How to normalize reflexive evaluation? Navigating between legitimacy and integrity

Lisa Verwoerd
VU University Amsterdam, The Netherlands
PBL Netherlands Environmental Assessment Agency, The Netherlands

Pim Klaassen
Barbara J. Regeer
VU University Amsterdam, The Netherlands

Abstract
While hybrid evaluation practices are increasingly common, many Western countries continue to favor modernist evaluation logics focused on performance management—hampering the normalization of reflexive logics revolving around system change. We use Normalization Process Theory to analyze the work evaluators from a policy assessment agency undertook to accomplish the alignment between the prevailing and proposed logics guiding evaluation practice, while implementing a reflexive evaluation approach. Ad hoc alignment strategies and insufficient investment in mutual sense-making regarding reflexive evaluation hindered normalization. We conclude that alignment requires developing reflexive evaluation legitimacy in the context of application and guarding reflexive evaluation integrity, while contextual structures and cultures and reflexive evaluation components are being negotiated. Elasticity (of contextual structures and cultures) and plasticity (of reflexive evaluation components) are introduced as helpful concepts to further understand how reflexive evaluation practices can become normalized. We reflect on the use of Normalization Process Theory for studying the normalization of reflexive evaluation.

Keywords
Evaluation practice, institutional logics, normalization, Normalization Process Theory, policy evaluation, reflexive evaluation

Corresponding author:
Lisa Verwoerd, Athena Institute, Science Faculty, VU University Amsterdam, Boelelaan 1085, 1081 HV Amsterdam, The Netherlands. Email: l.verwoerd@vu.nl
Evaluation 27(2)

Introduction

Contemporary policy processes increasingly occur in complex, multi-actor, and multi-level governance contexts. This has led to a proliferation of views on the purposes and roles of policy evaluation and suitable approaches and methodologies. Evaluation literature often distinguishes schools in policy evaluation along the lines of “modern” versus “post-modern” science (Funtowicz and Ravetz, 1993), “technical” versus “deliberative” models (Owens et al., 2004), “technocratic” versus “participatory” approaches (Chouinard, 2013) or “modernist” versus “reflexive” logics (Kunseler and Vasileiadou, 2016). We endorse the rejection of such distinctions as being overly rigid, siding with scholars who point out a trend in evaluation practices toward a tailored choice of functions, methodologies, and tools in which elements of allegedly opposed schools are combined (see, for example, Green et al., 2015; Van Hemelrijck and Guijt, 2016). However, evaluation practices can hardly be expected to be free from the influence of institutional, political, and societal conventions and preferences, and we observe that there are often limits to the extent to which evaluations can be tailored to the policy issue at hand. This is confirmed by the existence of hybrid evaluation practices that manifest a continued privileging of so-called modernist evaluation approaches in many Western countries (Chouinard, 2013; Fitzpatrick et al., 2008)—that is, evaluations largely based on assumptions about the linearity of policy processes, emphasizing objectivity, accountability, and performance management (Nieminen and Hyytinen, 2015) at the cost of, for example, inclusivity, usefulness, or learning.

As the (older) modernist knowledge tradition has provided the technocratic script for the science–policy interface for decades, norms and conventions describing the appropriate function and form of evaluations often derive from modernist foundations. Policy researchers are hence faced with a dilemma: in seeking ways to adequately inform policy processes and to address complex societal problems, they are drawn toward more systemic and reflexive evaluation approaches, while not being able to entirely elude historically entrenched organizational, political, and societal expectations of modernist approaches to policy evaluation.

This article explores an attempt by evaluators from the Netherlands Environmental Assessment Agency (PBL), with a long-standing modernist tradition of policy evaluation, to normalize a more reflexive evaluation practice. To do so, we adopt a practice-perspective and combine the concept of institutional logics (Berg Johansen, 2017; Friedland and Alford, 1991) with Normalization Process Theory (NPT), a theory developed for studying the normalization of innovations of organizational practice (May and Finch, 2009). Our case comprises a series of steps in the process of normalizing a novel approach to evaluation which was manifest both in the organizational context of PBL, and in context of the nature policy program that was evaluated. We specifically focus on what normalization of a reflexive evaluation practice entails when undertaken in contexts more readily amenable to so-called modernist evaluation logics. What we found is that this process of normalization is best described as a trajectory in which evaluators continuously negotiate and navigate between two different logics, modernist and reflexive.

Background

“Logics” of evaluation and their implications

Contemporary societal issues, including climate change, global poverty, and loss of natural resources, are increasingly understood as complex or “wicked problems” (Rittel and
Webber, 1973). These require reconsideration of (in)formal rules, dominant ways of thinking and doing, problem solving and resource management, as these are in many ways part of the problem (Beck et al., 1994). Scholars have made a case for adopting a systems approach in the governance of these problems (e.g. Geels, 2004) and it is argued that such approaches should include a “reflexive perspective” as well. This means that things that are usually taken for granted are scrutinized in ways that challenge their long assumed ‘self evident’ status, thereby creating possibilities for system change (Loeber et al., 2007: 84). To support the design and analysis of policy or interventions for system change toward sustainable development, evaluation approaches have emerged that also include such a reflexive perspective (e.g. Van Mierlo et al., 2010).

Reflexive evaluation diverges from modernist evaluations, which in their essence present an instrumental tool for warranting accountability and compliance. Following a linear inputs-outcome-outputs-impact framework, such approaches tend to overlook complexity (Nieminen and Hyytinen, 2015) and fall short in drawing attention to systemic properties that delimit the issue at hand (Arkesteijn et al., 2015). Modernist evaluation logics seem to dominate policy and program evaluation practices of many Western countries (Chouinard, 2013; Fitzpatrick et al., 2008; Nieminen and Hyytinen, 2015), despite rising calls in academic literature for more complexity-oriented and reflexive evaluation in research and attempts to do so in practice (Patton, 2010; Van Mierlo et al., 2010).

To gain understanding in the limited uptake of reflexive evaluation, we draw inspiration from social practice theorists, including Giddens (1984), Schatzki (2002) and Shove (2010). Rather than individual people and institutions and their opinions or intentions, or the surrounding social structures, we take practice itself as the basis focus of inquiry: the observable, collective and organized behaviors and actions that people purposively and routinely perform and consider to be “normal” ways of doing (Nicolini, 2012; Reckwitz, 2002). In this view, the limited uptake of reflexive evaluation approaches is not seen as the result of individuals’ intentions or beliefs being hampered by contextual barriers, but as a manifestation of institutionalized social practices (Warde, 2005). Scholars have argued that different evaluation practices can be viewed as the material embodiment of different “institutional logics” (Dahler-Larsen and Schwandt, 2012; Kunseler and Vasileiadou, 2016), which can broadly be understood as sets “of material practices and symbolic constructions [that] constitute organizing principles” (Friedland and Alford, 1991: 248) within particular institutions. Such logics operate as behavioral guidance, as they supply actors with the “rules of the game” (Jones et al., 2015). These logics are not static but are practiced and shaped at nested, interacting levels: among individuals and teams within organizations, at the organizational level and within their wider societal context. Scholars have demonstrated how—sometimes contradictory—institutional logics may be at play and create friction, giving rise to plurality and potential space for institutional change, and the emergence or development of novel practices (Berg Johansen, 2017).

Scholars have rejected an empirically discernable dichotomy between modernist and reflexive evaluation approaches (or, for example, “modern” vs “post-modern” science; Funtowicz and Ravetz, 1993) and have pointed out that such schools are rarely practiced in their pure form (Owens et al., 2004; Vaidya and Mayer, 2014). Indeed, theoretical understandings of evaluation are often equally hybrid, such as Beck et al.’s (1994) ideal of reflexive modernization which predominantly veers toward the reflexive side of the spectrum but which also contains modernist elements. While we endorse this view, for the purpose of this study, we distinguish between a modernist and a reflexive institutional logic of evaluation. Each is...
observable in the enactment of distinctive material practices and the discursive and non-discursive traces of complexes of beliefs, norms, and rules about what evaluation is, “should do,” and how this is best achieved (Kunseler and Vasileiadou, 2016). We propose to understand these logics as useful—though overstated—generalizations, much like Weberian “ideal-types” (Shils and Finch, 1949), to conceptualize the logical space within which evaluation practices can exist and from which practitioners may draw. A clear-cut conceptual distinction between modernist and reflexive evaluation is used as an epistemic tool to gain analytical depth and better understand the influence of these implicit logics on policy evaluation practice. In their hypothetical and ideal-typical form, the institutional logics of evaluation are rooted in fundamentally different epistemologies and political theories. We argue that analyzing the significance of these underlying differences helps to understand the challenges practitioners experience when attempting to practically reconcile them.

The modernist logic can be traced back to 18th-century Enlightenment ideals and 19th-century ideas of technocracy and rationality, as well as even older ideas on scientific objectivity, value-free science, and impartiality (Kunseler and Vasileiadou, 2016; Lentsch and Weingart, 2011). Built around epistemologies that present reality as objectively knowable (positivism, empiricism) and as substitute for religion and tradition, scientific knowledge was argued to be the best foundation for public decision-making. The highest quality of scientific facts could be obtained through an independent and bounded scientific process, free from social values. Modernist logic treats science and policy as strictly separated domains and the resulting science was believed to linearly advance progress and public welfare (Dahler-Larsen, 2012; Jasanoff, 2011). This modernist outlook on scientific knowledge and its proper relations with policy was further institutionalized with the rise of new public management in the 1980s as Western governments increasingly called for accountability and performance measures to enhance the performance of the public sector (Chouinard, 2013; Pollit et al., 2007). In this context, evaluation primarily focuses on regulation and compliance, while serving as a management instrument to assure accountability. Ideally, evaluators provide empirically confirmed and logically consistent statements to enable evidence-based decision-making, keep the domains of science and policy separated, and prevent the infringement of scientific fact with values (Sternlieb et al., 2013). Gold-standards for evaluators based on a modernist logic consist of objectivity, scientific autonomy, impartiality, and detachment from societal values and politics. Its preferred methods include technical-analytical processes that follow logic-model thinking, such as impact assessments, cost-effectiveness analyses, and modeling (Kunseler and Vasileiadou, 2016; Verwoerd et al., 2019) (Table 1). Second, the article considers a reflexive logic of evaluation, which gained increasing traction over the past 20 to 30 years. Linearly advancing public welfare had been demonstrated to have unintended (negative) side-effects, thereby disclosing the complex and uncertain interdependencies of ecological, social, economic, political, and institutional processes that modernist approaches are arguably poorly equipped to understand (Duijnhoven and Neef, 2016). Drawing on social-constructivist epistemologies (pluralism, relativism, pragmatism), reflexive logic builds on the premise that scientific knowledge is not produced in isolation, but is deeply intertwined with cultural understandings of socio-economic and socio-ecological relations (Jasanoff, 2011). Science and policy are considered inevitably entangled domains. Ideally, their relations are open, transparent, and deliberative, enabling social learning and the co-production of socially robust answers to complex problems. Evaluation approaches explicitly geared toward engaging with these complex problems take a system perspective (Moore et al., 2019; Patton, 2010) and aim
to support policies or interventions by stimulating their reflexivity on their relationship to these systems (Arkesteijn et al., 2015) and enhancing their ability to challenge various systemic (societal, institutional, political) contexts to allow for system change (Beers and Van Mierlo, 2017). Given the limitations of their ability to produce ultimate “truths” to dictate public decision-making, evaluators working from the reflexive logic take on an attitude of humility and organize inclusive and productive interactions on the science–policy (and society) interface to facilitate interactive learning and learning-by-doing. Norms and principles include inter-subjectivity (moving from “one objective knowable truth” to “understanding the world together”; De Jaegher et al., 2017), acknowledgment of perspective plurality and uncertainty, and transparency. Reflexive logic is partial to deliberative-analytical methods that are context-sensitive (Rog et al., 2012), real-time (Marjanovic et al., 2017), and include joint fact finding, knowledge co-creation, system analyses, and reflexive monitoring.

A myriad of evaluation practices have been derived from these logics by evaluators to suit the policy issue at hand. However, when practices rooted in different institutional logics are combined, something that appears self-evident within one logic may be highly problematic from within the other. For instance, where interaction between science and policy during evaluation may be considered inherent and crucial according to reflexive logic, modernist logic

| Table 1. Modernist and reflexive evaluation logic (adapted from Kunseler, 2017; Verwoerd et al., 2019). |
|-------------------------------------------------|-------------------------------------------------|
| **Modernist logic** | **Reflexive logic** |
| “Speaking truth to power by bridging the gap” | “Enhancing reflexivity for system change by organizing productive interactions and interactive learning” |
| **Epistemological foundations** | **Perspective on science–policy interface** | **Purpose of evaluation** | **Role of evaluators** | **Norms, principles, and disposition** | **Methodology and tools** |
| Positivism, empiricism, scientism | Technocratic and bureaucratic relations between science and policy; science and policy as strictly separated domains | Evaluation is as instrument for accountability assurance, performance assessment, and compliance | Evaluators provide empirically confirmed and logically consistent statements to inform evidence-based decision-making. They mediate domains of science and policy and work to keep them apart | Objectivity; scientific autonomy; impartiality and detachment from societal values and politics | Technical-analytical processes that follow logic-model thinking such as impact assessments, cost-effectiveness analysis, modeling |
| Social constructivism, pluralism, pragmatism | Open, transparent, and deliberative relations between science and policy; science and policy as ultimately entangled domains | Evaluation is a mechanism to enhance the reflexivity of policy or interventions for system change to deal with complex societal problems | Evaluators facilitate processes of interactive learning, learning-by-doing and reflection on systemic properties that hamper or facilitate system change. They organize productive interactions on the science–policy interface to develop socially robust answers to complex societal problems | Inter-subjectivity, acknowledgment of perspective plurality, and uncertainty; transparency; humility | Deliberative-analytical processes that include a system thinking such as joint fact finding, knowledge co-creation, system analyses, reflexive monitoring |
would argue this to be detrimental to the scientific quality of the research because its objectivity might be compromised (Turnhout et al., 2013). Furthermore, the modernist logic, being the historically more established of the two, predominately guides evaluation practice in many Western countries, giving rise to a culture of accountability in their policy and program contexts (Fitzpatrick et al., 2008). Consequently, it is hard for both evaluators and policymakers to move away from this dominant logic toward more reflexive approaches, as this requires facing unaccommodating beliefs, norms, and rules.

Thus, although the two logics are rarely practiced in their pure form, that does not mean that it is straightforwardly clear how to combine them well. Nor how those beliefs, norms, and rules of the modernist logic that are particularly unaccommodating toward reflexive evaluation can be engaged with in such a way that reflexive forms of evaluation become more normalized. It is our understanding that when a practice “out there” is newly introduced into an organization, a certain amount of modification is required for it to align to its practitioners and their context (Fullan and Pomfret, 1977). When the practice is taken up by the formal and informal structures of an organization in a way that the practice’s original integrity is maintained and viewed as legitimate, this is considered successful normalization (May and Finch, 2009). This article empirically investigates the process through which practitioners conduct a large-scale evaluation of a nature policy program, drawing on reflexive logic in an organizational and policy context that are partial to the modernist logic. We thereby aim to make recommendations on how normalization of reflexive evaluation practices can be encouraged.

**Institutional evaluation logic at the PBL: The Natuurpact reflexive evaluation**

The empirical material on which we draw comprises the evaluation of a Dutch nature policy program called the Natuurpact, conducted by the PBL. The PBL is a public knowledge institute charged with independent, scientific policy assessments “in the fields of the environment, nature and spatial planning” (PBL, 2017). It has established itself within a technocratic, modernist paradigm and has an authoritative status on the science–policy interface (Halffman, 2009). Given its own modernist orientation, PBL actively strives to practice a more reflexive logic in giving scientific advice, by attempting to innovate its research repertoire with more deliberative and interpretative modes of research. The co-existence of different logics at the PBL and its endeavors toward a more reflexive practice are well studied (Kunseler and Tuinstra, 2017; Kunseler and Vasileiadou, 2016; Petersen et al., 2011). The Natuurpact evaluation is the organization’s first large-scale longitudinal evaluation for which the entire approach was designed following reflexive principles.

In 2013, the Dutch national government, governments of the 12 provinces, and various societal organizations signed the Natuurpact. Through this pact, nature policy became the responsibility of provincial governments, and its signatories formulated and agreed upon high shared ambitions for nature policy. It was also recorded that PBL would conduct an “ex durante” evaluation to allow for timely policy adjustments. As we discuss below, the program presented a welcome opportunity for PBL to further its reflexive aspirations.

The Natuurpact evaluation will run until 2027 and comprises three-yearly evaluation cycles. The first cycle ran from 2014 to 2017; the second (2018–2020) has just concluded. The current authors were involved as external academic experts for our knowledge and skills in participatory research, reflexive monitoring, and evaluation research. We also performed the
role of reviewers of the impact of the Natuurpact reflexive evaluation after the first cycle had
ended. We discuss our role in more detail below.

NPT

To study how institutional logics shape evaluation practice, we draw inspiration from the field
of implementation science and view the implementation of reflexive evaluation—and with it,
an underlying reflexive logic—as an innovation of PBL’s organizational practice. Already in
the early 1970s, scholars in education innovation have pointed out the importance of conceiv-
ing implementation as a process that occurs within an institutional and wider context to under-
stand why some innovations succeed to become standard practice, and why others do not
(Fullan and Pomfret, 1977; Havelock, 1970, 1971; Huberman and Miles, 1984). May and
Finch (2009) develop this further and with NPT consider the mechanisms that facilitate or
hamper the normalization of an innovation into its “host context.” NPT focuses on the work
actors do for implementation to gain an understanding of the complex dynamics of imple-
menting and institutionalizing new technologies or innovations in organizational practice
(McEvoy et al., 2014; May et al., 2018). The theory includes a practice-perspective: rather
than individual people or institutions and their opinions or intentions, the basic focus of inquiry
is the observable, collective, and organized behaviors and actions that people purposively
perform (Nicolini, 2012). As such, NPT arguably allows for an analysis of how a new practice
becomes normalized into its social context as those involved make sense of it, buy into it,
agree on how it’s done, and appraise how it has value, and potentially provides a useful theo-
retical lens to understand how context influences normalization.

Central to NPT are four core mechanisms that comprise the work implementers of an inno-
vation do to normalize it: sense-making, engagement, enactment, and appraisal work. Further-
more, May et al. (2016) underline the importance of context for implementation pro-
cesses and argue how these are the emergent outcomes of interactions and negotiations
between components of the innovation and elements of the host context. Actors who aspire a
novel practice are thus required to negotiate a level of “fit” between the prevailing standards
and ways of working, and the aspired ideals and procedures, in order for the latter to have
viability and workability. In other words, the work belonging to the four core mechanisms
largely takes the shape of what we call alignment work (Figure 1).

We hypothesize that NPT is an appropriate lens through which to study how institutional
logics at play in the contexts in which the Natuurpact evaluation was conducted, shape eval-
uation practice. Below we elaborate the NPT mechanisms in more detail and show how each
applies to the case in question, and how the need for alignment shaped the work that was done
by the research team.

Methodology

Case study

Our case study concerned the Natuurpact evaluation’s first cycle, executed by researchers
from PBL and partner Wageningen Environmental Research (WER), supported by the authors
(Athena Institute). The interdisciplinary team consisted of six researchers (including two pro-
ject leaders). Few had prior experience with deliberative or reflexive evaluation approaches.
The team met bi-weekly to discuss the evaluation’s progress and research activities.
The evaluation’s primary participants consisted of provincial and national policy actors responsible for the development and implementation of nature policy. The main form of interaction with the participants consisted of bi-monthly meetings with a selected working group of 12 representatives from provincial governments, and eight multi-stakeholder workshops that occurred over the course of the evaluation. The purpose of these workshops differed according to the evaluation phase, and included making an inventory of the evaluation needs of public actors, joint interpretation of preliminary evaluation findings, and drawing joint conclusions for action and change. The team also met with the evaluation’s commissioners (administrators from national and provincial government) twice a year to ensure the evaluation was still on track in terms of timing and budget. Figure 2 presents a schematic overview of the evaluation’s first cycle.

**Material and methods**

*Participatory action research.* The first and third authors were members of the project team as participatory action researchers: they supported the design and execution of the reflexive evaluation, and simultaneously studied this process. They were tasked with developing a theoretical framework for reflexive evaluation based on academic literature including process principles and ideal outcomes. During project meetings, the authors would draw from this framework and their previous experiences with reflexive research to support the team with
operationalizing these principles, and organize critical reflection by the team on the evaluation’s progress. The authors monitored the challenges the team encountered conducting reflexive evaluation on a so-called Dynamic Learning Agenda (DLA; Van Mierlo et al., 2010). When the first evaluation cycle had concluded, the authors also assessed the policy impact of the reflexive approach, as commissioned by the project coordinators.

**Material.** All observations made as part of the project team were recorded in field notes, including the DLA. Additional field notes were kept during evaluation activities, including workshops, and seminars that were held at the PBL to inform the organization on the project’s progress. Data were also collected through 17 in-depth interviews with members of the team at the start, after the first year and after the finalization of the first cycle. All members were interviewed at least twice over the course of 3 years. The interviews focused on their experiences with reflexive evaluation, and were flexible and open-ended in order to gain in-depth understanding of the rationale behind their actions regarding the implementation of reflexive evaluation. The interviews were audio-recorded and transcribed.

**Data analysis.** All data were analyzed by the first and second authors, corroborated by the third author. We used content analysis (Hsieh and Shannon, 2005), making use of sensitizing concepts derived from the NPT framework (the core mechanisms). The second author had not previously been involved with the Natuurpact case and brought a different perspective to help sharpen the analysis.

**Results**

The findings showed that all four of NPT’s core mechanisms could be identified and several themes emerged for which alignment work was necessary, which in turn determined how the work for each core mechanism took shape. Time and again, the team had to negotiate its reflexive approach in order to align it sufficiently with the prevailing modernist norms and customs regarding policy evaluation within their home organization and nature policy.
practice. In the following, we discuss each NPT mechanism and how each mechanism became manifest in the form of alignment work undertaken by the team.

**Sense-making work**

The first core mechanism of the NPT framework is *sense-making work*: the work practitioners do to develop a shared understanding of a new practice and why it is important in relation to other practices (May, 2015). Coinciding with PBL’s prior interest in reflexive research, the agreement in the Natuurpact that its evaluation would have an *ex durante* character led to the decision to adopt a reflexive evaluation approach. The findings show that during the first stages of the project, the work the team did to make sense of what reflexive evaluation entailed beyond a general idea of its purpose and general process was limited. In what followed, the team set the approach apart from other approaches primarily with reference to its timing and the level of interaction with its intended participants. This is illustrated by the project leader’s explanation of reflexive evaluation:

> What is really different, is that we work with policy actors immediately from the start of the evaluation: there is a lot more interaction than there would be during a regular policy evaluation. (PBL project leader, personal communication, early 2015)

Furthermore, the team differentiated reflexive evaluation on the basis of its purpose, explaining it as an approach that searches to “reconcile evaluation purposes of policy learning and accountability” to allow for transformative learning and change, as opposed to exclusively focusing either on learning or on accountability (as during responsive evaluation and impact assessment, respectively; Evaluation Plan, 2015).

As part of sense-making, the team commissioned the authors to develop a theoretical framework for reflexive evaluation (Table 2). The framework served to guide the team in their evaluation design and implementation. The idea was that its principles required further operationalization to make them work in the particular project context. However, initially the framework gained relatively little traction with the team; in a way, the team acted as if it had outsourced the work of sense-making by involving the authors as experts in reflexive research. After being given the formal go-ahead, the size and scale of the project caused the team to quickly become more occupied with “doing” the evaluation than with “thinking” about it.

**Engagement work**

We observed that the amount of work done for sense-making of reflexive evaluation increased when the team was faced with the demands of *engagement work*. NPT’s second core mechanism concerns the work actors undertake to ensure engagement and commitment of others with the new practice, for instance, through stressing its urgency and orchestrating managerial endorsement, but also through arranging sufficient capacity for others to become involved.

To encourage participation from nature policy actors as well as support from the PBL community—both vital for normalization—the team first needed to make the purpose and value of reflexive evaluation explicit in a way that spoke to the prevailing conventions on useful and high-quality evaluation, All of this constituted “sense-making” work. To do so, the team employed a strategy of reframing: developing two co-existing, mutually inclusive narratives that they used in different contexts. The first narrative emphasized the evaluation objective as...
allowing for mutual learning [between policy actors and researchers] from experiences with nature policy to timely inform policy processes and benefit the progress that is made on the nature policy ambitions [and was mostly used when communicating with policy actors]. (Evaluation plan, 2015)

The second narrative describes the benefits of reflexive evaluation in terms of “increased research quality and impact” (PBL supervisor, start-up seminar, early 2015) and was mostly confined to internal communications within the PBL organization. In both narratives, the team framed the benefits of reflexive evaluation in a way that aligned with modernist ideals on the purpose of policy evaluation within their respective contexts: to enhance program performance (context of policy practice) and to promote research quality and impact (organizational context).

This demonstrated how the need for engagement work gave rise to sense-making, and resulted in growing mutual understanding of reflexive evaluation among the team members, as well as the policy actors and the PBL community. This could be observed through a gradual decline in critical inquiry from actors from both host contexts after the approach was adopted, and by an increase in the team’s confidence in explaining what reflexive evaluation is and does. Importantly, how sense-making and engagement work were done was determined by the need to align with the dominant frames within the intended contexts to implement the reflexive approach.

The project team reflected that developing commitment from both host contexts required a significant investment in engagement work: “Convincing my constituency of the value of reflexive evaluation really required some work. I had a lot of informal chats with them, all to make them shareholders of this new approach” (PBL project leader, personal communication, 2016).

Another example of how engagement work was informed by the need to align with modernist norms may be perceived in the team’s decision to involve the authors. While the original argument for our involvement was substantive, the team also mobilized our involvement strategically: to the PBL and policy communities they presented the purpose of our involvement as “guarding the scientific rigor of the reflexive evaluation approach.” Through this strategic argument, the team used our involvement to underline the scientific integrity of the approach, adhering to the technocratic belief in truth claims that predominated within the PBL. The PBL project leader explained: “I really used [the author’s] involvement to show: look, we are serious, this is also a scientifically sound method. This was critical to convince [PBL colleagues] that this was a valid evaluation approach.”

How the engagement work to convince actors to buy into reflexive evaluation was done—by emphasizing the scientific integrity of the approach—was strongly shaped by the need to create alignment.

**Enactment work**

The third mechanism consists of *enactment work* and comprises all actions of the team to “do” reflexive evaluation. A starting point for this enactment work was the operationalization of the four reflexive evaluation principles outlined in Table 2. We observed that operationalization was again shaped by the need to align the evaluation design with both modernist customs and reflexive ideals. To explore this, we draw on the operationalization of the first principle as an example: *There is multi-stakeholder participation in the design and conduct of the evaluation.* This principle is based on two premises. First, that complex problems require various stakeholders with different perspectives to become involved with a social learning process (Patton, 2010) and, second, that participation in evaluation design and
Table 2. Excerpt from the theoretical framework for reflexive evaluation developed to guide the Natuurpact evaluation (Van Veen et al., 2016).

| Evaluation element | Characteristics |
|--------------------|-----------------|
| Evaluation phases  | 1. Developing a joint evaluation framework; |
|                    | 2. Joint research and interpretation; |
|                    | 3. Drawing shared conclusions and dissemination. |
| Guiding process principles | 1. There is multi-stakeholder participation in the design and conduct of the evaluation; |
|                     | 2. The evaluation addresses the stakeholders’ evaluation needs in light of the policy program’s ambitions and draws attention to systemic properties that hamper or facilitate change; |
|                     | 3. The evaluation facilitates productive interactions among participants and experts; |
|                     | 4. Within this multi-stakeholder context, the evaluation simultaneously meets demands for accountability and performance management, and interactive learning, and facilitates system change. |
| Outcomes           | 1. Socially robust knowledge and a knowledge-enriched policy practice; |
|                    | 2. Identification and challenging of systemic barriers and opportunities to increase the potential for system change. |
|                    | 3. Concerted action by involved stakeholders |

conduct positively influences the relevance and shared ownership over the findings in light of the program’s ambitions (O’Sullivan, 2012). Our example focuses on the latter. The idea of involving stakeholders throughout the entire evaluation evoked a strong response from both the team and the PBL community. Objections centered on the potential compromise of the team’s objectivity and independence, and it was felt the credibility of the institute was on the line. These concerns were put out formally not only during the project’s start-up seminar but also in bilateral interactions with the project leadership. As a result, participation was confined to specific evaluation phases, and the topics about which participants would have a say were limited. For instance, policy actors were involved during the first phase, during which the scope and the main research questions of the evaluation were determined. Also, during the third phase, interactive stakeholder sessions were organized during which preliminary findings were collaboratively analyzed and interpreted. The participants were not, however, allowed to discuss research methods during the second phase, nor given a say on the substance of the conclusions in the final report. Regarding research methods, one team member explained,

> It is up to us to decide what methods are most appropriate. It would do our independent judgement no good if we let policy actors decide how they want to be evaluated. We are, in the end, the experts.

The PBL project leader confirmed, “Letting them co-decide on methods, I can’t account for that. It would be like allowing butchers to inspect their own meat.” The participants considered this no issue at all: “Research methods, that is really a topic for the researchers. I would have no idea,” one policymaker reflected. Modernist conventions—within both host contexts—on the distinct and separate roles of researchers and policymakers appeared beyond the scope of negotiation, regardless of the reflexive ideal to enable participation during all evaluation phases. Instead, this principle of reflexive evaluation was stretched in order for it to retain legitimacy in the eyes of the PBL and policy communities. In its renegotiated
interpretation, the principle came to mean something akin to “participation not during ‘all’ but during ‘most’ evaluation phases.”

A similar discussion occurred concerning the dissemination of the findings. The team suggested to co-publish the final report with the participants, to underline their joint efforts. But because they required a visibly independent evaluation report for the recommendations in the evaluation to have strategic-political value, the policy actors strongly opposed co-publication. Interestingly, this was the same argument that had prevented them from co-deciding on research methods earlier on, namely that, to their constituencies and the public, their participation might be perceived as compromising scientific autonomy; objectivity; and, ultimately, the credibility of the evaluation.

In these examples, the confinement of participation to particular research phases had shaped the reflexive principle in a way that neatly fitted in with modernist ideals of scientific autonomy and political distance. At the same time, it initially appeared to hold true to the principle’s reflexive ideal: it still allowed for sufficient participation to facilitate social learning processes and generate shared ownership over its findings. However, during appraisal work, to which we turn shortly, it became evident that this principle had been stretched too far.

While the norms and procedures for objectivity and scientific autonomy were quite rigidly maintained at times, there were also moments when there was more room to contemplate how to adhere to these norms while allowing for on-going science–policy deliberations. Notably, the procedures for guarding objectivity broadened over time. Rather than maintaining literal (physical) distance at all times, the team members distributed roles (some interacted with the participants while others focused on desk research and running models) to prevent researcher bias. Furthermore, for the final report, the peer-review community was extended to policy and societal actors. In doing so, the team implicitly expanded the norm of objectivity to inter-subjectivity. While these broadened norms and procedures initially elicited critical remarks during project seminars, these remarks toned down as the evaluation progressed. Over time, convictions about the roles of evaluators and policy actors as strictly separated became less and less articulated, suggesting that actors from both contexts became more used to the changed relationship. Illustrative of this development is the wide uptake by the PBL community of the term “requester” to replace “principal” when indicating the commissioner of a project, suggesting a more equal-level footing between researchers and policy actors while acknowledging researchers’ autonomy.

Appraisal work

The fourth NPT mechanism concerns appraisal work and comprises activities that judge the value and effectiveness of a new practice during and after its enactment. In this light, the current authors were tasked with assessing the different ways the reflexive approach had had an impact on nature policy practice. This was regarded with serious formality and weight, which in itself is illustrative for the degree of buy-in to the approach by the team and its supervisors. Notably, the review implicitly served a dual purpose: first, to learn from the experiences to improve the following cycle’s execution and, second, to legitimize the continuation of the reflexive approach within the PBL community by demonstrating its value. This was manifest in our task to assess the approach to institutionally shared beliefs on what would constitute such value: enhanced research quality and policy impact. While our assignment was commissioned on the basis of a modernist logic, we strived for a more reflexive and deliberative
approach, and brought to light effects more characteristic of reflexive evaluation. These included, for instance, enhanced policy learning, a strengthened nature policy community, and a knowledge-enriched policy practice (Verwoerd et al., 2017, 2020). Such effects beyond traditional linear ideas on policy impact have found uptake within the organization and are sometimes used to discuss potential impact of new studies.

The review was critical on the operationalization of some reflexive principles, which was argued to be ineffective in some respects. We return to the example of the operationalization “multi-stakeholder participation,” which had been demarcated to specific phases due to concerns about autonomy and the credibility of the PBL at large. The review identified that findings’ relevance and usability was limited: a mismatch was observed halfway through the project between the scale at which the public actors’ evaluation needs transpired (regional) and the scale at which the computational model that had been used provided findings (national). As a result, the participants felt the findings were of limited use to inform their nature policy plans as they provided few perspectives on regional action. Regarding this mismatch, the WER project leader noted:

That this model would be used was decided upon before we had even started. [. . .] It cost us a lot of time and effort to explain and repair this mismatch to policy actors. The decision which model to use should have been informed by the demands of our intended end-users.

In retrospect, by not involving the participants in the discussion on research methods, their evaluation needs were initially only partly met. Limiting the first principle thus also compromised the second, and reduced the usability of the evaluation for policy learning and change. To remedy the mismatch, the project team undertook a significant amount of work to produce findings on a more relevant scale.

The team later reflected that while opening up the determination of the methods to policy actors had been regarded as being beyond discussion, some members in addition had defaulted into working in a “research-driven” as opposed to a “practice-driven” fashion. In hindsight, the initial shared understanding at the start of the project of what reflexive evaluation is and how “it is done” appears superficial. Different understandings of reflexive evaluation emerged during its actual implementation and materialized in the form of conflicting modes of working. The team’s preoccupation with “doing” reflexive evaluation absorbed time from thinking and discussing the approach, which allowed different modes of enacting reflexive evaluation to remain untouched and deeply embedded routine ways of working unchallenged.

Despite these difficulties, after the first evaluation cycle had concluded, the overall feeling among those involved was one of enthusiasm, and the reflexive approach was continued during the second cycle. Moreover, at the time of writing, several other large-scale reflexive evaluation projects have been initiated at the PBL, and notions such as “reflexive thinking” have found some uptake in the organization’s vocabulary.

**Discussion and conclusion**

This article aimed to empirically investigate the process through which evaluators attempt to conduct and, in so doing, normalize an evaluation practice that draws on a reflexive evaluation logic in a context partial to modernist logic. We reflect on our findings and, starting off from the idea that developing legitimacy without compromising integrity is vital to successfully
normalize reflexive evaluation, make some suggestions for evaluators seeking to conduct and normalize reflexive evaluation. Before we present these, we briefly reflect on the value of using NPT as theoretical lens for understanding and facilitating such normalization.

Over the past decades, a wide variety of views on the purposes and roles of policy evaluation have proliferated in evaluation literature, accompanied by myriad evaluation approaches and methodologies (Stern et al., 2015). Some more recent approaches share that they are stakeholder-oriented and search to better address the complexities of the phenomenon under study while contributing to its goals and ambitions (e.g. Arkesteijn et al., 2015; Verwoerd et al., 2020). Much of the literature on these emerging practices concerns their theoretical underpinnings, experiences with applications in specific cases, or practical or theoretical differences between distinct approaches (Fetterman et al., 2015; Moore et al., 2019; O’Sullivan, 2012; Rolfe, 2019). There is a dearth of literature on challenges involved with implementing novel practices in institutional settings that are not necessarily conducive to it (Chouinard, 2013; Guijt, 2010; Kunseler and Vasileiadou, 2016; Petersen et al., 2011). An increased understanding of the processes through which evaluators address these challenges may support evaluators aspiring to reflexive logic to create room for such work in contexts where modernist approaches are privileged and to do so without compromising the reflexive ideal. In response, this study analyzed the work that was done by an evaluating project team to normalize reflexive evaluation, using NPT’s core concepts.

Our findings confirmed that the core mechanisms for normalization do not occur linearly, but rather in conjunction: each mechanism caused iterations in the others, and vice versa, continuously strengthening each other (McEvoy et al., 2014). In our case, many of the challenges in engagement, enactment, and appraisal work could be traced back to lack of mutual in-depth understanding and agreement among evaluators (and participants) on how (not) to “do” reflexive evaluation. Our findings suggest that as much of the work that was undertaken toward normalization occurred relatively ad hoc and in response to urgent unaccommodating structures or aspects of political and organizational culture—such as beliefs about who has a say in determining research methods—implicit conventions and evaluation routines factually remained unchallenged. Our findings resonate with challenges identified for the introduction of participant-oriented evaluations to development projects. A lack of participatory sense-making about an appropriate evaluation approach was found to be conducive to defaulting into modernist approaches (Van Hemelrijck and Guijt, 2016). Nieminen and Hyytinen (2015) suggest this is reinforced due to limited methodological repertoires that evaluators who seek to practice more systemic and reflexive approaches tend to draw on, consequential to the institutionalization of modernist logic and the epistemological differences that hinder different logics’ hybridization. Although unrelated to reflexive evaluation per se, scholars of implementation science also underline the importance of inclusive and mutual sensemaking prior to implementation for a new practice to become successfully normalized (Mair et al., 2012).

As illustrated by the continuation of the Natuurpact reflexive evaluation, and the initiation of various other reflexive evaluation projects, normalization has been at least partly successful. Despite the ad hoc character of their actions, the team succeeded in establishing alignment, or a “fit,” between the two evaluation logics. Crucially, this fit did not appear as a fixed state, but rather as a negotiated, emergent, and dynamic accomplishment. Indeed, the alignment work for normalizing reflexive evaluation encompassed navigating and negotiating reflexive evaluation legitimacy on one hand and reflexive evaluation integrity on the other. Specifically, our study has shown that alignment work comprises various strategies, including...
reframing the purpose of reflexive evaluation for it to make sense from the point of view of the dominant institutional logic. Another strategy concerned emphasizing the scientific rigor of the approach to demonstrate its validity as a peer-reviewed research approach. Such strategies “work” as they ensure the legitimacy of the reflexive evaluation approach from the modernist logic perspective. At other times, such strategies proved insufficient to acquire legitimacy: regardless of framing or appeal to scientific rigor, in our case, the involvement of policy actors in deciding on research methods, drawing conclusions or co-publication, were topics largely beyond discussion both for evaluators and participants. Here, the dominant modernist logic’s norms could not be negotiated to develop legitimacy.

May et al. (2016) have studied the degree to which contextual structures and cultures may be negotiated to normalize a novel practice and refer to contextual elasticity: the ease with which contexts accommodate new ways of working. They propose that the greater the elasticity of contextual structures and cultures, the less work is required from practitioners of a new practice to develop legitimacy. In relation to evaluation logics, this implies that the more the modernist logic is institutionalized and embedded, the more work is required from evaluators to normalize a reflexive practice—that is, to develop understanding, buy-in, agreement on enactment and appraisal. From our findings, it appears that when the roles of researchers and policy makers seemed to become too intertwined, the reflexive logic was disciplined by actors from either context, or both.

We also observed that some normative structures became more accommodating over time, particularly when the value of the reflexive approach became materially tangible, as happened when the lack of participation in deciding about research methods resulted in only limitedly useful findings for policy learning. Such observations confirm that normative structures more accommodating to the reflexive logic can be developed (Verwoerd et al., 2020). More in-depth studies of the elements that constitute contextual elasticity and of how more accommodating contextual structures and cultures may be developed, may be fruitful to further the understanding of normalization of reflexive evaluation. For instance, past research has shown that participation—as a constitutive element of reflexive evaluation—can be obstructed by diverse political, administrative, and social barriers and power inequalities (Engel and Carlsson, 2002; Gregory, 2000; Lehtonen, 2014). It would be very useful to study the role of such contextual elements in the context of attempts to normalize reflexive evaluation logic.

We observed that when contextual structures and cultures were unaccommodating, evaluators navigated these by altering components—principles, procedures, or purpose—of the reflexive approach. May et al.’s (2016) concept of innovation plasticity comes in useful to understand this, as plasticity refers to the malleability of components and the discretion practitioners have to develop alignment between the contexts and the innovation. The greater the degree of plasticity of innovation components, the less work is required from practitioners to normalize it (May et al., 2016). Already in 1977, Fullan and Pomfret discuss that a level of on-site modification may be required for innovations to effectively meet context-specific needs. However, adaptation beyond the “zone of drastic mutation” (Roitman and Mayer, 1982: 3) may compromise the fidelity to, or integrity of, the original intentions. Indeed, in our case, some reflexive components were stretched too far: not involving participants in making methodological decisions led to a decreased chance of policy learning. It appears that there are certain parameters that determine the plasticity of reflexive evaluation components (i.e. the zone of drastic mutation) and the leeway enactors have to retain reflexive evaluation integrity. These parameters were obscure (it was not immediately clear if or when a line was crossed)
and, as for contextual elasticity, what precisely determines these parameters, and thus the plasticity of reflexive evaluation components, requires additional inquiry.

Contextual elasticity and innovation plasticity are useful concepts to think about the required alignment work for negotiating and navigating reflexive evaluation legitimacy and integrity. Using both concepts in conjunction may shine new light on the explanation as to why it is so difficult for practitioners to innovate toward a more reflexive evaluation practice. This is important as scholars have previously problematized successful participation (Felt et al., 2012; Gregory, 2000; Roitman and Mayer, 1982), and it has been pointed out that, despite ideals for new roles for science and policy, novel repertoires often deviate little from their technocratic counterparts and traditional ideas on the relationship between scientists and policy makers (Reinecke, 2015; Turnhout et al., 2013). New “in vogue” terms, in our case reflexive evaluation, may even cover up this reality, promoting legitimacy for “innovative” approaches while actual practices remain unchanged. As Van Der Hel (2016) points out, limits to elasticity and plasticity seem a pertinent explanation of why researchers are inclined to do “more of the same under a different name” (p. 173). As researchers are pushed back by the contexts they operate in, they cannot help but default into modernist logic routines, and end up (unwillingly) greenwashing, tokenistic behaviours, and technocratizing participation (Chilvers, 2008). This is relevant, as such occurrences may promote participants’ subjugation rather than their empowerment (Aarts and Leeuwis, 2010; Turnhout et al., 2013), thereby reinforcing the modernist status quo.

We conclude with a reflection on the use of NPT for studying normalization of reflexive evaluation. Our findings confirmed the theory’s potential: applying NPT provided analytical depth to investigate how reflexive evaluation was operationalized and helped identify the evaluators’ challenges with implementation. Consistent with others’ experiences, applying NPT was not instantly intuitive (McEvoy et al., 2014). Although its numerous (sub)concepts make the theory versatile in use, these also required significant translation effort to render them applicable to the specific context of our study, an issue also addressed by Finch et al. (2012). For the purpose of our study, we decided early on to center our analysis around NPT’s core concepts, which, following the example of McNaughton et al. (2020) and with the aim to better engage with them, we re-labeled into more accessible language (e.g. “coherence” was re-labeled as “sense-making”).

Others have pointed out NPT’s undue emphasis on individual and collective agency, disregarding the influence of organizational and relational contexts in which implementation occurs (Clarke et al., 2013). We engaged with this shortcoming by including the concept of institutional logics, which drew attention to the interacting and nested contexts (team, organization, societal interactions) in which these logics were enacted and through which they shaped agency.

Overall, we conclude that timely investment in mutual sense-making can be recommended for the normalization of reflexive evaluation to occur and seems especially pertinent for joint understanding of the components of reflexive evaluation, their plasticity, and the parameters that safeguard reflexive evaluation integrity. In doing so, the methodological repertoires evaluators draw from may be increased and strengthened. In a similar vein, we conclude that considering unaccommodating contextual structures and aspects of political and organizational culture (and their elasticity) from the onset may help evaluators anticipate and orchestrate in a timely fashion alignment strategies that develop evaluation legitimacy. For both cases, using NPT prospectively and in action (De Brún et al., 2016) may be fruitful to guide
structural reflection and learning-by-doing and hence can facilitate the effective navigation of reflexive evaluation legitimacy and integrity in the normalization process.

Acknowledgements
The authors would like to thank all respondents, the Natuurpact team in particular, for their participation to this study. Special thanks goes out to Deborah Eade for editing and Evelien de Hoop for her contributions to the final version of this article.

Declaration of conflicting interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: After the research presented in this paper had concluded, but during the time of writing, Lisa Verwoerd has taken on a part-time position at the PBL methodology department.

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

ORCID iD
Lisa Verwoerd https://orcid.org/0000-0002-3474-5452

Notes
1. In earlier versions of the Normalization Process Theory (NPT) framework, enactment was the core focus, as it comprises the acts and behaviors of performing a practice, and therefore the practice itself (in alignment with other social practice theorists who consider enactment and practice two sides of the same coin, see, for example, Giddens (1984) or Reckwitz (2002)). For sake of clarity, in this article, we use the term “enactment” only in relation to the third NPT mechanism, which—in concordance with sense-making, engagement, and appraisal—may lead to the normalization of innovative practices.
2. While the inventors of NPT expand each mechanism with multiple sub-concepts, for the purpose of our study, we have chosen to focus our analysis around the four core mechanisms of NPT. We refer readers to May and Finch (2009) for a seminal account.

References
Aarts N and Leeuwis C (2010) Participation and power: Reflections on the role of government in land use planning and rural development. The Journal of Agricultural Education and Extension 16(2): 131–45.
Arkesteijn M, van Mierlo B and Leeuwis C (2015) The need for reflexive evaluation approaches in development cooperation. Evaluation 21(1): 99–115.
Beck U, Giddens A and Lash S (1994) Reflexive Modernisation: Politics, Tradition and Aesthetics in the Modern Social Order. Stanford, CA: Stanford University Press.
Beers PJ and Van Mierlo B (2017) Reflexivity, reflection and learning in the context of system innovation: Prying loose entangled concepts. In: Elzen B, Augustyn AM, Barbier M, et al. (eds) AgroEcological Transitions: Changes and Breakthroughs in the Making. Wageningen: Wageningen University & Research, Applied Arable and Vegetable Research, 243–56.
Berg Johansen C (2017) What are Institutional Logics—And where is the perspective taking us? In: Krücken G, Mazza C, Meyer RE, et al. (eds) New Themes in Institutional Analysis—Topics and Issues from European Research. Cheltenham: Edward Elgar Publishing, 12.
Chilvers J (2008) Deliberating competence: Theoretical and practitioner perspectives on effective participatory appraisal practice. *Science, Technology, & Human Values* 33: 155–85.

Chouinard JA (2013) The case for participatory evaluation in an era of accountability. *American Journal of Evaluation* 34(2): 237–53.

Clarke DJ, Godfrey M, Hawkins R, et al. (2013) Implementing a training intervention to support caregivers after stroke: A process evaluation examining the initiation and embedding of programme change. *Implementation Science* 8: 96.

Dahler-Larsen P (2012) *The Evaluation Society*. Stanford, CA: Stanford University Press.

Dahler-Larsen P and Schwandt TA (2012) Political culture as context for evaluation. *New Directions for Evaluation* 135: 75–87.

De Brún T, O’Reilly-de Brún M, O’Donnell CA, et al. (2016) Learning from doing: The case for combining normalisation process theory and participatory learning and action research methodology for primary healthcare implementation research. *BMC Health Services Research* 16(1): 346.

De Jaegher H, Pieper B, Clénin D, et al. (2017) Grasping intersubjectivity: An invitation to embody social interaction research. *Phenomenology and the Cognitive Sciences* 16(3): 491–523.

Duijnhoven H and Neef M (2016) Disentangling wicked problems: A reflexive approach towards resilience governance. In: Masys A (ed.) *Advanced Sciences and Technologies for Security Applications*. Cham: Springer, 91–106.

Engel PGH and Carlsson C (2002) Enhancing learning through evaluation: Approaches, dilemmas and some possible ways forward. In: Annual conference of the European Evaluation Society: Three movements in contemporary evaluation: Learning, theory and evidence, Sevilla, 10–12 October.

Felt U, Igelsboeck J, Schikowitz A, et al. (2012) Challenging participation in sustainability research. *The Journal of Deliberative Mechanisms in Science* 1(1): 4–34.

Fetterman DM, Wandersman A and Kaftarian S (2015) Empowerment evaluation is a systematic way of thinking: A response to Michael Patton. *Evaluation and Program Planning* 52: 10–4.

Finch TL, Mair FS, O’Donnell C, et al. (2012) From theory to “measurement” in complex interventions: Methodological lessons from the development of an e-health normalisation instrument. *BMC Medical Research Methodology* 12: 69.

Fitzpatrick J, Christie C and Mark MM (2008) *Evaluation in Action: Interviews with Expert Evaluators*. Thousand Oaks, CA: SAGE.

Friedland R and Alford RR (1991) Bringing society back in: Symbols, practices and institutional contradictions. In: Powell WW and DiMaggio PJ (eds) *The New Institutionalism in Organizational Analysis*. Chicago, IL: University of Chicago Press, 232–63.

Fullan M and Pomfret A (1977) Research on curriculum and instruction implementation. *Review of Educational Research* 47(1): 335–97.

Funtowicz SO and Ravetz JR (1993) Science for the post-normal age. *Futures* 25: 739–55.

Geels FW (2004) From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy* 33: 897–920.

Giddens A (1984) *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley, CA: University of California press.

Green J, Roberts H, Petticrew M, et al. (2015) Integrating quasi-experimental and inductive designs in evaluation: A case study of the impact of free bus travel on public health. *Evaluation* 21(4): 391–406.

Gregory A (2000) Problematizing participation: A critical review of approaches to participation in evaluation theory. *Evaluation* 6(2): 179–99.

Guijt I (2010) Accountability and learning: Exploding the myth of incompatibility between accountability and learning. In: Ubels J, Acquaye-Baddoo N-A and Folwer A (eds) *Capacity Development in Practice*. Washington, DC: Earthscan, 277–91.

Halfman W (2009) Measuring the stakes: The Dutch planning bureaus. In: Lentsch J and Weingart P (eds) *Scientific Advice to Policy Making: International Comparison*. Farmington Hills, MI: Barbara Budrich Publishers, 1–25.
Havelock RG (1970) *A Guide to Innovation in Education*. Ann Arbor, MI: University of Michigan.
Havelock RG (1971) *Planning for Innovation: Through Dissemination and Utilization of Knowledge*. Ann Arbor, MI: University of Michigan.
Hsieh H-F and Shannon SE (2005) Three approaches to qualitative content analysis. *Qualitative Health Research* 15(9): 1277–88.
Huberman AM and Miles MB (1984) *Innovation Up Close: How School Improvement Works*. New York: Springer.
Jasanoff S (2011) Quality control and peer review in advisory science. In: Lentsch J and Weingart P (eds) *The Politics of Scientific Advice: Institutional Design for Quality Assurance*. Cambridge: Cambridge University Press, 19–35.
Jones C, Boxenbaum E and Anthony C (2015) The immateriality of material practices in institutional logics. In: Boxenbaum E and Lounsbury M (eds) *Institutional Logics in Action, Part A (Research in the Sociology of Organizations)*. Vol. 39. Bingley: Emerald Group Publishing, 51–75.
Kunseler E-M (2017) *Government Expert Organisations In-Between Logics*. Amsterdam: VU University Amsterdam.
Kunseler E-M and Tuinstra W (2017) Navigating the authority paradox: Practising objectivity in environmental expertise. *Environmental Science & Policy* 67: 1–7.
Kunseler E-M and Vasileiadou E (2016) Practising environmental policy evaluation under co-existing evaluation imaginaries. *Evaluation* 22(4): 451–69.
Lehtonen M (2014) Evaluating megaprojects: From the “iron triangle” to network mapping. *Evaluation* 20: 278–95.
Lentsch J and Weingart P (eds) (2011) *The Politics of Science Advice: Institutional Design for Quality Assurance*. Cambridge: Cambridge University Press.
Loober A, Van Mierlo B, Grin B, et al. (2007) The practical value of theory: Conceptualising learning in the pursuit of a sustainable development. In: Wals AEJ (ed.) *Social Learning towards a Sustainable World*. Wageningen: Wageningen Academic Publishers, 83–98.
McEvoy R, Ballini L, Maltoni S, et al. (2014) A qualitative systematic review of studies using the Normalization Process Theory in research implementation processes. *Implementation Science* 9(2): 1–22.
McNaughton R, Steven A and Shucksmith J (2020) Using Normalization Process Theory as a practical tool across the life course of a qualitative research project. *Qualitative Health Research* 30(3): 217–27.
Mair FS, May C, O’Donnell C, et al. (2012) Systematic reviews—Factors that promote or inhibit the implementation of e-health systems: An explanatory systematic review. *Bulletin of the World Health Organization* 90: 357–64.
Marjanovic S, Cochrane G, Robin E, et al. (2017) Evaluating a complex research capacity-building intervention: Reflections on an evaluation of the African Institutions Initiative. *Evaluation* 23(1): 80–101.
May C (2015) Applying normalization process theory to the implementation of complex interventions. In: Richards DA and Hallberg IR (eds) *Complex Interventions in Health—An Overview of Research Methods*. London: Routledge, 239–47.
May C and Finch T (2009) Implementing, embedding, and integrating practices: An outline of normalization process theory. *Sociology* 43(3): 535–54.
May CR, Cummings A, Girling M, et al. (2018) Using Normalization Process Theory in feasibility studies and process evaluations of complex healthcare interventions: A systematic review. *Implementation Science* 13: 80.
May CR, Johnson M and Finch T (2016) Implementation, context and complexity. *Implementation Science* 11(1): 141.
Moore GF, Evans RE, Hawkins J, et al. (2019) From complex social interventions to interventions in complex social systems: Future directions and unresolved questions for intervention development and evaluation. *Evaluation* 25(1): 23–45.
Nicolini D (2012) Practice Theory, Work, and Organization: An Introduction, 1st edn. Oxford: Oxford University Press.

Nieminen M and Hyytinen K (2015) Future-oriented impact assessment: Supporting strategic decision-making in complex socio-technical environments. Evaluation 21(4): 448–61.

O’Sullivan RG (2012) Collaborative evaluation within a framework of stakeholder-oriented evaluation approaches. Evaluation and Program Planning 35(4): 518–22.

Owens S, Rayner T and Bina O (2004) New agendas for appraisal: Reflections on theory, practice, and research. Environment and Planning A: Economy and Space 36: 1943–59.

Patton MQ (2010) Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. New York: Guilford Press.

PBL (2017) About PBL. Available at: https://www.pbl.nl/en/aboutpbl

Petersen AC, Cath A, Hage M, et al. (2011) Post-normal science in practice at The Netherlands Environmental Assessment Agency. Science, Technology, & Human Values 36(3): 362–88.

Pollitt C, van Thiel S and Homburg V (2007) New Public Management in Europe: Adaptation and Alternatives. Available at: https://link.springer.com/book/10.1057%2F9780230625365

Reckwitz A (2002) Toward a theory of social practices: A development in culturalist theorizing. European Journal of Social Theory 5(2): 243–63.

Reinecke S (2015) Knowledge brokerage designs and practices in four European climate services: A role model for biodiversity policies? Environmental Science & Policy 54: 513–21.

Rittel HWJ and Webber MM (1973) Dilemmas in general theory of planning. Policy Sciences 4: 155–69.

Rog DJ, Fitzpatrick JL and Conner RF (2012) Context: A framework for its influence on evaluation practice. New Directions for Evaluation 135: 89–105.

Roitman DB and Mayer JP (1982) Fidelity and reinvention in the implementation of innovations. In: 90th annual convention of the American Psychological Association, Washington, DC, 23–27 August, 21. Washington, DC: National Science Foundation.

Rolfe S (2019) Combining theories of change and realist evaluation in practice: Lessons from a research on evaluation study. Evaluation 25(3): 294–316.

Schatzki TR (2002) The Site of the Social: A Philosophical Account of the Constitution of Social Life and Change. University Park, PA: Pennsylvania State University Press.

Shils EA and Finch HA (1949) Max Weber on the Methodology of Social Sciences, 1st edn. Glengoe, IL: The Free Press.

Shove E (2010) Beyond the ABC: Climate change policy and theories of social change. Environment and Planning A: Economy and Space 42(6): 1273–85.

Stern E, Saunders M and Stame N (2015) Standing back and looking forward: Editors’ reflections on the 20th Anniversary of Evaluation. Evaluation 21(4): 380–90.

Stermlieb F, Bixler RP, Huber-Stearns H, et al. (2013) A question of fit: Reflections on boundaries, organizations and social–ecological systems. Journal of Environmental Management 130: 117–25.

Turnhout E, Stuiver M, Klostermann J, et al. (2013) New roles of science in society: Different repertoires of knowledge brokering. Science and Public Policy 40(3): 354–65.

Vaidya A and Mayer AL (2014) Use of the participatory approach to develop sustainability assessments for natural resource management. International Journal of Sustainable Development & World Ecology 21(4): 369–79.

Van Der Hel S (2016) New science for global sustainability? The institutionalisation of knowledge co-production in Future Earth. Environmental Science & Policy 61: 165–75.

Van Hemelrijck A and Guijt I (2016) Balancing inclusiveness, rigour and feasibility: Insights from participatory impact evaluations in Ghana and Vietnam. CDI Practice Paper (14). Brighton: Institute of Development Studies. 1 February 2016.

Van Mierlo B, Regeer BJ, van Amstel M, et al. (2010) Reflexive Monitoring in Action: A Guide for Monitoring System Innovation Projects. Wageningen; Amsterdam: Wageningen University & Research.
Van Veen SC, Verwoerd L and Regeer BJ (2016) *Characteristics of Reflexive Evaluation—A Literature Review Conducted in the Context of the Natuurpact (2014-2027) Evaluation*. Amsterdam: Athena Instituut, Vrije Universiteit.

Verwoerd L, De Wildt-Liesveld R and Regeer BJ (2017) *The Value of Reflexive Evaluation—A Review of the Natuurpact Evaluation (2014-2017)*. Amsterdam: Athena Instituut, Vrije Universiteit.

Verwoerd L, Klaassen P and Regeer BJ (2019) *The Value of a Reflexive Evaluation Approach in the Eyes of Researchers*. Amsterdam: Athena Institute, VU University.

Verwoerd L, Klaassen P, van Veen SC, et al. (2020) Combining the roles of evaluator and facilitator: Assessing societal impacts of transdisciplinary research while building capacities to improve its quality. *Environmental Science & Policy* 103: 32–40.

Warde A (2005) Consumption and theories of practice. *Journal of Consumer Culture* 5(2): 131–53.

Lisa Verwoerd works at the Athena Institute (VU University Amsterdam). Her research centers on evaluating and the monitoring-in-action of reflexive policy research projects in complex multi-actor and multi-level governance settings.

Pim Klaassen is an Assistant Professor at the Athena Institute (VU University Amsterdam) and focuses on governance arrangements and (reflexive) research methodologies conducive to responsible research and innovation.

Barbara J. Regeer is an Associate Professor at the Athena Institute (VU University Amsterdam). Her work revolves around studying and normalizing transdisciplinary and reflexive strategies for sustainable development and system change.