

The value of a reflexive evaluation approach in the eyes of researchers

A review of the Natuurpact evaluation (2014-2017)



L. Verwoerd, P. Klaassen & B.J. Regeer

The value of reflexive evaluation in the eyes of researches – a review of the Natuurpact evaluation (2014-2017)

L. Verwoerd

P. Klaassen

B.J. Regeer



Athena Institute for Research on Innovation and Communication in Health and Life Sciences

The value of reflexive evaluation in the eyes of researchers – a review of the Natuurpact evaluation (2014-2017), Athena Institute, VU University Amsterdam, commissioned by the PBL Environmental Assessment Agency, The Hague

Authors

Lisa Verwoerd (l.verwoerd@vu.nl)

Pim Klaassen

Barbara Regeer

All affiliated to the Athena Institute, VU University Amsterdam

Reviewers

Bas Arts

Femke Verwest

Rob Folkert

Eva Kunseler

With special regards to:

Our respondents

Tara White for editing

All nature pictures in this report are courtesy of *Unsplash*: <https://unsplash.com/>

The Athena Institute

The Athena Institute is a research and education department at the Faculty of Beta Sciences, VU Amsterdam. The Athena Institute develops and applies transdisciplinary approaches for joint reflection and action on innovation processes among a wide variety of stakeholders who have different interests, values, visions in order to address complex problems in stagnating systems (society or organizations) in a sustainable and equitable way.

April 2019

The authors bear full responsibility for the content of the review but accept no liability whatsoever for any damages that may occur as a consequence of using the information in this publication. The Creative Commons license 4.0 for the Netherlands (attribution, non-commercial, no derivate works) is applicable to this work. *The value of reflexive evaluation in the eyes of researchers – review of the Natuurpact evaluation (2014-2017)* by Lisa Verwoerd, Pim Klaassen & Barbara Regeer, is licensed under a creative commons attribution-noncommercial-noderivatives 4.0 INTERNATIONAL LICENSE <https://creativecommons.org/licenses/by-nc-nd/4.0>



Contents

<i>Nederlandse samenvatting</i>	5
1. Introduction	10
2. Background	13
2.1 The theory of reflexive evaluation	13
2.2 Two ideal-types of policy evaluation	14
2.3 In practice: the Natuurpact reflexive evaluation	16
3. Methodology	19
3.1 Data collection	19
3.2 Data analysis	20
3.3 Validation	21
3.4 Our role	21
4. Results	23
4.1 Researchers perspectives on quality, usability and impact	23
4.2 Navigating logics during the Natuurpact evaluation	30
5. Conclusions & discussion	36
5.1 Conclusion: understanding the perspectives on the value of a reflexive evaluation approach	36
5.2 Discussion: towards a reflexive practice	37
6. Final recommendations	40
<i>References</i>	42

Nederlandse samenvatting

Het Natuurpact, opgesteld in 2013 is de laatste stap geweest van de decentralisatie van het natuurbeleid van het Rijk naar de provincies. Hierin hebben zij met elkaar afspraken gemaakt over hun ambities voor het natuurbeleid – zoals het voltooiën van het Natuur Netwerk Nederland en het versterken van de maatschappelijke betrokkenheid bij natuur – met 2027 als tijdshorizon. Daarnaast is afgesproken om de voortgang met deze ambities in beeld te brengen middels een ‘lerende evaluatie’, uitgevoerd door het Planbureau voor de Leefomgeving (PBL) en Wageningen University and Research (WUR), specifiek in opdracht van het voormalig ministerie van Economische zaken (nu: Landbouw, Natuur en Voedselkwaliteit) en het Interprovinciaal overleg (IPO). Afgesproken is elke drie jaar over de voortgang te rapporteren. Het huidige rapport betreft een beschouwing van de eerste evaluatie periode (2014-2017).

De opdrachtgevers en de onderzoekers hebben samen besloten om een lerende evaluatie uit te voeren. Daarmee is afgeweken van een meer traditionele evaluatie aanpak, omdat voorzien is dat hiermee beter recht gedaan wordt aan de multi-level en multi-actor context waarin natuurbeleid zich bevindt, de complexe opgave voor biodiversiteit en de verbrede ambities om ook de samenleving meer bij het natuurbeleid te betrekken. Centraal bij deze lerende aanpak staat de samenwerking tussen onderzoekers en de partijen die verantwoordelijk zijn voor de ontwikkeling en uitvoer van het Nederlands natuurbeleid—hoofdzakelijk de provincies. De evaluatie richt zich op het gezamenlijk trekken van conclusies en lessen, om zo het beleid tijdig te kunnen verbeteren en daarmee de kans op doelbereik te vergroten. In tabel i schetsen we de belangrijkste karakteristieken van lerend evalueren. Het is de hypothese van de opdrachtgevers en het PBL en de WUR, dat een lerende evaluatie de kwaliteit, bruikbaarheid en de impact van de tijdens de evaluatie tot stand gebrachte kennis, vergroot.

Tabel i. Karakteristieken van lerend evalueren, naar een studie door Van Veen, Verwoerd en Regeer (2016) in opdracht van het PBL en uitgevoerd in het licht van de lerende evaluatie Natuurpact.

Karakteristieken van lerend evalueren	
Doel van de evaluatie	Recht doen complexe multi-level governance en multi-actor contexten door het vergroten van de reflexiviteit van relevante stakeholders en het optimaliseren van hun adaptief vermogen
Rol van relevante stakeholders	Relevante stakeholders nemen actief deel aan het evaluatieonderzoek en dragen bij aan het evaluatieontwerp en de uitvoering, om afstemming van het onderzoek op hun evaluatiebehoeften te bewerkstelligen; relevantie van stakeholders geniet hierbij een breed begrip (bijvoorbeeld niet alleen beleidsmakers, maar ook zij die de effecten van het beleid ervaren)
Rol van onderzoekers	Onderzoekers vervullen een dubbele rol: zij zijn zowel beoordelaar van beleidsvoortgang (verantwoording van beleid) en facilitator van leerprocessen (leren van beleid). Onderzoekers beschikken hiertoe niet alleen over goede onderzoeksvaardigheden, maar ook over meer ‘zachte’ vaardigheden, zoals faciliteren en mediëren.
Relatie tussen beleidspraktijk en evaluatieonderzoek	Het evaluatieonderzoek is afgestemd op en vindt parallel plaats naast de beleidspraktijk, om zo beleidsprocessen tijdig te kunnen informeren met relevante kennis
Type geproduceerde kennis	Sociaal robuuste kennis: gedragen, legitiem, (wetenschappelijk) geloofwaardig en relevant voor maatschappelijke en beleidsopgaven

Lerend evalueren – zeker op de schaal van het Natuurpact – was voor het PBL en de WUR een nieuwe aanpak. Om deze evaluatiemethode wetenschappelijk te borgen heeft het PBL onderzoekers van het Athena Instituut (VU) gevraagd de eerste periode van de evaluatie te reviewen en de waarde daarvan voor zowel de beleids- als onderzoekspraktijk in beeld te brengen. In het eerste deel van deze review hebben we daartoe de diverse wijzen onderzocht waarop de lerende evaluatie volgens de beleidsbetrokkenen die deel hebben genomen aan de evaluatie van waarde is geweest (Verwoerd et al., 2017). Onze bevindingen toonden destijds aan dat de lerende evaluatie Natuurpact bijgedragen heeft aan vijf waarden, zowel op het niveau van individuele provincies, als over overheidsgrenzen heen. Ten eerste heeft de evaluatie bijgedragen aan inhoudelijk, of cognitieve, waarde; beleidsbetrokkenen ervaren een beleidspraktijk verrijkt met nieuwe kennis en inzichten om beleidsbeslissingen mee te informeren, en een meer gedeeld begrip en visie op de ambities van natuurbeleid tussen overheden. Ten tweede was er sprake van instrumentele waarde; de opgedane kennis en inzichten was ook daadwerkelijk toepasbaar in provinciale beleidsplannen. Ten derde bleek de evaluatie van sterke strategische waarde, daar het provinciale – en in beperktere mate ook nationale – beleidsbeslissingen legitimeerde richting de maatschappij. Deze eerste drie typen waarde zijn in meer of mindere mate te verwachten, ook als resultaat van regulier evaluatieonderzoek. In aanvulling hierop hebben we ten slotte nog twee waarden geïdentificeerd, die kenmerkend zijn voor lerend evalueren: netwerkwaarde en affectieve waarde. Zo spraken de deelnemers over de waarde van de evaluatie voor hun netwerk, waarbij bestaande relaties zijn versterkt en nieuwe relaties zijn ontstaan, en observeerden we tot slot dat er tussen de provincies onderling een gevoel van saamhorigheid is ontstaan, dat bijzonder is gewaardeerd. Daarbij is er, alhoewel voorzichtig, ook sprake van een groeiend vertrouwen tussen Rijk en provincies.

Dit rapport presenteert het tweede en tevens laatste deel van de review van de eerste periode van de lerende evaluatie Natuurpact. Centraal hierin staan de onderzoekspraktijk en de perspectieven van betrokken onderzoekers op de waarde van lerend evalueren voor het vergroten van de kwaliteit, bruikbaarheid en impact van het onderzoek.

Onderzoeksvragen

Het PBL heeft ons gevraagd te onderzoeken hoe betrokken onderzoekers de lerende evaluatie hebben gewaardeerd in het licht van de kwaliteit, bruikbaarheid en impact van de geproduceerde kennis voor en op de beleidspraktijk. We hebben daartoe de volgende drie hoofdvragen geformuleerd:

- 1. Welke perspectieven hebben onderzoekers op de waarde van de lerende evaluatie voor de kwaliteit, bruikbaarheid en impact van de kennis die tijdens de evaluatie is geproduceerd?*
- 2. Hoe kunnen we deze verschillende perspectieven begrijpen?*
- 3. Welke aanbevelingen voor de volgende evaluatie periode volgen uit het beter begrijpen van de verschillende perspectieven?*

Theoretisch kader: modernistisch en responsief ideaaltype

Om onze hoofdvragen te kunnen beantwoorden zijn we vertrokken vanuit een theoretisch kader waarin we twee ideaaltypische evaluatiemethoden als extremen tegenover elkaar hebben geplaatst: het modernistische en het responsieve ideaaltype. Studies naar het proces van kennisproductie hebben aangetoond dat onderzoekers voor het ontwerp en de uitvoering van hun onderzoek putten uit verschillende 'logica's' ten aanzien van wetenschappelijk onderzoek. Deze logica's omvatten

normen voor de functie van onderzoek en kennis in de samenleving en gouden standaarden waaraan ‘goed’ onderzoek dient te voldoen. Om de perspectieven op de waarde van de lerende evaluatie die onderzoekers uitten te begrijpen, gebruiken we in dit rapport de modernistische en responsieve ideaaltypische logica’s, welke we uiteenzetten in tabel ii.

Tabel ii. Functie en gouden standaarden van ‘goed’ evaluatieonderzoek, volgens de modernistische en responsieve logica (naar Kunseler, 2017).

	Modernistische logica	Responsieve logica
Doel van de geproduceerde kennis	Evaluatie is een mechanisme ter verantwoording van beleid door het beleid te beoordelen op (kosten)effectiviteit	Evaluatie is een mechanisme om meervoudigheid van kennis te vatten en wederzijdse leerprocessen te voeden door reflectie op instrumenten, strategieën, doelen en normen en waarden
Principes voor kwaliteit	Nadruk op wetenschappelijke geloofwaardigheid, norm van onafhankelijkheid, afstand, objectiviteit, wetenschappelijke autonomie	Nadruk op legitimiteit en draagvlak; norm van inclusie van de pluraliteit aan perspectieven en interactie met actoren ter co-creatie van kennis
Principes voor bruikbaarheid	Onderzoekers-gedreven: onderzoek ontworpen om beleid te kunnen beoordelen via systematische data verzameling ten aanzien van input, output en outcomes	Stakeholder-gedreven: onderzoek gezamenlijk ontworpen met actoren om recht te doen aan de beleidscomplexiteit en leerprocessen daarover te voeden
Principes voor impact	Impact is gerealiseerd op basis van de mate waarin het onderzoek succesvol is in het objectief beoordelen van beleid. Op basis van deze beoordeling kunnen overheden kiezen beleid aan te passen	Impact is gerealiseerd op basis van de mate waarin het onderzoek succesvol is in het voeden van wederzijdse leerprocessen ten aanzien van de werking van beleid

Methodie

Om de hoofdvragen te beantwoorden hebben we een kwalitatieve onderzoeksapproach gehanteerd, waarin zeven semi-open diepte-interviews gehouden zijn met nauw betrokken onderzoekers van de Natuurpactevaluatie. De respondenten verschilden in thuisorganisatie (PBL dan wel WUR), (hoofd)discipline (ecologie, bestuurskunde, economie) en functie (plaatsvervangend projectleider, deelprojectleider, onderzoekers). In aanvulling op conventionele maatregelen om de kwaliteit van kwalitatief onderzoek te borgen (zoals onderzoekertriangulatie en *member checks*), hebben we onze conceptresultaten voorgelegd tijdens een interactieve workshop waarin onze respondenten en een extra ring aan betrokkenen (projectleiding en onderzoekers die betrokken zijn bij de nieuwe evaluatie ronde) aanwezig waren om de validiteit van onze bevindingen te toetsen en deze te verrijken.

Conclusies

De voornaamste conclusie van dit rapport is dat de perspectieven van de onderzoekers op de kwaliteit, bruikbaarheid en impact van de kennis consistent zijn met de geformuleerde hypothese, gezien de onderzoekers van mening zijn dat de lerende aanpak van de evaluatie zowel de kwaliteit van de geproduceerde kennis als haar bruikbaarheid en impact positief heeft beïnvloed. Voortbordurend op onze eerste review concluderen we daarmee dat de lerende evaluatie Natuurpact van waarde is geweest voor zowel de beleids- als onderzoekspraktijk.

Het breed gedeelde beeld over de waarde van lerend evalueren gaat hand in hand met een aanzienlijke variatie in toelichtingen die onderzoekers uiten om hun oordeel over de waarde van lerend evalueren te onderbouwen. Onderzoekers navigeren daarbij steeds tussen de functies en gouden standaarden van zowel de modernistische als de responsieve logica. Dit navigeren lijkt kenmerkend voor de lerende evaluatie Natuurpact. Bij de aanvang van dit grootschalige evaluatiearrangement was nog niet volledig helder wat lerend evalueren precies zou betekenen en

hoe dit in deze context in de praktijk kon worden gebracht. In verschillende arena's – namelijk het projectteam met de betrokken onderzoekers, hun institutionele omgeving, en de beleidsarena met daarin de opdrachtgevers en beleidsbetrokkenen – is gedurende de evaluatie onderhandeld om te komen tot een evaluatieaanpak die in grote lijnen aan de wensen van alle partijen, en daarmee aan de voorwaarden van zowel modernistische en responsieve aanpakken, voldeed. De (impliciet) aanwezige perspectieven op 'goed' evalueren hebben daarbij de samenwerking tussen onderzoekers gelijktijdig gefrustreerd – in ieder geval wat betreft het proces van inter- en transdisciplinaire kennisintegratie – als verrijkt. Verklaringen voor de variatie in perspectieven zijn te vinden in de disciplinaire en, in grotere mate, de institutionele achtergrond van onderzoekers, alsmede de maatschappelijke en politieke verwachtingen ten aanzien van de evaluatie. In zowel de institutionele achtergrond als de politiek-maatschappelijke context waarin de evaluatie plaatsvond is de modernistische logica van oudsher stevig ingebed en genormaliseerd. De ingesleten routines die daarin hun oorsprong vinden maken dat het voor onderzoekers soms lastig is om de responsieve kant op te bewegen. Tegelijkertijd stellen we dat de onderzoekers in hun praktijk van lerend evalueren voorbijgaan aan de zogenaamde dichotomie tussen de modernistische en responsieve logica. Gezamenlijk geven ze daarbij gezamenlijk vorm aan een meer reflexieve evaluatiepraktijk waarbij de beleidsopgaven en de politiek-maatschappelijke context waarin deze zich afspeelt de uitgangspunten zijn voor het bepalen van een geschikte evaluatieaanpak. In deze reflexieve praktijk worden onderhandelingen ten aanzien van een wenselijk evaluatieontwerp en -uitvoering vanuit verschillende kanten belicht en daarmee verrijkt, waardoor men kan komen tot een vorm die zowel wetenschappelijk als maatschappelijk robuust is. Onze aanbevelingen richten zich dan ook op het versterken en aanmoedigen van deze reflexieve praktijk in de eerste plaats in de context van het Natuurpact. Tevens zijn deze door te vertalen naar het institutionele vlak, namelijk binnen het PBL en de WUR.

Aanbevelingen

Ondersteun het navigeren van verschillende logica's ten aanzien van evalueren, door een gedeelde reflexieve praktijk aan te moedigen en in te bedden in institutionele structuren en werkwijzen

Een gedeelde reflexieve praktijk voorziet onderzoekers van een flexibel kader waarbinnen verschillende logica's op 'goed' evaluatieonderzoek kunnen worden verkend en effectief genavigeerd, om zo tot een wetenschappelijk en maatschappelijk robuuste evaluatieaanpak te komen. Dit navigeren kan worden ondersteunt door met regelmaat op gestructureerde en methodische (professioneel begeleide) wijze interactief te reflecteren (bijvoorbeeld via *frame reflection*). Zo kunnen logica's, verwachtingen en belangen ten aanzien van het evaluatieonderzoek expliciet worden gemaakt, ten goede van wederzijds begrip en afstemming tussen logica's, en binnen de diverse arena's waarin de evaluatieaanpak wordt onderhandeld.

Bed kennisintegratie structureel in het evaluatieontwerp in

Aangezien verschillende logica's (impliciet) in het spel zijn is inter- en transdisciplinaire kennisintegratie tijdens lerend evalueren een extra uitdaging waar onderzoekers voor staan. Om dit proces te borgen loont het om in het evaluatieontwerp hier structureel tijd voor vrij te maken om grip te krijgen de verschillende deelprojecten, hoe deze in elkaar haken en welk verhaal ze gezamenlijk vertellen. Ook hiervoor dragen we aan frequente ontmoetingen tussen

onderzoekers in te richten, zowel van formelere als meer informele aard. Belangrijk is hierbij dat deze ontmoetingen gericht zijn op kennisintegratie, bijvoorbeeld middels geschikte methodieken en werkvormen.

Borg de afstemming tussen het evaluatieonderzoek en de natuurbeleidspraktijk, door een praktijk aan te moedigen waarin coproductie de norm is

Een juiste afstemming tussen evaluatieonderzoek en beleidspraktijk houdt in dat de evaluatie aansluit op de evaluatiebehoefte van de eindgebruikers, en dat zij begrijpen op welke wijze de evaluatie daarin beoogt te voorzien. Zowel onderzoeker- als deelnemer-gedreven elementen zijn daarbij legitiem in de evaluatie, mits het voor de deelnemers helder en transparant is hoe tot deze beslissingen zijn gekomen. Voeling houden met deze evaluatiebehoefte is daarbij aan te bevelen om te zorgen dat deze afstemming van duurzame aard is. Coproductie impliceert openheid, transparantie en afstemming met de deelnemers over alle beslissingen ten aanzien van het evaluatieonderzoek, vergezeld van duidelijke verwachtingen van de onderzoekers en de rol die zij beogen te vervullen.

Maak de meervoudige doelstelling van lerend evalueren – namelijk zowel het verantwoord en leren van beleid – expliciet en maak de deelnemers deelgenoot en eigenaar van deze doelen

Verantwoord en leren van evaluatieonderzoek wordt zowel door onderzoekers, hun organisaties, en door de deelnemers veelal als dichotomie beschouwd. Liever dan vasthouden aan deze zogenaamde tegenstelling en deze te benadrukken en problematiseren raden wij aan de verrijking van het samenbrengen van deze doelen in één evaluatieontwerp in het voetlicht te plaatsen. Deelnemers mede-eigenaar maken van de verschillende overwegingen die onderzoekers hierbij ervaren helpt beide doelen te borgen en deze effectief samen te brengen. Op deze gezamenlijke werkwijze wint het evaluatieonderzoek aan belangrijke kwaliteitscriteria, zoals geloofwaardigheid, legitimiteit en relevantie.

1. Introduction

This report presents the second and final part of our review of the first period of the Natuurpact reflexive evaluation (2014-2017). Together with Wageningen University Research (WUR), the PBL Dutch Environmental Assessment Agency (Dutch: Planbureau voor de Leefomgeving, PBL) has conducted a reflexive evaluation of Dutch decentralized nature policy (PBL & WUR 2017), commissioned by the Ministry of Economic Affairs (Dutch: Ministerie van Economische Zaken, EZ) and the Association of the Provinces of the Netherlands (Dutch: Interprovinciaal Overleg, IPO). It has been decided to take a non-traditional *reflexive* evaluation approach geared towards enhancing the learning processes of those actors involved with nature policy. This decision was motivated by the three ambitions for nature policy for 2027 that have been recorded in the Natuurpact agreement (2013) – halting biodiversity decline, strengthening societal engagement with nature and the relation between nature and economy – and the complex multi-level governance and multi-actor settings in which these have to be attained. Furthermore, the latter two other ambitions have not been translated into concrete attainable goals: the policy process for these ambitions is largely of a goal-seeking nature, to which the evaluation intended to contribute.

What we precisely mean by reflexive evaluation will be explained in section 2, but for now it is important to note that the approach was largely new for the PBL and the WUR. Especially for the PBL, reflexive evaluation approaches entail a significant departure from their conventional approach to evaluation research. Moreover, what is unique about the Natuurpact reflexive evaluation is that it is the first of its scale and duration in the Netherlands; the PBL and the WUR have been commissioned to evaluate every three years up to 2027. To ensure the scientific rigor of the approach, researchers from the VU University Amsterdam Athena Institute were assigned with the task of supporting the Natuurpact project team with the evaluation's design and execution by informing these processes with scientific knowledge and expertise with reflexive research methodology.

The resources – time, funding, capacity – allocated to the Natuurpact reflexive evaluation were significant and the expectations with regards to its outcomes – increased quality, usability and impact on policy practice – were high. This created a need amongst members of the project team to inquire after the Natuurpact evaluation's merit. This need was even greater due to the novelty of the reflexive evaluation approach: have the evaluation's purposes been attained and have the hypothesized outcomes concerning quality, usability and impact been achieved? Through such a

Textbox 1.1 - Value of reflexive evaluation according to participants

Our findings showed that the Natuurpact evaluation has established five categories of values that manifest on both the provincial level, and on a collective level (between governmental levels). First, the evaluation had conceptual value. This means that policymakers felt that the evaluation enriched policy practice with new insights and knowledge and contributed to the development of a more shared understanding and joint vision on the ambitions of nature policy. Secondly, the evaluation resulted in affective value, as it improved trust between EZ and the provinces and incited a strong sense of togetherness amongst the provinces, which they highly valued. Thirdly, the evaluation has a strong strategic value, as it allows for legitimization of policy decisions. Fourthly, the evaluation had network value: the participants' network was expanded and relations were strengthened. Finally, we observed the evaluation to have instrumental value – insights were directly used to inform policy decisions (Verwoerd et al., 2017).

review the project team sought to draw lessons from bottlenecks and successes, in order to improve the implementation of the evaluation for the second evaluation period (2018-2020).

In order to determine whether the aforementioned expectations of the evaluation were met, we were asked to review how the reflexive evaluation approach has been of value to policy and research practice. As a first part of this review, we conducted an analysis of the perspectives of actors who participated with the Natuurpact evaluation, namely representatives from national government, the provincial governments and (to a lesser extent) societal organizations involved with nature policy (Verwoerd et al., 2017). In Textbox 1.1 we provide a brief summary of the primary conclusions of this first part of our review. For the second part of our review we have focused on the research practice in which the evaluation took place. We studied the perspectives of the researchers who were involved with the design and execution of the Natuurpact evaluation, on the value of the reflexive approach for increasing the quality, usability and impact of the knowledge produced during the evaluation. The current report presents the results of this second part of our review.

The purpose of this study was thus to collect the perspectives of the researchers on whether and how the reflexive approach has contributed to the quality, usability and impact of the produced knowledge. Additionally, we sought to understand the differences in perspectives and use these insights to make recommendations for the continued implementation of the Natuurpact reflexive evaluation. As such, this report seeks to answer the following research questions:

1. What perspectives on the quality, usability and impact of the knowledge produced can be identified in researchers' reflections on the value of the reflexive evaluation approach?
2. How should we interpret the differences in perspectives?
3. What recommendations follow from understanding these differences for the continued implementation of the Natuurpact reflexive evaluation?

To answer these questions, we have adopted a qualitative research design. We have interviewed a selection of researchers from the PBL and the WUR who were involved with first evaluation period on the reflexive approach and criteria such as quality, usability and impact of knowledge. Through the lens of ideal-typical 'logics' on evaluation research we have analyzed the data and sought to understand the differences in perspectives that we identified.

As mentioned previously, this current study should be seen as a continuation of our previous work and, although we provide some information in section 2 (Background) of this report, we refer the readers to Verwoerd, De Wildt-Liesveld, & Regeer (2017) for an elaborate account on the Natuurpact evaluation itself. Here you will find further information about the Natuurpact evaluation's context, its commissioners and participants, and how the participants perceive the value of reflexive evaluation for their practice. Moreover, Van Veen, Verwoerd and Regeer et al. (2016) offer an extensive theoretical framework of reflexive evaluation that forms the basis of our reviews and that interested readers may find useful. In the remainder of this report section 2 provides some contextual and theoretical background to help situate the Natuurpact reflexive evaluation; this may be especially useful for those who are unfamiliar with this sort of work. Section 3 presents the research methodology we have applied, and results are featured in section 4. Section 5 presents our conclusions and discussion. In section 6, finally, we provide recommendations for the continuation of the Natuurpact reflexive evaluation.

2. Background

As mentioned in the introduction of this report, the commissioners (the Ministry of Economic Affairs and the provinces) of the Natuurpact evaluation decided, together with the researchers, on a reflexive evaluation approach. This approach was chosen to optimally deal with the multi-level governance and multi-stakeholder complexities of nature policy and its ambitious goals. Exactly *what* this approach entailed and how it may be implemented was not fully crystallized at the outset but rather developed during a subsequent iterative and emergent process. To support this process, the Athena Institute was commissioned to conduct an extensive literature review in order to capture lessons from relevant studies about reflexive evaluation. Using this literature, we were also asked to develop a conceptual framework showing what such an approach would look like. This framework was subsequently used to inform the researchers during the implementation of the Natuurpact reflexive evaluation. In this section, we first give a concise description of what reflexive evaluation is *in theory*¹ and subsequently describe how the Natuurpact reflexive evaluation (period 2014-2017) was organized *in practice*.

Scholars have demonstrated that the process of operationalizing a conceptual research ideal into a real-life research is affected by different logics on what ‘good’ research entails from which researchers draw. In an attempt to understand this process, we begin section 2.2 by explaining two predominant logics (the modernist and responsive) on evaluation research, which we have conceptualized into two ideal-typical evaluation approaches. We use these logics and respective ideal-types in the rest of this report to explain the different perspectives we found on the value of reflexive evaluation for realizing quality, usability and impact of knowledge.

2.1 The theory of reflexive evaluation

In academic literature, participatory and transdisciplinary research is argued to produce socially robust knowledge (Nowotny, 2000). This sort of is better suited to inform problems of a ‘wicked’ (Rittel & Webber, 1973) or ‘intractable’ (Hisschemöller & Hoppe, 1995) nature, than is non-participatory and mono-disciplinary research. By opening up research to society, the produced knowledge becomes enriched with societal perspectives, thereby increasing its legitimacy, (scientific) credibility and overall salience for addressing societal needs (Cash et al., 2002). By nature, reflexive evaluation comprises elements of such a participatory approach. Reflexive evaluation addresses the multi-level governance and multi-actor complexities of the issue at hand by striving to increase the *reflexivity* of the involved stakeholders. This is accomplished by contributing to their understanding of underlying causal mechanisms of the problem and their role within it, and thereby optimizing their capacity for adaptive change. Reflexive evaluation accomplishes these objectives by balancing the two primary functions of evaluation: accountability and learning (Van der Meer & Edelenbos, 2006). This implies that evaluation is undertaken both *during* and parallel to policy processes, and that relevant stakeholders (such as policymakers) are actively involved with the research to ensure the evaluation

¹ For interested readers, we refer to our previous work *Scientific justification of the reflexive evaluation and its application during the Natuurpact (2014-2027) program evaluation*’ by Van Veen et al. (2016) for a more elaborate theoretical description.

design aligns to their evaluative needs. This allows for the evaluation to inform policy processes in a timely manner (Edelenbos & van Buuren, 2005; Michael Quinn Patton, 2000). Ideally, reflexive evaluation is designed in an emergent and iterative manner, thereby allowing it to continuously adapt in order to optimally align with the evaluative needs of its participants. In doing so, we can maximize the impact of the reflexive evaluation on the policy process under scrutiny.

Table 2.1. Characteristics of reflexive evaluation (after Van Veen et al., 2016)

Characteristics of reflexive evaluation	
Evaluation purpose	Addressing highly complex multi-level governance and multi-actor contexts by increasing reflexivity of relevant stakeholders and optimizing their capacity for adaptive change
Role of relevant stakeholders	Relevant stakeholders are active participants who engage with evaluation design and conduct, to ensure alignment of the research to their evaluative needs; broad take on who relevant stakeholders are (e.g. not just policymakers, but also those affected by policy)
Role of researchers	Researchers perform a dual role: they function both as assessor and judge of policy progress (for accountability purposes) and as facilitator of learning processes (for learning purposes). As such, they require not just good research skills, but also 'soft' process skills (facilitating, mediating, etc.)
Relation between policy practice and evaluation research	The evaluation research takes place aligned and parallel to policy practice, to optimally inform current policy processes
Type of knowledge produced	Socially robust knowledge: legitimate, (scientifically) credible and relevant to societal needs

Table 2.1 shows the central theoretical characteristics of reflexive evaluation (after Van Veen, Verwoerd, & Regeer, 2016). However, these characteristics provide little guidance to researchers on *how* exactly to go about putting a reflexive evaluation approach to practice. The Natuurpact researchers were required to operationalize these 'principles' of reflexive evaluation within the political, societal and institutional contexts at hand in order to realize the Natuurpact reflexive evaluation.

2.2 Two ideal-types of policy evaluation

In order to relate reflexive evaluation with other approaches to evaluation, we here present two 'extreme' ideal-typical approaches to evaluation. More precisely, we present two ideal-types of evaluation, each with different understandings of what quality, usability and impact of knowledge means and, therefore, of what 'good' evaluation practice comprises. Of course, ideal types are not supposed to correspond one-on-one to empirically observable phenomena. As Max Weber, who first coined the concept, put it, "[a]n ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytical construct" (Shils & Finch, 1949:90). Indeed, although the theoretical framework assumes ideal-types as extremes, or as opposites in a dichotomy, we wish to emphasize that, in the real-world, approaches to evaluation research tend to encompass elements of both ideal-types simultaneously. Therefore, what one can expect to find 'in the wild', are generally not ideal-typical instances of views on evaluation, but rather more or less successful practices of navigating the logical space that stretches across the extremes.

Different scholars have demonstrated that when progressing from 'ideal-types' to 'real-types', various logics are at play that affect this process of operationalization (Felt, Igelsböck, Schikowitz, & Völker, 2016; Kunseler, 2017; Van der Hel, 2016). These logics comprise perspectives and expectations on what it entails to do 'good' research. In her work, Kunseler distinguishes between the modernist and the responsive² logic, which we have adopted in this report as heuristic to explain the researchers' various takes on the value of reflexive evaluation for increasing the quality, usability and impact of the produced knowledge. In the remaining sections of this section, we explain these logics and their respective ideal-types of policy evaluation.

The modernist ideal-type is related to the historically earlier research tradition. According to this logic, rooted in positivism, reality is singular and objectively knowable. As such, science is able to produce value-free and neutral knowledge to guide human action and inform policy (Jasanoff, 2004). Here, policy evaluation may be seen as a tool for scientists to produce evidence-based knowledge on the effectiveness and (cost)efficiency of policy instruments and on strategies for attaining pre-set policy goals. As a result, the modernist ideal-type evaluation primarily serves the purpose of holding governmental bodies (or any other organization for that matter) accountable for their given tasks. Based on the insights derived from the evaluation research, these governmental bodies may or may not choose to adapt their policies. This ideal-type subsumes a modernist take on the relation between science and policy, and understands the domains of science and policy as being strictly separated (Kunseler & Vasileiadou, 2016). In this relation, researchers perform the role of scientific experts and their expertise enjoys a higher status in relation to lay-knowledge. As such, '*speaking truth to power*' is evaluation's main function.

At the other end of the hypothetical spectrum, we find the responsive logic. This logic is rooted within a social constructivism discourse and views knowledge as being pluralistic. Especially in complex contexts, where a diversity of stakeholders and perspectives on the issue (and its solutions) are involved, knowledge on effectiveness and (cost)efficiency is insufficient (and actually, in some cases, unattainable, for instance due to high levels of uncertainty). It presents only one of the multiple types of knowledge that may be used to adequately inform policy development. According to the responsive ideal, a plurality of knowledges (held by different actors who are involved with the issue) should be included in order to fully understand and justice to the inherent complexity of the issue. In this light, researchers are not authoritative experts, but rather stakeholders with an equal stake and an equally valid knowledge base as the other actors. Through this fully collaborative approach to knowledge production, researchers facilitate and engage in processes of mutual learning by interactive reflection on policy instruments, strategies and goals, and underlying norms and values. In this logic, the domains of science and policy are considered 'intertwined' and inseparable (Kunseler & Vasileiadou, 2016); with actors from both domains functioning simultaneously as both knowledge *producers* and *users*.

For these two logics, the *purpose* of the knowledge produced by evaluation research differs. As mentioned, the modernist logic strives to provide a mechanism for ensuring accountability, whereas the responsive logic primarily aims to facilitate mutual learning processes amongst stakeholders.

² In her work, Kunseler (2017) actually uses the term 'reflexive logic', but for sake of coherency and to omit confusion we use the term 'responsive' to maintain the distinction with the reflexive evaluation approach.

Naturally, the different primary aims of these different logics require different outcomes and indicators for determining whether the objective has been met. These differences, explained in terms principles for quality, usability, and impact are summarized in Table 2.2.

Table 2.2 – Principles via which purposes of evaluation is attained according to both logics

	Modernist logic	Responsive logic
Purpose of the produced knowledge	Evaluation is a mechanism for ensuring accountability (assessment of policy performance against set goals)	Evaluation is a mechanism for informing mutual learning processes (reflection on instruments, strategies, goals, norms and values)
Quality principles	Emphasis on scientific credibility; norms of independence, distance and objective research conducted with scientific autonomy	Emphasis on legitimacy and salience; norms inclusiveness of plurality of perspectives; extensive interaction with relevant actors in order to co-create knowledge
Usability principles	Researcher-driven: evaluation designed to assess policy performance by systematic data collection on inputs, outputs and outcomes	Stakeholder-driven: evaluation collaboratively designed by relevant actors (including researchers) to grasp policy complexity and assess policy in responsive manner to inform learning
Impact principles	Impact is established by the degree to which evaluation holds the evaluated government accountable for their progress on set goals (upwards accountability), which may be used to inform policy decisions	Impact is established by the degree to which the evaluation is able to inform mutual learning processes (reflection on instruments, strategies, goals, norms and values) and thereby policy decisions

2.3 In practice: the Natuurpact reflexive evaluation

2.3.1 Motivation for reflexive evaluation

In the Natuurpact agreement, national government, the provinces and societal organizations involved with nature policy finalized the decentralization of nature policy to the provinces. Through this agreement, the provinces became fully responsible for the development and execution of policy to realize the ambitions of Natuurpact, while national government remained responsible for providing the governmental frameworks and retained its responsibility towards the European Commission for attaining European biodiversity goals (recorded in the Bird –and Habitat Directives and the Water Framework Directive). Through these agreements, nature policy has become increasingly characterized by multi-level governance structures. Furthermore, the ambitions in the Natuurpact are vast and go beyond biodiversity conservation alone. Improving societal engagement with nature has also become an important goal of nature policy. Consequently, the provinces are continually searching for ways to develop and implement policy in collaboration with societal actors, resulting in nature policy becoming an increasingly multi-actor and multi-level endeavor. This multi-level governance and multi-actor character of nature policy was one of the reasons that a reflexive evaluation was chosen. Additionally, the different perspectives on nature and societal involvement and the difficult relations between national and provincial governments and societal organizations were reasons for the commissioners and the researchers to choose for a reflexive evaluation. They hypothesized that through such a participatory approach, the inclusion of experiential knowledge of policy actors would significantly improve the quality of the evaluation. Furthermore, this approach was expected to increase the usability and impact of the evaluation in order to directly improve nature policy and attain the set ambitions by 2027.

2.3.2 Implementation

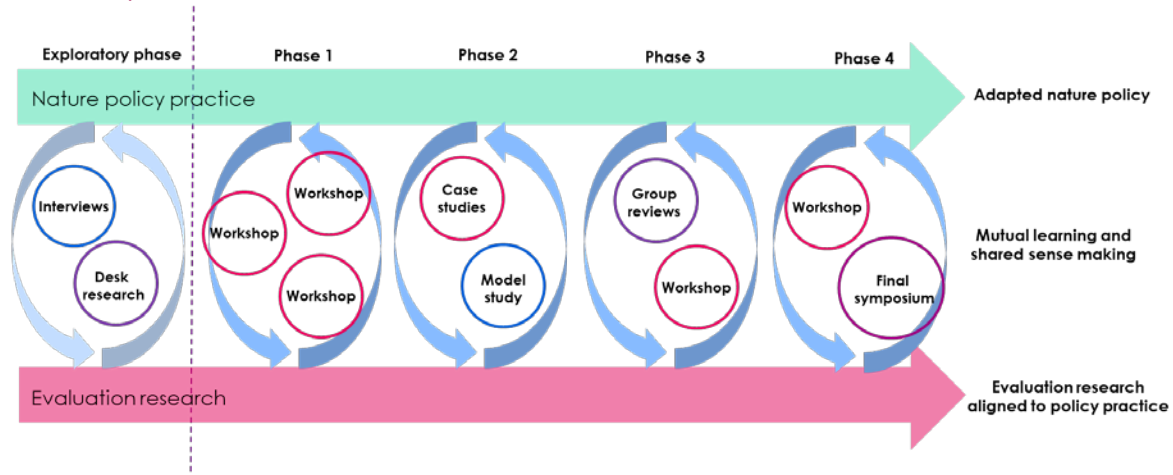


Figure 2.1 – Schematic overview of the interactions between policy practice and the evaluation research during the different evaluation phases

Together with the WUR and the VU, researchers from PBL designed the evaluation as an interactive process between nature policy practice and the evaluation research. The evaluation research may be perceived as consisting of four major phases:

- 1) collectively determining the evaluation scope and its demarcations;
- 2) joint collection and analysis of data;
- 3) joint sense making and interpretation of findings, and;
- 4) formulating action perspectives for policy improvement and dissemination.

Throughout these different phases, frequent interaction sessions were organized with representatives from nature policy practice. These sessions intended to simultaneously inform both the research and policy processes, thereby ensuring alignment of the two domains. These interactions occurred at three different levels. First, interaction occurred collectively, during dedicated learning workshops. During these sessions national and provincial governments and societal organizations joined with the researchers to discuss overarching topics. These topics included: the evaluation scope and demarcations, evaluation results about the provinces' renewed policy strategies and, in the final phases of the evaluation, the action perspectives resulting from the evaluation's preliminary conclusions. Secondly, interaction occurred through the formal IPO workgroup Nature Policy, who guided the evaluation's content and conduct. The workgroup Nature Policy meets monthly and, in this group, each province is represented by one policymaker. These representatives were responsible for organizing and communicating the evaluation process and progress within their respective province. As such, they formed an important link between their province and the evaluation research. Third and finally, interaction between policy practice and evaluation research occurred as part of the *research* activities. Examples of these interactions include, data collection during interviews, group discussions during case studies on policy innovations, and group reviews on the progress of provincial policies.

At the onset of this study we expected that the logics researcher's draw from in operationalizing the reflexive evaluation, discussed in the previous section, to be at play in all phases and at all

aforementioned levels. In the following sections, we first discuss our methodology (section 3) and then we review the perspectives of the researchers on the quality, usability and impact of the produced knowledge and on the challenges they encountered in conducting the Natuurpact reflexive evaluation (section 4).

3. Methodology

In this section we discuss the methods used to answer our main research questions. We adopted an exclusively qualitative research approach, as this is most appropriate for obtaining in-depth insight into personal opinions and perspectives as well as for gaining a holistic understanding of the subject under scrutiny (Creswell, 2009; Creswell & Miller, 2000).

3.1 Data collection

3.1.1 Document analysis

Data was collected by means of document analysis and semi-structured interviews. For document analysis, the report of the self-evaluation (internal report) conducted by the evaluation researchers after publication of the first evaluation report was used. The self-evaluation was analysed for what the evaluation researchers perceived as successes and mishaps of the project, as well as their perspectives on the factors that caused these. Furthermore, we made use of our previous review on the value of reflexive evaluation for participants and factors that contributed to this value (Verwoerd et al., 2017). As such, the self-evaluation and the review informed the design of the semi-structured interviews.

3.1.2 Interviews

A total of seven evaluation researchers involved in the reflexive evaluation of the Natuurpact were interviewed — see Table 3.1³. They were selected based on their roles in the evaluation process. Most of the interviewees were leaders of sub-projects or performed otherwise crucial roles. The selected researchers also reflected the diversity of the researchers involved in the evaluation, in terms of disciplinary backgrounds, range of sub-projects researchers were involved with, and station. Researchers with economical, ecological and public-administrative backgrounds, both from PBL and WUR, who were involved with the different sub-projects within the reflexive evaluation were interviewed. More evaluation researchers were considered for selection, but since data saturation occurred after six interviews it was decided not to continue after the seventh interview.

Table 3.1 – Interviewees and their basis characteristics

Researcher	Organization	Discipline	Primary function
1	PBL	Public-administration	Researcher
2	WUR	Public-administration	Sub-project leader
3	WUR	Public-administration	Sub-project leader
4	WUR	Public-administration	Project leader
5	PBL	Ecology	Sub-project leader
6	PBL	Ecology	Researcher
7	WUR	Economy	Sub-project leader

³ What complicates matters is that several researchers switched roles during the evaluation period and that some researchers have multi-disciplinary backgrounds. For the sake of simplicity, Table 3.1 does not attempt to capture this, but rather emphasizes the main backgrounds and roles of the interviewees.

The interviews were semi-structured, meaning a topic list was loosely used (based on the document analysis) which allowed sufficient room for divergence should the respondent offer other more pressing topics. Table 3.2 displays the topic list used.

Table 3.2 Overview of the interview guide and its topic list

Topic list	Suggested questions
Quality of knowledge	<ul style="list-style-type: none"> - To what extent has the reflexive evaluation approach resulted in higher quality evaluation outcomes? <ul style="list-style-type: none"> o How do you understand 'quality'? o Which factors affected quality? o ...
Usability of knowledge	<ul style="list-style-type: none"> - To what extent has the reflexive evaluation approach resulted in higher usability of evaluation outcomes? <ul style="list-style-type: none"> o How do you understand 'usability'? o On what do you base your conclusion? o Which factors affected usability? o ...
Impact of knowledge	<ul style="list-style-type: none"> - To what extent has the reflexive evaluation approach resulted in higher impact of evaluation outcomes? <ul style="list-style-type: none"> o What do you consider 'impact'? o On what do you base your conclusion? o Which factors affected impact? o ...
Interrelations quality, usability and impact	<ul style="list-style-type: none"> - How do you perceive quality, usability and impact of knowledge interrelate (reinforcing, trade-offs, etc.)?
External evaluation process (transdisciplinary)	<ul style="list-style-type: none"> - How do you look back on the external process; e.g. the interaction with the participants)? <ul style="list-style-type: none"> o Highlights/dilemmas/lessons
Internal evaluation process (interdisciplinary)	<ul style="list-style-type: none"> - How do you look back on the internal process; e.g. the interaction between the evaluation researchers and the sub-projects)? <ul style="list-style-type: none"> o Highlights/dilemmas/lessons
Learning and accountability	<ul style="list-style-type: none"> - How do you perceive the evaluation functions of learning and accountability were reconciled this evaluation period?
The role of societal partners	<ul style="list-style-type: none"> - What is your perspective on the role of societal partners; what are reasons for (not) including them? <ul style="list-style-type: none"> o How should an increase in evaluation participants be anticipated (also regarding data bulk)?

3.2 Data analysis

All interviews were transcribed and subsequently analysed. Open coding (with ATLAS.ti, 8.0.33.0) was used in combination with framework-based sensitizing concepts (e.g. Bowen, 2008) similar to the topic list in Table 3.1. As such, our coding strategy may be considered simultaneously predetermined and emergent. The sensitizing concepts included for instance knowledge quality, interdisciplinary collaboration, interaction with evaluation participants. Emergent concepts included differentiations of the predetermined concepts (e.g., specification of the understanding of knowledge quality), organisational differences between researchers, different understandings of participant-oriented evaluation approaches.

Researchers coded transcripts both individually and in collaboration. This was done by one researcher who has been intensively involved with the Natuurpact evaluation for the past two years, and one researcher who was new to the Natuurpact case. Wherever possible, data, researchers, and

methodological (i.e., observations and interviews) triangulation was applied for cross-reference, to guard against bias and to ensure completeness.

3.3 Validation

Participation in the interviews was voluntary. All interviews were audio-recorded and transcribed with permission of the respondents. After each interview member checks were done by sending back to the interviewees selected highlights of the interview. This allowed us to ensure correct interpretation for quality purposes. Quotes used in this report have been approved of by the respective respondents.

The preliminary findings of this study were also shared during a validation workshop. During this workshop, a number of respondents from this study as well as researchers and supervisors involved with the second evaluation period were present to assess the validity of our interpretations and conclusions and to reflect upon the implications of these findings for the future of the Natuurpact reflexive evaluation⁴. Based on their feedback and reflections we have adapted and added to the interpretations in our Results and Conclusion sections.

3.4 Our role

As it would be somewhat dubious, if not inappropriate, if we would not devote at least *some* words to our dual role during the Natuurpact reflexive evaluation, we end this methodology section with a brief reflection hereon. Two of the authors of this current report (Verwoerd and Regeer) have been involved with the Natuurpact evaluation from its onset. Our assignment, commissioned by PBL, included feeding the evaluation design and process with the necessary scientific theoretical and conceptual building blocks for reflexive evaluation. As such, we had a role in how the evaluation process was designed. As part of our assignment, we were also given the task to review the evaluation's value for nature policy practice. Here our dual role becomes visible, as this implied we would have to review an evaluation that we ourselves helped shape. Although being so close to the evaluation process gives significant advantages, as it allowed us a deeper and accurate understanding of the researchers' practice and their perspectives than an 'outsider' could reach, it of course is not without tension. Indeed, this dual role and close interaction also brings a potential risk in terms of unintentional researcher bias. Keeping an open mind and maintaining a wide breadth of interpretive possibilities may be difficult when a research project and its researchers are so familiar. The validation activities discussed in section 3.3 were done to help manage this tension. To further help us refrain from prejudicial interpretations or jumping to conclusions, our research team was expanded with a researcher who had no prior involvement with the Natuurpact evaluation (Klaassen). His role, as a relative outsider, included asking critical questions, putting assumptions to the test, formulating alternative hypotheses or interpretations and bringing in fresh (theoretical) perspectives. The data analysis was done by multiple researchers and through frequent reflection on interpretations (also with other researchers from our department) we aimed to reduce inadvertent researcher bias.

⁴ These reflections specifically, have landed in a separate document, titled *Process principles for reflexive evaluation*. This contains a shared operationalization of principles for reflexive evaluation and is used by researchers in designing and executing the second evaluation period (2018-2020).

Furthermore, we sought to guard internal validity through member checks with the respondents and the validation workshop with respondents and additional researchers from the second evaluation period, as explained above.

4. Results

The working hypothesis behind the Natuurpact evaluation was that a reflexive evaluation approach would result in increased quality, usability and impact of the knowledge produced. In this section we present researchers' perspectives on quality, usability and impact of the produced knowledge of the Natuurpact evaluation and explain the differences in understanding by use of the logics presented earlier (section 4.1). We then continue with a reflection on the different arenas in which the researchers draw from these logics to negotiate the design and execution of the reflexive evaluation amongst themselves and other involved stakeholders (section 4.2).

4.1 Researchers perspectives on quality, usability and impact

4.1.1 Quality

Understanding quality as Scientificity vs. Usability

We found two perspectives on value of reflexive evaluation for the quality of the produced knowledge. The first understands quality as to mean *scientificity* (being scientific). In this perspective, quality is first and foremost associated with objectivity and independence. The fact that the reflexive nature of the evaluation entailed close interaction with provincial policymakers is, in this perspective, viewed as a threat to the quality of the knowledge produced. What stands out is that researchers who expressed this perspective, simultaneously expressed the upsides of producing knowledge in interaction with those for whom it is intended, as its benefits the recognisability and usability of the knowledge. This perspective understands quality as the extend with which it is found usable by those for whom it is intended (Table 4.1, #1 & #2). In this perspective, the interdisciplinary and transdisciplinary character of the produced knowledge is a measure for its quality.

Table 4.1 – Quotes on quality

#	Quotes
1	<i>'You could tell we really aligned to their [the policymakers', ed.] frames of reference; they were able to quickly translate the results to their own practice and to discuss these with us. This for me really demonstrates quality.'</i> (ER4)
2	<i>'By linking the biodiversity and administrative sub projects, I really started to speak their [the provincial policymakers'] language, I think. They understood how we produced the results and recognized how it was relevant to them.'</i> (ER1)

Whether interaction with participants was regarded as potential threat to quality, thus which perspective is expressed, depended on which research phase the researchers were discussing. For relatively early phases such as formulating the evaluation research questions, data collection (Table 4.2, #3) and data analysis, the researchers were more inclined to value interaction with the participants, as this promotes the alignment of the evaluation research to the participants' frames of reference. Researchers appreciate that this accommodates the recognisability and relevance of the findings and, as such, their quality. However, when discussing other phases, including method determination and formulating conclusions (#4), most researchers tended to perceive interaction as

less appropriate and possibly obstructive for the scientificity of the research. As a result, there was a tendency and desire among several researchers to limit interaction with the provinces during the mentioned phases. Which perspective researchers are inclined to draw from thus correlates with which evaluation phase was at hand — rather than seeing a one-on-one relationship between individual researchers and specific (elements of) evaluation logics, one sees researchers navigating the logical space. We return to this in section 4.2.

Table 4.2 – Quotes on quality

#	Quotes
3	<i>‘Collecting data via bilateral consultations [in-depth interviews, ed.] allowed us to understand how we should interpret their policy plans. We never would’ve been able to do so correctly without their explanations. So yes, that benefited quality.’ (ER5)</i>
4	<i>‘Regarding scientific quality, when we were reporting our findings some provinces really pushed for alternative framings. The pressure was high: we also didn’t want them to disregard the findings as false, because they find them unrecognizable. It is a tension inherent when combining evaluation for learning and for accountability.’ (ER2)</i>

This schematic divide – scientificity and usability – on the meaning of quality, did not entail a comparable divide as regards what value researchers give to inter- and transdisciplinarity. We somewhat expected this due to the implications both approaches have on interaction with participants throughout the evaluation. Rather, despite the dual and simultaneous understanding of how research quality is obtained – via scientificity or usability – the researchers were unanimous that the inter- and transdisciplinary elements of the reflexive evaluation approach positively contributed to the quality of the outcomes. They also all agreed that the quality of the produced knowledge could have been even higher. Suggestions on *how* higher quality knowledge might have been obtained, however, opinions do diverge — in ways consistent with the schematic divide. In short: some researchers argued that to this end scientificity should be improved, while others considered improving transdisciplinarity to be instrumental in this regard.

Quality in light of the ideal-types

According to the modernist ideal-type, quality is something to safe-guard. This can best be done by conducting distant, objective and independent research. Researchers who drew from this logic, operationalized quality as *having no interaction* (beyond data collection and dissemination of the findings) with the provinces. However, following the responsive ideal-type, quality is something to pursue. This is done by producing findings that are recognizable, applicable and, thus, usable by the participants. On this operationalization of quality, *having interaction* (during all evaluation phases) with the provinces is requisite to attain quality.

As was mentioned before, in order to navigate the tension between these two operationalizations of quality, most researchers were inclined to delineate interaction with the provinces to specific evaluation phases for which they deemed interaction appropriate. Thus, in practice, the researchers leaned more towards the modernist logic. In addition, they appointed specific researchers who would take it upon them to ‘do’ the interaction, and clearly demarcated and communicated what the purpose of the interaction was and the ‘rules to play by’. For instance, when presenting preliminary conclusions, the researchers clearly communicated to the provinces that interaction about the findings was only intended for checking factual inaccuracies and to adjust the conclusions in light of

these. They made clear that whether or not the participants agreed with the conclusions that were drawn, or *how* these were drawn in terms of methods, would not affect their content – to clearly communicate and guarantee their independent judgement would not be affected (Table 4.3, #5).

A number of researchers expressed the tension between the different logics’ quality (#6). In contemplating this, they articulated a new operationalization of quality altogether: *having equal interaction* with all relevant stakeholders (not just the provinces, but also e.g. societal organisations) and being *transparent* about these interactions. One researcher put forward that such a shared operationalization would have allowed for both independence *and* interaction, thereby navigating both logics (#7).

Table 4.3 – Quotes on quality

#	Quotes
5	<i>‘We asked them only to comment on factual inaccuracies and to test whether the conclusions were recognizable. We communicated clearly we wouldn’t alter our findings based on their comments.’ (ER5)</i>
6	<i>‘I spoke with two to three provinces a day I think; some really pushed hard for us to frame things their way. Staying objective and true to our results was paramount, as was aligning to their frames of reference.’ (ER1)</i>
7	<i>‘A well-thought out strategy on how to deal with pushy actors and also ensuring not just one actor gives feedback on results, but all of them, would’ve been nice to have at the start, I guess. Then you can test the quality of your findings and ensure independence at the same time.’ (ER2)</i>

4.1.2 Usability

Understanding usability as Recognizability, understandability and applicability

We found that the perspectives on usability were consistent with what we previously established, namely that the perspectives on usability are intricately intertwined with that of quality (we refer back to quotes #1 and #2 from Table 4.1, quotes which both illustrate how usability of findings is used as a metric for quality). One perspective on usability we found emphasizes the extent to which the produced knowledge is recognisable (i.e. by relevant topics, relevant research questions) and understandable (i.e. aligning to frames of references, using shared language). All researchers expressed this perspective and were, in general, positive towards it. Another perspective we found emphasized the applicability of the produced knowledge as a measure for usability –where usability here is the extent to which policymakers could readily apply the knowledge to inform policy decisions. Specifically, when emphasizing usability as applicability, some researches felt that the usability of the produced knowledge could have been greater.

Mostly the researchers with a public administration background held that applicability could have been improved if the evaluation would have been better aligned to the ‘evaluative needs’ of the provinces. The main culprit for suboptimal applicability, according to these researchers, was that parts of the evaluation research were decided on the basis of routine and without thinking through what the evaluative needs of participants meant for the type of research that was required. Importantly, these researchers believed that this resulted in a mismatch between the national level, on which level the conclusions were applicable, versus the provincial level, which is the level of the

provinces evaluative needs (Table 4.4, #8). An important element in this mismatch was the model that was chosen for the biodiversity assessments, which loses scientific rigour when used on more local or regional scales. This model – called the MetaNatuurplanner – was decided upon by the researchers responsible for the ecological research. To overcome this mismatch, it was decided to add so-called ‘fact sheets’ that provided more insight into individual provinces’ ecological results. This exercise proved highly time-consuming and, although beneficial for usability according to most researchers (#9), some maintained the scale mismatch was a major issue for usability.

Table 4.4 - Quotes on usability

#	Quotes
8	<i>‘I think this is the biggest issue for usability, and it still hasn’t been tackled. PBL believes it should deliver national analyses, a national picture – it’s their core business, so to them it was just evident that specific models would be used to provide this national image. But for provinces, their true policy questions transpire on the provincial level. The evaluation doesn’t provide answers to these questions and you can’t translate the national picture to provincial action perspectives, so usability will always be less than optimal. This cannot be fundamentally solved if the researchers routinely continue to give national picture precedence.’ (ER4)</i>
9	<i>‘That we wouldn’t provide findings on provincial scale came as a surprise to the provinces. To us this was only logical, the model we’ve used simply doesn’t give such output. To still meet their needs we decided to make fact sheets for each province with all relevant information on their policies, etc. It comprised the largest part of my work in the final six months. But we were really responsive and it really improved usability in my view.’(ER1)</i>

Usability in light of the ideal-types

According to the modernist ideal-type, evaluation research should be *researcher-driven*. This means that, although the evaluation’s commissioners usually provide general questions and the goal for the evaluation (for instance, assess the effectiveness of a policy program), expert scientists ultimately decide *how* this goal may best be achieved. What are feasible research questions, demarcations and the appropriate methodology? Researchers independently conduct the analysis, draw conclusions and disseminate the findings. More specifically, in policy evaluation, experts focus on assessing whether the policy under scrutiny succeeds in attaining its intended goals. In this perspective, usability is operationalized as evaluation results that draw policymakers’ attention to policy failures (and successes), thereby informing policy agendas. On the ideal-type of responsiveness, however, researchers ideally work in a purely *participant-driven* fashion. They are facilitators and the participants of the evaluation are involved in all major evaluation decisions; the research questions, scope, interpreting data and drawing joint conclusions. Usability is operationalized here as evaluation that is fully salient to the participants’ evaluative needs (e.g. in scope, language, time-frame, the way findings are published, *et cetera*).

For the Natuurpact evaluation, decisions on its content and scope were made in consultation with the provinces and were based on the inventory of evaluative-needs – from the onset, the research team intended to work participant-driven. However, these decisions did not all find their way in the eventual evaluation research that was conducted (Table 4.5, #10). When executing the evaluation research, some researchers – mostly from the PBL – seemed to default into working in a researcher-driven manner and worked on the basis of what they considered to be important for evaluation. Indeed, this is underlined by the PBL-researchers who felt the evaluation rather was too participant-

driven; in their point of view the perspectives of participants on certain decisions has little merit and it is the prerogative of the researchers to decide on matters such as design and scope (#11).

Table 4.5 – Quotes on usability

#	Quote:
10	<i>'I had hardly started thinking about the project, when I understood other researchers already finalized their lists of interview questions for the provinces – I thought we'd first inventory more what their needs were, and then decide on demarcations, on which policies we are focusing... You run the risk of collecting way too much data without knowing what to use them for.'</i> (ER2)
11	<i>'It is our prerogative as researchers to, from our independent position, signal and put evaluation topics on the agenda we feel are relevant, of course while considering the political context of the issue. [...] Of course, you want to take into account their needs, but we are also allowed to say 'no' at some point, there needs to be a line. In trying to accommodate all their needs, there just was so much extra work.'</i> (ER6)

Indeed, different ideas on whether to work researcher- or participant-driven can be identified amongst the researchers, and this gave rise to discussions within the project team. The most notable we touched upon earlier (i.e., the 'mismatch' of scales). The use of the MetaNatuurplanner for the ecological research of the evaluation was a researcher-driven decision. According to some researchers, this decision was made on the basis of a specific understanding about the role of PBL as a national institute, and part of PBL and WUR's in-house expertise and the institutes' élan. To them, it was logical and self-explanatory that this method would be used, while others held different opinions (Table 4.6, #12 and #13).

Table 4.6 – Quotes on usability

#	Quote:
12	<i>'That we would use the MetaNatuurplanner was self-evident. So, we decided how there would be measured, that is part of the expertise we offer. It was known from the start there would be no conclusions on the provincial scale. And the national picture is really of value too: it allows the provinces to gain a sense of where they and their policies fit in the larger scheme of biodiversity.'</i> (ER6)
13	<i>'Even before it all started, it was decided the MetaNatuurplanner would be used! But those who made that decisions, you'd expect them to also be the ones who'd knew that the model is not suitable for provincial scale. It caused us so much time and effort to explain this to the provinces, and we had to repair this with the fact-sheets. It was really an issue, the decision which model to use should have been informed by the demands of the end-users.'</i> (ER1)

Some researchers (mostly those with a background in public administration, from the WUR) reflected that the evaluation was too researcher-driven for other issues as well. To give one example, these researchers explained how the ecologists from PBL felt the international Convention on Biological Diversity (CBD) biodiversity targets were an important international government framework to incorporate into the evaluation. Even though the provinces hardly mentioned the CBD targets when inventorying their evaluative needs at the start of the evaluation, the framework remained part of the evaluation research up until its final phase. In the end, the findings on the CBD targets were excluded from the final report as they were not considered to be of interest to the provinces, despite the amount of work that was already put in (Table 4.7, #12).

Table 4.7 – Quotes on usability

#	Quotes
---	--------

12 *'At the start we really worked to develop this shared evaluation framework; what are we evaluating, what are the goals? What are provinces doing, what do they want to know? But even up to the final stages of writing the report, some researchers insisted on saying something about the CBD [Convention on Biological Diversity] targets – while the provinces never even mentioned these. You can't just use a target that's not considered relevant for those under evaluation as benchmark in a reflexive evaluation setting. The evaluation framework should be recognizable to those for whom the evaluation is intended.'* (ER4)

We observe that these issues on working researcher- or participant-driven mainly concerned the ecological sub-research project of the evaluation. Notably, the inclination to work researcher-driven is mostly seen in the ecologists, whereas the public administration researchers were more familiar with the participant-driven perspective. The conflicting ideas on how research 'should' be initiated remained mostly implicit and led to frustration and to misunderstandings of each other, thereby highlighting the importance of explicating early on the expectations and underlying norms on what is considered 'good' research.

Interestingly, in their reflections, the researchers (from both organizations) searched for reconciliation of both ideal-types. Rather than working either researcher- or participant-driven, they seemed to contemplate how to work within what we understand as 'coproduction driven'. They underlined not only the prerogative of their expert-positions, but also the importance of accommodating the participants' evaluative needs. One of the researchers explained that a more open dialogue on the chosen models and methods – what these can and cannot do and what their underlying assumptions are – and attuning these to the evaluative needs of the participants would have met the standards from both logics, thereby benefitting overall usability (Table 4.8, #15).

Table 4.8 – Quotes on usability

#	Quotes
15	<i>'It would have helped greatly, I think, if the methods had been discussed more. Maybe not in terms of co-deciding on which methods to use, but at least to have more of a discussion on how these methods work and what insights they provide.'</i> (ER4)

4.1.3 Impact

Understanding impact as Learning vs. Accountability

The perspectives on impact share that impact is understood as the extent to which the evaluation affects nature policy practice. *How* this effect is achieved is where the perspectives differ. The first perspective views impact as the extent to which participants have *learned* from the evaluation process and its findings in order to enrich their policy practice. According to this perspective, the uptake of findings in the participants' vocabulary when discussing nature policy and in their provincial Nature Visions⁵ is indicative of informed learning processes (Table 4.9, #16, #17 & #18). Once more, in this perspective there is a close relation between the concepts of quality, usability and impact.

Table 4.9 – Quotes on impact

#	Quotes
---	--------

⁵ In their Nature Visions, the provinces formulate their vision on their provincial nature for the coming years.

16	<i>'The concepts we proposed, such as "policy plan potential" and "implementation potential" – they started using these themselves. That's an important success, I think.'</i> (ER1)
17	<i>'I think usability is especially visible in how the evaluation findings seem to have informed some of the provinces' Nature Visions.'</i> (ER7)
18	<i>'That the evaluation has had impact is evident I think, findings are recognized and accepted by the provinces. [...] You really see the evaluation served as inspiration for the Nature Vision, it is really usable in giving shape to their own provincial policy.'</i> (ER3)

Another perspective views impact as something that is achieved via the route of accountability. Impact here is understood as the extent to which evaluation allows for the provinces to be held accountable to the national government (i.e. their commissioner) in regards to their progress on the ambitions agreed upon in the Natuurpact. In this perspective, evaluation has impact when it legitimatizes policy change.

All researchers acknowledged both perspectives on impact and considered learning and accountability as equally valid and important routes by which it may be achieved. Both purposes were pursued during the Natuurpact evaluation. In practice, this created tensions as the two routes appear to be at odds with one another. It asks from researchers to function as distant, neutral judges on the one hand (accountability) and engaged facilitators (of learning processes) on the other. This tension between accountability and learning was further complicated by the provinces' different purposes for the evaluation. Some required the report to primarily fill the purpose of accountability and, therefore, to contain sharply formulated conclusions so the report would have more legitimizing value to them. Other provinces desired a more positive framing of the findings to underline the learning character of the evaluation (Table 4.10, #19). This tension between perspectives on the purpose of the evaluation led to elaborate discussions on the framing of the conclusions between the researchers and some of the provinces (#20 and #21, both on the final phases of writing the evaluation report).

When inquiring after whether the researchers felt the reflexive approach had benefitted the impact of the produced knowledge, they found it harder to assess impact compared to quality or usability, especially when they considered the perspective of accountability as impact. They felt it was too soon to draw any conclusion on this matter, as the first evaluation period had only just been concluded.

Table 4.10 – Quotes on impact

#	Quote:
19	<i>'One wanted us to be nicer and more positive for the conclusions to find uptake with his colleagues, while another wanted us to be very sharp so he could use it as a strategic message to his supervisor. It was impossible to accommodate all their wishes at the same time.'</i> (ER5)
20	<i>'Our external independent judgement could have been much sharper. But it was about balancing impact and endorsement, and this was really difficult. In the final phases of writing, I spoke with two provinces a day who kept calling and pushing for different phrasing, or removal of findings altogether.'</i> (ER1)
21	<i>'The interactive process when reporting was tricky – I wonder if perhaps some of the conclusions have been less sharply formulated because of it, at least for the case study research that was done on innovative provincial strategies. It was a major glass half-full, half-empty discussion. But it's about navigating impact and scientific quality.'</i> (ER2)

Impact in light of the ideal-types

The two purposes of evaluation, rooted in the perspectives we found, correspond to the two ideal-types. The underlying logics also explain why the two purposes cause friction in practice. From a modernist point of view, evaluation for accountability benefits from scientificity; the study's objectivity and independency. The second route conforms to the responsive point of view, namely that in order to encourage learning processes and thus have impact, it is necessary to engage with the participants to align to their frames of reference and needs, and build trust. The issue corresponds to our earlier discussion on the operationalization of quality in terms of (no) interaction with participants.

Regardless of the routes via which impact may be achieved, the researchers all understood impact as something that is attained or becomes visible after a certain amount of time has passed. This in itself is a more modern understanding of the concept. From a responsive standpoint, impact is viewed as a more incremental process that manifests in implicit and indirect ways. It may, for instance, be observed in changes in people's narratives and frames and the language they use. Impact is thus something that occurs continuously and is difficult to establish in a relatively short time frame. From a modernist standpoint, impact is viewed as something that is achieved at the end of a research project when major decisions are made based on its final conclusions. It is more easily observed, but only after sufficient time has progressed for the relevant actors to make their decisions and act on the final conclusions.

So far, we have discussed the perspectives of the researchers on the quality, usability and impact of the produced knowledge. We have attempted to explain the different perspectives expressed by demonstrating how these relate to two ideal-typical views on evaluation research that researchers draw from when designing and conducting evaluation research. The question then rises of what it is that determines or predisposes researchers to one or the other ideal-typical view of evaluation. Also, if logics are employed by researchers when reflecting on the value of reflexive evaluation in retrospect, it is rather likely these same logics played a role *during* the evaluation's design and execution. We discuss both these matters in the next section.

4.2 Navigating logics during the Natuurpact evaluation

In the previous paragraph, we established that, in designing and conducting the reflexive evaluation, researchers draw from different logics to negotiate the 'right' course of action. The question then arises of which logic is favored and why. To gain understanding in this matter, we look at the three main arenas in which the negotiations on the reflexive evaluation took place. The first is within the research team itself, comprising the project team and other researchers that were closely involved. The second arena may be perceived between the research team and their home organizations (PBL and WUR). Finally, the third arena concerns the team and the 'outside world'; those actors from nature policy practice, both participants and non-participants of the evaluation. Naturally, these three arenas stand in close interaction and mutually affect each other – we use the distinction here for analytical purposes.

4.2.1 Negotiating reflexive evaluation within the research team

The research team was the primary arena in which collaborations and interdisciplinary knowledge integration – the main process of a successful evaluation, according to the researchers – occurred.

The researchers were unanimously positive about their collaborative efforts, although most also regard it as a challenging process. From their reflections, different expectations and preferred ways of working became evident, which explain the challenges they experienced and which correspond, in part, to the logics we have found to be at play.

In the previous paragraphs we already touched upon moments where different logics inspired discussion between the researchers. For example, the matter of the intensity of interaction with the participants and during which phases of the evaluation this was appropriate. Or, when some researchers defaulted into working researcher-driven while others felt researcher-driven was in order. Our findings suggest that there is some relationship between disciplinary background and affinity with either ideal-type. We also observe, however, that this is not conclusive. Researchers whose primary expertise is in one of the natural sciences appeared less comfortable with the responsive ideal-type than those who are more at home in a gamma-discipline like public administration. However, researchers' solidified routines appeared to be of equal importance. For example, despite formulated reflexive ambitions and the intent to align to the provinces' evaluative needs, the ecologist initiated a researcher-driven approach in 'their' parts of the research. The know-how, or skill, to break with routines and researcher-driven research seemingly lacked. Furthermore, their normative views on how government – or society as a whole for that matter – should interact with nature, seem an important determinant for how researchers understand their legitimate researcher role in relation to policy processes. This, in turn, determines the type of evaluation approach they will find appropriate. This in itself is of little surprise as the logics on ideal-typical evaluation are rooted in and intertwined with logics on society and how it, and its governance, should function. These logics on society and governance help to place the researchers' negotiations in a broader perspective.

When discussing the collaboration and knowledge integration within the research team, several challenges also arose that cannot be readily appointed to a conflict in logics. Rather, they were a matter of project management. We discuss these here in brief. To start, researchers commented that time pressure was a major compromising factor throughout the evaluation. High workloads and pressing deadlines caused them to focus on their own tasks first, rather than prioritizing knowledge integration. This was especially the case towards the end of the evaluation period, when the final reports were written (Table 4.11, #22 & #23). Most researchers shared the view that the way collaboration and knowledge integration was organized and prioritized were the main culprits. For example, integration was intended to occur during bi-weekly meetings of the subproject leaders. However, those meetings frequently filled up with other topics that were felt to be more pressing at the time, manifesting the familiar dynamic in which urgent matters take precedence over important ones (that are perceived as less urgent). Although all researchers agreed that interdisciplinary knowledge integration was important, in practice it was secondary to completing their individual tasks. Some researchers reflected that there was limited ownership in regards to the integration of the various elements of the report (#24). The underlying issue seems to be that integration was not structurally embedded within the evaluation design (#25). This complicated the process of finding connections and relations between the different subprojects, which is necessary for knowledge integration. Lack of embedding and of reward structures for integration made it difficult for researchers to prioritize integration. Taken together, these findings help explain why integration was not generally perceived as urgent (even though it was perceived as important).

Table 4.11 – Quotes on the research team arena

#	Quotes
22	<i>'It really was the intention and expectation that the integration of all subparts would happen, in particular within the ecological part. And even though we all tried to connect, in the end we didn't succeed. The time pressure became really high for all of us, and we all fell back in our default mode: prioritizing our own tasks and forgetting integration. (ER5)</i>
23	<i>'In the end it meant the project leaders took it upon them to integrate the ecological and public administrative findings in the final report. Not all researchers feel this was done correctly, not all recognize their work in it, I think. I'm not sure if it's a bad thing, or that it means anything for the quality... But it would've been better if we'd have a discussion on integration before actually writing things done, then the conclusions might have been more shared.'</i> (ER4)
24	<i>'It was like, we all knew it had to happen, but everyone waited for someone else to actually do it.'</i> (ER3)
25	<i>'When pressure became high, we all drew back on our own separate islands. Integration should have occurred from the start, and should have been structurally embedded in the entire evaluation design, I think.'</i> (ER5)

4.2.2 Negotiating reflexive evaluation with researchers' home institutions (WUR and PBL)

The second arena where the design and execution of the Natuurpact evaluation was negotiated was with and within the researchers' home institutions. The dominant culture and structures of a researcher's home organization seem to be important determinants for the type of research approach that is adopted and executed. Both the PBL and WUR are organizations of which the culture and structures are rooted within the modernist tradition. This encompasses specific ideas on the appropriate role and position of a knowledge institute within the policy arena, namely at distance from politics. Especially for the PBL, a public knowledge institute that is organizationally part of national government, the modernist culture and its emphasis on the independency of the organization as one of its core values are dominant. This dominance was exerted by the organization's hierarchical structure. For all their 'top-projects', PBL researchers are required to account for and adapt their research design and progress at several moments during the course of the project. The Natuurpact's innovative approach was heavily scrutinized by colleagues, management and direction for its scientific rigor and objectivity (Table 4.12, #26 and #27). This hierarchical structure made it more difficult for PBL researchers to diverge from the institute's modernist tradition. Although experienced as a burden at times, especially when deadlines were pressing, researchers also valued the endorsement of management and direction for their approach. They perceive it to be elemental for its success even though it meant their reflexive aspirations were at times bounded. With regards to the WUR, their larger distance to policy processes and the more horizontal structure of their institute makes that the researchers have more leeway in the design of their studies. To the WUR researchers, the dominant logic of their home institution seems less of an issue.

Table 4.12 – Quotes on the institutional arena

#	Quotes
26	<i>'The organizations are just structured so differently. We are more flat—a real project-oriented organization. To a large degree, you are your own boss when you are doing a project. Whereas PBL has a much more hierarchical structure that affects how a project progresses, with much more internal</i>

accountability. That's not necessarily a bad thing – it also makes the PBL a strong organization – but it was really slow going and frustrating sometimes.' (ER4)

27 *'The project start-up seminar, the mid-term seminar, the final seminar... All these things at the PBL where you continuously have to explain and account for your evaluation approach. It is good that it is critically reflected upon, but also makes it slow going and just costs a lot of time from the project leader.'* (ER2)

As for the first arena, we also perceive factors to be at play in the institutional arena that affected the Natuurpact evaluation. These factors, although related to the organizations' cultures, structures and 'ways of working', do not readily correspond to the logics. For instance, practical matters such as, a number of researchers reflected they found it difficult to actually interact with their colleagues from the other organization: the distance between PBL and the WUR is considered large (Table 4.13, #28). Also, researchers from the WUR work more project-based. This means that they have a specific number of hours for a specific task and that they must manage their time more strictly. Consequentially, they have less time available to take on additional work or to diverge too far from the formalized plans. This is in contrast with the PBL where it is more common to take on additional tasks and change plans, until the work is found sufficiently satisfactory by all involved. At the PBL, employees are able to do this, as the organization has more financial leeway than the WUR. This difference in 'work ethic', so to speak, caused some frustration at times (#29). Furthermore, the institutes both hold specific understandings on their role and fields of expertise, which affected the decisions that were made in regards to specific methods. Finally, the institutes differences were also perceived in their use of language and conceptual understanding. An illustrative example is the differences in understanding of concepts such as *policy* and the *object of evaluation* (which was understood as the state of nature by PBL and the state of nature policy by the WUR). That these understandings differed frustrated parts of the collaboration between the researchers, but remained mostly implicit (or, at least, were not outspoken during project meetings).

Table 4.13 – Quotes on the institutional arena

#	Quotes
28	<i>'You don't see each other very often, so it's hard to keep track what everyone is doing. The distance was just quite large to keep in touch.'</i> (ER3)
29	<i>'The number of hours you put into a study, that is also differently arranged between PBL and WUR. For the WUR researchers, they have this pre-set number of hours and when these are used, it's done regardless whether more work is needed. Whereas at PBL it very common to just keep on going until you are satisfied with the quality of your study. In the final stages, I just really needed their help sometimes, and this was a tense discussion.'</i> (ER5)

4.2.3 Negotiating reflexive evaluation with policy practice and society

The third and final arena where the evaluation design and execution were negotiated is where the evaluation research interacts with policy practice and society. Not only the researchers, but also the evaluation's participants and some onlookers hold specific ideas on what evaluation is and should do, and how it should be executed. Most of these actor groups were not (or were hardly) familiar with reflexive, participatory evaluation. Initially this made some of them apprehensive and withholding,

and meant that the researchers at times had to dial back on their reflexive aspirations in order to guarantee the involvement of the intended participants of the evaluation.

For instance, at the outset of the evaluation, the researchers aspired to include all relevant actors who are involved with nature policy as participants of the evaluation, befitting the rationale behind reflexive evaluation (as proposed in literature). However, the primary end-users of the evaluation, the provincial policymakers, were not enthusiastic about the idea of discussing their preliminary policy plans with societal organizations. They anticipated that the evaluation research would put them in a vulnerable position – by the judgement of their policy progress – and, therefore, found the involvement of societal organization inappropriate and a risk. Their expectations and assumptions on the involvement of societal organizations demonstrate the modernist logic at play. To ensure the continued participation of the provinces, it was decided to give the societal organizations a smaller role than originally intended. Most researchers felt that this compromised part of the quality of the knowledge produced, as it limited data triangulation. Also, the researchers were concerned that a less inclusive evaluation may risk losing legitimacy and societal support for its findings (Table 4.13, #30). As the involved actors became more familiar with the reflexive evaluation approach, however, they seemed to shift in logic. Increasingly, they showed trust that the evaluation is intended not just for accountability purposes, but also for mutual learning.

Table 4.13 – Quotes on the policy and societal arena

#	Quotes
30	<i>'I regretted that during these final phases we lost the value of polyphony: confronting the comments of the provinces with the perspectives of societal partners is an effective instrument to guard for becoming biased towards just one party. And I believe it is befitting a reflexive evaluation to accommodate and confront a plurality of perspectives, not just hear one party.'</i> (ER2)

Furthermore, the researchers reflect that the different expectations provinces had regarding the evaluation affected the roles they were allowed to perform. For instance, a few researchers explained how some provinces did not trust the evaluation would not be used for benchmarking and were concerned for severe financial consequences should progress with the nature policy goals appear too slim. As a result, these provinces were resistant to sharing data and were overall highly critical towards the researchers during the evaluation. Researchers were pushed back to only perform the role of traditional expert, and could not act as facilitators of learning processes or as critical friends. Also, during the final stages, it became evident that the provinces needed the researchers to perform a clear, independent role in order for the evaluation report to have strategic value to them (Table 4.14, #31). On the other hand, there were also provinces who, after the final report was published, leaned more towards the other direction and rather asked the researchers to advise them on their policies (#32). The researchers agree that the different expectations of the participants regarding the evaluation's purpose and the researchers' roles requires the researchers to be 'role-steady' in order to ensure independency and credibility.

Table 4.14 – Quotes on the policy and societal arena

#	Quotes
31	<i>'But sometimes they needed us as expert; they didn't want to collaborate to the degree with we suggested. The final report had to be ours, for instance. Otherwise they could never have used it to legitimize their policy decisions.'</i> (ER1)

32 *'It was quite intense, they [referring to some provincial policymakers] kept asking us for more and well, pretty much asking us for advice on how to design their nature policies. And it sometimes was so subtle – you really have to be role-steady, we are not an advisory agency, we provide all relevant information but the decision always lays with them. But where to draw the line, when to stop facilitating them?'*
(ER5)

Beyond the participants of the Natuurpact evaluation, researchers were also confronted with diverging ideas and expectations regarding evaluation from other ministerial departments and governmental bodies. For instance, the project leaders were requested to present their approach to the Policy quality and Evaluation commission⁶. The members of this commission were highly critical of the reflexive approach adopted and questioned the PBL's and WUR's legitimacy and independence. The project leader frequently had to explain the approach and account for its scientific rigor and added value over more traditional approaches. As he phrased it: *'Paradoxically, when you conduct a learning-oriented evaluation you end up having to account yourself to almost everyone.'* The commissioned work by the researchers from the Athena Institute (i.e. the authors of current report) on writing a scientifically grounded conceptual framework for reflexive evaluation and our review of the first evaluation period (of which the current report is part) may also be seen in light of buttressing the scientific robustness and therefore the credibility and legitimacy of the reflexive evaluation approach.

The three arenas demonstrate that the negotiations on reflexive evaluation are not limited to those researchers charged with its design and execution. Rather, a wide variety of actors and institutions, who all have their own perspectives on adequate, feasible and 'good' evaluation research affected the Natuurpact reflexive evaluation. For a large part these perspectives may be understood through the modernist logic that is historically more embedded than the responsive logic. The negotiations in all arenas are further fueled by logics on society and modes of governance, in which the logics on evaluation are rooted. Researchers are required to navigate these logics, as well as their own, in all three arenas in order to develop an evaluation design and conduct that is valued and endorsed by all these involved actors. Even though these diverging logics at times stood in the way of the researchers' reflexive aspirations, we would rather see it in a more positive light. Through their effective navigational skills, a predominantly reflexive evaluation was designed and executed, while still holding sufficient modernist elements to be recognizable to more traditionally oriented actors, therefore resulting in their endorsement and overall acknowledgement of the research's authority.

⁶ In Dutch: Beleidskwaliteit –en Evaluatie Commissie (BEC), a department within the Ministry of Agriculture, Nature and Food Quality (previously Ministry of Economic Affairs) which advises on evaluation design and is comprised of scholarly experts on evaluation research.

5. Conclusions & discussion

5.1 Conclusion: understanding the perspectives on the value of a reflexive evaluation approach

5.1.1 Perspectives on quality, usability and impact

At the onset of this study, we set out to capture how the researchers assessed the quality, usability and impact of the Natuurpact reflexive evaluation. However, due to the rich diversity of meanings researchers ascribed to these concepts, we decided to first unpack these different perspectives and their properties (see paragraph 4.1) before proceeding to draw conclusions on how the researchers assessed the Natuurpact evaluation. Interestingly, despite that researchers were observed to hold plural understandings of quality, usability and impact, all were predominantly positive about the value of the reflexive evaluation approach. For instance, *quality* was understood as both scientificity and usability, but, regardless of which meaning researchers ascribed to the concept of quality, they agreed that the inter- and transdisciplinary components of the reflexive evaluation positively contributed to the quality of the outcomes. Something similar applies to the concept of usability, however, some researchers held the opinion that usability could have been improved by adopting a more co-productive approach, which they assume could have been aligned better to the participants' evaluative needs. Finally, concerning impact, the researchers were also generally positive, perceiving the Natuurpact evaluation as contributing both to participants' learning processes and to the purpose of accountability. Some, however, withheld from commenting on impact, as they considered it too early to draw any conclusions on impact (implicitly demonstrating a dominantly modernist understanding of impact).

This conclusion resonates with the findings from our previous review about the perspectives of participants on the value of the reflexive evaluation approach. This previous review focused on the policy and societal arena in which the evaluation was negotiated, while the current brought into view the research team and their home institutions as arenas. From both these reviews combined, conclude that the reflexive approach was highly valued by participants and researchers alike.

5.1.2 Interpreting the different perspectives: the ideal-types revisited

With our second research question we sought to interpret the differences in perspectives. We have done so using the framework composed of the ideal-typical modernist and responsive approaches to evaluation that was introduced in section 2. Each ideal-type embodies its own specific logic on evaluation, the purpose of evaluation and its conviction on how evaluation research *ought* to be done. In section 4.1 we have used these ideal-types to explain the different perspectives researchers exhibit on the value of the reflexive approach for the quality, usability and impact of the evaluation findings. In section 4.2 we have used the logics of these ideal-types as a lens through which to understand the negotiations on the evaluation's design and conduct that took place in the three different arenas (i.e., the arena of research team, the institutional arena and the policy and society arena). Our findings suggested that the modernist logic tended to be more dominant during the negotiations. This was especially true in the institutional and policy arenas, as the modernist logic is historically the earlier and more institutionalized logic. This institutionalization makes it difficult for researchers to diverge too far or break off from solidified modernist routines. Although this may be

perceived as a negative consequence to the reflexive aspirations of the researchers, we would argue that the diversity in logics that the researchers were confronted with added value to the reflexive evaluation. This diversity and the researchers' skills in navigating between logics have allowed for a reflexive evaluation that simultaneously met the demands of "sound evaluation" stipulated by a diversity of actors. Rather than perceiving the presence of disparate logics as a risk, we underline its merit and benefit for a broadly supported, legitimate and credible evaluation design. This is especially true if the capacity to navigate between logics is seen as an asset rather than as, for instance, a sign of inconsistency.

5.2 Discussion: towards a reflexive practice

What stands out from our study is that despite, or perhaps due to, the disparate logics at play, a shared practice of reflexive evaluation is increasingly emerging amongst the research team. To answer our final research question, on the lessons that may be drawn for the continuation of the Natuurpact reflexive evaluation, we first further unpack this 'reflexive practice', its added value and its barriers. The answer to our final research questions is subsequently provided in section 6, where we suggest several recommendations for the Natuurpact reflexive evaluation.

5.2.1 The emergence of a reflexive practice

Our findings on the negotiations about the reflexive evaluation in the different arenas where the disparate logics are navigated, point towards the emergence of a shared reflexive practice amongst the members of the research team. This shared practice encompasses notions on the purpose of reflexive evaluation and how quality, usability and impact may be attained, while adhering to the different demands and standards of actors involved via different arenas. As such, rather than being a logic in itself, this reflexive practice may be understood as a shared framework that demarcates the researchers' navigational space to satisfy both the modernist and responsive logics. Table 5.1 presents this emerging reflexive practice synthesized from the understandings expressed by the researchers on the purpose of reflexive evaluation and how its quality, usability and impact may be secured. In the remainder of this section, we substantiate this view on reflexive evaluation through use of relevant literature from the science-policy nexus and evaluation domains.

Purpose of the produced knowledge

In a reflexive evaluation practice, evaluation is oriented towards the support of policy practice in light of complex and dynamic societal problems. As such, evaluation fills both the purposes of accountability and of informing learning processes. This may be captured by the concept *multi-directional accountability* (Van Veen et al., 2016). This means that, in addition to hierarchical accountability (upwards; Edwards & Hulme, 1996) the evaluation allows for horizontal (towards those affected or otherwise involved with the policy; Regeer, de Wildt-Liesveld, van Mierlo, & Bunders, 2016) and internal (towards those responsible for designing and implementing the policy; (Ebrahim, 2009) accountability. These latter two forms of accountability in particular may be understood as processes of interactive learning about the policy instruments, strategies, goals, underlying values, and accompanying norms.

Based on this view on the purpose of a reflexive practice, we can formulate principles for the quality, usability and impact of the knowledge that reflexive practice helps construct.

Table 5.1 – Description of a reflexive practice that supports the navigation of modernist and responsive logics on evaluation.

	Modernist logic	In a reflexive practice	Responsive logic
Purpose of the produced knowledge	<i>Speaking truth to power</i> ; evaluation is a mechanism for ensuring accountability (assessment of policy performance against set goals)	<i>Multi-purpose</i> ; evaluation is a mechanism that enhances multi-directional accountability which includes informing learning processes to improve policy performance in light of complex societal problems	<i>Co-create highly contextualised solutions</i> ; evaluation is a mechanism for informing learning processes and problem solving (reflection on instruments, strategies, goals, norms and values)
Principles for quality	Emphasis on scientific robustness; credible, independent, distant, objective research conducted with scientific autonomy	Balancing credibility, legitimacy and salience in order to produce knowledge that is both scientifically and socially robust; equal interaction with all relevant actors and openness and transparency to underline scientific autonomy	Emphasis on social robustness; legitimate, inclusiveness of plurality of perspectives; extensive interaction with relevant actors in order to co-create knowledge
Principles for usability	Researcher-driven: evaluation designed to assess policy performance by systematic data collection on inputs, outputs and outcomes	Co-production driven: researchers and participants co-decide on an evaluation design that optimally informs participants' evaluative needs enriched by expert perspectives	Participant-driven: evaluation design fully informed by relevant actors to grasp policy complexity and assess policy in responsive manner to inform learning
Principles for impact	Impact is established by the degree to which evaluation holds the evaluated government accountable for their progress on set goals (hierarchical accountability)	Impact is established by attaining the multiple evaluation purposes; by allowing for accountability, including hierarchical, horizontal and internal accountability and informing the respective learning processes	Impact is established by the degree to which the evaluation is able to inform learning processes of the evaluated government (reflection on instruments, strategies, goals, norms and values) and thereby improve policy performance

Principles for quality

Regarding quality, scholars have suggested principles such as credibility, legitimacy and salience to understand the effectiveness of evaluation research (Cash et al., 2002). During a reflexive practice, these principles are met not just via a scientific route of scientific robustness (via independency and objectivity), but are met also via a societal route and socially robust knowledge. Credibility, legitimacy and salience of the research are balanced by engaging with the relevant actors in an equal, open and transparent fashion to underline scientific autonomy (van der Hel & Biermann, 2017)

Principles for usability

Regarding usability, a theme some researchers touched upon concerns the role of co-production in reflexive practice. The researchers seek to align their expert opinion on the evaluation design to the evaluative needs of the participants in a way that allows for a feasible evaluation design in which modes of working researcher- or participant driven are combined. This approach to knowledge co-production is often discussed in works on transdisciplinary research (Mauser et al., 2013). Elements such as effective and clear communication, openness to other perspectives and developing mutual understanding are key to fruitful deliberations (Edelenbos & van Buuren, 2005; Lehtonen, 2014). It is also necessary to collaboratively translate the evaluative needs of participants into research questions

and, in order to manage expectations, communicate clearly when this is not possible or when matters fall outside the scope of evaluation (Patton, 2010; Patton, 2008; Stake & Abma, 2005).

Principles for impact

Finally, a reflexive practice may be considered to have impact when it is successful in attaining its multi-purpose. The challenge here is the reconciliation between evaluating for accountability and evaluating for learning, as was also evident in the researchers' reflections on their roles and on societal expectations of the evaluation. A promising strategy here includes openness and transparency, and making the participants part of the considerations that underlie the tension. Thereby, ownership and responsibility for attaining the evaluation's multi-purpose becomes shared and will likely then reduce some of the aforementioned tension.

5.2.2 Institutionalizing a reflexive practice

The reflexive practice that is emerging amongst the researchers gives them a framework within which they can effectively navigate the disparate logics in a way that satisfies both. In support of this practice, we hold that structural opportunities for researchers – and participants for that matter – to reflect on the logics, expectations and routines that are at play would be beneficial. In literature it is often argued that the different perspectives between researchers in interdisciplinary teams benefit from interactive frame reflection (Schön & Rein, 1994) to bridge epistemic and normative differences (Regeer & Bunders, 2009). By explicating expectations and assumptions about evaluation and the underlying logics which people draw from, mutual understanding can be expected to increase. This mutual understanding can also lead to the development of a more shared practice and to the reframing of the issues at stake. Such reflection would be valuable beyond the context of specific projects with a reflexive character and would serve any organization with interdisciplinary or transdisciplinary ambitions.

We established that the logics were navigated not just within the research team, but also within the researchers' home institutions and within the policy and societal arena. Our findings demonstrate that the reflexive ambitions of the researchers were, at times, hampered by the more dominant modernist logic that is at play in these other two arenas. The modernist logic is deeply embedded within institutional structures and cultures of the researchers' background organizations, especially as the PBL is concerned. The modernist perspective similarly seems to dominate the outside world (i.e., the provinces, national government and societal organizations). Institutionalization of a reflexive practice, at least amongst the researchers' home institutes, would normalize and legitimize the navigation of logics. Within such a reflexive practice the modernist and responsive logic may be perceived as equally valid approaches to evaluation research rather than as opposing ends of a dichotomy (Kunseler & Verwoerd, in progress). Within this line of thinking, it becomes a matter of fit and appropriateness regarding the problem, and the societal and political context in which it transpires, that decides which logic (or functional hybrid) to adopt. Institutionalization of a reflexive practice may prevent the dominant modernist logic from hampering more reflexive ambitions of researchers, their organizations, participants and commissioners of the evaluation alike.

6. Final recommendations

Based on our findings we make the following recommendations for the continuation of the Natuurpact's reflexive evaluation, along the lines already sketched in section 5.2. With these recommendations we propose to encourage the emerging reflexive practice within which different logics may be effectively navigated. Although the recommendations are primarily intended in support of the second Natuurpact evaluation period, they also have wider relevance to encourage a more reflexive practice at the organizational level.

Support the navigation of disparate logics on evaluation and the emergence of a shared reflexive practice

The shared reflexive practice provides a framework within which disparate logics on evaluation may be effectively navigated by researchers. This is accomplished by upholding a flexible understanding of reflexive evaluation that satisfies both logics. Navigation may be supported through frequent, structured (and guided) interaction and (frame) reflection. This helps to make logics, expectations and interests (from individual researchers and their home organizations, and participants and commissioners of the evaluation) explicit, thereby increasing mutual understanding and alignment between the different logics and within the different arenas.

Structurally embed knowledge integration into the evaluation research design

From the start of the evaluation, researchers need to be familiar with each other's sub projects and how these align to their own; what are important links between the projects and what story do they tell together? For interdisciplinary knowledge integration it is essential that researchers are up to date with each other's think- and work-processes and with relevant developments within their respective projects. To achieve this, structured and frequent interaction is required. This interaction can take the form of formal meeting, but less formal, bilateral meetings can also be fruitful in this regard. What matters here is that interactions should be supported and methodical, not *ad hoc*, casual encounters.

Ensure alignment between evaluation research and nature policy practice by working towards a co-production-driven practice

Appropriate alignment means that researchers are in tune with the evaluative needs of the end-users, and that the end-users understand how the evaluation is intended to inform these evaluative needs. The evaluation may comprise both researcher- and participant-driven elements, as long as these are understandable and transparent for the end-users. Monitoring the evaluative needs of the end-users is recommendable in order to ensure continuous alignment throughout the evaluation.

Co-production implies openness, transparency and deliberation *with* the participants on all evaluation decisions with clear communication on the expectations and the roles of researchers. When participants are not willing or do not see the use of such deliberations,

transparency and recording of decisions made may be adequate strategies to prevent discussions later on.

Make the multi-purpose of reflexive evaluation explicit and make the participants of the evaluation shareholders

Make it explicit that reflexive evaluations can serve multiple purposes and organize the process such that participants of the evaluation become co-responsible for identifying its purpose(s) and for realizing this or these purpose(s). By making intended end-users shareholders of the different considerations that underlie any research approach, the transparency of the research increases. This transparency is beneficial for achieving important quality criteria for evaluation research such as credibility and legitimacy.

References

- Bowen, G. (2008). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, 5(September), 12–23. <https://doi.org/Article>
- Cash, D., Clark, W., Alcock, F., Dickson, N., Eckley, N., & Jäger, J. (2002). Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making. *KSG Working Paper Series, RWP02-046*. <https://doi.org/http://dx.doi.org/10.2139/ssrn.372280>
- Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (3rd ed.). Thousand Oaks: SAGE Publications Inc.
- Creswell, J. W., & Miller, D. L. (2000). Determining Validity in Qualitative Inquiry. *Theory Into Practice*, 39(3), 124–130. https://doi.org/10.1207/s15430421tip3903_2
- Ebrahim, A. (2009). Placing the Normative Logics of Accountability in “Thick” Perspective. *American Behavioral Scientist*, 52, 885–904. <https://doi.org/10.1177/0002764208327664>
- Edelenbos, J., & van Buuren, A. (2005). The learning evaluation: a theoretical and empirical exploration. *Evaluation Review*, 29(6), 591–612. <https://doi.org/10.1177/0193841X05276126>
- Edwards, M., & Hulme, D. (1996). Too close for comfort? the impact of official aid on nongovernmental organizations. *World Development*, 24(6), 961–973.
- Felt, U., Igelsböck, J., Schikowitz, A., & Völker, T. (2016). Transdisciplinary Sustainability Research in Practice: Between Imaginaries of Collective Experimentation and Entrenched Academic Value Orders. *Science Technology and Human Values*, 41(4), 732–761. <https://doi.org/10.1177/0162243915626989>
- Hisschemöller, M., & Hoppe, R. (1995). Coping with intractable controversies: the case for problem structuring in policy design and analysis. *Knowledge and Policy*, 8(4), 40–60.
- Jasanoff, S. (2004). Science in Culture and Politics. In *States of Knowledge. The co-production of science and social order*. (pp. 25–98). London, New York: Routledge.
- Kunseler, E.-M. (2017). *Government expert organisations in-between logics*.

- Kunseler, E.-M., & Vasileiadou, E. (2016). Practising environmental policy evaluation under co-existing evaluation imaginaries. *Evaluation*, 22(4), 451–469. <https://doi.org/10.1177/1356389016668099>
- Kunseler, E.-M., & Verwoerd, L. (n.d.). *Van reflex naar reflexiviteit in publieke kennisorganisaties - kennis voor beleid en samenleving in verandering*. PBL The Hague, Athena Institute VU University Amsterdam.
- Lehtonen, M. (2014). Evaluating megaprojects: From the “iron triangle” to network mapping. *Evaluation*, 20(3), 278–295. <https://doi.org/10.1177/1356389014539868>
- Mauser, W., Klepper, G., Rice, M., Schmalzbauer, B. S., Hackmann, H., Leemans, R., & Moore, H. (2013). Transdisciplinary global change research: the co-creation of knowledge for sustainability. *Current Opinion in Environmental Sustainability*, 5(3–4), 420–431. <https://doi.org/10.1016/j.cosust.2013.07.001>
- Patton, M. Q. (2000). Utilization-Focused Evaluation. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation Models* (pp. 425–438). Boston: Kluwer Academic Publishers.
- Patton, M. Q. (2008). *Utilization-Focused Evaluation*. Sage Publications, Inc.
- Patton, M. Q. (2010). *Developmental Evaluation: Applying Complexity concepts to Enhance Innovation and Use*. Guilford Press.
- Regeer, B. J., & Bunders, J. F. G. (2009). *Knowledge co-creation : Interaction between science and society*. Advisory Council for Spatial Planning, Nature and the Environment (RMNO).
- Regeer, B. J., de Wildt-Liesveld, R., van Mierlo, B., & Bunders, J. F. G. (2016). Exploring ways to reconcile accountability and learning in the evaluation of niche experiments. *Evaluation*, 22(1), 6–28. <https://doi.org/10.1177/1356389015623659>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in general theory of planning. *Policy Sciences*, 4, 155–169.
- Schön, D. A., & Rein, M. (1994). *Frame Reflection*. New York: Basic Books Inc.
- Shils, E. A., & Finch, H. A. (1949). *Max Weber On The Methodology of Social Sciences* (First edit). Glengoe, Illinois: The Free Press. Retrieved from <http://fs2.american.edu/dfagel/www/Class>

Readings/Weber/weber_on_methodology_of_social_sciences.pdf

Stake, R. E., & Abma, T. A. (2005). Responsive Evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation*.

Van der Hel, S. (2016). New science for global sustainability? The institutionalisation of knowledge co-production in Future Earth. *Environmental Science & Policy*, *61*, 165–175.

<https://doi.org/10.1016/J.ENVSCI.2016.03.012>

Van der Hel, S., & Biermann, F. (2017). The authority of science in sustainability governance: A structured comparison of six science institutions engaged with the Sustainable Development Goals. *Environmental Science & Policy*, *77*, 211–220.

<https://doi.org/10.1016/J.ENVSCI.2017.03.008>

Van der Meer, F.-B., & Edelenbos, J. (2006). Evaluation in Multi-Actor Policy Processes: Accountability, Learning and Co-operation. *Evaluation*, *12*(2), 201–218.

<https://doi.org/10.1177/1356389006066972>

Van Veen, S. C., Verwoerd, L., & Regeer, B. J. (2016). *Characteristics of reflexive evaluation - a literature review conducted in the context of the Natuurpact (2014-2027) evaluation*.

Verwoerd, L., De Wildt-Liesveld, R., & Regeer, B. J. (2017). *The value of reflexive evaluation - a review of the Natuurpact evaluation (2014-2017)*.