actors involved. That is, both the LIP and the NEQO are set up and implemented in the evolving network society. Therefore, I do not only pay attention to how the state governs the environment and environmental actors and institutions from a governmental and state-centered approach. I also apply network governance theory (developed below) to assess the role and implications of evaluations (and evaluation systems) in environmental policy making and implementation.

Network governance

This framework adopts Sørensen’s and Torfing’s notion of network governance. They define network governance as –

... a stable articulation of mutually dependent, but operationally autonomous actors from state, market and civil society, who interact through conflict-ridden negotiations that take place within an institutionalized framework of rules, norms, shared knowledge and social imaginaries; facilitate self-regulated policy making in the shadow of hierarchy; and contribute to the production of “public value” in a broad sense of problem definitions, visions, ideas, plans and concrete regulations that are deemed relevant to broad sections of the population. (2009: 236)

This definition reflects that governance in the network society is constituted by actors from different spheres and that the state is but one key actor.

The same authors have also developed a framework to assess successful network governance (Sørensen and Torfing, 2009). According to this, effective network governance starts by first defining a relevant policy problem (e.g., how to promote SD) and ends by “building capacity for future cooperation.” Sørensen and Torfing put forward six potential outcomes from a successful governance network, three of which are relevant for examination of environmental evaluations (2009: 242):

- (1) produces a clear and well-informed understanding of the often complex and cross-cutting policy problems and policy opportunities at hand;
- (2) generates innovative, proactive, yet feasible policy options that match the joint perception of the problems and challenges facing the network actors; and
- (3) ensures relatively smooth policy implementation based on continuous coordination and a high degree of legitimacy and program responsibility among all relevant and affected actors, including target groups, client advocacy groups, stakeholder organizations, public administrators, and politicians.

In this thesis, Sørensen’s and Torfing’s network governance approach is used to analyze environmental evaluations (and evaluation systems) at the national level from a stakeholder perspective, and to help understand how key actors use the evaluations in their environmental work.

The framework supports analysis of evaluation systems against two governance models, MBO and network governance. Both these governance modes are in use in Sweden and elsewhere.

Policy instruments

Sweden, like other nation states, has applied a range of “new” environmental policy instruments in addition to regulation (Jordan et al., 2004). These new instruments consist of market-based instruments such as eco-taxes, voluntary agreements, and eco-labeling.

A distinction can be made between proactive and reactive policy instruments. One way of conceptualizing proactive policy instruments is by using carrots, sticks, and sermons (Bemelmans-Vidéc et al., 2003). “Carrots” refer to economic policy instruments that either take away or hand out material resources, “sticks”

refer to different types of regulations, and “sermons” denote various kinds of information, for example information campaigns.

Monitoring and evaluation has been depicted as one policy instrument in the policy toolbox (Johansson, 2006). It is mainly seen as a tool to assess the effectiveness or efficiency of the outcome of an intervention or of the intervention process, that is, a *reactive* policy instrument (Johansson, 2006). But evaluation knowledge can also be used as a *proactive* policy instrument, as sermons intended to affect actors’ behavior and incite dialogue with the aim to take action towards SD, for example. National environmental evaluations are in this framework conceived and analyzed as both a proactive (sermon) policy instrument and a reactive policy instrument. Paper III reflects on the use of evaluations as a proactive and reactive policy instrument by exploring various measures to enhance usability.

Critical program theory analysis

To analyze the evaluation systems from a government perspective I employ a critical PT analysis. (Hanberger, 2014; Leeuw, 2003). There are numerous ways of conducting a PT analysis; and PT analysis is both an analytical approach and a method (Brousselle and Champagne, 2011). The underlying principle encompassing all PT approaches is to assess the design of a program or policy to find out if it can achieve its intended effects. In this thesis, Leeuw’s (2003) policy-scientific approach of PT is applied to analyze national environmental evaluations (systems) in Sweden. The PT is assessed against relevant theories, and, as in paper II, against MBO and network governance. The critical PT approach is developed further in the Materials and Methods section.

In sum, the framework conceives and analyzes national evaluations and evaluation systems from a governance perspective and as policy instruments (proactive and reactive). It combines a national government (MBO) and network governance perspective and applies critical PT analysis. The analytical framework also takes a critical approach, meaning that environmental evaluations (systems) can, in addition to the intended effects, have indirect, unpredictable, and negative or detrimental effects. To more fully represent, and come to terms with, the consequences of evaluations and evaluation systems and their subsequent contribution to SD, I also focus on the wider influence of environmental evaluations and evaluation systems, which will be described below.

Exploring the validity and relevance of evaluations

The governance context, as described above, frames how the evaluations’ validity and relevance are conceived and analyzed within this framework. A precondition for evaluations to be of value and to be used is that they have acceptable quality.

This includes, among other things, validity and relevance. How these two concepts are understood in this framework is described in the following. This part of the framework, together with the analysis of the contribution of evaluations to SD (see below), is mainly applied in paper I and paper IV.

Validity

Validity is often understood to determine to what extent a description or conclusion corresponds to reality and is measured by statistical methods. House (1980) expands this notion of validity to include the concepts coherence and justice. In addition, he suggests that truth should not primarily hinge on designs or correlations, but rather, on “the validity of the arguments that supported the use of the designs and correlations” (House, 2014). That is, the arguments put forward for choosing a specific design or method are as important for the validity of the evaluation as the validity of the method itself. The thesis follows this broader understanding of validity.

The framework also helps explore the validity of evaluative knowledge compiled from many evaluations with this broad understanding of validity. In paper I we developed 10 criteria for assessing the quality of evaluations to be included in synthesizing knowledge of intervention effects. Six of the criteria reflects sound design and valid information and four criteria reflect how to analyze program effects. If the evaluations score high on these 10 criteria the evaluative knowledge is considered valid. If the evaluations also provide convincing arguments for the chosen methods and evaluation design (House, 2014), the validity is considered higher.

Relevance

Relevance of evaluations is a key condition for evaluation use. The analysis of the relevance of evaluative knowledge in this framework is made from a stakeholder perspective and pays attention to four factors deemed as significant mediators for relevance: knowledge needs, stakeholder involvement, proximity, and timeliness.

To explore whether an evaluation meets different *knowledge needs*, a key relevance criterion, Ekblom's (2002) and Nutley et al.'s (2007) depiction of five types of knowledge is applied. These five types have been developed in the context of crime prevention and are here adapted for environmental knowledge needs. (1) Know about problems: this type refers to knowledge concerned with the problem at hand. In an environmental context, this relates to how environmental degradation of social and ecological systems are conceived and described, for example, including relations between actions and environmental degradation. (2) Know what works: i.e., what policies, strategies, or specific interventions will bring about the desired outcomes. This can refer, for example, to knowledge

about what measures are most effective in reducing greenhouse gases. (3) Know how to put something into practice: knowing what should be done is not the same as being able to do it effectively. This can, for example, refer to knowledge about how to conduct sustainable forestry. (4) Know who to involve: this knowledge includes information about different actors' needs as well as which stakeholders and experts should be communicated with and which should be involved in an urban development project, for example. (5) Know why: knowledge about why action is required, for example by stating the value of the environment to human welfare through ecosystem services (Farber et al., 2002). An environmental actor can require one or several of these knowledge types and the evaluation may or may not provide this knowledge.

Stakeholder involvement is also of importance when analyzing evaluation relevance because environmental evaluations often intend to influence a plurality of actors and previous research shows that stakeholder involvement promotes use (Brandon and Fukunaga, 2014). However, enhancing use by stakeholder involvement is not a problem-free enterprise. As Mickwitz and Birnbaum state,

... the wide set of stakeholders, in relation to any environmental issue, may result in such broad involvement that nobody fully buys into the evaluation; the focus gets blurred and use becomes hindered. Unfortunately, we have seen this happen too often in the environmental field. But still, evaluation approaches without any involvement of key stakeholders with contesting claims, interests, and beliefs are far more problematic. (2009:109)

In addition, stakeholder involvement can mean different things, such as power to influence the evaluation and the evaluation question, or legitimatizing the policy under way including its evaluation.

Proximity is another potential central factor for perceived relevance, particularly within evaluation systems. The idea behind evaluation systems is that placing the evaluation knowledge closer to the policy cycle will increase both relevance and use (Kusek and Rist, 2004; Leeuw and Furubo, 2008). When the evaluation system is intended to be used and presumed to influence a number of different actors there can be a difference in use between actors who are closely connected to the policy cycle and actors who are far removed. In this framework, the actors are therefore allocated to different categories (central, proximal, and peripheral) depending on their proximity to the national policy cycle.

The factor *timeliness* refers to how timely the evaluation results are disseminated to decision makers and other actors, which may also affect how their relevance is perceived. An evaluation system is intended to ensure timeliness by producing

recurring knowledge that feeds directly into the policy cycle. However, since the research on this factor primarily has been on stand-alone evaluations, not ongoing streams of evaluative information, it remains to be seen whether evaluative knowledge produced in evaluation systems are also perceived as produced in a timely manner.

Intended users of evaluations produced in evaluation systems may have different knowledge needs, be more, or less, involved in the evaluation, have different proximity to the decision-making process, and operate within different time frames. Hence, to be relevant to different actors, evaluations must meet their different needs and roles in the governance structure.

Exploring the usability, use, and influence of evaluations

This part of the framework is mainly applied in paper III (analysis of translation activities) and paper IV (analysis of key actors' use and non-use of evaluations); I also discuss further implications and consequences of evaluation in the final section of this summary (kappa).

Usability

An important factor for understanding and analyzing the value of evaluations and their implications for intended actors is how knowledge comes across or reaches its potential targets. This particular feature of evaluation utilization research has been labeled “brokering,” “transfer,” “dissemination,” and “translation,” among other things (Prewitt et al., 2012).

“Usability” of evaluation refers to how the “design of an evaluation – both its output and the way it is undertaken – maximizes, facilitates or disables its potential use” (Saunders, 2012: 422). Usability is therefore made up of two key dimensions: “the design of the vehicle of the message to maximize engagement ... but also the way in which the design of the evaluation lends itself to communicability ...” (Saunders, 2012: 429). The first dimension of usability refers to aspects of communication beyond the evaluation itself, aimed to improve evaluation use, such as presentations, press releases, summaries, discussion fora, and so on, also referred to as “translations” in this framework. The second refers to the form of the evaluation itself and the extent to which it communicates with its intended users.

It should be noted that an evaluation may be appraised to have high usability, when in fact it is not used at all, and vice versa, since how the evaluation is used depends on more factors than just the design and dissemination of the evaluation. The validity and relevance as discussed above, and the institutional context, potential user's belief system, and other mitigating factors also need to be taken

into account when assessing usability and use. The framework helps explore whether activities intended to increase usability, such as press releases and presentations, actually do increase usability. Whether prevailing dissemination activities have improved evaluations' usability, and, if so, which activities, is investigated in paper III.

Use

For environmental evaluations to be able to contribute to SD and environmental policy, they need to be used. So far, usability and factors facilitating use have been put forward. In the evaluations literature, it is recognized that evaluation use can take many forms. How different types of use are analyzed in this thesis is expanded below.

This thesis modifies Vedung's (2015) definition of evaluation use as described above, leaving out constitutive use (see Figure 1). This is because constitutive use or effects, as depicted by Dahler-Larsen (2011, 2014), are a structural rather than an actor-centered phenomenon and therefore fall in the category of influence, which is analyzed separately to better determine the wider influence of the evaluations (see 'Influence' below).

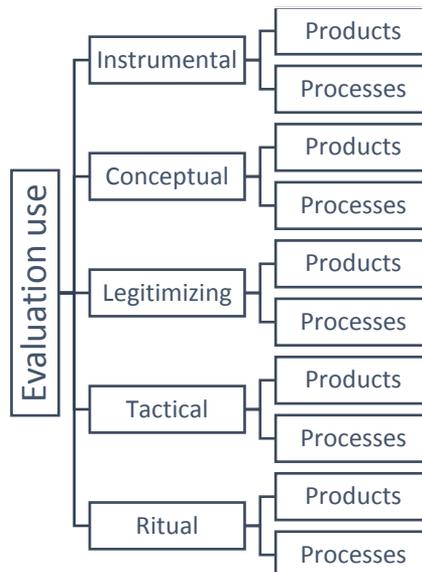


Figure 1. Types of use (omitting constitutive use), divided into processes and products.

The framework applies Vedung's (ibid.) five types of use and his distinction between process and product use is applied in the framework. As discussed above,

use is also explored in its political and social setting. The framework therefore also directs attention to the wider influence of evaluations and their contribution to actors' environmental work and SD, which goes beyond actors' actual use of evaluations.

Influence

Some researchers have argued that the concept of use only partly reflects how an evaluation affects its surroundings. Use implies an actor-oriented approach whereas evaluations can have more structural impacts on society (Dahler-Larsen, 2011). Kirkhart (2000) suggests that in order to grasp the broader impact of evaluations, influence is a more fitting concept to apply. Her framework of evaluation influence is developed "... to examine effects that are multidimensional, incremental, unintentional, and non-instrumental, alongside those that are unidirectional, episodic, intended and instrumental" (Kirkhart 2000: 7). Henry and Mark (2003) have also developed a framework based on the notion of influence. They suggest that influence can take place at different levels: the individual, interpersonal, and collective level. Evaluation influence at the individual level may entail attitude change or skill acquisition whereas influence at the interpersonal level may involve persuasion or changes in social norms. Finally, influence at the collective level, as focused in this thesis, can for example lead to changes in the political agenda or policy-oriented learning.

In the context explored in this thesis, evaluation influence may appear as a consolidation of environmental work or it may provide a common knowledge ground of interaction and learning for different actors, two examples of influence that goes beyond actors' use of evaluations. To analyze the contributions of environmental evaluations to SD, both the reported use by various actors, and the wider influence of the evaluations are accounted for.

In sum, the framework employs the concepts of validity (with specified demands), relevance (knowledge needs, stakeholder involvement, proximity, and timeliness), usability (communicative aspects of the evaluation itself and dissemination activities), use (Vedung's five types omitting constitutive use), and influence, as described above, in the empirical analysis of environmental evaluations and evaluation systems. In addition, to analyze the contribution of evaluations to SD, the concept of SD requires further conceptual clarification.

Exploring the contribution of evaluations to sustainable development

The contribution of evaluations to SD and this part of the framework was mainly applied in paper I and also in paper IV.

Notions of sustainable development

Sustainable development has been conceptualized in a number of ways. To investigate the contribution of environmental evaluations to SD, the concept of SD was scrutinized. As outlined in the Introduction, SD and the transition towards SD can be viewed in a number of ways (Hopwood et al., 2005). The thesis and framework follow the definition of SD as being built on the three pillars (ecological, social, and economic) (WCED, 1987) of sustainability and distinguishes between weak and strong notions of SD (Cabeza Gutiérrez, 1996; Nourry, 2008).

To be able to empirically analyze the contribution of environmental evaluations to the three pillars of SD, the capital approach described earlier (see “Research for environmental evaluation”) is applied (Ekins and Medhurst, 2006; Larsson and Hanberger, 2015). Contributions to ecological sustainability are conceived and measured by either halting degradation of ecological capital (e.g., species and habitats) or restoring ecological capital (e.g., forests and wetlands). Social sustainability is conceived and measured by either increasing human capital (e.g., awareness or behavioral change) or increasing social capital (e.g., democratic participation or social mobilization). Economic sustainability is understood and measured by either increased manufactured capital (e.g., environmentally friendly technologies) or cost-efficient interventions (accounting for environmental costs). An evaluation can compile knowledge of the contribution of various interventions to all or some of the three SD dimensions (ecological, social, and economic), dependent on the content of the programs in the first place.

The philosophical understanding of SD is related to how its location on a spectrum from anthropocentric to ecocentric SD is conceived (Baker, 2016). An anthropocentric view holds that the natural environment exists to service human needs, whereas an ecocentric view holds that the natural environment has intrinsic value and that the relationship between humans and nature should rest on reciprocity.

Similarly, closer to the context of environmental programs and policy, there is the division between strong and weak SD. Policies and programs that promote weak SD have economic growth as their principal goal, although environmental costs are taken into consideration (Cabeza Gutiérrez, 1996). The natural environment is seen as a resource that is readily interchangeable with other resources; in other words, natural and economic resources can restore each other. This viewpoint is best illustrated through the concept of emission rights, that is, you can exchange your economic resources for the right to pollute.

Strong SD emphasizes that protection of the environment is a prerequisite for economic development, rather than the other way around. This view also gives natural resources a higher status as they are conceived as not interchangeable, meaning that they must not be depleted. In addition, it underscores that the complexity of the natural environment calls for a more cautionary approach compared to weak SD, which assumes that depleted natural resources can be restored or replaced. Strong and weak SD also differ in their view on economic growth. Strong SD does not permit any economic growth that depletes natural resources; by contrast, weak SD has a more flexible view and accepts some environmental degradation.

A concept connected to SD is ecological modernization. As addressed in the Introduction, this corresponds to the status quo and reformist notion of SD transition. Ecological modernization leans on the idea that technological advancement can bring about SD and should therefore be seen as a central part of environmental policy. Economic and environmental goals can be integrated within the context of industrial modernity. The possibility to decouple ecological degradation from economic growth is therefore at the center of the theory. Subsequently, the state should steer business towards eco-efficient practices without undermining competitiveness. Langhelle (2000) makes a point of not conflating SD with ecological modernization, claiming that ecological modernization is less ambitious than SD and therefore a necessary aspect of SD, but not sufficient. However, in this thesis, ecological modernization is treated as a part of the weak notion of SD whereas Langhelle's version of SD is understood as a strong notion of SD.

Whether Swedish environmental policy and evaluation is guided by the weak or the strong notion of SD, and which pillar of SD is being promoted is treated as an empirical question. It is also possible that Swedish environmental policy and evaluation includes both notions, of strong and weak SD. This framework pays attention to whether a program or a policy is evaluated from a viewpoint of weak SD or one of strong SD because this impacts whether that program or policy is appraised to be successful.

The main focus in papers I and IV is on the contribution evaluations make to the three dimensions of SD. In addition, I also investigate whether the programs are evaluated from a viewpoint of a strong or a weak SD. Finally, in the concluding discussion of this summary, the relation of the evaluations to the status quo, reformist, and transformationist understandings of transition towards SD (see Introduction) is discussed.

Exploring the consequences of environmental evaluations

This part of the framework is included in the analysis of all papers and further developed later on in this summary (kappan).

To provide a wide understanding of the contributions and consequences of environmental evaluations, the framework also draws attention to both intended and unintended effects including any indirect, unpredictable, and negative aspects that may follow from evaluation and evaluation systems.

An important part of evaluating SD effects is to distinguish effects of the evaluated object (e.g., a program or policy), referred to as “intervention effects,” from other effects in order to capture the relative significance of the intervention. The framework follows Rossi and Freeman’s (1993: 221) conceptualization of intervention effects: outcomes (gross outcomes) are caused by intervention effects (net outcomes), by effects of other processes (confounding factors), and sometimes by the study itself (design effect). This is how the framework deals with the “isolation problem,” distinguishing intervention effects/SD effects from other possible effects.

The framework also accounts for unintended consequences from evaluations. One such consequence is crowding-out effects, which can, for example, mean that evaluation activities take time and other resources from other activities that could possibly contribute to SD in a more direct way. This consequence has been recognized and discussed in the evaluation literature (Dahler-Larsen, 2014) and in paper I of this thesis (Larsson and Hanberger, 2015).

Another concern related to assessing consequences and effects of environmental evaluation is, as discussed above, that different stakeholders may value different aspects of natural resources, which can lead to conflict (Scolobig et al., 2008). How stakeholders value natural resources and other SD dimensions will differ, and environmental evaluations must take this into account.

In sum, the framework has been developed to explore environmental evaluation systems and evaluations from different and complementary viewpoints. These include a governance perspective; consideration of the validity and relevance of environmental evaluations; investigation of the usability and use of evaluations; and the contribution to SD and any consequences of evaluations and evaluation systems. These viewpoints and components of the framework have been applied to address the research questions and the specific aim of the thesis, which was to explore the contributions and consequences of national environmental evaluations (systems) to environmental policy and SD in Sweden. By extension, the framework could also be applied to the examination of environmental

evaluation in other countries nationally or at the local level, or at the EU level, as a whole or in part.

Institutional setting

This section briefly describes Sweden's political system in which national environmental policies and environmental evaluations are developed, as well as the institutional setting for the two evaluation systems set up to support Sweden's national environmental policy and the transition towards sustainability. The unit of analysis is national environmental evaluation and evaluation systems in their institutional settings.

Sweden is a nation state with 21 county administrative boards and 290 municipalities which enjoy extensive autonomy with regard to implementing national policy and legislation (OECD, 2014). In the Swedish political system, environmental policy is mainly a state policy domain, meaning that the state develops the policy and uses different policy instruments to achieve what it sets out to achieve. It decides the direction by setting up environmental objectives and targets, while the municipalities and other actors are expected to implement the government's agenda. While it is recognized that municipalities can themselves also develop environmental policy, the focus in this thesis is on national environmental policy. The state can also involve the municipalities, as well as the private sector and environmental movements, in network governance and the implementation of national environmental policy.

Sweden has long had a reputation for being at the forefront of innovative environmental policies (Eckerberg, 2000; Kronsell, 1997). Even though Sweden has been at the forefront of environmental work, questions were raised around the organization and pace of the environmental work nationally in the 1990s and an overhaul of Sweden's national environmental policy was carried out (Vail, 2008). It started with the new strategy called "Ecologically Sustainable Sweden," which was launched in 1996 and which entailed three principal steps (Bäckstrand, 2010). Firstly, the existing environmental protection laws were consolidated into a single environmental code (Ekblad, 2015). Secondly, the largest environmental program in Swedish history, the Local Investment Programs (LIP), was established. Finally, the National Environmental Quality Objectives (NEQO) system was established to consolidate environmental work by organizing Sweden's national environmental policy with 16 objectives. The LIP and the NEQO system were both accompanied by a comprehensive evaluation plan. This thesis investigates the evaluation systems set up for evaluating two of the principal steps taken under the Ecologically Sustainable Sweden strategy, the LIP and the NEQO. Although the evaluations conducted as part of the LIP program do not share all elements of an evaluation system (Leeuw and Furubo, 2008) the program eventually was equipped with an evaluation system. The NEQO system was equipped with an inbuilt evaluation system from its inception.

The Local Investment Program and subsequent evaluations

This section provides a brief description of the evolution of the LIP and its evaluation system, outlining how the first explored evaluation systems appeared in its institutional setting. In 1998, the LIP, the largest and most comprehensive environmental program in Swedish history, was launched by the government to promote a major shift towards SD (Wandén, 2005; Hanberger et al., 2002). Between 1998 and 2003, the LIP allocated 4.4 billion SEK (16 billion SEK with co-financing) to 211 local LIP programs (SEPA, 2009). By 2008, all resources had been allocated, and in 2012, all LIP projects were finalized. The allocation of funds was first decided on through a special division within the Department of Environment rather than the traditional way of letting national agencies allocate funds (Hanberger et al., 2002). This can be seen as a way for the national government to remain in control over the process of allocation of funds; but the responsibility to decide which local LIP programs to fund was turned over to the Swedish Environmental Protection Agency (SEPA) after 1–2 years. The core idea of the LIP program can be described in terms of its program theory, outlining how the intended effects were to be achieved. The PT for LIP is described in paper I (Larsson and Hanberger, 2015: 95):

If the state financially supports municipalities to shift current practices towards SD and if they compete for funding, then local actors mobilize, and the most effective local LIPs are developed and funded, and then local environmental work is coordinated and improved and SD enhanced. If knowledge from best practices is disseminated, then environmental work will be improved and SD enhanced in the whole country.

The LIP explicitly focused on SD, especially the ecological dimension, and was meant to support the local level to meet the seven environmental objectives at the time (Wandén, 2005), which were to: (1) reduce the impact on the environment; (2) increase efficacy in use of energy and natural resources; (3) favor the use of renewable raw materials; (4) increase recycling and reuse; (5) contribute to strengthen biological diversity and defend cultural heritage; (6) contribute to improving the circulation of plant nutrients in a cycle; and (7) improve the indoor environment in buildings. In addition, the LIP aimed to create new jobs (economic sustainability), support participation, behavioral changes, and local capacity building, and generate other components of social capital (social sustainability) (ibid.). Promoting social sustainability was aimed to facilitate or supplement the overarching goal of ecologically SD. In this sense, the social aspects and social capital were seen as drivers of the ecological objectives, and were largely regarded as an intermediate variable (to support ecological sustainability). Examples of initiatives that received funding in the LIP

framework are housing renewal projects, establishing wetlands, and installing district heating, among many others.

The national program should be seen as distinct from the 211 LIPs, developed by each municipality in charge, with financial support from the central LIP. The national program was based on a set of assumptions and ideas borrowed from NPM (Rhodes, 1997). Local and regional actors in the public and private sphere were to compete to develop the most effective LIPs. It was also assumed that the application process would encourage local actors from different municipal departments and sectors to develop coordination of local environmental work into comprehensive regional LIPs. Knowledge and experience from the best LIPs would then be disseminated to municipalities that failed to get their applications funded.

Development of the LIP was followed by an extensive evaluation process involving at least 16 national evaluations that focused on different aspects of the disparate interventions of the programs and projects. The evaluations were coordinated but lacked some of the attributes of an evaluation system at the outset. A unified epistemological perspective is hard to find in the evaluations and since the program was time-bound the evaluations did not have a level of permanence. Most of the evaluations were commissioned by SEPA, which indicates a systems trait following the criteria that an institution or organization should be responsible for organizing and/or carrying out the evaluation activities. To a large extent, SEPA filled this role as central coordinator for LIP evaluations. Finally, the evaluations were to be linked to decisions and implementation in the organization. To what extent results of evaluations of LIPs were taken on board by SEPA has not been studied. Mainly knowledge from evaluations and successful LIP projects was displayed on a website as well as disseminated through the Council of Sustainable Development (Hållbarhetsrådet) (SOU, 2004:79) with the intention to inspire other actors/organizations to follow the example. The Council had the main responsibility for spreading knowledge from and experiences of the LIP. The LIP evaluation system was set up after the LIP grants were received; consequently, experiences of and learning from the evaluation system were not available from the beginning of the process.

The evaluations used disparate methods and covered different aspects of LIP interventions, such as wastewater management, wetland manufacturing, district heating, and resident renewal projects, among others (see paper I; Wandén, 2005).

The National Environmental Quality Objectives system and evaluations

The LIP was followed by a reconfiguration of Sweden’s national environmental goals. This was the third principal step in the new strategy. The national environmental objectives before the NEQOs were seen as disparate and in need of consolidation. The government adopted the NEQO policy framework in 1999 to make environmental work more effective, transparent and action-directed. This resulted in the development of an overarching goal and national environmental objectives, which, in turn, led to the local and regional levels developing their own objectives structured around the national objectives. The overarching goal, also called the “generational goal” (SEPA, 2015), is to “hand over to the next generation a society in which the major environmental problems have been solved, without increasing environmental and health problems outside Sweden’s borders.”

The core of the NEQO framework comprises the 16 environmental quality objectives (EQOs) shown in Table 1. These objectives indicate the desired condition of the Swedish environment and when these are achieved the generational goal is also achieved. Directly linked to the 16 EQOs, milestone targets are continuously being developed to provide direction and priorities for the ongoing environmental work in various sectors. The milestone targets are intended to direct action towards the changes needed in order to achieve the NEQOs. Regional and local environmental goals are expected to be developed by local actors in line with the national objectives and milestone targets.

1. Reduced climate impact	9. Good quality groundwater
2. Clean air	10. A balanced marine environment, and flourishing coastal areas and archipelagos
3. Natural acidification only	11. Thriving wetlands
4. A non-toxic environment	12. Healthy forests
5. A protective ozone layer	13. A varied agricultural landscape
6. A safe radiation environment	14. A magnificent mountain landscape
7. Zero eutrophication	15. A good built environment
8. Flourishing lakes and streams	16. A rich diversity of plant and animal life

Table 1. The 16 environmental quality objectives (EQOs) of the National Environmental Quality Objective (NEQO) framework. Source: SEPA (2015).

The NEQO policy framework “expresses the environmental dimension of efforts to promote sustainable development and concretizes the Swedish environmental code to promote a sustainable development that ensures a healthy and good environment for the present and coming generations” (Government Communication, 2013). Furthermore, the NEQO “signals to societal actors at large what the parliament and government want to achieve by means of environmental policies and in this way guides environmental work in all levels

and sectors of society” (Government Communication, 2013). In the NEQO system, a parliamentary committee, the Environmental Objectives Committee (EOC), is given the role of proposing strategies and milestone targets in collaboration with societal actors. This is intended to “provide good prerequisites for the broad acceptance of the strategies and milestone targets” (Government Communication, 2013).

An evaluation system was integrated into the NEQO policy framework in the early 2000s (Gov. Bill 2000/01:130, 2001; Gov. Bill 2009/10:155, 2010) to enable continuous follow-up and evaluation of the extent to which the NEQOs are met (Lundqvist, 2004). The evaluation system and its accompanying committees have been subject to changes since adoption of the NEQO. This thesis focuses on the NEQO after the overhauls made in 2010, especially the in-depth evaluation of 2015 and the annual follow-up report (see below) of 2016.

The evaluation system entails producing annual reports on environmental conditions as well as conducting in-depth evaluations every 4 years. These evaluations do not measure the direct effects of national environmental policies, but, rather, track the status of the environment; the assumption is that feasible and effective environmental policies and effective environmental work in all sectors support NEQO achievements. The Swedish Environmental Protection Agency has overarching responsibility for evaluating the NEQOs, including special follow-up of seven of the NEQOs, while seven additional central government agencies are tasked with doing a special follow-up of the rest of the NEQOs. Since 2010, annual follow-up reports have been published and two in-depth evaluations (SEPA, 2012; 2015) have been published.

The evaluations are intended to –

... provide a foundation for the governments’ politics and priorities as well as for central agencies’ planning and development of their activities. They should also provide guidance for all actors’ environmental work in Sweden. Taken together the evaluations should contribute to increasing the pace in the efforts to reach the national environmental quality objectives. (SEPA, 2015: 3)

The evaluations are intended to contribute to SD in three different ways: (1) inform the Swedish government’s office on environmental issues and thus assist them in developing sound environmental policies; (2) provide suggestions for how to carry out their environmental work to the national agencies; (3) consolidate and support all actors’ environmental work (ibid.).

The NEQO evaluations display many of the traits of evaluation systems. They have a clear epistemological perspective, produce recurring reports, and show permanence. They also have clear intended users and how the evaluation should be carried out is clearly defined.

The Swedish Environmental Protection Agency is responsible for overseeing the evaluation system while the aforementioned parliamentary committee plays an advisory and strategic role vis-à-vis the government. The EOC comprises both government and opposition ministers of parliament to ensure broad-based agreement at strategic level. In addition, the EOC includes experts and representatives of industry, central government agencies, ministries, environmental non-governmental organizations (NGOs), and local government. This brief description shows that the settings for the two evaluation systems are political and involve many actors with different roles and stakes.

Following this representation of the LIP and NEQO evaluations and their respective institutional settings, it is clear that evaluation systems can take different shapes and forms, as has been highlighted in the evaluation literature as well (Kusek and Rist, 2004; Leeuw and Furubo, 2008). A point of departure for this thesis is that the way evaluation systems are organized, steered, and managed will affect what consequences they have. Every evaluation system is based on a set of implicit and explicit assumptions of how it is intended to function. The PT of LIPs is examined in paper I and described above and the program theory behind the NEQO policy framework is discussed in paper II. Whether the evaluation systems function in accordance with these assumptions, and whether there might be other consequences of the evaluation system is an empirical question addressed in this thesis.

Materials and Methods

To answer the research questions, a mixed-method approach was applied. Below, the materials and methods used in each paper are broadly described, followed by a detailed description of the methods.

Paper I included data based entirely on documents, mainly LIP evaluations. The paper employed a narrative synthesis method to describe the knowledge derived from the LIP evaluations. A reconstruction of the PT of the LIP was also carried out.

Paper II was based on interviews and documents (mainly evaluations and policy documents). The interviews (n=9) were conducted with the aim to support the reconstruction of the PT of the NEQO evaluation system. Three of the nine interviews were conducted face to face². The documents include evaluations of the NEQOs and official government documents that were used to reconstruct the PT of the NEQO system. The paper employed a “critical PT analysis” to investigate NEQO evaluations’ and the evaluation system’s intended effects and assess their likelihood to be achieved in an MBO system.

Paper III was based on interviews (N=21), documents, and observations. The interviews were conducted by telephone. All interviews were conducted and transcribed by the author. The documents used included the NEQO evaluations as well as materials from a website (the Environmental Objectives Portal), an electronic newsletter (Environmental Objectives News), and press releases about newly published NEQO evaluations, which were analyzed as knowledge translations. In addition, observations at two Environmental Objective days, which are official conferences on the NEQO system and recorded presentations from other Environmental Objective days were observed and analyzed. The paper explored the role of knowledge translations to enhance usability of evaluations by applying a direct content analysis and a critical PT analysis.

Finally, paper IV was based on the same interviews (N=21), documents, and observations as in paper III. The interviews were conducted by telephone. All interviews were conducted and transcribed by the author. The observations were carried out at two Environmental Objective days. In addition, recorded presentations from other Environmental Objective days were analyzed. The study

²Telephone interviews are often described as a lesser good than face-to-face interviews; however, this view has been questioned in recent years (Novick, 2008). One study, which conducted the same number of telephone interviews and face-to-face interviews, found no significant difference in the data between the two interview modes (Sturges and Hanrahan, 2004).

employed direct content analysis and first and second order interpretation to analyze the contribution of the NEQO evaluations to SD.

	Documents	Interviews	Observations
Paper I	16 evaluations: Berglund & Hanberger, 2003; Rehnlund et al., 2004; Roth et al., 2004; Byman et al., 2004; Sköllerhorn & Hanberger, 2004; Svensson et al., 2004; Kärrman et al., 2004; Vredin Johansson, 2004; Forsberg, 2005; Kåberg & Jurgensen, 2005; Eckerberg et al., 2005; Stenberg et al., 2005; Birath et al., 2005; Starberg, 2005; Byman et al., 2005; Kempinsky et al., 2008		
Paper II	Policy documents: Gov. Bill 1997; Gov. Bill 2001; Gov. Bill 2010; Gov. Communication 2013; Ministry of the Environment, 2012 Evaluations: SEPA 2011; 2012; 2013; 2014	3 (I), 6 (T)	
Paper III	Evaluations: SEPA 2015; 2016 Translations: Press releases; Newsletters; Website	21 (T)	EOD, Umeå (2016) EOD, Växjö (2016)
Paper IV	Evaluations: SEPA 2015; 2016	21 (T)	EOD, Umeå (2016) EOD, Växjö (2016)

Table 2. Summary of the papers in this thesis. Complete references to the documents can be found in the individual papers. EOD = Environmental Objective day; I = face-to-face interviews; T = telephone interviews. The same interviews and observations were used for papers III and IV.

Narrative synthesis

Knowledge gained from many evaluations, as in paper I, can be synthesized in a number of ways. The method to apply depends on the evaluative knowledge in the evaluations and the purpose of the synthesis. Meta-analysis is an established way to synthesize knowledge in a number of fields (Barnett-Page and Thomas, 2009). It can be used to synthesize knowledge from evaluations providing quantitative results from primary studies to form robust and universal evidence for a single research question. This approach requires homogenous studies that are more or less identical in their methodological implementation³.

When a sufficient number of homogenous studies cannot be found, a narrative synthesis can be employed (Scott-Little et al., 2002). This approach does not statistically synthesize the results of the individual studies; instead, it narrates the findings of each study; however, it can still have a quantitative focus.

³The most notable examples of such types of studies can be found in the Cochrane Collaboration and the Campbell Collaboration. These are also for the most part based on primary studies that use randomized controlled trials (RCTs).

The benefits of applying a narrative synthesis, as opposed to meta-analysis, are usually put forward as follows: firstly, narrative reviews can provide the reader with a more comprehensible and deeper understanding of a given program and can therefore provide a more intelligible understanding of the outcomes of said program for the user (Pawson, 2002). Secondly, narrative reviews should be deployed when the studies included in the review lack methodological characteristics that would qualify them for a meta-analysis (Scott-Little et al., 2002). In other words, if the studies' outcome measures cannot be aggregated statistically. This argument is closely connected to the discourses surrounding Evidence-Based Practice (EBP) and the question whether the credibility of evidence can be hierarchically ordered (Schwandt, 2008). Narrative reviews should, following this line of argument, be conducted solely when a meta-analysis is not viable on account of the methodological characteristics of the primary studies. More often than not, when synthesizing knowledge of SD effects, the different studies do not provide enough homogeneity to perform a meta-analysis. Following this line of argument, narrative synthesis is seen as second best, not having the validity of a meta-analysis. However, from a pluralistic perspective, as adopted in this thesis, narrative synthesis compiles equally valid knowledge. In addition, meta-analysis as well as narrative synthesis can vary in quality.

Paper I used a narrative synthesis because the evaluations consist of both quantitative and qualitative data and the results from the evaluations are expressed in different ways and reflect the three dimensions of SD. Meta-analysis can therefore not be used. The paper used an established narrative synthesis method and refined it to synthesize knowledge on SD effects. It comprises a meta-evaluation, that is, an evaluation of the quality of evaluations (Stufflebeam, 2001) to single out evaluations that merit a review of SD effects. A narrative synthesis (Scott-Little et al., 2002) of SD effects is thereafter employed, based on those evaluations that have acceptable quality.

The meta-evaluation focused on the ten criteria specified in box 1. The quality assessment ranged from "not acceptable" to "excellent."

Sound design and valid information

1. Methodology: relevant for assessing intervention effects of environmental programs
 - description of applied methods/methodology, e.g., RCTs, before and after studies, comparative case studies
 - description of how the methods/methodology are applied
2. Outcome measures:
 - output measures (implemented interventions) and/or outcome measures (effect variables) provided
 - outcome data before and after the intervention provided
3. Quality of data:
 - description of type of data (primary, secondary, register)
 - description of study limitations provided
 - assessment of quality of data (primary, secondary, biased/unbiased)
4. Content of the intervention
 - description of the content of the intervention(s), e.g., information; infrastructure; renewable fuels
 - specification of content
5. Scope of the intervention
 - size of the intervention (SEK, €, \$)
 - size of the intervention, measured in the evaluation (SEK, €, \$)
 - spatial scope (within/between municipalities)
 - duration of the intervention (e.g., months)
6. Intended effects
 - ecological, social, and/or economic sustainability
 - interplay of SD dimensions
 - other intended effects

Analyses of program effects

7. Gross outcomes
 - change in outcome variables after the intervention
 - change in outcome variables before and after the intervention
8. Intervention effects
 - positive effects (e.g., process effects, short- and long-term effects) accounted for
 - negative effects (e.g., unintended/crowding-out effects) accounted for
9. Study design effects
 - positive effect (e.g., stakeholders' involvement in evaluation) accounted for
 - negative effects (e.g., skewing effects, exaggeration of intervention effects) accounted for
10. Control for spurious effects
 - control for other interventions that may affect outcome variables
 - control for changes in outcome variables caused by nature or humans
 - control for media effects, e.g., behavioral changes promoted by the media

Scoring

- = not acceptable, relevant, or absent
- 1 = partly acceptable, but can be criticized for incompleteness or vagueness
- 2 = acceptable in terms of comprehensiveness and clarity; only minor criticisms can be raised
- 3 = excellent in terms of comprehensiveness and clarity

Box 1. Criteria for assessing the quality of evaluations for the purpose of synthesizing knowledge on intervention effects. RCT = randomized controlled trial; SD = sustainable development; SEK = Swedish krona.

Sustainable development effects were synthesized in terms of the three SD dimensions analyzed using Ekins and Medhurst's (2006) capital approach (see box 2). Using narrative synthesis, the study described which parts of SD the program contributed to.

1. Ecological sustainability

- 1a. stop degradation of ecological capital (e.g., non-renewable resources, species, habitats)
- 1b. restore ecological capital (e.g., forests, biodiversity, wetlands, oceans)
(degradation/negative effects)

2. Social sustainability

- 2a. increase human capital (e.g., knowledge, awareness, behavioral change)
- 2b. increase social capital (e.g., social mobilization/public democratic awareness and participation, networks, trust)
(decreased human or social capital)

3. Economic sustainability

- 3a. increased manufactured capital/employment (e.g., renewable resources, environmentally friendly technologies)
- 3b. cost-efficient intervention (accounting for environmental costs)
(decreased capital/cost-inefficient)

+ = positive effect; 0 = no effect; - = negative or crowding-out effect.

Box 2. Criteria for synthesizing sustainable development (SD) effects of environmental programs.

The narrative synthesis applied here has the advantage that it can include evaluations using different methods and approaches. It can also account for all three dimensions of SD and provides transparency regarding the quality of the evaluative knowledge included in the synthesis, so that potential users can appraise the synthesized knowledge for themselves. How the narrative synthesis and the meta-evaluation were carried out is further described in paper I.

Program theory analysis

As mentioned, to analyze whether an evaluation system achieves what it is intended to do, I employ a critical PT analysis. Program theory evaluation and PT analysis have been developed by a number of evaluation researchers (Funnell and Rogers, 2011; Hanberger, 2014; Larsson and Hanberger, 2016; Leeuw, 2003; Rogers et al., 2000). They are often referred to as a key contribution of evaluation research. The notion of PT has been applied on policies, programs, projects, and, as in this case, evaluation systems. There are numerous ways of conducting a PT analysis (cf. Brousselle and Champagne, 2011). The underlying principle in all approaches is to assess how the program or policy is assumed to achieve its intended effects. In this thesis, I apply Leeuw's (2003) policy-scientific approach.

Reconstructing a PT using the policy-scientific approach entails using argumentative analysis (ibid; Toulmin, 1969) and probing the assumptions against relevant theories. Employing an argumentative analysis means to first reconstruct the arguments put forward in the program and then to analyze the validity of these arguments. Therefore, policy-scientific PT analysis involves two basic steps. In the first step, the PT is reconstructed with the aim to find out the underlying assumptions of a program or a policy. In other words, how the program/policy is intended to function to achieve its intended effect. The second

step is to assess the reconstructed PT to investigate possible missing links and unwarranted assumptions. This step identifies warrants in the argumentation (Toulmin, 1969). The warrant can be described as the “because” or the “since” part of an argument, as in: *A gives B because/since X*. The reconstructed PT can be used to describe the core of a program/policy or an evaluation system. The first step has a value in its own right, but a critical PT analysis includes the second step. The second step entails three parts: (1) determining the logical consistency of the PT; (2) determining the support for the PT in relevant empirical studies; and (3) determining to what extent the PT focuses on factors that are possible to manipulate. These three parts can, in other words, be described as different ways of determining the validity of the PT.

The thesis uses PT in two distinct ways: as an analytical tool in papers I and III (step 1) and as the central methodology in paper II (steps 1 and 2). In paper I, PT was used to reveal the logic behind LIP, that is, the assumptions as to how LIP can accomplish intended outcomes. In paper III, PT was applied to demonstrate the logic of how the translation activities are intended to increase the usability of the evaluation knowledge in the NEQO system. In both these studies, the PT analysis involved step 1. In paper II, a PT analysis of Sweden’s NEQO system was conducted following Leeuw’s (2003) policy-scientific approach as adopted by Hanberger (2014). This consists of reconstructing the PT (step 1) and assessing its consistency and evidence support (step 2). The PT analysis also included an empirical assessment of the extent to which the NEQO system translated into practice is congruent with the assumptions of the NEQO’s PT (a third step). This part of the analysis corresponds to Funnell and Rogers’ (2011: 473) “causal analysis of congruence.”

Directed content analysis

Content analysis is one of the several qualitative methods available for analyzing data and interpreting its meaning (Elo et al., 2014). It contains several analytical approaches with different theoretical and methodological points of departure. Hsieh and Shannon (2005) differentiate between three distinct approaches: conventional, directed, and summative. This thesis employs directed content analysis in papers III and IV.

Directed content analysis sets out to validate or extend a theoretical framework or theory. Existing theory or research is applied to help focus the research question (Hsieh and Shannon, 2005). Directed content analysis was applied by letting the research questions, theory, and previous research findings guide the interview guide, coding process and analysis of the interviews, documents, and field notes from the observations. The interview guide, further described in paper III and IV was structured around a series of open-ended questions followed by a

targeted question. The directed content analysis in paper III was carried out to identify, characterize, and appraise the contribution of knowledge translations. In paper IV, directed content analysis was carried out, primarily based on interviews. In addition, a first and second order interpretation was applied (Yanow, 1999). “First order interpretation” refers to how interviewees interpret the event they are asked about. Interviewees provide statements, interpretations, and arguments that help answer the purpose and research questions. “Second order interpretations” are the researcher’s interpretations of the first order interpretations. They also include the researcher’s own observations and reflections informed by previous research. Although second order interpretations were made in all papers they have mainly been made while analyzing the wider influence and consequences of the evaluations, in paper IV and in this summary (kappa). These methods and the study materials are further described in detail for the individual papers.

Results – summary of papers

Paper I. Effects on sustainable development from large environmental programs: a review of 16 evaluations

Published in the Journal of Integrative Environmental Science, 2015

Authors: Magnus Larsson and Anders Hanberger. ML conceived the topic of the paper. ML and AH jointly developed the methodological and theoretical framework. ML and AH jointly prepared and revised the manuscript.

The purpose of this paper is to develop criteria for meta-evaluation and narrative synthesis of effects of large environmental programs on SD. First, the paper develops the criteria and then it demonstrates how these can be applied in the case of the Swedish LIP targeting SD, through a review of 16 national evaluations.

The review shows that the program's effects on SD are mixed. Although the positive effects dominated, the program also generated some negative and crowding-out effects. Overall, the program contributed to halting the degradation of ecological capital and also, in small measure, to restoring ecological capital. The program also had a significant positive effect on human capital and on social capital in municipalities that already had strong local environmental work before the program was implemented. Furthermore, the LIP contributed to increased manufactured capital and enhanced cost-efficiency.

However, the LIP also had negative effects on local environmental work in municipalities that failed to get their programs financed. In addition, the LIP had a crowding-out effect on innovation. The program's focus on high ecological effectiveness had diminished the possibilities for innovation. Hence, the LIP promoted the use of established, rather than innovative, technologies.

A methodological conclusion was that the quality of evaluations in terms of judging SD effects varied, but all evaluations lacked some important information for undertaking sound evaluation of effects, for example concerning the scope of intervention and how effects were assessed. More than half of the evaluations scored low on intervention effects, a key criterion in the review. Knowledge had to be compiled from evaluations using varying quality and methods, and with different perspectives on SD. If a very high quality standard had been adopted, there would be no knowledge to compile. One way of managing the inherent challenges of grasping and assessing SD effects on diverse evaluations is to apply a systematic and transparent methodology (Gysen et al., 2006). This also allows the reader to consider limitations and shortcomings and take them into account

when assessing the conclusions of the review. The knowledge generated from this review can be used when synthesizing knowledge from other large environmental programs, and can support the planning of future evaluations and facilitate a well-informed discussion about SD effects on large environmental programs. In summary, reviews of SD effects need to compile knowledge from evaluations that use different approaches with varying quality. What is defined as acceptable quality for evaluations to be included in the review determines the basis for, and comprehensiveness of, the conclusions drawn on SD effects.

Paper II. Evaluation in management by objectives: A critical analysis of Sweden's national environmental quality objectives system

Published in Evaluation, 2016

Authors: Magnus Larsson and Anders Hanberger. ML conceived the topic of the paper and collected and analyzed the data. ML and AH jointly developed the methodological and theoretical framework. ML and AH jointly prepared and revised the manuscript.

This paper investigates what can be achieved by one particular MBO system – the Swedish NEQO system and its evaluation function. A critical PT analysis is first developed to reconstruct the PT of the NEQO system. Next, the robustness of the PT is analyzed in terms of internal consistency, theoretical support, and empirical support.

The results indicate that the validity of the NEQO system can be questioned on several counts because it includes insufficient activities involving all intended users and provides insufficient knowledge relevant to, or actionable for, all actors. Although the NEQO's PT has some validity at the national level, its consistency (i.e., internal validity) is low when it comes to the assumption of what the evaluative knowledge can achieve. The evaluative knowledge produced in the evaluation reports is intended to be useful for actors operating at different levels and in different sectors, but the knowledge is not relevant to, or actionable for, all actors, particularly not local and private sector actors.

Furthermore, the NEQO system is not aligned with key demands of the applied theories, MBO, SD, and network governance, and is therefore not likely to achieve its objectives. The activities intended to build political consensus around the NEQO system and address the long time frames of environmental processes have some support in MBO research and SD theory. However, many of the activities for achieving the objectives lie outside the NEQO system, as some of the involved problems are primarily global whereas others are best addressed at the local

governing level and in the private sector where the NEQO system has low relevance. Furthermore, the NEQO system is unable to achieve its objectives within its system because it has no mandate to reallocate resources.

The empirical assessment of how the NEQO system works in practice (i.e., assessment of congruence) further supports the conclusion that the system and its inbuilt evaluation system do not operate according to the demands of MBO or network governance. Furthermore, the consensus-making activities, the EOC, and the collaboration group are inadequate, and do not reflect the essence of network governance. Applying the MBO model in the public sector is bound to fail because it is unrealistic and seems unfeasible in this context, as the conditions for public environmental governance differ radically from those of corporate governance.

Moreover, though the NEQO system contributes to somewhat more effective decision making at the national level, it does not do so at all levels, for several reasons: excessive time and resources are spent developing objectives; objectives and resources are separated; and evaluations provide little knowledge on what constitutes effective environmental policy and work. The monitoring and evaluation system built into the NEQO system matches the policy process in Swedish public administration, but does not match ecological processes. Environmental changes supporting SD are measured indirectly and all environmental changes are assumed to result from all actors' policy work. This, however, does not show which policy work is effective and which is not effective.

Paper III. Translations of knowledge from an environmental evaluation system – do they contribute to usability of evaluation knowledge?

Submitted.

Author: Magnus Larsson.

To promote use of evaluation knowledge, translations are often applied. This paper sets out to investigate different types of knowledge translations and how they affect the usability of evaluation knowledge. The object of investigation is five translation activities of evaluation knowledge: exchanges of knowledge at one national and one local Environmental Objective day, a website, press releases, and an electronic newsletter. The perceptions of translation activities were explored through interviews (N=21) with key actors, and participant observations and analysis of documents.

The paper explores whether translations have enhanced the usability of evaluations within an evaluation system (the Swedish NEQO system). An overall conclusion is that the assumptions underpinning translations in the NEQO system, referred to as “the PT of translations,” have limited support. Three additional conclusions are drawn. Firstly, the translations have highlighted different knowledge in the evaluations; most have omitted or altered some knowledge from the evaluations, mainly knowledge about what works and who to involve.

Secondly, the contribution of the translations to increasing usability was perceived as minor by most actors and varied depending on the actors’ position in relation to the policy cycle. Peripheral actors, from the business sector and environmental movements, at whom the translations are primarily aimed, perceived the translations as having less value compared to central actors connected more closely to the policy cycle. Timeliness of evaluation knowledge and the translations increased the usability for central actors in the policy cycle, but not for other actors. In addition, many actors were unaware of the translations or disregarded them for other reasons.

Finally, the limited value of translations regarding usability of evaluations relates to the perceived lack of relevance of evaluations in the first place and also poor communication quality.

The peripheral actors stated that one of the reasons for not engaging with (reading or participating in) the translations was that they perceived them to lack relevance for their own environmental work. This reflects the problems with attaining the objective of the evaluation systems, that is, reaching disparate actors with different knowledge needs.

An additional problem has to do with the relevance of the evaluations. It seems a waste of time to try to create interest in evaluations through translations if the evaluations, or the system they are embedded in, are seen as providing knowledge of no practical use for the targeted actors’ environmental work.

It is recognized that the translations also had other values, such as keeping actors informed about the NEQO system in general. The translations also worked as a “push notification,” which kept some actors up to date on the activities within the NEQO. The local and national Environmental Objective days were to a large degree well received by the actors, but not primarily for increasing usability of the evaluations. Instead, the opportunity to network and exchange experiences was perceived as the primary virtue of the Environmental Objective days.

Paper IV. Stakeholders' use of environmental evaluations (system) and their contribution to sustainable development

Submitted.

Author: Magnus Larsson.

This paper investigates to what extent, and how, a national evaluation system can contribute to SD. The paper draws on previous research on evaluation use and SD. The data consists of interviews with key actors, the author's observations at seminars where the evaluations were discussed, and analysis of evaluation reports. The actors are divided into three categories (central, proximal, and peripheral) to indicate their closeness to the national policy cycle. The results show that most interviewed actors used the evaluations to some extent. The most striking difference in use was that only one out of five peripheral actors used the evaluations, while all central actors and all but one proximal actor used them to some extent. The evaluations were used in a number of ways, with conceptual and instrumental use being the most common. Product use was generally found to be more common than process use.

The reasons why the evaluations were used or not used differed between the actors. Central actors primarily found timeliness to be an important factor in using the evaluations, while proximal actors identified stakeholder involvement as a primary reason for using the evaluations. One factor mentioned by central, proximal, and peripheral actors alike as a reason for using the evaluations was credibility, which extended beyond the credibility of the evaluation approach and analysis. The actors cited the importance of political consensus around the NEQO system, which made actors (especially politicians) take notice and use the evaluations more than they would have done otherwise.

Finally, the contribution of the evaluations to SD was mainly their ability to consolidate and create continuity around the environmental work and ecological SD. Without the NEQO system and its built-in evaluation system, the environmental work would be less coordinated, because many activities and working groups are centered on producing and discussing the evaluations. In addition, the evaluation system provides continuing attention to the progress towards SD in two ways: firstly, it provides recurrent information on developments in the Swedish environment over time. Secondly, it continuously focuses attention on the state of the environment, environmental objectives, and environmental work. This can be compared to the function of an alarm that recurrently reminds actors to think about and act to address what are claimed to be the most pressing SD issues.

However, the evaluation system also has some drawbacks. Several actors noted that producing the evaluation reports takes a lot of time and resources that could perhaps be used more effectively. If we link this criticism to the statement, made by several actors, that decision makers already have enough knowledge to serve as a basis for action and that the real problem is not that we do not know what to do, we could argue that the emphasis should be placed on implementing what we already know. Seen in this light, the evaluations should focus more on evaluating implementation of prioritized actions.

Conclusions

The conclusions and subsequent discussion are guided by the two research questions posed in this thesis:

Under what conditions, in the context of evaluation systems, can environmental evaluations contribute to sustainable development?

Have environmental evaluations (and evaluation systems) contributed to national actors' environmental work and SD and, if so, how have they contributed?

Regarding the first research question, the overall conclusion drawn is summarized as follows, and further explained below:

To contribute to SD in the context of evaluation systems and network governance, environmental evaluations need to be of sufficient quality and meet different stakeholders' knowledge needs.

Paper I addresses research question 1 by compiling SD effects from environmental evaluations carried out within a large environmental program, the LIP. To manage the inherent challenges of grasping and assessing SD effects from diverse evaluations, a narrative synthesis was conducted. This is a systematic and transparent methodology (Gysen et al., 2006) that also allows the reader to consider limitations and shortcomings and take those into account when assessing the results of the review. When synthesizing knowledge on intervention effects the quality of the evaluations must first be assessed. Following House's (1980; 2014) wider notion of validity, invoking transparency in the methodology enables a "generosity principle" to be applied when including evaluations in a synthesis. Subsequently, trustworthiness of the results from a review increases with increased transparency. The review also becomes more usable since the user is able to consider the limitations of the review more clearly.

The quality assessment of the evaluations showed that all evaluations lacked some important information for undertaking a sound evaluation of the effects, e.g., concerning the scope of intervention and how the effects are to be assessed. For example, more than half of the evaluations scored low on intervention effects, a key criterion in the review. Hence, the meta-evaluation may result in identifying few evaluation studies that merit for compiling knowledge about the program's effects on SD from evaluations.

Paper II found that the relevance of the evaluation system and of the evaluations developed in the NEQO system was perceived as low by peripheral actors. The evaluations failed to provide sufficient knowledge for all actors. The evaluations partly met central and proximal actors' knowledge needs, but the evaluation reports were not relevant to, or actionable for, environmental movements and private sector actors.

Furthermore, the paper recognized that applying the MBO model in the public sector is bound to fail because it is unrealistic and seems unfeasible in this context, as the conditions for public environmental governance differ radically from those for corporate governance. The evaluation system is more likely to be valid, trustworthy, and usable if MBO is replaced with some kind of network governance and if conditions for effective network governance are created (Sørensen and Torfing, 2009). The NEQO system and the evaluation reports do not produce "a clear and well-informed understanding of the often complex and crosscutting policy problems and policy opportunities at hand;" nor do they create "a high degree of legitimacy and program responsibility among all relevant and affected actors." Consequently, they fail to meet important prerequisites for network governance (Sørensen and Torfing, 2009: 242).

Paper III also showed that the translations did not contribute to increase usability for peripheral actors because they were of little relevance to them. The translations largely failed to meet these stakeholders' knowledge needs and did not create interest in the evaluations themselves. Paper IV showed that actors closer to the national policy cycle found the evaluations more relevant; by contrast, actors further removed from the center were less inclined to appraise the evaluations as relevant.

In sum, validity and relevance (i.e., knowledge need, proximity, stakeholder involvement and timeliness) are key quality criteria and conditions for evaluations to contribute to SD. However, the validity of the LIP evaluations, assessed against the quality criteria in paper I, indicates uneven and low quality. The relevance of NEQO translations and of the evaluation system and evaluations was perceived as high among central and proximal actors, but not among peripheral actors (paper IV). Hence, only some LIP evaluations have the potential to contribute to SD, and the NEQO evaluations mainly contributed to central and proximal actors' sustainability work.

The main conclusion that can be drawn from the second research question is summarized as follows, and further explained below:

The main value of national environmental evaluations and translations is that they reinforce the national objectives, provide a recurrent report on

achievement of objectives, and push/urge actors to take responsibility to improve their environmental work. Furthermore, the NEQO evaluations (and evaluation system) contribute to support the environmental policy work of central and proximal actors towards ecological SD.

As mentioned, paper III shows that the actions taken to increase the usability of the NEQO evaluations through translations have some positive effects for central and proximal actors but largely fail to impact the principal target groups, i.e., the stakeholders farthest removed from the national policy cycle. This in turn limits their contribution to SD since they are not successfully disseminated to important stakeholders such as the business sector and environmental movements. It is, however, recognized that the translations do have other values, such as keeping actors informed about the NEQO system in general. The translations also worked as a “push notification” for some actors, keeping them up to date on the activities within the NEQO system. The local and national Environmental Objective days were largely well received by the actors, though not primarily for increasing usability of the evaluations. Instead, the opportunity to network and exchange views and experiences was perceived as the primary virtue of the Environmental Objective days.

The peripheral actors stated that one of the reasons for not engaging with (i.e., reading about or participating in) the translations was that they perceived them as having little relevance to their own environmental work. They also described lack of relevance of the evaluations themselves for their environmental work. An additional explanation given for not engaging with the translations is that the communication quality was poor. Actors from the business sector, in particular, stated that the communications of the evaluation knowledge were unappealing and failed to create interest in the evaluations. Therefore, lack of relevance of the translations and the evaluations themselves, and poor communication quality explain why peripheral actors did not think the translations helped improve usability. It seems a waste of time to try to create interest in evaluations through translations if the evaluations, or the system they are embedded in, are seen as providing knowledge of no practical use for the targeted actors’ environmental work.

Paper IV shows that the NEQO evaluations have mainly been used by central and proximal actors and that they meet their knowledge needs the most. The most striking difference in use is that only one out of five peripheral actors used the evaluations, whereas all central actors and all but one proximal actor used the evaluation to some extent. This difference is closely aligned with the findings of how translations were perceived to enhance usability.

The evaluations were used in a number of different ways. Instrumental and conceptual use were the most common types of use, but other types of use (tactical, legitimizing, and ritual) were also found. It is therefore clear that the evaluations are indeed used in several different ways, with the exception of peripheral actors who use the evaluations to a very limited extent. Product use was generally found to be more common than process use; however, proximal actors used the evaluations process more compared to central and peripheral actors. The evaluations mainly contributed by informing and facilitating central and proximal actors' environmental work. This can be interpreted as a contribution to maintain the NEQO system, which also supports decision makers' environmental policy work and their monitoring of achievement of environmental objectives.

Why the evaluations were used, or not used, was associated with different factors and conditions. Central actors primarily found timeliness to be an important factor for using the evaluations. Proximal actors pointed to stakeholder involvement as a primary reason for using the evaluations. One factor that was present in the reasons given by central, proximal and peripheral actors', for using the evaluations was credibility of the evaluations and how they were made. By and large, the majority of the actors interviewed found the evaluations credible and trustworthy. In addition, the credibility of the evaluations extended beyond the approach and analysis. The interviewed actors also cited the importance of political consensus around the NEQO system. The parliamentary consensus in this respect made actors (especially politicians) take notice and use the evaluations to a larger extent.

Reasons for limited use and non-use also varied among the actors. Central actors primarily cited lack of interplay between the three dimensions of SD as a reason for non-use, whereas proximal actors primarily said that the evaluations evaluated known facts. The annual follow-up report was criticized for not providing new knowledge, which was due to ecological systems showing few changes over a year. Peripheral actors mentioned other knowledge sources as the principal reasons for not using the evaluations. Other knowledge sources were described to be more accessible and to better meet their own knowledge needs.

The framework pays attention to the wider influence of evaluations. It seems that the main contribution of the evaluation system to SD should be looked for in terms of its coordinating effects. Without the NEQO system and its built-in evaluation system, the environmental work would be less well coordinated because many activities and working groups are centered on producing and discussing the evaluations. In addition, the evaluation system provides continuity in two regards. Firstly, it provides knowledge about the developments in the Swedish environment over time. Secondly, it continuously places focus on the

national environmental objectives and prevailing environmental work in relation to the state of the environment. This can be compared to the function of an alarm clock that recurrently reminds actors to think and act on the most pressing issues regarding SD. However, paper II also illustrates that the evaluation system and the NEQO system lead to ineffective decision making and public management (MBO) because excessive time and resources are spent on developing interim and milestone objectives and on paperwork, and there is a potential conflict between achieving environmental objectives and resources spent on evaluation activities.

In sum, the national environmental evaluations and the two evaluation systems provided knowledge about ecological SD and limited evaluation knowledge concerning social and economic sustainability. The LIP evaluations provided some knowledge on these two dimensions of SD, but not the NEQO evaluations. The results on the NEQO system demonstrate the inability of the evaluations to engage with and contribute to peripheral actors' environmental work. While business actors, environmental movements, and local actors are peripheral in terms of decision making at the national level, they can be considered key actors in SD work. Both evaluation systems have unintended negative effects and crowding-out effects. Negative or crowding-out effects of LIP were identified in the LIP evaluations. Whether these evaluations eventually contributed to manage some of these negative and crowding-out effects of LIP is not known. The NEQO evaluation system's main contribution was to consolidate environmental work and provide guidance towards an ecological SD transition at the national level. Further, it contributed by continuously reminding stakeholders about the NEQOs and everyone's responsibility to implement them.

Concluding discussion

The wider question of whether evaluation knowledge can inform environmental policy and work and help bring about sustainability can only partly be answered by this thesis. Although the contribution of the evaluation systems and evaluations is shown to have limited relevance and value for some actors working in the environmental policy field, their value and contribution to SD should not be entirely dismissed. The main contribution of the NEQO evaluations and NEQO evaluation system is that they reinforce a common vision and provide recurrent reminders of the national environmental objectives and the achievement of the objectives. These recurrent reminders demonstrate the government's expectations and that different actors should take their share of the responsibility for the transition towards sustainability. A limitation is that the evaluation knowledge produced in the NEQO system mainly serves central and proximal actors' (but not peripheral actors') knowledge needs.

Still, the NEQO evaluation system follows many of the ten steps set up by Kusek and Rist (2004) for a sound monitoring and evaluation system. It partially engages stakeholders, sets up baseline performances and milestone targets, and monitors the state of the environment, measured against the objectives. However, the NEQO evaluation systems fails in accommodating all potential users' knowledge needs and only partially provides "feedback on the outcomes and consequences of government action" (ibid., 12). It does not achieve the third objective of the evaluation system, to "consolidate and support all actors' environmental work" (SEPA, 2015: 9).

Whether the reason for this failing is due to poor implementation and organization of the evaluation system or to expectations of what the NEQO evaluation system can achieve being unrealistically high can be answered here. As shown in paper II, the evaluation system's PT is only partly consistent and shows little support in the applied research. This indicates theory failure or a combination of theory and implementation failure (Weiss, 1998a). This implies that, to avoid theory failure and to better contribute to SD, the knowledge needs and involvement of peripheral actors should be integrated in the evaluation system in accordance with the prerequisites for network governance (Sørensen and Torfing, 2009) rather than with the prerequisites of MBO or a rational model for an evaluation system (Kusek and Rist, 2004). To revise the NEQO system to better correspond to the ten steps put forward by Kusek and Rist or to MBO would demand additional resources and could potentially create unintended and crowding-out effects. Again, to avoid theory failure it needs to be developed to meet the knowledge needs in network governance. Furthermore, it is recognized that evaluations are always value-laden and will therefore adhere to a certain SD

perspective. Developing an M&E system relevant for supporting SD from different perspectives is most likely unattainable.

Regarding the evaluative knowledge produced by the two systems, there is an advantage with the LIP evaluations compared to the NEQO evaluations. The LIP evaluations evaluate interventions (the programs and projects included in LIP) directly whereas the NEQO evaluations primarily measure the state of the environment, with only limited links to intervention effects. The NEQO evaluation system is based on the assumption that changes in the state of the environment are caused by environmental policy and work, but this is an assumption that could be questioned. National environmental policy and different policy instruments do not alone change the environment, and, as shown, not all action improves the environment. Even though the LIP evaluations also have limitations they have the possibility to provide knowledge that is more actionable or usable as they measure direct effects of interventions, and some also indirect or crowding out effects.

This thesis has explored two evaluation systems at the national level, both of which have shown shortcomings in their design and achievements. Evaluation knowledge, produced with other, more practice-oriented evaluations, can possibly be more valid and relevant for informing environmental policy and work and helping bring about SD. For example, evaluations closely connected to organizations' daily operations may produce more actionable knowledge. Perhaps evaluations generating "evidence-based knowledge," a knowledge requested by some actors in the environmental field today, could be more useful in the transition towards a sustainable environment? This knowledge is less likely to be relevant to other actors as there is little evidence-based knowledge that can be used to guide environmental work in different contexts. The referred study by Dahler-Larsen (2006) shows that this knowledge is questioned by actors on the grounds that they do not find it relevant, and it can therefore not be expected to be used extensively to support SD. There is also the question of what is regarded as evidence-based knowledge.

The evaluations and evaluation systems investigated in this thesis include disparate types of evaluative knowledge, supporting different interpretations and aspects of SD. As established in the Conclusion, most of the evaluative knowledge concerns the ecological dimension of SD while to some extent omitting the two other dimensions of SD; this is especially apparent in the NEQO evaluations. Both the NEQO evaluations and the LIP evaluations primarily support weak sustainability over strong sustainability, even though exceptions can be found. Returning to the discussion in the Introduction, the evaluative knowledge from the two evaluation systems primarily supports *status quo* and *reform* perspectives on SD and only provides limited support for the *transformationist*

perspective. The majority of the evaluations do not question the prerequisites of the programs and policies or the existing power relations and governance structures which, according to the transformationist perspective, need to be altered to bring about SD. One exception is the LIP evaluation “In the shadow of LIP” (Forsberg, 2005), where focus is placed on the municipalities that were not given funding, and how their environmental work regressed on account of not being included in the LIP. This is an example where one prerequisite of the program, competition for funding, was critically explored in the evaluation.

In the NEQO evaluations, several actors noted that developing the evaluation system takes a lot of time and resources that perhaps could be used more efficiently elsewhere. If we link this criticism to the statement by a number of actors, that there is already enough knowledge on hand for decision makers to act upon and that the real problem is to implement what we already know needs to be done, we might argue that the evaluations should rather focus more on implementation of interventions known to be effective. However, such interventions still need evaluation because they could be poorly implemented and thus demonstrate implementation failure (Weiss, 1998a). Furthermore, evaluations of new innovative environmental interventions (policies or programs) could possibly contribute with new knowledge and ultimately more to SD than current evaluations.

Contributions of the thesis

This thesis contributes methodologically and empirically to research on evaluation systems in general, and to evaluation systems in the environmental field in particular. Methodologically, the framework contributes to studying environmental evaluation systems and evaluations from different viewpoints to arrive at a many-sided representation of evaluations in the governance context. The framework pays attention to the governance context with embedded evaluation systems and evaluations that highly affect how well the system performs. Focusing on the governance context contributes to a better understanding of evaluation as a political and social conduct and of how the evaluation system works in practice, and how evaluations are used and contribute to SD. It also addresses and accounts for potential side effects or detrimental effects of evaluations and evaluation systems. Leaving these out would not give a fair picture of the consequences and influence of the evaluations and evaluation systems.

A narrative synthesis was developed to compile knowledge from evaluations of the contribution of environmental programs to SD (paper I), which accounts for both positive and negative and crowding-out effects. A PT analysis was developed (paper II) to investigate what an evaluation system can actually achieve. The

framework supports a critical analysis of the assumptions underpinning the evaluation system and of the system's validity against different governance theories. Attention was also given to how to explore whether translation activities can improve the usability of evaluations (paper III), and to reasons for use and non-use of evaluative knowledge as well as the wider influence of evaluations and the evaluation system (paper IV). All these analytical and methodological tools help provide novel insights into evaluation systems.

The framework also contributes by addressing that evaluations are value-laden and are tools for different kinds of transition. Policies and actors that adopt the status quo perspective on SD transition, assuming that no fundamental changes to society are needed and that novel knowledge is unnecessary, will not need new evaluative knowledge. They would be happy with existing knowledge that can justify small steps towards a sustainable society. By contrast, actors viewing SD transition from a transformative perspective will demand further evaluation knowledge that can be used to radically and fundamentally change behaviors and practices that are causing current environmental problems. A few LIP evaluations serve as examples of evaluations that could be used this way.

Empirically, the thesis contributes to knowledge about which factors and conditions promote and restrain evaluation use in evaluation systems. Evaluations and evaluation systems have the potential to contribute to SD, but the actual contribution is dependent on numerous factors and conditions, such as the quality of the evaluations (paper I), the evaluation systems' governance context (paper II), and whether the evaluations are usable (paper III) and, ultimately, used in ways that contribute to SD (paper IV). Moreover, the evaluative knowledge must support action that alleviates environmental problems and other action towards sustainability. The contribution made by evaluations and evaluation systems to SD is therefore highly dependent, not only on the quality and relevance of evaluations, but also on how they are acted upon by the environmental actors. Generally, if evaluations contribute to replacing ineffective policies, programs, and measures and give way to more environmentally friendly and sustainable practices they will be part of the solution. If not, they may be a waste of time.

The empirical material in this thesis includes interviews, documents, and observations that together provide a many-sided representation of evaluation systems and evaluations. If more, or different, actors would have been interviewed, the results could have differed somewhat. Interviews with other politicians, government officials, actors from national agencies, the business sector, and environmental movements could have strengthened or modified the findings. Perhaps some of the translation activities could have created more interest in the evaluations among local actors, for example. As described, the

evaluation systems are moving targets and in change. If the current evaluation system is revised and studied in, say, 3 or 5 years, a forthcoming study may not arrive at the same results, because the evaluation system and upcoming evaluations may be more in line with and support a transformative perspective on SD, for example.

Research on environmental evaluation systems has so far been scarce. This thesis has begun to fill a knowledge gap. The theoretical framework was applied in an analysis of two national environmental evaluation systems from multiple viewpoints in a Swedish governance context in change. The framework was developed along with the papers and the thesis, but also with the intention to be used for exploring other environmental evaluation systems. Further research on environmental evaluations and evaluation systems at the national, regional, and local level in other countries, and at the EU and international level, would expand the knowledge on environmental evaluation systems. Compiling knowledge from other environmental evaluations by using the narrative synthesis developed in paper I could also include compiling knowledge of the interplay between the different pillars, or dimensions, of SD. Understanding the interplay (synergies as well as goal conflicts) between social, economic, and ecological aspects is crucial to the understanding of a transition to SD.

References

- Alkin MC (2013) *Evaluation roots : a wider perspective of theorists' views and influences*. Thousand Oaks: SAGE Publications.
- Allen M, Barros V, Broome J, et al. (2014) *IPCC Fifth Assessment Synthesis Report - Climate Change 2014 Synthesis Report*. Intergovernmental Panel on Climate Change (IPCC).
- Argyris C and Schön DA (1978) *Organizational learning : a theory of action perspective*. Reading, Mass.: Addison-Wesley.
- Baker S (2006) *Sustainable Development*. 1 edition. London: Routledge.
- Baker S (2016) *Sustainable Development*. 2 edition. London: Routledge.
- Baker S and Eckerberg K (2008) *In Pursuit of Sustainable Development: New Governance Practices at the Sub-national Level in Europe*. London: Routledge.
- Barnett-Page E and Thomas J (2009) Methods for the synthesis of qualitative research: a critical review. *BMC Medical Research Methodology* 9: 59.
- Bemelmans-Videc M-L, Rist RC and Vedung E (2003) *Carrots, sticks & sermons : policy instruments and their evaluation*. Comparative policy analysis. New Brunswick, N.J. ; Transaction.
- Brandon PR and Fukunaga LL (2014) The State of the Empirical Research Literature on Stakeholder Involvement in Program Evaluation. *American Journal of Evaluation* 35(1): 26–44.
- Brousselle A and Champagne F (2011) Program theory evaluation: Logic analysis. *Evaluation and Program Planning* 34(1): 69–78.
- Brunsson KH (2002) Management or Politics – or Both? How Management by Objectives May be managed: A Swedish Example. *Financial Accountability & Management* 18(2): 189–209.
- Bruyninckx H (2009) Environmental evaluation practices and the issue of scale. *New Directions for Evaluation* 2009(122): 31–39.
- Bäckstrand K (2010) *Environmental politics and deliberative democracy: examining the promise of new modes of governance*. Cheltenham: Edward Elgar.
- Cabeza Gutés M (1996) The concept of weak sustainability. *Ecological Economics* 17(3): 147–156.

- Caplan N (1979) The Two-Communities Theory and Knowledge Utilization. *American Behavioral Scientist* 22(3): 459–470.
- Cousins JB and Leithwood KA (1986) Current Empirical Research on Evaluation Utilization. *Review of Educational Research* 56(3): 331–364.
- Crabb A and Leroy P (2012) *The Handbook of Environmental Policy Evaluation*. London: Earthscan.
- Dahler-Larsen P (2006) Organizing knowledge: evidence and the construction of evaluative information systems. In: *From studies to streams: managing evaluative systems*. Comparative policy evaluation series; 12. New Brunswick, N.J.: Transaction Publishers, pp. 65–80.
- Dahler-Larsen P (2011) *The Evaluation Society*. Stanford, California: Stanford University Press.
- Dahler-Larsen P (2012) Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable. *Scandinavian Journal of Public Administration* 16(3): 29–46.
- Dahler-Larsen P (2014) Constitutive Effects of Performance Indicators: Getting beyond unintended consequences. *Public Management Review* 16(7): 969–986.
- Du Pisani JA (2006) Sustainable development – historical roots of the concept. *Environmental Sciences* 3(2): 83–96.
- Eckerberg K (2000) Sweden: progression despite recession. In: *Implementing sustainable development: strategies and initiatives in high consumption societies*. Oxford: Oxford University Press, pp. 209–244.
- Ekblad G (2015) *Ett kompendium om miljöbalken (SFS 1998:808): för en hållbar utveckling* [The environmental code: for a sustainable development]. Stockholm: Lagtexten.
- Ekblom P (2002) From the source to the mainstream is uphill: The challenge of transferring knowledge of crime prevention through replication, innovation and anticipation. *Crime Prevention Studies* 13: 131–203.
- Ekins P and Medhurst J (2006) The European Structural Funds and Sustainable Development A Methodology and Indicator Framework for. *Evaluation* 12(4): 474–495.
- Elo S, Kääriäinen M, Kanste O, et al. (2014) Qualitative Content Analysis: A Focus on Trustworthiness. *SAGE Open* 4(1): 1–10.

- Farber SC, Costanza R and Wilson MA (2002) Economic and ecological concepts for valuing ecosystem services. *Ecological Economics* 41(3): 375–392.
- Fitzpatrick JL, Sanders JR and Worthen BR (2012) *Program evaluation: alternative approaches and practical guidelines*. Upper Saddle River, N.J.: Pearson Education.
- Forsberg B (2005) *I skuggan av lokala investeringsprogrammet: kommunerna som inte beviljades stöd samt synen på staten i LIP:s efterföljd [In the shadow of LIP]*. Evaluation reports / Umeå Centre for Evaluation Research, 1403-8056; 16. Umeå: Umeå Centre for Evaluation Research.
- Funnell SC and Rogers PJ (2011) *Purposeful program theory: effective use of theories of change and logic models*. San Francisco: Jossey-Bass.
- Furubo J-E, Rist RC and Sandahl R (eds) (2002) *International Atlas of Evaluation*. New Brunswick, U.S.A: Transaction Publishers.
- Government Bill 1997/98:45 (1997) Miljöbalk [The environmental code]. Available at: <http://www.regeringen.se/sb/d/108/a/1352> (accessed 13 November 2014).
- Government Bill 2000/01:130 (2001) Svenska miljömål - delmål och åtgärdsstrategier [Sweden's National Environmental Quality Objectives – milestone targets and implementation strategies]. Available at: <http://www.regeringen.se/sb/d/108/a/1197> (accessed 13 November 2014).
- Government Bill 2009/10:155 (2010) Svenska miljömål - för ett effektivare miljöarbete. [Sweden's National Environmental Quality Objectives – for a more effective environmental work]. Available at: <http://www.regeringen.se/sb/d/12166/a/142456> (accessed 13 November 2014).
- Government Communication 2013/14:145 (2013) Svenska miljömål visar vägen! [Sweden's Environmental Quality Objectives shows the way!] Available at: http://www.riksdagen.se/sv/Dokument-Lagar/Forslag/Propositioner-och-skrivelser/Svenska-miljomal-visar-vagen_H103145/?html=true (accessed 13 November 2014).
- Guba EG and Lincoln YS (1989) *Fourth Generation Evaluation*. Newbury Park, Calif.: SAGE.
- Gysen J, Bruyninckx H and Bachus K (2006) The Modus Narrandi A Methodology for Evaluating Effects of Environmental Policy. *Evaluation* 12(1): 95–118.

- Hajer MA (2003) *Deliberative Policy Analysis: Understanding Governance in the Network Society*. Cambridge: Cambridge University Press.
- Hanberger A, Eckerberg K, Brännlund R, et al. (2002) *Lokala investeringsprogram: en förstudie inför utvärderingen* [The local investment program: a pilot study before the evaluations]. Evaluation reports / Umeå Centre for Evaluation Research, 1403-8056; 10. Umeå: Umeå Centre for Evaluation Research.
- Hanberger A (2011) The real functions of evaluation and response systems. *Evaluation* 17(4): 327–349.
- Hanberger A (2014) What PISA Intends To and Can Possibly Achieve: a critical programme theory analysis. *European Educational Research Journal* 13(2): 167-180.
- Hanberger A (2016) Evaluation in Local School Governance: A Framework for Analysis. *Education Inquiry* 7(3): 217-236.
- Hanberger A and Lindgren L (2018) Evaluation systems in local eldercare governance. *Journal of Social Work*: 1468017318760788.
- Hanberger A, Carlbaum S, Hult A, et al. (2016) School evaluation in Sweden in a local perspective: A synthesis. *Education Inquiry* 7(3): 349-371.
- Henry GT and Mark MM (2003) Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions. *American Journal of Evaluation* 24(3): 293–314.
- Hildén M (2009) Time horizons in evaluating environmental policies. *New Directions for Evaluation* 2009(122): 9–18.
- Hildén M, Jordan A and Rayner T (2014) Climate policy innovation: developing an evaluation perspective. *Environmental Politics* 23(5): 884–905.
- Hood C (1995) The “new public management” in the 1980s: Variations on a theme. *Accounting, Organizations and Society* 20(2): 93–109.
- Hopwood B, Mellor M and O'Brien G (2005) Sustainable development: mapping different approaches. *Sustainable Development* 13(1): 38–52.
- House ER (1980) *Evaluating with validity*. Beverly Hills, Calif.: SAGE.
- House ER (1993) *Professional evaluation: social impact and political consequences*. Newbury Park, Calif.: SAGE.
- House ER (2014) Origins of the Ideas in Evaluating with Validity. *New Directions for Evaluation* 2014(142): 9–15.

- Hsieh H-F and Shannon SE (2005) Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15(9): 1277–1288.
- Huitema D, Jordan A, Massey E, et al. (2011) The evaluation of climate policy: theory and emerging practice in Europe. *Policy Sciences* 44(2): 179–198.
- Højlund S (2014a) Evaluation use in evaluation systems – the case of the European Commission. *Evaluation* 20(4): 428–446.
- Højlund S (2014b) Evaluation use in the organizational context – changing focus to improve theory. *Evaluation* 20(1): 26–43.
- Jasanoff S (1990) *The fifth branch: science advisers as policymakers*. Cambridge, Mass.: Harvard Univ. Press.
- Johansson V (2006) *Tillsyn och effektivitet: statliga inspektörers yrkesroller och strategival* [Inspection and efficiency: state inspectors professional roles and strategies]. 1 edition. Umeå: Boréa.
- Johnson K, Greenesid LO, Toal SA, et al. (2009) Research on evaluation use a review of the empirical literature from 1986 to 2005. *American Journal of Evaluation* 30(3): 377–410.
- Jordan A, Wurzel RKW and Zito AR (2004) *New Instruments of Environmental Governance? National Experiences and Prospects*. London: Routledge.
- Jordan A, Wurzel RKW and Zito A (2005) The Rise of ‘New’ Policy Instruments in Comparative Perspective: Has Governance Eclipsed Government? *Political Studies* 53(3): 477–496.
- Khakee A (2003) The Emerging Gap between Evaluation Research and Practice. *Evaluation* 9(3): 340–352.
- Kirchhoff CJ, Lemos MC and Dessai S (2013) Actionable Knowledge for Environmental Decision Making: Broadening the Usability of Climate Science. *Annual Review of Environment and Resources* 38(1): 393–414.
- Kirkhart KE (2000) Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation* 2000(88): 5–23.
- Kronsell A (1997) *Greening the EU: power practices, resistances and agenda setting*. Lund political studies, 0460-0037; 99. Lund: Lund Univ. Press.
- Kusek JZ and Rist RC (2004) *A handbook for development practitioners : ten steps to a results-based monitoring and evaluation system*. Washington, DC: World Bank.

- Lafferty WM and Eckerberg K (eds) (1998) *From the earth summit to local agenda 21: working towards sustainable development*. London: Earthscan.
- Langhelle O (2000) Why ecological modernization and sustainable development should not be conflated. *Journal of Environmental Policy & Planning* 2(4): 303–322.
- Larsson M and Hanberger A (2015) Effects on sustainable development from large environmental programmes: a review of 16 evaluations. *Journal of Integrative Environmental Sciences* 12(2): 85–105.
- Larsson M and Hanberger A (2016) Evaluation in management by objectives: A critical analysis of Sweden's national environmental quality objectives system. *Evaluation* 22(2): 190–208.
- Ledermann S (2012) Exploring the Necessary Conditions for Evaluation Use in Program Change. *American Journal of Evaluation* 33(2): 159–178.
- Leeuw FL (2003) Reconstructing Program Theories: Methods Available and Problems to be solved. *American Journal of Evaluation* 24(1): 5–20.
- Leeuw FL and Furubo J-E (2008) Evaluation Systems What Are They and Why Study Them? *Evaluation* 14(2): 157–169.
- Lehtonen M (2005) OECD Environmental Performance Review Programme Accountability (f)or Learning? *Evaluation* 11(2): 169–188.
- Liefferink D and Skou Andersen M (1997) *The innovation of EU environmental policy*. Oslo: Scandinavian University Press.
- Lindvall J and Rothstein B (2006) Sweden: The Fall of the Strong State. *Scandinavian Political Studies* 29(1): 47–63.
- Liverani A and Lundgren HE (2007) Evaluation Systems in Development Aid Agencies: An Analysis of DAC Peer Reviews 1996–2004. *Evaluation* 13(2): 241–256.
- Lundqvist L (2004) *Sweden and Ecological Governance: Straddling the Fence*. Manchester: Manchester University Press.
- Mickwitz P and Birnbaum M (2009) Key insights for the design of environmental evaluations. *New Directions for Evaluation* 2009(122): 105–112.
- Nourry M (2008) Measuring sustainable development: Some empirical evidence for France from eight alternative indicators. *Ecological Economics* 67(3): 441–456.

- Novick G (2008) Is there a bias against telephone interviews in qualitative research? *Research in Nursing & Health* 31(4): 391–398.
- Nutley SM, Walter I and Davies HTO (2007) *Using evidence : how research can inform public services*. Bristol: Policy Press.
- OECD (2014) OECD Environmental Performance Reviews: Sweden 2014. Available at: http://www.keepeek.com/Digital-Asset-Management/oecd/environment/oecd-environmental-performance-reviews-sweden-2014_9789264213715-en (accessed 9 November 2017).
- Owens S (2015) *Knowledge, Policy, and Expertise: The UK Royal Commission on Environmental Pollution*. Oxford: Oxford University Press.
- Owens S, Petts J and Bulkeley H (2006) Boundary work: knowledge, policy, and the urban environment. *Environment and Planning C: Government and Policy* 24(5): 633 – 643.
- Patton MQ (2008) *Utilization-Focused Evaluation*. Thousand Oaks, Calif.: SAGE Publications.
- Pawson R (2002) Evidence-based Policy: The Promise of ‘Realist Synthesis’. *Evaluation* 8(3): 340–358.
- Persson Å, Weitz N and Nilsson M (2016) Follow-up and Review of the Sustainable Development Goals: Alignment vs. Internalization. *Review of European, Comparative & International Environmental Law* 25(1): 59–68.
- Pierre J (2000) *Debating governance: authority, steering, and democracy*. Oxford: Oxford University Press.
- Pierre J and Peters BG (2000) *Governance, politics and the state*. Basingstoke : Macmillan.
- Pollitt C (2013) The logics of performance management. *Evaluation* 19(4): 346–363.
- Premfors R (1998) Den komplexa staten. In: *Stater som organisationer*. Stockholm: Nerenius & Santérus.
- Prewitt K, Schwandt T and Straf M (2012) *Using Science as Evidence in Public Policy*. Available at: </catalog/13460/using-science-as-evidence-in-public-policy> (accessed 28 November 2014).
- Reed MG and Bruyneel S (2010) Rescaling environmental governance, rethinking the state: A three-dimensional review. *Progress in Human Geography* 34(5): 646–653.

- Rhodes RAW (1997) *Understanding governance : policy networks, governance, reflexivity and accountability*. Public policy and management, 99-1698973-7. Buckingham: Open University Press.
- Rist RC and Stame N (2006) *From studies to streams: managing evaluative systems*. Comparative policy evaluation series ; 12. New Brunswick, N.J.: Transaction Publishers.
- Rockström J, Steffen W, Noone K, et al. (2009) Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecology and Society* 14(2).
- Rogers PJ, Petrosino A, Huebner TA, et al. (2000) Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation* 2000(87): 5–13.
- Rossi PH and Freeman HE (1993) *Evaluation: a systematic approach*. Newbury Park, Calif.: SAGE.
- Rossi PH, Lipsey MW and Freeman HE (2003) *Evaluation: A Systematic Approach*. Thousand Oaks, Calif.: SAGE Publications.
- Rowe A (2012) Evaluation of Natural Resource Interventions. *American Journal of Evaluation* 33(3): 384–394.
- Sandin S, Neij L and Mickwitz P (2017) Review of evaluations of policy instruments for energy efficiency in buildings in Sweden. In: *ECEEE SUMMER STUDY PROCEEDINGS*. Available at: https://www.ecee.org/library/conference_proceedings/ecee_Summer_Studies/2017/8-monitoring-and-evaluation-building-confidence-and-enhancing-practices/review-of-evaluations-of-policy-instruments-for-energy-efficiency-in-buildings-in-sweden/ (accessed 8 May 2018).
- Saunders M (2012) The use and usability of evaluation outputs: A social practice approach. *Evaluation* 18(4): 421–436.
- Schwandt T (2015) *Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice*. Stanford, California: Stanford Business Books.
- Schwandt T, Ofir Z, Lucks D, et al. (2016) Evaluation: a crucial ingredient for SDG success. IIED Briefing. Available at: <http://pubs.iied.org/17357IIED/> (accessed 18 March 2018).
- Schwandt TA (2008) Educating for Intelligent Belief in Evaluation. *American Journal of Evaluation* 29(2): 139–150.
- Scolobig A, Broto VC and Zabala A (2008) Integrating Multiple Perspectives in Social Multicriteria Evaluation of Flood-Mitigation Alternatives: The

- Case of Malborghetto-Valbruna. *Environment and Planning C: Government and Policy* 26(6): 1143–1161.
- Scott-Little C, Hamann MS and Jurs SG (2002) Evaluations of After-School Programs: A Meta-Evaluation of Methodologies and Narrative Synthesis of Findings. *American Journal of Evaluation* 23(4): 387–419.
- Scriven M (1991) *Evaluation thesaurus*. Newbury Park, Calif.: Sage.
- SEPA (2009) *Effekter av investeringsprogrammen LIP och KLIMP* [Effects of the investment programmes LIP and KLIMP]. Stockholm: Naturvårdsverket. ISBN 978-91-620-6357-3.
- SEPA (2012) *Steg på vägen* [Steps on the way]. Stockholm: Naturvårdsverket. Available at: <http://www.naturvardsverket.se/Om-Naturvardsverket/Publikationer/ISBN/6500/978-91-620-6500-3/> (accessed 13 November 2014).
- SEPA (2015) *Styr med sikte på miljömålen* [Steer towards the environmental objectives]. Stockholm: Naturvårdsverket. Available at: <http://www.naturvardsverket.se/Om-Naturvardsverket/Publikationer/ISBN/6600/978-91-620-6666-6/> (accessed 30 March 2017).
- SEPA (2016) *Miljömålen* [The Environmental Quality Objectives]. Stockholm: Naturvårdsverket. Available at: https://www.miljomal.se/Global/24_las_mer/rapporter/malansvariga_myndigheter/2016/au2016.pdf (accessed 30 March 2017).
- Shaw I, Shaw IGR, Greene JC, et al. (2006) *The SAGE Handbook of Evaluation*. Thousand Oaks, Calif.: SAGE.
- Shulha LM and Cousins JB (1997) Evaluation Use: Theory, Research, and Practice Since 1986. *American Journal of Evaluation* 18(3): 195–208.
- SOU (2004:79) *Allt ljus på storstadspolitikens lokala utvecklingsavtal? : förslag till nationellt utvärderingsprogram* [Focusing on metropolitan areas development agreements? A proposition for a national evaluation program]. Statens offentliga utredningar, 0375-250X; 2004:79. Stockholm: Fritzes offentliga publikationer.
- Steffen W, Richardson K, Rockström J, et al. (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347(6223): 1259855.
- Stufflebeam DL (2001) The Metaevaluation Imperative. *American Journal of Evaluation* 22(2): 183–209.

- Sturges JE and Hanrahan KJ (2004) Comparing Telephone and Face-to-Face Qualitative Interviewing: a Research Note. *Qualitative Research* 4(1): 107–118.
- Sørensen E and Torfing J (2009) Making governance networks effective and democratic through metagovernance. *Public Administration* 87(2): 234–258.
- Toulmin S (1969) *The uses of argument*. London: Cambridge Univ. Press.
- Vail B (2008) Ecological Modernization at Work? *Scandinavian Studies* 80(1): 85–108.
- Vedung E (1997) *Public policy and program evaluation*. New Brunswick, N.J.: Transaction Publishers.
- Vedung E (2010) Four Waves of Evaluation Diffusion. *Evaluation* 16(3): 263–277.
- Vedung E (2015) Six Uses of Evaluation. In: *Nachhaltige Evaluation? Auftragsforschung Zwischen Praxis und Wissenschaft. Festschrift zum 60. Geburtstag von Reinhard Stockmann*. New York: Waxmann Verlag, pp. 187–210.
- Wandén S (2005) *Vad vi kan lära av LIP : samlade erfarenheter av de lokala investeringsprogrammen utifrån utvärderingar gjorda under åren 2003-2005* [What can we learn from LIP - experiences from evaluations of LIP between 2003-2005]. Stockholm: Naturvårdsverket.
- WCED (1987) *Our common future*. Oxford: Oxford Univ. Press.
- Weiss CH (1979) The Many Meanings of Research Utilization. *Public Administration Review* 39(5): 426–431.
- Weiss CH (1998a) *Evaluation: Methods for Studying Programs and Policies*. Upper Saddle River, N.J.: Prentice Hall.
- Weiss CH (1998b) Have We Learned Anything New About the Use of Evaluation? *American Journal of Evaluation* 19(1): 21–33.
- Wildavsky AB (1979) *Speaking truth to power: the art and craft of policy analysis*. Boston, Mass.: Little, Brown and Co.
- World Bank (2012) *Turn Down the Heat: Why a 4°C Warmer World Must Be Avoided*. Washington D.C.: World Bank.
- Yanow D (1999) *Conducting Interpretive Policy Analysis*. Thousand Oaks, Calif.: SAGE Publications.