faculteit der Aard- en levenswetenschappen

# Characteristics of reflexive evaluation - a literature review conducted in the context of the Natuurpact (2014-2027) evaluation

2016
Saskia C. van Veen
Lisa Verwoerd
Barbara J. Regeer

Athena Instituut

*vrije* Universiteit *amsterdam*

# Characteristics of reflexive evaluation - a literature review conducted in the context of the Natuurpact (2014-2027) evaluation

Saskia C. van Veen, Lisa Verwoerd, and Barbara J. Regeer

Version 0, 2016

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.

We are interested in you experiences and any feedback you may have: Barbara Regeer b.j.regeer@vu.nl and Saskia van Veen s.c.van.veen@vu.nl

# Table of Contents

# Preface

There is an increased call for new types of evaluation approaches that are more congruent with the complexities, ambiguities and uncertainties of contemporary policy practice. These calls appeal for approaches that are participative, responsive, and integrative. Such approaches not only generate knowledge about the performance of the policies under evaluation in achieving its intended outcomes (through which the accountability function of evaluation is established), but also contribute to learning processes of the involved stakeholders along the way in order to ultimately contribute to improvement of policy practice and thereby increased likelihood of goal attainment. Examples of these types of approaches are fourth generation evaluation, responsive evaluation, participative evaluation, utilisation-focused evaluation, and learning evaluation. This document discusses primarily the features of a 'reflexive evaluation' based on existing scientific literature.

The PBL Netherlands Environmental Assessment Agency (Dutch: Planbureau voor de Leefomgeving) (PBL) and partners Alterra (Wageningen University), LEI (Wageningen University), and the Athena Institute (VU University Amsterdam), have begun to experiment with a 'reflexive evaluation' in the context of a longitudinal evaluation of the Dutch nature conservation policy programme Natuurpact (2013-2027). When employing new and emerging methods of evaluation it is important to establish its scientific basis. The goal of this document is to provide such a scientific justification, by investigating the underlying assumptions and methodological principles as put forward in scientific, peer-reviewed articles about reflexive evaluations and adjacent evaluation methodologies. The result is an integrated framework synthesised from scientific literature, containing the most important criteria regarding the inclusion of stakeholders, the functions of the evaluation, the process of the evaluation, the outcomes, and the evaluation team that guides the evaluation.

The preliminary research for this large-scale programme evaluation was initiated in 2014. This preliminary phase resulted in an evaluation framework and a plan for the evaluation of the Natuurpact up to 2016, and the evaluation research is presently being conducted by PBL and partners, with a first report expected for publication in 2016. The proposed approach outlined in the evaluation plan is a *reflexive* evaluation. The role of the Athena Institute in the evaluation team is to design, implement and reflect on how learning processes are integrated in the reflexive evaluation in balance with the evaluation of goal attainment. This overview of the key features of a reflexive evaluation as currently available in scientific literature is part of that. Monitoring of the team's implementation of the reflexive evaluation has simultaneously been initiated. Accordingly, the VU will publish its final report in December 2016, in which a reflection will be provided on the evaluation team's approach to implementing the criteria and recommendations for a reflexive evaluation, suggestions for improvement, and an assessment of the added value of a reflexive evaluation in relation to more traditional evaluation approaches. For more step-by-step information on executing or participating in reflexive evaluations, PBL will produce a handbook for professionals in policy and policy evaluation, based on the experiences of the Natuurpact evaluation and other evaluation projects conducted by PBL.

Thus, the primary aim of this document is to provide a scientific justification for the reflexive evaluation approach as employed in the evaluation of the Natuurpact. The secondary aim of the document is to serve as a basis for reviewing the application of the method in practice in order to identify points for improvement, and in order to establish the promised added value of this type of evaluation.

# Nederlandse samenvatting

In de afgelopen jaren is het natuurbeleid gedecentraliseerd en de afspraken hierover zijn vastgelegd in het Bestuursakkoord Natuur (2011/2012) en het Natuurpact (2013). Twaalf Nederlandse provincies maken én implementeren natuurbeleid om samen met het Rijk deze afspraken te realiseren in 2027. De provincies en het Rijk richten zich volgens dit akkoord op de realisatie van het Natuurnetwerk Nederland, het halen van de internationale doelen (de Vogel en Habitat richtlijn, VHR, en de Kaderrichtlijn Water, KRW) en het versterken van de betrokkenheid van de samenleving bij de natuur. Het Planbureau voor de Leefomgeving (PBL) is gevraagd door het Ministerie van Economische Zaken (EZ) en het Interprovinciaal Overleg (IPO) om eens in de drie jaar de voortgang van de afspraken uit het Natuurpact te evalueren. De decentralisatie van het natuurbeleid heeft veel veranderingen teweeg gebracht; traditionele partijen worden verwacht nieuwe rollen te vervullen, en nieuwe partijen, zoals maatschappelijke partijen, bedrijven en burgers, die oude én nieuwe taken op zich nemen zijn ten tonele verschenen. Bovendien zijn een aantal natuurambities uit het Natuurpact generiek opgesteld; de provincies werken met de betrokken partijen deze ambities voor 2027 uit naar concrete natuurdoelen en bijbehorende beleidsstrategieën.

Het staat buiten kijf dat deze ontwikkelingen in het natuurbeleid voor alle betrokken partijen vernieuwingen met zich mee brengen; het Natuurpact is geen *business as usual*. De opdrachtgevers hebben PBL daarom gevraagd om in plaats van een reguliere impact-evaluatie, een *lerende* evaluatie uit te voeren.  De eerste rapportage van deze evaluatie staat gepland in 2016.

**Wat is een lerende evaluatie?**

De naam zegt het al: in een lerende evaluatie komen leren en evalueren samen. In ‘klassieke’ evaluaties worden successen en mislukkingen van beleid in beeld gebracht en voorzien van een oordeel. Een oordeel achteraf kan echter niet meer helpen om een proces bij te sturen of te verbeteren. De lerende evaluatie komt voort uit onvrede over de bruikbaarheid van de resultaten van klassieke evaluaties. In een lerende evaluatie vindt de evaluatie al tijdens het proces plaats. Er is veel ruimte om te leren en bij te sturen. Dit leren vindt plaats op twee niveaus: (1) de beleidspraktijk en (2) de onderliggende opvattingen en veronderstellingen over doelen, instrumenten en problemen (beleidstheorie). Ook is de oordeelsvorming dynamisch: er wordt niet keihard afgerekend op vooraf gestelde doelen, want de doelen kunnen tijdens het proces veranderen. In de woorden van bestuurskundigen Edelenbos en Van Buuren (2006), “een lerende evaluatie is niet zozeer het systematisch langs een van tevoren uitgezet traject een gewenst doel bereiken, als wel het vanuit een open grondhouding, vertrekkend vanuit globale onderzoeksdoelen, op een intelligente en creatieve wijze omgaan met de onzekerheid en onvoorspelbaarheid in een voortdurend veranderende werkelijkheid”. In een lerende evaluatie komen evaluator en geëvalueerden of gebruikers gezamenlijk tot betekenisgeving en verbetering. De evaluator anticipeert, reageert en reflecteert voortdurend op de bij de evaluatie betrokken actoren.

*Bron: Plan van aanpak. Lerende evaluatie Natuurpact (PBL, 2015)*

De aanpak van een lerende evaluatie is relatief nieuw – ook voor PBL.  Het stapt daarmee af van zijn rol als louter klassieke evaluator en zal meer interactief betrokken zijn bij de partijen wiens beleid geëvalueerd wordt. Niet alleen voor PBL is deze aanpak nieuw;  het vraagt van alle betrokken partijen een andere rol en een andere kijk op beleidsevaluaties dan de gebruikelijke. Vanzelfsprekend brengt dit spanningen en vragen met zich mee. Hoe behoudt PBL bijvoorbeeld zijn onafhankelijkheid in deze nieuwe rol? Wordt het evalueren van de doelrealisatie niet uit het oog verloren als er zoveel nadruk wordt gelegd op samenwerken en interactie? En hoe wordt er recht gedaan aan alle verschillende perspectieven die spelen rondom het natuurbeleid?

Om voorbereid te zijn op deze vragen en potentiële spanningen heeft PBL het Athena Instituut van de Vrije Universiteit Amsterdam gevraagd alles wat bekend is in de wetenschappelijke literatuur over lerende evaluaties op een rijtje te zetten. Naast een wetenschappelijke uiteenzetting van een lerende evaluatie dient deze review ook om geïnteresseerden inzage te geven in het achterliggende gedachtegoed, de voorwaarden en de beoogde meerwaarde van een dergelijke evaluatie aanpak. In een volgend rapport – gepland voor 2016 – zal er gereflecteerd worden op de toepassing hiervan door PBL, inzicht gegeven worden in wat deze aanpak concreet heeft opgeleverd om te bezien in hoeverre beoogde verwachten realiteit zijn verworden alsmede om zo de evaluatie aanpak te verbeteren. Om concrete handvatten te bieden voor het uitvoeren van, en deelnemen aan, lerende evaluaties, zal PBL een handreiking lerende evaluaties maken voor professionals in de praktijk, gebaseerd op de ervaringen van de Natuurpact evaluatie en andere evaluatieprojecten van PBL.

**Waarom lerend evalueren?**

Recente veranderingen in de context van beleid hebben ervoor gezorgd dat een nieuwe blik op het maken van beleid – en daardoor ook de evaluatie daarvan – gewenst is. Deze ontwikkelingen zijn onder andere de toegenomen complexiteit van maatschappelijke problemen, zoals bijvoorbeeld in voedselveiligheid, klimaatverandering en duurzame energie. Dergelijke problemen zijn niet eenvoudig op te lossen; ze worden veroorzaakt door een wirwar van ecologische, sociale en economische aspecten. Hierdoor zijn ook mogelijke oplossingsrichtingen complex. Kenmerkend is dat besluitvorming in beleid daarover meer en meer plaatsvindt in samenwerking tussen overheden, burgers, maatschappelijke partijen en bedrijven. Bovendien is er interactie tussen beleidsprocessen op verschillende niveaus (EU, Rijk, provincie, regio). Men spreekt dan over *multi-actor* betrokkenheid en *multi-level governance.*

Tegelijkertijd blijft er in beleid een grote vraag naar het meetbaar maken van resultaten en het koppelen daarvan aan geld en andere middelen die geïnvesteerd zijn – in andere woorden, in het verantwoorden van beleid. Deze trends – het toegenomen multi-actor en multi-level governance karakter van beleid, én de vraag naar verantwoorden – lijken met elkaar in conflict. Immers, wie waar verantwoordelijk voor is wordt steeds moeilijker te bepalen naarmate er meer partijen betrokken zijn bij het maken en implementeren van beleid. Elke betrokken stakeholder heeft zijn eigen relevant (ervarings-)kennis die belangrijk is in besluitvorming en elke stakeholder zal duidt uitkomsten van beleid op zijn eigen manier. Bovendien wordt het aantonen van oorzaak-gevolg relaties tussen enerzijds beleid en anderzijds maatschappelijke processen bemoeilijkt door onzekerheid en de wederzijdse afhankelijkheid van complexe maatschappelijke vraagstukken.

Evaluatiewetenschappers hebben beargumenteerd dat evaluatie in dit type multi-actor en multi-level beleidsprocessen een andere invulling krijgt. Zij benadrukken dat de verschillende actoren met elkaar en van elkaar zouden moeten leren - over de eigen en elkaars denk- en handelingskaders en hoe die de beleidspraktijk beïnvloeden. In dit leerproces is het belangrijk dat een ieders kennis gezamenlijk tot vernieuwende inzichten leidt. Nieuwe vormen van evaluatie kunnen daarbij een belangrijke rol spelen.

De afgelopen jaren zijn er nieuwe vormen van evaluatie ontwikkeld die precies dit als doel hebben; het stimuleren van leren over een praktijk of programma tijdens het evalueren daarvan. In de literatuur over deze nieuwe vormen van evaluatie worden verschillende argumenten voor de nieuwe aanpak genoemd. De ervaring met meer traditionele evaluaties leert dat de bevindingen daarvan marginaal gebruikt worden in de geëvalueerde werkpraktijk. Vaak vinden deze evaluaties plaats vóór of na het implementeren van beleid en zijn daardoor minder onderdeel van de beleidscyclus. Bovendien ligt bij dat soort evaluaties de nadruk vooral op óf de doelen behaald worden, in plaats van te verklaren hoe dat gebeurt en hoe dat proces eventueel ondersteund kan worden: er is dus vooral aandacht voor het verantwoorden van het gekozen beleid. Daarnaast richten reguliere beleidsevaluaties zich vaak op enkelvoudige beleidsprogramma's, terwijl complexe vraagstukken multi-level governance vergen. Tijdens een *lerende evaluatie* wordt het evaluatie onderzoek gebaseerd op de leerbehoeftes van partijen wiens praktijk of beleid geëvalueerd wordt, waarmee de bruikbaarheid van de bevindingen in beleidsbeslissingen wordt vergroot. De evaluatie vindt plaats *tijdens* het ontwikkelen en implementeren van beleid, zodat tijdig aanbevelingen gedaan kunnen worden om het doelbereik van het beleid te vergroten. Leren staat daarbij centraal; er is veel aandacht voor het onderzoeken van best practices, het onderling delen van ervaringen en gezamenlijke reflectie op de eigen werkpraktijk om te voorkomen dat elke partij het wiel opnieuw moet uitvinden.

De meeste literatuur over de lerende evaluatie gaat over het leren in enkelvoudige project evaluaties, zoals het evalueren van de bijdrage van één bepaalde beleidsstrategie aan een beleidsambitie. Maar net zo belangrijk is hoe dit gerealiseerd kan worden in een situatie waarin meerdere projecten, op verschillende niveaus, tegelijkertijd een bijdrage leveren aan het evalueren van een beleidsambitie. Men spreekt dan over een 'evaluatie arrangement' waarin de som van verschillende sub-evaluaties van beleid op verschillende niveaus leidt tot een uitspraak over het doelbereik gerelateerd aan een bepaalde beleidsambitie. Hierbij wordt, naast leren tussen de verschillende actoren (bijv. beleidsmedewerker, ondernemer, vertegenwoordiger van een natuurorganisatie) én bestuursniveaus (bijv. Rijk, provincie of gemeente), ook ingezet op het onderzoeken van beleidseffectiviteit en efficiëntie in relatie tot doelbereik. Hoewel de naam wellicht anders doet vermoeden, slaat een lerende evaluatie dus een brug tussen leren én verantwoorden (zie Tabel I).

***Tabel 1. Kenmerken van top-down, bottom-up en lerende evaluaties (geïnspireerd door Kuindersma en Boonstra 2005)***

| | Top-down evaluaties | Lerende evaluaties | Bottom-up evaluaties |
|---|---|---|---|
| **Relevante stakeholders** | Alleen de partijen wiens beleid gevalueerd wordt (e.g. | **Primair gericht op stakeholders die eigenaarschap hebben** | Beleidsmakers, de begunstigden en de slachtoffers van het |

| | | | |
|---|---|---|---|
| | beleidsmakers) | **over de beoogde verandering, c.q. degenen bij wie verandering van handelingsperspectieven plaats zal vinden.** | beleid |
| **Relatie evaluatieonderzoek en beleidspraktijk** | Evaluatie en beleidspraktijk zijn gescheiden | **Geïntegreerd: evaluatie is optimaal afgestemd op beleidspraktijk** | Gefuseerd: evaluatie is onderdeel beleidspraktijk |
| **Doelen en kaders van de evaluatie** | Vooraf top-down bepaald | ***Emergent design* met in achtneming van vooraf bepaalde doelen** | *Emergent design* |
| **Nieuwe rollen: stakeholders** | Passief: geven slechts informatie aan de evaluatoren | **Actief: geven input aan het evaluatieontwerp, reflecteren gezamenlijk op denk- en handelingskaders, interpreteren (tussentijdse) bevindingen, etc. Daarnaast óók verstrekken van informatie** | Actief: dragen bij aan ontwerpen van evaluatieonderzoek, doen mee aan reflectie momenten en zetten zich in voor leren |
| **Rol van de evaluator** | Onafhankelijk, afstandelijk, objectief | **Interdisciplinair evaluatie team dat aan meerdere eisen tegemoet komt; zowel onafhankelijkheid als faciliteren van leerprocessen** | Nauw betrokken bij de deelnemers van de evaluatie: responsief naar hun behoeften, speelt een actieve rol in het beleidsproces |
| **Type kennis** | Primair gebruik van expert kennis (nadruk op kwantitatief onderzoek) | **Kennis van zowel wetenschappelijke als maatschappelijke actoren wordt gecombineerd (mixed methods onderzoek)** | Primair gebruik van kennis uit het veld (nadruk op kwalitatief onderzoek) |
| **Doel** | Verantwoorden, evaluatie gericht op doelbereik (impact assessment) | **Verantwoorden én leren** | Leren, evaluatie gericht op verbetering van beleid (en daarmee vergroting doelbereik) |

**De lerende evaluatie van het Natuurpact**

Het natuurbeleid in Nederland is precies zo'n complex maatschappelijk probleem dat gekarakteriseerd wordt door multi-actor en multi-level governance en waar leren tijdens het implementeren van het beleid kan bijdragen aan het succesvol behalen van de doelen, maar waar ook een grote vraag is naar verantwoording. Om aan beide trends in beleid tegemoet te komen, is besloten het Natuurpact te evalueren door middel van een lerende evaluatie.

De evaluatie van het Natuurpact bestaat uit meerdere opeenvolgende fasen. In de eerste fase is het evaluatiekader – met daarin de onderzoeksvragen, de te evalueren beleidsstrategieën, en de doelen van de evaluatie – opgesteld in samenwerking met provinciale beleidsmedewerkers, maatschappelijke partners en vertegenwoordigers van het IPO en EZ. Onderzoeksvragen die in de eerste fase zijn geïdentificeerd, worden in de vervolg fasen, in verschillende deelprojecten onderzocht. Dit gebeurt wederom in nauwe interactie met de verschillende stakeholders. Daarnaast zijn er tijdens het evaluatie onderzoek van het Natuurpact gezamenlijke leermomenten ingebouwd. Door regelmatige interactie tussen de onderzoekers en de stakeholders worden leervragen en kennis uit de praktijk verbonden aan onderzoek. Hiermee wordt een ander soort kennis ontwikkeld dan bij reguliere evaluaties. Naast kennis over doelbereik en efficiëntie wordt er namelijk ook transformationele kennis ontwikkeld; kennis over het proces dat leidt tot de gewenste uitkomsten.

**Stakeholders evaluatie Natuurpact**
Tijdens de evaluatie van het Natuurpact worden beleidsbetrokkenen op nationaal en provinciaal niveau, de maatschappelijke partners en andere belangrijk stakeholders actief betrokken. Zij leveren niet alleen de gegevens die nodig zijn voor de evaluatie, maar denken ook mee over welke informatie voor hun werkpraktijk relevant is.

**Wat is lerend evalueren?**
Omdat we met een lerende evaluatie een brug proberen te slaan tussen effectiviteitsonderzoek enerzijds en wederzijds leren anderzijds, plaatsen we de lerende evaluatie tussen klassieke, top-down evaluaties (zoals bijvoorbeeld impact assessments) en bottom-up evaluaties (waar responsieve evaluaties een voorbeeld van kunnen zijn). In Tabel 1 staan voor een aantal aspecten de verschillen tussen deze evaluatie types beschreven.

De zeven karakteristieken die we hebben gevonden in de literatuur zijn:

- Een lerende evaluatie is een multi-stakeholder proces;
- Het evaluatieproces is optimaal afgestemd op praktijk;
- De doelen en kaders van de evaluatie ontplooien zich met de tijd (*emergent design*)
- Nieuwe rollen: de deelnemende stakeholders;
- Nieuwe rollen: de evaluatoren;
- Kennis en perspectieven van verschillende stakeholders worden geïntegreerd;
- Er wordt geleerd én verantwoord.

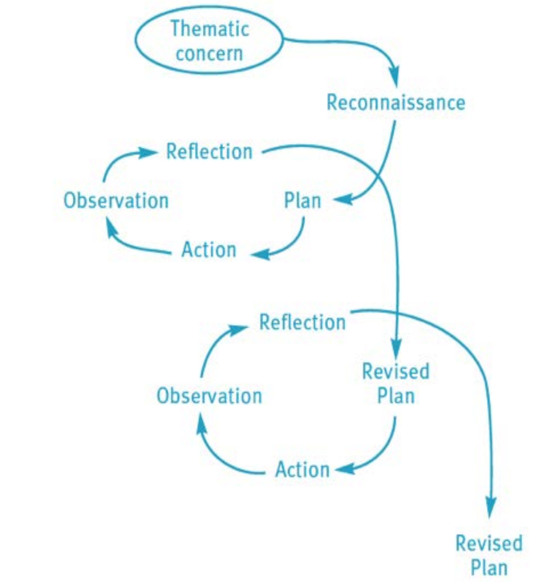
*Multi-stakeholder proces*
In de literatuur wordt vaak beschreven dat tijdens een evaluatie de ‘relevante’ stakeholders betrokken moeten worden. Het begrip relevantie verschilt daarbij per evaluatieperspectief. Tijdens een lerende evaluatie worden niet alleen beleidsmakers (wiens beleid geëvalueerd wordt) relevant gevonden, ook stakeholders die beïnvloed worden door het beleid – zoals maatschappelijke partners, bedrijven en burgers – kunnen daarbij gevraagd worden input te leveren voor de evaluatie.

Een brede kijk op stakeholder betrokkenheid zorgt dat de bevindingen uit de evaluatie sterk gefundeerd zijn op de werkpraktijk. Hierdoor krijgen beslissingen gebaseerd op deze bevindingen maatschappelijk draagvlak. Om dit te realiseren wordt vanuit de literatuur het regelmatig uitvoeren van stakeholder analyses aangedragen, om er zo zeker van te zijn dat tijdens elke stap in de evaluatie juist díe partijen aanwezig zijn die belangrijk zijn voor de inhoud.

*Het evaluatieproces is optimaal afgestemd op de praktijk*

Om tegemoet te komen aan zowel verantwoorden als leren, wordt het evaluatieproces zo ontworpen dat het optimaal kan bijdragen aan de ontwikkelingen in de beleidspraktijk. Uit de literatuur blijkt dat deze afstemming te realiseren is op een aantal manieren, door:

- Het evaluatiekader – met daarin de doelen, de te evalueren beleidsstrategieën en de afbakening van de evaluatie – gezamenlijk op te stellen met alle relevante stakeholders, zodat de bevindingen hun leerbehoeften beantwoorden. Daarbij worden ‘harde’ doelstellingen uit Nationale en Internationale afspraken ook meegenomen.
- Daarnaast bestaat het evaluatieproces uit opeenvolgende cycli van plannen, handelen, observeren en reflecteren (leren). Op deze manier wordt de evaluatie gebruikt om beleidskeuzes op te baseren en beleid te ontwikkelen en uit te voeren. De ervaringen uit de praktijk vormen vervolgens weer input voor het evaluatieproces.



*Afbeelding 1. Schematische weergave van evaluatiecycli.*

*De doelen en kaders van de evaluatie ontplooien zich in de tijd* (emergent design)

Zoals blijkt uit de vorige karakteristiek van lerend evalueren, wordt het evaluatie onderzoek continu aangepast aan de beleidspraktijk. Dit betekent dat voorafgaand aan het evaluatieproces lastig bepaald kan worden welke onderwerpen specifiek onderzocht gaan worden, welke methodieken daar van toepassing zijn en welke resultaten dat op zal leveren. Flexibiliteit – van het evaluatieontwerp, maar ook van de evaluatoren en de deelnemende stakeholders – is daarbij van groot belang, zodat ingesprongen kan worden op zaken die mogelijkerwijs op onverwachte momenten boven tafel komen.

*Nieuwe rollen: stakeholders*

Flexibiliteit is één van de eisen die een lerende evaluatie stelt aan zowel de deelnemende stakeholders als de evaluatoren. De deelnemende stakeholders geven niet – zoals tijdens een klassieke evaluatie – passief informatie aan de evaluatoren, maar nemen ook actief deel aan het evaluatieproces. Zij:

- geven input voor het evaluatiekader (met daarin de doelen, de te evalueren beleidsstrategieën en de afbakeningen van de evaluatie);
- interpreteren gezamenlijk (tussentijdse) bevindingen;

- reflecteren op elkaars en eigen denk- en handelingskaders.

Evaluatiewetenschappers schrijven bovendien dat het cruciaal is dat de stakeholders open staan voor de nieuwe evaluatieaanpak en voor de perspectieven van andere partijen, die mogelijk kunnen conflicteren. Bovendien is het belangrijk dat ze het belang en de urgentie van leren herkennen en dus *willen* leren. Als dit niet het geval is, verliest de evaluatie al snel zijn kracht.

*Nieuwe rollen: evaluatoren*

Niet alleen voor de deelnemende partijen is de rol anders dan bij meer reguliere evaluaties, aan de evaluatoren worden ook andere eisen gesteld. Zo zijn zij niet alleen objectieve beoordelaars van beleid, maar hebben ze ook een actieve rol in het faciliteren van dialogen en interactiemomenten tussen de deelnemers. Ze stimuleren daarbij de leerprocessen van deze partijen. Daarnaast verzorgen ze de afstemming van het evaluatie onderzoek op de beleidspraktijk en hebben ze aandacht voor de bereidheid om te leren van de deelnemende partijen, zoals beschreven bij de vorige karakteristiek. Dat laatste kunnen ze bijvoorbeeld aanmoedigen door responsief te zijn naar zorgen die de deelnemers uiten met betrekking tot de evaluatie.

Vanuit de literatuur wordt beschreven dat een interdisciplinair evaluatieteam aan al deze eisen kan voldoen; door kennis van verschillende achtergronden samen te brengen kan aan de verschillende rollen tegemoet gekomen worden. Verder kan een interdisciplinair team verschillen in epistemische culturen overbruggen wat de transparantie binnen de evaluatie ten goede komt. Een onafhankelijke externe review op het proces kan de wetenschappelijke kwaliteit en onafhankelijkheid waarborgen.

**Het belang van vertrouwen**

Een gebrek aan wederzijds vertrouwen tussen partijen kan leren in de weg staan. Wantrouwen is niet ondenkbaar in situaties waar machtsrelaties (tussen overheden onderling, maar ook tussen overheden en maatschappelijke partners, bedrijven en burgers) zijn. Om in te zetten op het realiseren van vertrouwen tussen partijen, worden in de literatuur twee strategieën aangedragen:

1. Het samenbrengen van alle partijen in een professioneel gefaciliteerde dialoog in een veilige omgeving, waarbij er veel ruimte is voor het uitwisselen van verschillende perspectieven en belangen;
2. Vooraf potentiele machtsrelaties inzichtelijk maken en daar waar mogelijk afspraken te maken, waardoor een basis gecreëerd kan worden voor wederzijds vertrouwen.

**Hoe blijft de evaluator onafhankelijk?**

Een vraag die – terecht – frequent de kop op steekt bij lerende evaluaties, is de vraag hoe de evaluerende partij onafhankelijkheid behoudt zodat de gevonden resultaten voldoende valide en betrouwbaar zijn. Een oplossing hiervoor is het inzetten van een interdisciplinair team; meer reguliere evaluatietrajecten en meer faciliterende leerprocessen kunnen verdeeld worden over de verschillende teamleden. Daarnaast is een belangrijke rol weggelegd voor (onafhankelijke) externe review van het proces en het product. Bij het Natuurpact is bijv. aan een externe partij gevraagd, het Athena Instituut om de implementatie van de lerende evaluatie te analyseren. kan voor wetenschappelijke borging door een partij de implementatie van de lerende evaluatie geanalyseerd worden – het Athena Instituut is in het geval van het Natuurpact de partij die deze analyse voor PBL uit gaat voeren in 2016 en externe reviewers zijn ook betrokken.

*De kennis en perspectieven van verschillende stakeholders worden geïntegreerd*

Tijdens een lerende evaluatie is er veel aandacht voor de verschillende typen kennis en perspectieven van de stakeholders die betrokken zijn. Verschillende typen kennis komen op deze manier samen, zoals wetenschappelijke kennis, maar ook ervaringskennis sterk gefundeerd in de beleidspraktijk. Als de stakeholders deze typen integreren ontstaat er kennis die robuust is, die breed gedragen wordt door alle deelnemende partijen en waarop oplossingen gebaseerd kunnen worden die in de verschillende werelden betekenis hebben.

*Verantwoorden en leren*

Zoals eerder besproken, is het kenmerkend voor een lerende evaluatie dat deze zowel tegemoet komt aan de behoefte om te evalueren of vooraf gestelde doelen wel gehaald worden (verantwoorden), als de behoefte om te leren tijdens nieuwe omstandigheden en een veranderde beleidsomgeving, waarbij doelen zich ontwikkelen. In veel literatuur worden deze twee doelen beschreven als onverenigbaar; beide vragen om een geheel andere aanpak en insteek. Bovendien, omdat de verantwoordingsvraag vaak van bovenaf komt (e.g. de opdrachtgever van de evaluatie) is het niet ondenkbaar dat deze voorrang krijgt in het uiteindelijke evaluatieproces.

Om aan deze spanning tegemoet te komen, adviseren evaluatiewetenschappers om: i) tijdens het evaluatieonderzoek gericht sub-evaluaties in te bouwen die één van beide doelen ondersteunen en die vervolgens aan elkaar te linken; ii) regelmatig leermomenten in te plannen, waar tussentijdse resultaten gezamenlijk geïnterpreteerd worden en er gereflecteerd wordt op eigen en elkaars werkpraktijk; en iii) opdrachtgever en beleidspraktijk met elkaar in dialoog te brengen zodat er ruimte blijft bestaan voor leren naast verantwoording.

De decentralisatie van het natuurbeleid zoals beschreven in het Natuurpact en de gevolgen daarvan voor de werkpraktijk van de betrokken partijen maakt dat een lerende evaluatie zich uitstekend leent om én de voortgang van het natuurbeleid te evalueren, én ondersteuning te bieden aan de betrokken partijen en de nieuwe rollen die zij gevraagd zijn op zich te nemen. De evaluatie biedt ruimte om 1) het natuurbeleid te onderzoeken op doelbereik van de Vogel- en Habitatrichtlijnen (VHR) en de Kaderichtlijn Water (KRW) – waarvoor het Rijk verantwoording aflegt aan de Europese Commissie, 2) bij te dragen aan de uitwerking door provincies van de doelen met bijbehorende beleidsstrategieëen voor de andere ambities op gebied van natuur en samenleving en natuur en economie en 3) te leren over verschillende mogelijke handelingsperspectieven om deze doelen dichterbij te brengen.

De betrokken partijen bij natuurbeleid – de provinciale beleidsmedewerkers, de maatschappelijke partners (zoals terreinbeheerders), maar ook agrariërs en zelfs burgers - hebben verschillende waarden en kijken anders aan tegen het voorliggende probleem en de mogelijke oplossingen. Het is waarschijnlijk dat zij ook verschillen in hun visie op welke kennis hiervoor nodig is. Deze onzekerheden – Hoe kan het probleem worden gedefinieerd? Wat is de juiste oplossing? Wat is überhaupt ‘juist’? Wat moeten we daarvoor evalueren? – en de verschillen in perspectieven en waarden zijn inherent aan complexe maatschappelijke problemen. Tijdens een lerende evaluatie wordt dit herkend, erkend én geaccepteerd. De lerende evaluatie behelst niet alleen het beter kunnen realiseren van de doelen van het Natuurpact, maar ook het gezamenlijk vormgeven van deze lerende evaluatie om zo de weg naar 2027 zoveel en zo goed mogelijk te ondersteunen.

# Summary

In recent years, nature policy in the Netherlands has become decentralised of which the agreements are recorded in Administrative Agreement Nature (2011/2012) and the Natuurpact (2013). The twelve Dutch provinces develop and implement nature policy in order to, together with the national government, realise these agreements by 2027. They aim to realise the Dutch Nature Network, attain the international goals as stated in the European Bird and Habitat Directives and the Water Framework directive, as well as strengthen societal engagement with nature. The Ministry of Economic Affairs (EZ) and the Association of Provincial Authorities (IPO) have commissioned the Netherlands Environmental Assessment Agency (PBL) to evaluate the progress of the agreements every three years. The recent decentralisation of nature policy has brought about many changes for the involved stakeholders: many traditional parties are taking on new roles, and new parties (including societal partners, such terrain management organisations, private businesses and citizens) have appeared on the proverbial nature policy stage, and will take on both new and old tasks. In addition, the nature ambitions for 2027 as described in the Natuurpact are of a more generic character, and thus the provinces will collaborate with the aforementioned involved parties to translate these ambitions into concrete nature goals and associated policy strategies.

Without a doubt, these developments in nature policy entail changes in practice for all involved parties, meaning that the Natuurpact is far from business as usual. In response, instead of a traditional/classical impact assessment, the commissioners have asked PBL to conduct an evaluation where *learning* has a more prominent position. The first report of this evaluation is planned for 2016.

**What is a reflexive evaluation?**

As the name already suggests, a reflexive evaluation process is centred upon reflecting and learning from existing practices. In 'classical' evaluations the success or failure of a policy in realising its intended purpose is determined and connected to (aspects of) the policy strategy. Such a retrospective assessment, however, cannot serve to timely adjust and improve the evaluated policy, and generally its use in informing future policy decisions is limited due to ever developing contexts. Reflexive evaluation is a response to the low user-value of classical evaluation approaches. In contrast, a reflexive evaluation occurs during the policy development and implementation, giving sufficient space and time for learning processes and making adjustments. These learning processes take place on two levels: 1) policy practice, and 2) the underlying beliefs and assumptions regarding the problems, the policy goals and the policy instruments. Additionally, the assessment is of a dynamic character; there is no 'hard' assessment on predetermined goals, as the goals may develop over time, taking into account levels of uncertainty and unpredictability that surrounds a problem that cause its context to be ever changing. During a reflexive evaluation, evaluators and evaluated come together for collective interpretation of evaluation findings and how this contributes to the overall improvement of policy practice. The evaluator thereby continuously anticipates on, responds to and reflects with the participating stakeholders.

*Source: translated from Dutch from the Plan van aanpak. Lerende evaluatie Natuurpact (PBL, 2015)*

The approach of a reflexive evaluation is relatively new for PBL. Throughout the process PBL steps away from its role as a purely classical evaluator and will become more actively involved with the parties whose policy practice is under evaluation. Thus, the approach is not only novel for PBL; it demands from all involved parties a divergence from their traditional role, and a new outlook on policy evaluations. Naturally, such a new approach brings about tensions and questions. For instance, how will PBL maintain its independent status in this new role? Will the evaluation research lose focus on goal-attainment when there is so much emphasis on collaboration and interaction? And how will such an evaluation do justice to all the different perspectives that surround nature policy?

To be prepared for these and other questions, as well as potential tensions, PBL has commissioned the Athena Institute of the VU University Amsterdam to review the scientific literature regarding reflexive evaluation. Besides a scientific exploration of the theoretical basis for reflexive evaluation, this review also serves to provide insight into the conditions and intended added value of such an evaluation approach. Another report (planned for 2016) will provide a reflection on PBL's implementation of the reflexive evaluation, to assess whether the envisioned added value has been realised, as well as to provide recommendations for future improvement of the evaluation approach. To provide concrete tools for implementing as well as participating in a reflexive evaluation, PBL will publish a guidebook for policy professionals directly based on their experiences with the Natuurpact evaluation and other evaluation projects.

**Why reflexive evaluation?**

Recent developments on the context of contemporary policy have given reason for a new outlook on policy making, as well as evaluation thereof. These developments are, for instance, the increased level of complexity of societal problems, as seen in food security, climate change and the transition to sustainable energy. Such problems are difficult to solve as they are affected by a plethora of ecological, social and economic aspects. Consequently, identifying effective solutions is equally complex. Policy decisions regarding these unstructured problems increasingly take place in collaboration between governments, citizens, societal organisations and businesses. Moreover, there is interaction between policy processes at various levels (for instance, the European Union, national, provincial and local governments). Such societal problems are thus characterised by multi-actor involvement and multi-level governance.

At the same time, there is an increased demand for measuring the results of policy and linking these results to invested money and resources to account for policy decisions. These trends (that is, the increased multi-actor and multi-level governance character of policy, as well as the call for accountability) seem conflicting. After all, who is accountable for what becomes harder to determine when increasingly more parties are involved with developing and implementing policy. Each involved stakeholder has his own relevant experiential knowledge required for decision-making and each stakeholder will interpret the effect of policy in his own way. Moreover, demonstrating cause-effect relationships between policy on the one hand, and social processes on the other, becomes more complicated due to uncertainty and the interdependence of complex social issues.

Evaluation researchers have argued that evaluation in this type of multi-actor and multi-level governance policy processes has the potential to fill a new role. They emphasise that the different actors should learn *from* and *with* each other – about their own and each other's perspectives and practices and how these influence their shared policy practice. During such learning processes a desired outcome is that the knowledge shared among actors leads to innovative insights. Novel approaches to evaluation can play a significant role in realising this outcome.

In recent years, new evaluation types have been developed that have this exact aim; stimulating learning about a practice or programme, during the evaluation thereof. The literature provides several reasons for shifting towards these new approaches. Firstly, research demonstrates that findings gained through more traditional/classical evaluations are generally only marginally used in policy practice. As such evaluations occur either before or after the implementation of policy, they are therefore less likely to be effectively integrated into the policy cycle. Moreover, during traditional evaluations, there is generally a predominant focus on whether or not the goals are attained, rather than determining causes for failure and how these may be overcome. In addition, such regular policy evaluations often concern singular policy programmes, while complex problems require multi-level governance. In contrast, during a reflexive evaluation, the evaluation research is aligned with the needs of the parties whose policy is evaluated, thereby increasing the user-value of the findings. The evaluation occurs *during* the development and implementation of policy, by which timely adjustments may be made to policy and further increase the likelihood of goal-attainment. Learning has a central position; there is significant focus on analysing best practices, sharing experiences, and collectively reflecting on one’s own and each other’s practices to prevent each party from individually ‘reinventing the wheel’.

Literature on reflexive evaluation primarily concerns singular policy programmes or project evaluations, such as the evaluation of the contribution of one single policy strategy to the attainment of policy ambitions. Equally relevant, however, is the implementation of evaluation research in a context where multiple projects, on various levels, simultaneously contribute to the realisation of a policy ambition. In such situations we refer to an ‘evaluation arrangement’ consisting of multiple sub-evaluations on different operational levels, with the sum of these various sub-evaluations resulting in an assessment of goal-attainment. Here, in addition to learning by different actors (e.g. policy professionals, entrepreneurs, representatives of nature organisations), determining policy effectiveness and efficiency in relation to goal-attainment also takes a prominent role. A reflexive evaluation thus aims to meet the demand of evaluating for accountability purposes, as well as the call for collective learning processes.

### The reflexive evaluation of the Natuurpact

Nature policy in the Netherlands is a complex societal problem that is characterised by multi-actor involvement and multi-level governance. Learning during the course of implementing nature policy may greatly contribute to successful goal realisation, but simultaneously there is a demand for impact assessments to provide justification for invested finances and resources. To adhere to both calls, the decision was made to evaluate the Natuurpact through a reflexive evaluation approach.

**Stakeholders during the Natuurpact evaluation**

During the evaluation of the Natuurpact, policymakers on national and provincial level, societal partners and other important stakeholders are actively involved. Not only do they provide input for the evaluation research, but they also contribute by sharing their learning and informational needs relevant for their own work practice.

The Natuurpact reflexive evaluation comprises multiple subsequent phases. During the first phase, the evaluation framework – containing the research questions, the policy strategies up for evaluation and the goals of the evaluation – has been developed in collaboration with provincial policy professionals, societal partners and representatives of EZ and IPO. The research questions determined in the first phase will be answered during research in the following phases, during several sub-projects. During these phases, the stakeholders will also be in close interaction. Additionally, multiple collective learning occasions have been planned throughout the course of evaluation research. Through regular interaction between the evaluation researchers and the stakeholders, learning questions and knowledge obtained from practice are aligned with the evaluation research. Thereby, a different type of knowledge is developed than that of traditional evaluations; this includes not only the development of knowledge regarding goal-attainment and efficiency, but also transformational knowledge (i.e. knowledge about the processes required for realising the desired outcomes).

### Characteristics of reflexive evaluation

A reflexive evaluation focuses on collective learning processes in aims to bridge the gap between results-based evaluation on the one hand, and purely process-based evaluation. Hence, a reflexive evaluation may be positioned between more traditional, top-down evaluations (such as impact assessments) and bottom-up evaluations (such as, for instance, responsive evaluations). Table I describes the differences between these evaluation types.

The seven characteristics for reflexive evaluation found in the literature are as follows:

- A reflexive evaluation is a *multi-stakeholder process*;
- The *evaluation research* is optimally aligned with the *evaluated practice*;
- The evaluation objectives and its framework develop over time (it is an *emergent design*);
- Roles of participating stakeholders shift from passive to *active*;
- Roles of evaluators shift from outside observer to *facilitator of participation and learning*.
- Knowledge and perspectives from different stakeholders are *integrated*;
- Both demands for *learning* and *accountability* are satisfied.

*Table I. Differences between top-down, bottom-up and reflexive evaluation, inspired by Kuindersma & Boonstra 2006.*

| | Top-down evaluation | Reflexive evaluation | Bottom-up evaluation |
|---|---|---|---|
| Scientific discourse | System analysis | **Critical-theoretical** | Social constructivism |
| Objective | Accountability (impact assessment, and by extension policy improvement, though usually limited in practice) | **Accountability and learning** | Learning (policy improvement) |

| Policy perspective | Monocentric (government) | **Both** | Pluricentric (governance) |
|---|---|---|---|
| Relevant stakeholders | Only those parties whose policy is evaluated (e.g. policy makers) | **Primarily those stakeholders that have ownership regarding the intended changes; those groups whose practice will change as a result of the evaluation** | Policy makers, intended benificiaries and victims of policy |
| Evaluation framework | Formally predetermined (top-down, e.g. by government) evaluation goals | **Evaluation goals are interactively set up by involved stakeholders, but top-down predetermined goals also have a prominent place in the framework** | Evaluation goals are interactively set up by the involved stakeholders and may evolve over time |
| Relation evaluation research and policy practice | Seperated fields | **Integrated: evaluation research is optimally aligned with policy practice** | Fused: evaluation is merged with policy practice |
| Role of evaluator | Independent, distant and objective | **Interdisciplinary team that is actively involved and facilitates learning, but maintains independence** | Interacts with participants, is responsive to their needs, is actively involved |
| Role of stakeholders | Passive: stakeholders only provide data for the evaluation research. Typically only one stakeholder group. | **Active: develop the framework, use and learn from the evaluation findings, but also provide data for the evaluation research. Variety of (relevant) stakeholder groups** | Active: provide input, are involved in developing the framework. They also use and learn from the evaluation findings. Variety of stakeholder groups. |
| Type of knowledge | Expert knowledge, scientific knowledge | **Both expert knowledge, scientific knowledge and exepriental knowledge to develop new knowledge that is relevant for practice (mixed methods)** | Knowledge from practice and experience (qualitative) |

*Multi-stakeholder process*

In scientific literature it is often stated that during an evaluation the ‘relevant’ stakeholders should be involved in the evaluation process. Who is considered ‘relevant’ differs per evaluation scope. During a reflexive evaluation, this includes not only the policymakers (whose policy is being evaluated), but also stakeholders that are affected in any way by the policy programme (such as societal partners, businesses or citizens) who may be requested to contribute to the evaluation process. A broad outlook on stakeholder involvement ensures that the findings of the evaluation are strongly founded within policy practice. Thereby, decisions informed by these evaluation findings will enjoy solid social support. In order to realise this, in the literature researchers recommend regularly conducting stakeholder analyses, to ensure that during each evaluation phase all the right parties are involved.

**The importance of trust**

A lack of mutual trust between parties may compromise learning processes. It is

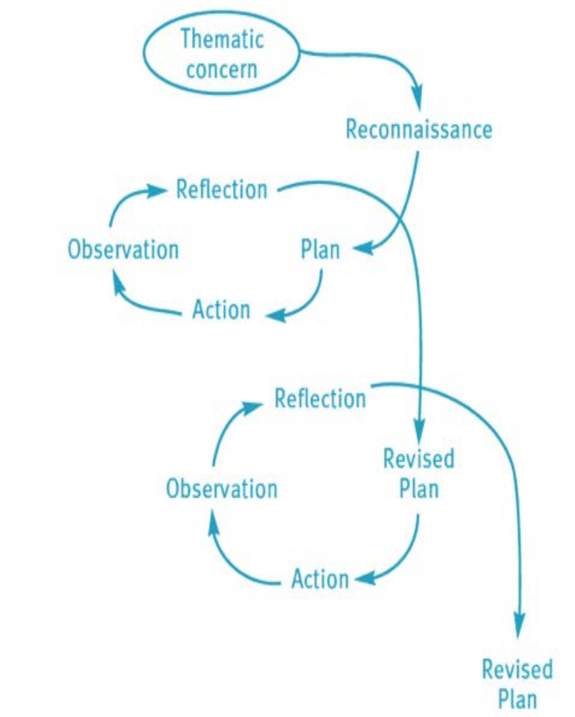


*Figure I. Schematic representation of evaluation cycles*

*The evaluation research is optimally aligned with the evaluated practice*

To meet the demands for both accountability and learning, the evaluation process is designed in a way that optimally contributes to the developments in policy practice. Researchers argue that such alignment may realised through a number of ways, including the following:

- The evaluation framework – comprising the objectives of the evaluation, the policy strategies up for evaluation, and the demarcations of the evaluation research – should be developed in collaboration with all relevant stakeholders. This ensures that the evaluation adheres to learning needs and thereby optimally informs the decision-making process.  Here, 'hard' objectives determined in national and international agreements also have prominent position on the evaluation agenda.
- The evaluation process features multiple subsequent cycles of planning, acting, observing and reflecting (learning). Evaluation is used to inform policy decisions and to develop and implement policy. And vice versa, experiences with policy practice inform the design of the evaluation process.

*The evaluation objectives and its framework develop over time (it is an emergent design)*

Apparent from the previous characteristic of reflexive evaluation, the evaluation research is continuously adapted to policy practice. This implies that prior to the evaluation process, it may be difficult to determine which subjects specifically will be researched, which methods will be appropriate and what kind of results will be produced. Flexibility – in terms of both the evaluation design and from the evaluators and participating stakeholders – is of utmost importance. By maintaining a flexible approach, it is possible to adequately deal with unexpected situations or problems that may arise at any given time.

*Participating stakeholders actively involved*

Flexibility is one of the requirements a reflexive evaluation asks from the participating stakeholders as well as the evaluators. The participating stakeholders do not – as during a classical evaluation approach – merely passively provide data to the evaluators. Rather, they actively participate in the entire evaluation process. They:

- provide input for the evaluation framework (comprising the goals, the to-be evaluated policy strategies and the demarcations of the research);
- collectively interpret the (preliminary) results of the evaluation;

- reflect on their own and each other's perspectives and practices.

Evaluation scholars additionally argue that it is crucial for the stakeholders to have an open mind for the new evaluation approach, as well as for the constructions of the other stakeholder parties, which may conflict with their own. Moreover, it is key the stakeholders recognise the importance and the urgency of learning; they thus must be *willing* to learn. If this is not the case, the evaluation will surely lose momentum and power.

*Evaluators shift from outsiders to facilitators*

Not only the participating parties fill a different role compared to more traditional evaluations, a reflexive evaluation also requires evaluators to diverge from their classical role. To start, they function as not merely objective assessors of policy, they also have an active role in facilitating dialogue and moments of interaction between the participants. Thereby they stimulate the learning processes of these actors. Additionally, they ensure the integration of evaluation research within policy practice, and have attention for the participants' willingness to learn, as described for the previous characteristic. They may encourage such willingness by, for instance, being responsive to the concerns participants share regarding the evaluation.

In literature it is also described that an interdisciplinary evaluation team is preferred to meet these demands; through connecting knowledge from different institutional backgrounds and different scientific fields, such a team is capable to fill all the different roles required for a reflexive evaluation. Moreover, an interdisciplinary team can bridge epistemic cultures, which positively affects the transparency of the evaluation process. An independent external review of the evaluation process can furthermore guarantee the scientific rigour and independency of the evaluation research.

**How does the evaluator maintain an independent status?**

A question that is understandably frequently asked, is how the evaluating party maintains an independent position, by which the evaluation findings may be considered valid and reliable. A solution to this is setting up an interdisciplinary team. Multiple regular assessments as well as the facilitation of learning processes can be divided across the diversity of team members. Moreover, there is an important role for an independent external review party to overlook the evaluation process and outcomes. In the case of the Natuurpact, an external party was asked to analyse the implementation of the reflexive evaluation, to contribute to the scientific rigour of the evaluation research. The Athena Institute of the VU University Amsterdam is the external party that will conduct this analysis for PBL in 2016. In addition, the theoretical framework they will use for analysis, will be reviewed by other external experts in the field reflexive evaluation.

*Knowledge and perspectives from different stakeholders are integrated*

During a reflexive evaluation there is attention for different types of knowledge and different perspectives of the involved stakeholders. By this approach, such different types of knowledge are brought together, such as scientific knowledge, but also experiential knowledge strongly founded in policy practice. When the stakeholders integrate these types of knowledge, knowledge is co-created that is robust and broadly supported by all participating parties. On this knowledge, solutions may be based that adhere to the different constructions of the participating stakeholders.

*Accountability and learning*

As discussed previously, it is characteristic for a reflexive evaluation to meet the demand for assessing whether predetermined goals are being attained by the current approach (accountability), as well as the demand for learning during a changing context in which goals are developed. A large body of literature argues these two functions of evaluation are irreconcilable with each function requiring a different approach and scope. Moreover, as the accountability function usually is enforced top-down (e.g. by the commissioner or funder of the policy programme) it is not uncommon this gains priority during the actual evaluation process.

To deal with this tension, evaluation researchers recommend to 1) build in sub-evaluations during the evaluation research that focuses on one of both function, and to subsequently link these sub-evaluations; 2) plan regular learning moments, where preliminary results are collectively interpreted and where there may be reflected on one own and each other's practice; and, 3) ensure continuous open dialogue between the commissioner and the evaluated policy practice to maintain that both demands for accountability and learning have prominent position on the evaluation research agenda.

The decentralisation of nature policy as described in the Natuurpact and its consequences for the practices of the involved stakeholders is such that a reflexive evaluation approach is optimally suited for assessing both the progress of nature policy, as well as for providing support for the involved parties and the novel roles they are required to fulfil due to this new situation. The evaluation allows 1) nature policy to be assessed on goal attainment regarding the Bird and Habitat Directives and the Water Framework Directive – for which the national government is accountable towards the European Commission, 2) a contribution to the development of provincial nature goals, and complementary policy strategies for the ambitions regarding societal engagement and the relation between nature and economy, and 3) to learn about various potential action perspectives available to bring these goals closer.

The involved stakeholders in nature policy – such as provincial policy professionals, societal partners (e.g. terrain management organisations), but also farmers and even citizens – all have different values and a different outlook on the present problem, as well as the appropriate solutions. It is likely that they also differ in their view on which knowledge is required to solve the matter. Such differences can include the following: How should we define the issue at hand? What is the right solution? What do we consider 'right'? What do we need to evaluate to come to the required findings? These uncertainties and the different perspectives and values are inherently related to complex societal issues. During a reflexive evaluation this is recognised, acknowledged and accepted. This reflexive evaluation thus does not only concern increasing the likelihood of goal attainment of the Natuurpact, but also collectively giving shape to the entire evaluation process to provide the best possible support on the way realising the ambitions by 2027.

# 1. Introduction

In recent years, Dutch nature policy has been decentralised and the 12 provinces have become responsible for not just for the development, but also the implementation thereof. The agreements on the decentralisation of nature policy are recorded in the Coalition Agreement Nature 2011/2012 (Bestuursakkoord Natuur 2011/2012) and the Natuurpact 2013 (Natuurpact 2013). With these agreements, national government and the provinces agreed to work together to develop the Dutch Nature Network (Natuurnetwerk Nederland) (NNN), achieve international nature goals (the Birds and Habitat Directives and the Water Framework Directive) and increase societal engagement with nature. These ambitions, part of the Natuurpact and representing an add-on to Coalition Agreement Nature 2011/2012, are to be realised by 2027.

The Ministry of Economic Affairs (Ministerie van Economische Zaken) (EZ) and the Association of the Provinces of the Netherlands (Interprovinciaal Overleg) (IPO), commissioners of the evaluation, have decided that evaluation of the Natuurpact should place emphasis on learning from experience with nature policy and on the progress in achieving the nature ambitions. PBL Netherlands Environmental Assessment Agency (Planbureau voor de Leefomgeving) (PBL), the national institute for strategic policy analysis in the fields of environment, nature and spatial planning, has been commissioned together with partner organisation Alterra (Wageningen University) to conduct this evaluation and will report on the progress every three years. Developments in the context of policy practice, the multi-actor character of nature policy and the often low impact of findings of more traditional evaluation approaches have encouraged PBL and Alterra to take on a novel evaluation approach that accurately reflects the complexities of contemporary policy practice – a reflexive evaluation approach. Alterra's earlier work on learning-oriented evaluation (Boonstra & Kuindersma, 2008) and evaluation arrangements (Kuindersma et al., 2006) provided productive building blocks for the current evaluation approach.

While the primary purpose of more traditional evaluation approaches are to monitor progress and instil accountability, conducting an evaluation that combines the accountability and monitoring functions with learning aspects is a new endeavour for PBL. Thus, PBL is keen to explore the potential added value of this approach and how it can be implemented in an effective and efficient way. This novel approach does not only entail a relatively new practice for PBL, it also asks from the participating stakeholders a new outlook on policymaking and evaluation thereof. Naturally, this new method produces tensions and questions – for instance, how will PBL maintain its independence as a policy evaluator when the evaluators frequently interact with the evaluated parties? Will goal attainment retain its important position on the evaluation agenda if interaction and learning is increasingly emphasised? And how will the reflexive evaluation do justice to all the different perspectives from the diversity of stakeholders involved in Dutch nature policy?

To prepare for these questions and other potential tensions produced by the new evaluation approach, PBL has commissioned the Athena Institute at the Vrije Universiteit (VU) Amsterdam to conduct an extensive scientific literature review on reflexive evaluations. Besides providing an elaborate scientific background on this new method, this review also serves as a work of reference for insight on the body of thought on which reflexive evaluation is based, the conditions recommended for its success, and its intended added value in relation to more traditional evaluation approaches. In a subsequent report (planned for December 2016) the Athena Institute will use this literature review to reflect on PBL's implementation of reflexive evaluation in the case of the Natuurpact, in order to come to recommendations to improve the evaluation approach. For a more step-by-step guide on executing or participating in reflexive evaluations, PBL is planning to publish a separate handbook for professionals in policy and policy evaluation, based on the experiences of the Natuurpact evaluation and other evaluation projects conducted by PBL.

Following this introduction, Chapter 2 presents a brief overview of the Natuurpact evaluation, the context in which this literature review took place. The methodology used for the literature review is presented in Chapter 3. Chapter 4 provides an overview of the key elements of a reflexive evaluation as described in the scientific literature, synthesised into a framework. Finally, Chapter 5 discusses the matters of quality control and scientific rigour in the context of multi-actor processes, as well as a plan for the reflection on PBL's implementation of the Natuurpact reflexive evaluation, planned for December 2016.

## 1.1 New forms of evaluation for policy practices

In the past decades, there have been new developments in the approach to policymaking. Due to internationalisation and the appearance of new citizen actors, modern societal problems have become increasingly complex, with high levels of uncertainty, widespread ecological and socio-economic impact, multi-actor involvement, and multi-level governance (Hajer 2003). Consequently, classical modernist political institutions cannot use the same strategies as before to solve today's complex societal problems in an effective and legitimate manner. In response, policymaking increasingly occurs in polycentric networks of governance in which power is dispersed over many involved actor and thus the traditional role of government has changed (Fischer 2006). Strategies that aim to solve societal problems, taking complexity and numerous stakeholders into account, similarly call for novel approaches to evaluation (Hajer 2003; Lehtonen 2014; Regeer et al. 2009).

Simultaneously, calls for more transparency and accountability in regards to policy and policy outcomes are apparent (Van der Meer & Edelenbos 2006; Guijt 2010). Proponents request clear policy goals and meticulous measurement of progress towards these goals, so that effectiveness and efficiency can be assessed and policymakers can be held accountable. In a time where competition for funding is particularly high, accountability is a major function of evaluation, and thus the presentation of hard facts as proof that a programme and/or policy is worth spending resources is most prioritised (Guijt 2010). Indeed, evaluations may also be used for political strategic ends, rather than its initial intent: the improvement of policy. When examining this, however, it is clear that this trend for added accountability appears to be contradictory with the multi-actor, interactive nature of policymaking described above. The accountability function of evaluation has become more complicated (Mayne 2003; Van der Meer & Edelenbos 2006), as proving causality between policy and societal developments has become problematic due to the increased complexity and interconnectedness of the policy context and societal processes. Moreover, due to the high number of stakeholders involved, consensus on clear policy goals and criteria is often hard to reach. Nature policy can thus be considered a so-called ‘unstructured’ or ‘intractable’ problem (Arkesteijn et al. 2015; Hoppe & Hisschemöller 1996; Rein & Schön 1993). These types of problems are characterised by uncertain knowledge (biodiversity measures are long-term and influenced by multiple external factors) and disagreements about normative elements (policymakers, NGOs, citizens, and industry disagree about the goals of nature policy). In other words, there is no agreement on facts or values (Douglas & Wildavsky 1982). For this reason, contemporary policy processes tend to be more goal-*seeking* rather than goal-*driven* (Van der Meer & Edelenbos 2006). For unstructured, intractable problems, there are no clearly defined pathways to solutions. To account for this, it is vital that the existence of intractable problems be considered in the design and implementation of evaluation approaches.

Traditional or ‘top-down’ evaluation types focus on assessing whether or not pre-determined policy goals have been achieved and, if not, it focuses on causes for lack of success (e.g. Steinmetz 1983; Kuindersma & Boonstra 2005). These types of evaluations are referred to as ‘top-down’ as they are generally commissioned by a hierarchically higher party, for instance the party that funds a programme or project such as the government. Through impact or performance assessment, such top-down evaluations primarily address the call for accountability and are thus less applicable to collaborative and interactive policy processes where goals evolve due to newly acquired insights (Kuindersma et al. 2006). There is thus a need for novel evaluation approaches that address the issue of accountability of policy outcomes while simultaneously taking into account the multi-actor and

**The epistemological position of reflexive evaluations**

New generations of evaluation approaches largely have their roots in a constructivist epistemological position – the idea that the development of knowledge is shaped (constructed) by an entanglement of social, natural and human aspects (Jasanoff, 2004). As Guba and Lincoln (1994) describe: *‘constructivism is […] the transactional/subjectivist assumption that sees knowledge as created in interaction among investigator and respondents’.* In general, new evaluation approaches thus signal a move away from the more positivistic (traditional) perspectives on evaluation, such as system analysis, but the stand-off between positivistic and constructivist positions in policy evaluation have not necessarily yielded fruitful results in evaluating large-scale policy programmes. Scholars have argued for a more pragmatic epistemological position, such as the approach Fischer describes in his book *Evaluating Public Policy* (1996) where he integrates empirical and normative evaluation into what he calls ‘practical deliberation’. Indeed, pragmatism is not bound to a system of philosophy, but rather focuses on ‘what works best when’ (Patton 1990) and on the possibilities of action rather than stringent recording of past experiences (Cherryholmes 1992). In pragmatic research, choices are driven by the objective of the research and its ‘anticipated consequences’ (Cherryholmes 1992, p. 13), and pragmatist researchers are free to choose from the whole range of research methods to conduct their studies (Creswell 2009). Hence, pragmatism is considered a philosophical support for the use of mixed methods approaches to obtain the best possible understanding of the research object, much aligned with the aim of a reflexive evaluation: to come to the highest possible quality of data to improve policy practice in the process of designing new systems (Arkesteijn et al. 2015). Rather than resigning to the proposed dichotomy between positivistic and constructivism worldviews, the reflexive evaluation approach is found in the pragmatic position and functions from the epistemologies and accompanying methods that work best at that given time, especially focusing on regular reflection on both functions (Arkesteijn et al. 2015).

multi-level governance character of contemporary policy practice – allowing an emergent design and changing policy objectives when addressing intractable problems such as nature policy.

At the other end of the 'evaluation spectrum' from an accountability to a learning focus (Kuindersma & Boonstra, 2005) we find 'bottom-up' evaluation approaches such as responsive evaluation, which primarily aims to facilitate interaction among stakeholders so that they come to new policy insights through mutual learning in order to improve policy (Abma 1996). 'Bottom-up' here signifies that more local parties - whose practice is determined or influenced by the programme or project - have a say in the evaluation scope and design. In complex multi-actor situations, learning becomes increasingly important as these situations are characterised by greater uncertainty and increasing ambiguity regarding the impacts and dynamics of the policies, requiring the co-construction of different types of knowledge (Van der Meer & Edelenbos 2006). Therefore, the value of more participative, inclusive types of evaluation has been recognised, as these aim to go beyond the monitoring and accountability aspect of evaluation findings, thereby improving policy practice and fostering learning at all levels of the policy process (Arkesteijn et al. 2015; Kirkhart 2000; Preskill & Torres 2000). A bottom-up approach is suitable for complex, unstructured policy processes where the policy goals may evolve over time due to newly acquired insights (Van der Meer & Edelenbos 2006), although the accountability goals of the evaluation are usually not met and may not even be the intention of the evaluation.

Table 1 shows the distinctions between evaluation types often made in the literature. Clear differences are seen in the objectives of the evaluation types, as well in how the evaluation framework is developed and the role of the stakeholders that are involved. Furthermore, the role of evaluator in a regular impact assessment (i.e. top-down) is drastically different from an evaluator during a bottom-up approach. During top-down evaluation, the evaluator is a distant and objective observer that has very limited interaction with the evaluated as to ensure his or her independence. During bottom-up evaluation, however, the evaluator is actively involved in the evaluation process and is in frequent contact with the stakeholders. Objectivity through independence is ensured through different ways, also found in qualitative research approaches; e.g. by researcher triangulation, where multiple researchers analyse data to check on selective perception (Patton 1990).

*Table 1. Features of top-down, bottom-up and reflexive evaluations (inspired by Kuindersma et al. 2006).*

| | Top-down evaluation | Reflexive evaluation | Bottom-up evaluation |
|---|---|---|---|
| Scientific discourse | System analysis | **Critical-theoretical** | Social constructivism |
| Objective | Accountability (impact assessment, and by extension policy improvement, though usually limited in practice) | **Accountability and learning** | Learning (policy improvement) |
| Policy perspective | Monocentric (government) | **Both** | Pluricentric (governance) |
| Relevant stakeholders | Only those parties whose policy is evaluated (e.g. policy | **Primarily those stakeholders that have** | Policy makers, intended benificiaries and victims of |

| | | | |
|---|---|---|---|
| | makers) | **ownership regarding the intended changes; those groups whose practice will change as a result of the evaluation** | policy |
| Evaluation framework | Formally predetermined (top-down, e.g. by government) evaluation goals | **Evaluation goals are interactively set up by involved stakeholders, but top-down predetermined goals also have a prominent place in the framework** | Evaluation goals are interactively set up by the involved stakeholders and may evolve over time |
| Relation evaluation research and policy practice | Seperated fields | **Integrated: evaluation research is optimally aligned with policy practice** | Fused: evaluation is merged with policy practice |
| Role of evaluator | Independent, distant and objective | **Interdisciplinary team that is actively involved and facilitates learning, but maintains independence** | Interacts with participants, is responsive to their needs, is actively involved |
| Role of stakeholders | Passive: stakeholders only provide data for the evaluation research. Typically only one stakeholder group. | **Active: develop the framework, use and learn from the evaluation findings, but also provide data for the evaluation research. Variety of (relevant) stakeholder groups** | Active: provide input, are involved in developing the framework. They also use and learn from the evaluation findings. Variety of stakeholder groups. |
| Type of knowledge | Expert knowledge, scientific knowledge | **Both expert knowledge, scientific knowledge and exepriental knowledge to develop new knowledge that is relevant for practice (mixed methods)** | Knowledge from practice and experience (qualitative) |

Reflexive evaluation is a new form of policy evaluation that builds upon the strengths of both top-down and bottom-up evaluation approaches, generating both learning and accountability and falling within in the middle of the evaluation spectrum. Through a deliberative and collaborative approach, a reflexive evaluation addresses the issue of accountability in the multi-actor setting (as in top-down approaches) while simultaneously facilitating collective learning to improve policy practice (as in bottom-up approaches). Although accountability and learning are often seen as opposing and incompatible concepts (e.g. Guijt 2010), a reflexive evaluation aims to reconcile these concepts by considering learning *as instrumental* in bringing about accountability, as it requires the adoption of a responsive and reflexive approach (Guijt 2010; Arkesteijn et al. 2015; Regeer et al. 2016). To illustrate this, accountability is generally understood as a programme answering to its funders and commissioners, to ensure money and resources are spent as originally intended and predefined results are attained (Guijt 2010). This is referred to as upwards or financial accountability (Ebrahim 2005; Regeer et al. 2016). To unite accountability and learning, however, we may expand the notion of accountability to other directions as well: a programme is not just accountable to its commissioner and funders, but also to parties otherwise involved or affected, which is called horizontal accountability (Regeer et al. 2016). Finally, a programme is accountable to itself; its own mission and goals, which Ebrahim (2005) has termed internal accountability. These latter two types of accountability may effectively be realised by constructing participative and responsive learning processes within the evaluation (Regeer et al. 2016). Thus, by learning about a programme's underlying mechanisms and its boundaries through the sharing of the experiential knowledge of its stakeholders, an increased understanding of the programme's functioning is realised. Simultaneously, establishing an accurate accountability track in the complex context of the evaluation becomes more feasible.

## 1.2 Evaluation arrangements

In large-scale complex policy programmes covering multiple domains, often multiple individual policy processes occur simultaneously. To deal with this complexity, an 'evaluation arrangement' may be implemented comprising multiple individual evaluations that, when analysed together, provide in-depth insights into the entire evaluation process (Teisman et al. 2002). In these individual evaluations, the sub-strategies of the complex policy programme are assessed individually, and these evaluations are connected to increase the overall coverage and quality of the arrangement's findings (Van der Meer & Edelenbos 2006). As the Natuurpact is a large-scale policy programme, its evaluation arrangement consists of multiple individual evaluations of sub-strategies that together make up the reflexive evaluation. Indeed, a reflexive evaluation supports the evaluation of large-scale programmes of an unstructured character, with many stakeholders and with multiple sub-policy processes. Optimal alignment between evaluation research and policy practice is sought in order to achieve the greatest possible improvement of policy practice (Van Mierlo et al. 2010; Regeer et al. 2009; Edelenbos & Van Buuren 2005).

Although all policy evaluations aim to contribute to learning in order to improve policy practice, a reflexive evaluation makes the learning ambitions explicit from the outset and integrates the learning process explicitly in the research design and chosen methods. There is no blueprint for a reflexive evaluation: choices regarding its design are dependent on the aims, ambitions and context of the evaluation. In order to ensure the quality of the learning evaluation of Natuurpact, this document describes the conceptual model that has been developed to guide the research, ensuring that the learning and accountability aspects will be effectively integrated throughout the evaluation process.

## 1.3 Origins of 'reflexive evaluation'

In the Dutch context of evaluation, the type of evaluation PBL is referring to is also called a 'learning evaluation'. However, literature search shows that the terms 'learning evaluation' and 'learning-focused evaluation' are not widely used in the English language scientific literature. The combination of the search terms 'learning' and 'evaluation' most often refers to evaluation of learning and education, and sometimes to learning as an element of evaluation. Dutch scholars Edelenbos and Van Buuren coined the term 'learning evaluation' as a specific approach to policy evaluation in 2005, but wider uptake of the concept has been limited. The term also suggests an emphasis on learning, while the evaluation as conducted by PBL strives to meet the demands for learning as well as for accountability. Moreover, the Natuurpact evaluation aims to instil reflection on the institutional settings of the programme that may inhibit transformational change. In literature, these types of evaluation, that acknowledge the uncertainty and disagreement that are inherent to complex problems, as well as the system structures that inhibit desired change, are referred to as 'reflexive evaluation' (Arkesteijn et al. 2015). Van Mierlo et al. argue a project may be regarded as reflexive '*if the network of those involved develops new ways of acting while the institutional context is changing too (and partly as a result of this)* (2010, p. 36). The goal of reflexive evaluation is to learn how there may be contributed to system innovation, by reflecting on the role of prevailing values and institutional barriers. As a result, collective learning processes throughout the network inspire changes in practices, as well as how these practices are embedded in institutions. A reflexive evaluation allows for developing goals and emergent design, and places emphasis on participation (Arkesteijn et al. 2015). In light of the recent decentralisation of Dutch nature policy, reflexive evaluation that has regard for the developments in the institutional context and its effect on policy is undoubtedly valuable, and it is for these reasons the Natuurpact evaluation is described as a reflexive evaluation approach.

Other forms of policy evaluations that originate from similar starting points are 'reflexive monitoring in action' (Van Mierlo et al. 2010), 'utilisation-focused evaluation' (Patton 2000; Patton 1984), 'developmental evaluation' (Patton 1994) and also 'practical deliberation of policy analysis' (Fischer 1996). These approaches to evaluation build on earlier work on 'fourth generation evaluation' (Guba & Lincoln 1989) and 'realist evaluation' (Kazi & Spurling 2002; Mark, Henry, & Julnes 1998). They share a similar outlook on evaluation; that by aligning evaluation research with the practice of those under evaluation, the evaluation findings are more likely to make a valuable contribution to the improvement of practice, and thereby increase the likelihood of goal attainment (regardless of whether goals were top-down/predetermined or bottom-up). Also, these forms of policy share an emphasis on interactive deliberation among the evaluation stakeholders, by which reflection and thereby learning may be promoted.

Beyond the evaluation literature, participative, integrative and interactive approaches to research have informed new evaluation approaches. Examples are 'deliberative policy analysis' in the field of policy sciences (Hajer 2003; Fischer 2006) and 'participative action research' (Heron & Reason, 1997) which has its roots in development studies. More generally, inclusive and participative forms of research, such as 'knowledge co-creation' (Regeer & Bunders 2009), fall under the banner of transdisciplinary research (Thompson Klein et al. 2001, Regeer & Bunders 2007) and, more recently, Responsible Research and Innovation (RRI) (Klaassen et al. 2014). Additionally, research on the science-policy interface has enriched insight in the process of knowledge transfer between the science and policy fields, similarly influencing approaches to evaluation (e.g. Watson 2005).

In this report, we have chosen to use the term reflexive evaluation, referring to the different evaluation schools and signifying an evaluation approach where there is specific attention for the learning processes of the participants and the evaluation team, without neglecting the call for accountability. It has similarities with other schools but, nevertheless, also differences, and is thus deserving of a distinct title. In our research, we have primarily consulted the literature on evaluation studies, while also integrating our own knowledge of participative and interactive approaches to research, and have integrated various conceptual and methodological cues and recommendations on how to design and execute a reflexive to ultimately come to our understanding of a reflexive evaluation approach.

# 2. The Natuurpact evaluation

In this chapter we will describe the design of the Natuurpact reflexive evaluation as formulated in the Evaluation Plan (2015), and the path that led to this design. As previously mentioned, a reflexive evaluation is especially appropriate for unstructured problems that may have large ecological and socio-economic consequences for which no clear pathway to a solution exists. Additionally, a reflexive evaluation may be greatly instrumental when there are numerous stakeholders involved or affected by a policy. A key benefit of a reflexive evaluation over traditional evaluation approaches is its ability to respond, through a combination of reflective learning processes and impact assessments, to the trends of increasing governance in multi-actor and multi-level policy settings.

In literature, a problem is argued to be unstructured or complex when it is characterised by two aspects: i) there is large uncertainty about the causes (and thereby the appropriate solutions) for the problem as many different factors (on multiple governance levels) affect the causal pathways. Cause and effect are thereby difficult to determine. Additionally, ii) the stakeholders involved differ in values regarding the problem and as a result there is no consensus among the stakeholders regarding the appropriate solutions. Nature policy is characterised by both dimensions. The fundamental ambitions that will be evaluated during the reflexive evaluation of the Natuurpact are 1) to improve biodiversity, 2) increase societal engagement with nature, and 3) strengthen the relationship between nature and the economy. For all three ambitions, the causal pathways are largely undefined, by which there is uncertainty how best to realise the ambitions. For improving biodiversity, the policy strategies are to an extent unpredictable, inherent to ecosystem functioning. The other two ambitions have initiated a relatively new policy field in the Netherlands, and strategies to best realise these ambitions are still largely in the research and development phase. Also, nature policy is characterised by a multi-stakeholder context, and additionally the decentralisation in 2011 have caused multi-level governance to become a key feature of nature policy. Many involved parties disagree on which aspects deserve priority, which knowledge is required and which solutions are thereby deemed most appropriate. These stakeholders include not just policy professionals, but also societal partners, businesses, farmers and citizens. In literature there is argued that to solve an unstructured problem, multi-stakeholder processes that inspire social learning are required. A reflexive evaluation may therefore be most beneficial, as it combines reflective learning processes and impact assessments. In such a policy climate, learning during the implementation of policy can contribute to improvement of practice, whereas at the same time attention for impact assessment for accountability purposes (e.g. towards the EU, enforcer of the international biodiversity goals) is of great importance. Therefore, the commissioners of the evaluation want to adhere to the need for learning and accountability in the evaluation by adopting a reflexive evaluation.

## 2.1 Phases in the Natuurpact evaluation

The Natuurpact evaluation consists of two phases. Phase I involves the development of the evaluation framework, which comprises the research questions, the policy strategies for assessment and the demarcation of the evaluation research. Subsequently, this evaluation framework has been formulated in the Evaluation Plan, in which the actual evaluation research is described. Based on the Evaluation Plan, an evaluation arrangement has been set up to adequately answer the research questions, which is executed in Phase II. Furthermore, several moments of interaction are build in to present and interpret intermediate findings towards the main goal of the evaluation. The regular interaction between the researchers of the evaluation and the involved stakeholders aim to connect emerging questions from practice to the evaluation research. Figure 1 gives an overview of the timeline of each phase of the reflexive evaluation of the Natuurpact 2016. Below we provide further information about the specific activities in Phases I and II.

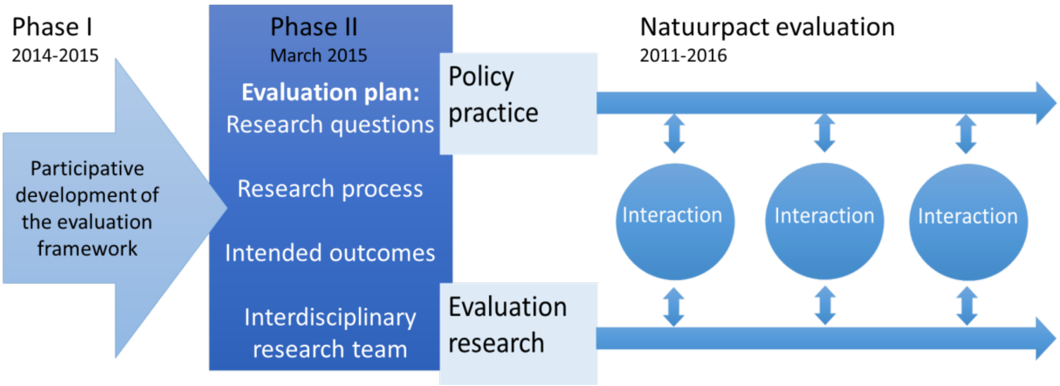


*Figure 1: Timeline of the reflexive evaluation of the Natuurpact 2016*

### *Phase I – Participative development of the evaluation framework*

During preparation of the evaluation, PBL and Alterra set out to develop the evaluation framework in collaboration with the relevant stakeholders. The first step in Phase I was to determine which stakeholders should be included in the participating group. At multiple instances, policy makers, representatives from the commissioners' parties, and societal and commercial parties were included in interviews and participative 'learning sessions' for the purposes evaluation agenda-setting and drafting the evaluation framework.

The process of developing the evaluation framework included three stages of interaction with the participants described above. First, interviews were conducted as the first moment of interaction with provincial delegates of all provinces, national delegates and societal partners, complemented by an extensive literature review on the current perspectives and views of these stakeholders regarding the three aforementioned ambitions of the Natuurpact. Second, during three learning sessions, the participants came together and jointly constructed the theory of change (i.e. the visualisation of the relationship between policy strategies and outcomes, and the assumptions underlying these relationships) describing what is known about the mechanisms of nature policy. The three sessions were designed to build upon previous work (interviews, document analysis, previous sessions) whereby an open dialogue and interactive reflection were stimulated. Table 2 gives more specific information about the aims of the three individual learning sessions. Finally, the theory of change constructed in the learning sessions was described in the 'PBL-notitie Evaluatie van het Natuurpact; Een voorstel voor een evaluatiekader' by Folkert et al. (2015)[1], which has been sent to the

[1] This publication can be retrieved from http://www.pbl.nl/publicaties/gezamenlijk-evaluatiekader-natuurpact-vastgesteld

commissioners, participants and the Parliament's second chamber to grant their approval and maintain the transparency and accountability of the evaluation.

*Table 2. The three learning sessions of Phase I, including their objectives, and the invited stakeholders.*

| Learning session | Objectives | Invited stakeholders **(i.e. participants)** |
|---|---|---|
| **I** | Prioritising nature goals and drafting a policy theory | National and provincial civil servants |
| **II** | Validation of policy theory and inventorying evaluation research questions | National and provincial civil servants<br>Societal partners<br>Agricultural representatives |
| **III** | Gathering information needs and indicators for answering research questions | National and provincial civil servants (policy makers and monitors)<br>Societal partners<br>Agricultural representatives |

### *Phase II – Translating the theory of change into an evaluation plan for 2011 - 2016*

In Phase II, the goals and related strategies depicted in the theory of change were further developed into an evaluation plan.  From the learning sessions it was apparent that the participants wanted to gain insight into the potential effectiveness and efficiency in attaining the biodiversity goals, specifically regarding the goals described in the European Bird and Habitat directives (in Dutch: Vogel en Habitat Richtlijn, VHR) and the European Water Framework Directive (in Dutch: Kaderrichtlijn Water, KRW) by 2027. For the other two themes – societal engagement, and nature and economy – the translation of these ambitions into tangible goals and theories of change is still high on the agenda and, consequently, the related strategies cannot be assessed in terms of effectiveness or efficiency regarding goal attainment. They shall therefore be studied for their potential effects on the ambitions and the required conditions for these effects. Results thereof may then contribute to the development of the theory of change of the two new policy themes. Furthermore, with regard to the decentralisation of Nature policy to the provincial level, participants wanted to obtain insight into the development of provincial nature policy in relation to governmental frameworks at the national level and international requirements. Thus, the role distribution among provinces, government and other stakeholders (e.g. societal partners, businesses, and citizens) was an important research subject.

Based on the above considerations, the following main goals of the evaluation are described in the Evaluation Plan: (i) obtaining insight on the (changes in) provincial nature policy and governmental frameworks, (ii) assessing and explaining the effectiveness and efficiency of provincial nature policy, (iii) providing policy recommendations and action perspectives for the provinces and the central government to increase effectiveness and efficiency of nature policy, and (iv) to contribute to learning on a policy system level. These goals have been translated to nine central research questions divided over three research clusters:

Cluster A: Policy development and role distribution (addressing main goal [i]):

1. How do the provincial goals, the policy strategies, the instruments and the organisational structures develop in light of biodiversity and the relation between nature, economy and society?
2. How do the policy and legal frameworks provided by the central government develop in relation to nature policy?
3. How does the role distribution within nature policy develop among the central government, the provinces, societal partners, businesses and citizens?

Cluster B: Effectiveness and efficiency (addressing main goals [ii] and [iii]):

4. Which (potential) contributions do the provincial policy strategies provide to the realisation of the VHR and KRW goals (effectiveness)?
5. How do the costs of these provincial policy strategies relate to the goal realisation (efficiency)?
6. How can the (potential) contribution to goal realisation be explained (what are success and fail factors)?
7. Which new action perspectives and improvements in the provincial policy strategies and the central governmental policy frameworks may further increase the effectiveness and efficiency of the strategies?

Cluster C: System learning (addressing main goal [iv]):

8. How do the provincial policy strategies, goals and central governmental frameworks and the role distribution between the central government, the provinces, societal partners, businesses and citizens relate to the ambitions in the field of biodiversity, societal involvement with nature and the relation between nature and economy?
9. What adjustments in nature policy, role distribution and the underlying policy theory are required for improving goal attainment in the field of biodiversity, societal involvement with nature and the relation between nature and economy?

### *The reflexive Natuurpact evaluation*

This section provides a more in-depth description of the reflexive aspect evaluation of the Natuurpact, based on the intentions described in the Evaluation Plan. The 2016 evaluation, after which a new evaluation will be executed every three years until 2027, is also referred to as the ex-ante evaluation, as it comprises the assessment of the potential contribution of nature policy strategies to the biodiversity goals prior to their implementation. The Natuurpact and the relating policy strategies are in different phases of development; thus the outcomes and effects described in the first (2016) report will still be limited.

*Table 3. The three ambitions of the Natuurpact in relation to evaluation activities for 2016 - 2027*

| Ambition: | Ultimate Goals: | Evaluation 2016: | Evaluation After 2016 (-2027): |
|---|---|---|---|
| **Improving biodiversity** | International goals, recorded in the VHR and KRW | a) Existing regular strategies transferred to provincial governments, assessed for potential contribution to goal attainment | a) These regular strategies will be evaluated for their actual contribution to goal attainment, starting from 2017 |

| | | | |
|---|---|---|---|
| | | b) New, innovative strategies are developed and explored, goal contribution is secondary | b) Depending on the process of developing these new strategies, these will be assessed for actual goal attainment |
| **Increasing societal engagement** | *Formulation of tangible goals is explored* | a) The same new, innovative strategies as described above, are explored for their potential to increasing societal engagement | a) Tangible goals are in the process of being developed; depending on this progress and the development of the novel strategies, contribution to goal attainment will be assessed |
| **Strengthening relation between nature and economy** | *Formulation of tangible goals is explored* | a) As well as for the relation between nature and economy | a) Similar as above |

Table 3 provides an overview of the three ambitions in relation to the evaluation activities that will be executed in 2016, and after 2016 (up until 2027). Regarding the biodiversity goals of the VHR and the KRW, the tasks for implementing existing, 'regular' policies have been transferred from national to provincial governments and their potential for goal attainment will be assessed. Furthermore, the reflexive evaluation will describe the initial experiences with the functioning of new, innovative policy strategies in nature conservation. For example, regarding new ways of nature management that may be more efficient for provincial management, but also experiments and pilot studies to increase societal engagement and the relation between nature and economy. These policy streams are currently developed in pilot programmes and vision documents. Therefore, learning is important on these issues and accountability is not possible yet.

The evaluations planned after 2016, until 2027, will be based upon monitored outcomes and effects of the executed policies. The actual contributions of the provincial policies to goal attainment will be assessed, as well as the complementary cost-effectiveness of these strategies. When the policies aimed at increasing societal engagement, and strengthening the relation between nature and economy have developed further, PBL will also evaluate these policy strategies on contribution to goal attainment regarding these themes.

An evaluation arrangement has been designed that includes all approaches required to answer the research questions described in the evaluation plan, which are based on the input of the participants during the evaluation framework development. However, the needs of the participants are likely to develop over time. To account for this, the reflexive evaluation has a flexible design that allows for adaptations according to the needs of the participants, but also to unexpected developments in the political administrative context. Furthermore, all the sub-research projects are intertwined and often address multiple research questions and involve multiple actors. The reflexive evaluation adopts smaller and larger cycles of action and reflection in its arrangement. Thereby, the evaluation is designed as a continuous iterative process of reflection and adaptation by both research and practice. The evaluation as a whole gives insight in the three-year cycle on the status of goal attainment and reflection on policy adaptation with relevant key stakeholders. In two collaborative learning sessions all stakeholders come together to give meaning to intermediate results. Furthermore, several reflection workshops are planned in the context of sub-research projects with relevant stakeholders (see Figure 1). Thereby, multiple stakeholders are involved and have an active role in the evaluation.

During the Natuurpact evaluation there is also frequent interaction with the political administrative context in which the programme and its evaluation function. This is mostly organised by the commissioners' parties; for instance, there are frequent meetings between the commissioners and the project team leaders of the evaluation. Also, from the commissioners' parties there is a representative who is more actively involved in the evaluation and its process, and has more regular contact with the project team. Simultaneously, the project team members are sensitive to the political administrative context the evaluation takes place, and is responsive to this by engaging the commissioners during the learning sessions and by being open and transparent in the evaluation process.

## 2.2 The intended deliverables of the Natuurpact reflexive evaluation

Several intended deliverables are explicated in the evaluation plan. These consist of intermediate outcomes and a final report of the evaluation 2016. First of all, the theoretical framework for the reflexive evaluation (this document) is one of the planned deliverables to ensure that the accountability and learning function of the evaluation both are adequately addressed. This will be complemented at the end of the evaluation by a report that shows the results of the operationalised framework, addressing the process and outcomes of the reflexive evaluation. Second, there are separate reports related to the learning sessions that show intermediate results of generated co-created knowledge. This knowledge can be used by different end-users to account for resources used in relation to outcomes realised. Third, there is a concept report and a final report of the reflexive evaluation. The concept report will provide all participants with the opportunity to give feedback on the presentation of the results. Based on these initial outcomes they can already make informed decisions in policy practice by which practice is no longer primarily informed by professional insights and experience, but also by insights from other stakeholders as well as by scientific inquiry. Thereby, the reflexive evaluation arrangement results not only in increased knowledge, but also in enhanced practice. The produced knowledge is not only scientifically informed but also socially and the enhanced practice is not only based on professional judgement but also on scientific insights.

## 2.3 The interdisciplinary evaluation team

An interdisciplinary project team coordinates all the different research activities and ensures that all intended outcomes are met.  PBL gives guidance to this team with researchers with backgrounds in policy research (Alterra/WUR), cost-effectiveness research (LEI/WUR), ecology (PBL), and transdisciplinary research (Athena/VU). The team members give guidance to sub-research projects in the evaluation arrangement. The evaluation is commissioned by the provinces through IPO, and by the Ministry of EZ.

In the interdisciplinary project team different roles can be explicated. The researchers from PBL, Alterra and LEI conduct the evaluation research on nature policy and the role distribution between provinces and national government, according to the research questions. Together they design the reflexive evaluation. PBL has final responsibility for the evaluation, but leads the evaluation together with Alterra. A large part of the evaluation's design is based on Alterra's expertise on learning evaluations and evaluation arrangements. Besides conducting the evaluation research, PBL and Alterra are also commissioners of the sub-projects that together shape the evaluation arrangement. They oversee the process of research, ensure that all research questions are covered, and are responsible for the integration of all information into a final report for the commissioners of the reflexive evaluation of the Natuurpact. Next to the organisation of the reflexive evaluation of the Natuurpact, they also have the task to report their experiences to their own organisations. PBL is in a transition from a solely on objectivity focused evaluation agency to an agency that adopts a participative approach in evaluation. Adopting a reflexive evaluation approach is a 'social experiment' for PBL wherein they are especially concerned about the scientific quality and rigour. In this process, PBL and Alterra are assisted by researchers from the VU University, by means of strategic advice and support. They help to employ methods differently, so that these inform transformative change. Also, the learning sessions are designed and facilitated by the VU researchers in cooperation with the other organisations. Furthermore, the VU researchers have the task to monitor and evaluate the learning processes in the reflexive evaluation. They thus have a double function that may interfere, and, therefore, an external review committee of experts in reflexive and learning-oriented evaluation approach will assess the quality of the theoretical framework described in this review, which the VU researchers have developed and will use for their evaluation. The management and Chief Scientist of PBL also follow this process closely and evaluate this experiment on its scientific quality and rigour. The different roles and responsibilities in the evaluation are visualised in figure 2.

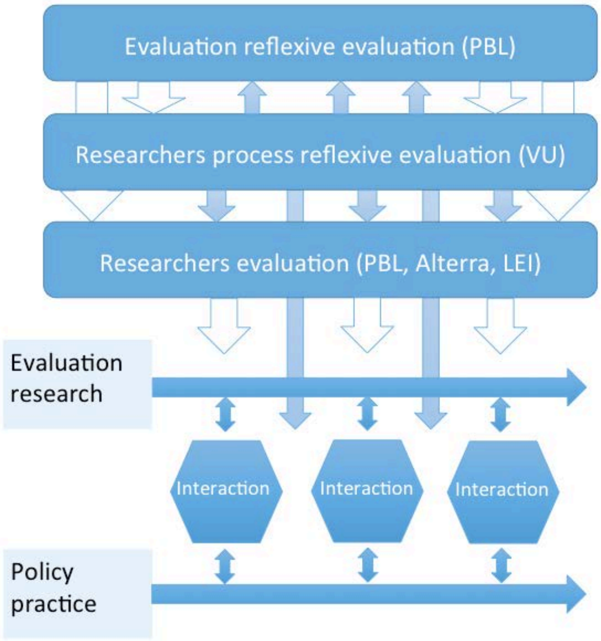

Evaluation reflexive evaluation (PBL)
Researchers process reflexive evaluation (VU)
Researchers evaluation (PBL, Alterra, LEI)
Evaluation research
Interaction
Interaction
Interaction
Policy practice

*Figure 2. Schematic overview of the relations between the different roles and responsibilities of the project team members.*

## 3. Methodology: Literature review

A structured literature search was used to determine the source of materials to review the concepts inherent to a reflexive evaluation. Peer-reviewed literature was the main source of information, however, a certain amount of grey literature was also included for more practical viewpoints on implementing a reflexive evaluation.

Search terms included different schools of evaluation studies such as ' learning evaluation', 'fourth generation evaluation', 'responsive evaluation', 'reflexive monitoring in action', 'utilisation-focused evaluation', 'developmental evaluation', 'realist evaluation' and 'evaluation arrangements'. These schools were selected because they all emphasise learning processes in order to come to sustainable solutions for policy practice, in most cases also in a multi-actor and/or multi-level governance context.

*Table 4. Selected evaluation schools that were used to obtain literature for analysis*

| **Evaluation schools**<br>***(Founders)*** | **Objective** | **Emphasises** |
|---|---|---|
| **Reflexive evaluation** ***(Arkesteijn, Mierlo, van & Leeuwis 2015)*** | Challenge systemic stability and support learning processes for institutional change | Reflection on current practice and systemic structures that inhibit desired system change |
| **Learning evaluation** ***(Edelenbos & Van Buuren 2005)*** | Improvement of goal attainment by learning among different stakeholders | Collective learning processes |
| **Fourth generation evaluation** ***(Guba & Lincoln 1989)*** | Improvement of policy practice by empowerment of stakeholders | Including different stakeholder groups throughout the evaluation |
| **Evaluation arrangements** ***(Meer, van der & Edelebos 2006)*** | Dealing with complexity of multilevel governance by connecting different sub-evaluations in a policy system | Optimal alignment of different sub-evaluations that focus on learning and accountability |
| **Reflexive monitoring in action** ***(Mierlo, van et al. 2010)*** | Promoting system innovation by learning from institutional bottlenecks | Collective reflection on practice, incorporating different system levels |
| **Responsive evaluation** ***(Stake 1991)*** | Improvement of policy practice by relying on knowledge of involved stakeholders | Responding to emerging needs of stakeholders in changing practices |
| **Utilisation-focused evaluation** ***(Patton 1984)*** | Improvement of policy practice by increasing usability of evaluation findings in policy decision making | Optimal alignment evaluation research and policy practice |
| **Developmental evaluation** ***(Patton 1994)*** | Supporting innovative initiatives through evaluation by responding to emerging needs of innovative practice | Collective reflection on process of development of initiative and intentional change |
| **Realist evaluation** ***(Pawson & Tilley 1997)*** | Development and improvement of practice by obtaining insight in what works best for whom | Identification of underlying mechanisms of policy outcomes |

Though there is extensive literature on multi-actor processes and multi-actor research, we have demarcated our search specifically to studies that focused on evaluation. No specific key words were required as inclusion criteria; a relatively small number of studies exist on the topic, so a 'bottom-up' search strategy was required. For each school of evaluation research, we selected one article from the founders of the approach and one article from its implementation in practice. We are aware that this is not an analysis of all literature related to the selected schools of evaluation, but it gives a sufficiently extensive overview of the main concepts addressed. From this literature selection we identified five key themes important for the understanding of reflexive evaluation. Subsequently, we expanded our search on topics within these key themes that, in our opinion, were insufficiently addressed by our primary literature selection. For instance, stakeholder selection in complex participative evaluation processes remained ambiguous and therefore we added literature on this for our analysis. Similarly, information on knowledge co-creation was added to enrich the theoretical framework. These articles are not included in the extensive analysis, since they do not specifically relate to schools of evaluation.

A total of 22 publications were selected for analysis, of which four were scientific reports, one book, and 17 peer reviewed articles (see Appendix 1). Articles that primarily focus on singular policy programme evaluation as well as publications that provide insight in the evaluation of more complex policy endeavours through evaluation arrangements were used. The selected publications were read thoroughly and, through consecutive cycles of open and structured coding, the main issues and underlying aspects were identified. The following five key themes that are important for the understanding of the reflexive evaluation were identified, presented in the order they will be discussed:

1. The intended outcomes of a reflexive evaluation
2. The accountability and learning functions of a reflexive evaluation
3. The identification of stakeholders involved in a reflexive evaluation
4. The process (and process-requirements) of a reflexive evaluation
5. The new role of the evaluators during a reflexive evaluation

In reviewing the literature, we concluded that many key themes were conceptually similar in the different evaluation schools, although they were articulated differently. Thus, for simplicity of exposition, rather than highlighting differences between evaluation schools, we focus here on specific insights proposed in the literature linked to each of the five key themes.

# 4. Framework for reflexive evaluation

Our analysis of the literature shows that a reflexive evaluation is especially appropriate when the policy programme or project addresses a complex societal problem, and is characterised by a large number of stakeholders and multi-level governance. Moreover, if the programme under evaluation comprises multiple sub-trajectories (e.g. an entire policy system) it may be so that an evaluation arrangement is required to adequately evaluate the associated complexities (Van der Meer & Edelenbos 2006). Evaluation arrangements comprise multiple interconnected evaluation studies that cover the sub-parts of the evaluated programme, aiming for optimal usability of the findings for informing decisions made by policy actors, and fulfilling accountability and learning purposes of the evaluation (Teisman et al. 2002). For this review we have combined literature on properties of both single reflexive evaluations and evaluation arrangements (which largely overlap) to come to a comprehensive understanding of the requirements for a successful reflexive evaluation.

Due to the inherent complexities of large multi-stakeholder policy programmes, the emergent character of both the programme and the evaluation arrangement, and the challenges associated with the multi-stakeholder character of the evaluation, a straightforward overview of steps to be followed to conduct a reflexive evaluation did not follow from our literature review. Nonetheless, we were able to identify a number of focus points that need to be addressed in a reflexive evaluation. How to exactly address these issues will depend on various contextual factors, including the scope of the evaluation, the institutional embedding of the evaluation, developments in the programme under evaluation, and the knowledge and skills of the involved researchers. The exact form and shape of a particular reflexive evaluation will evolve over time during its execution, or, paraphrasing Adam Kahane (2007): in cases of high complexity we have to *learn our way towards a solution.*

*We start at the end*

Based on our analysis of the selected literature we have identified a range of key characteristics and experts' recommendations for reflexive evaluation. Figure 3 depicts five focus points of reflexive evaluation: the interdisciplinary evaluation team, the process and its requirements, the stakeholders that are involved, the two functions of the evaluation (learning and accountability) and the intended outcomes of a reflexive evaluation. We continue with a discussion on each of these focus points in reverse order, starting with the outcomes of reflexive evaluation and ending with the interdisciplinary evaluation team.

## 4.1 Outcomes of a reflexive evaluation

With regard to the intended outcomes of a reflexive evaluation, it may be obvious that ultimately it aims to contribute to a more effective policy practice and thereby an increased likelihood of goal attainment (Edelenbos & Van Buuren 2005). Moreover, a reflexive evaluation intends to realise this ultimate aim through: 1) the generation of co-created knowledge, and 2) improved decision-making and policy practice through this co-created knowledge.

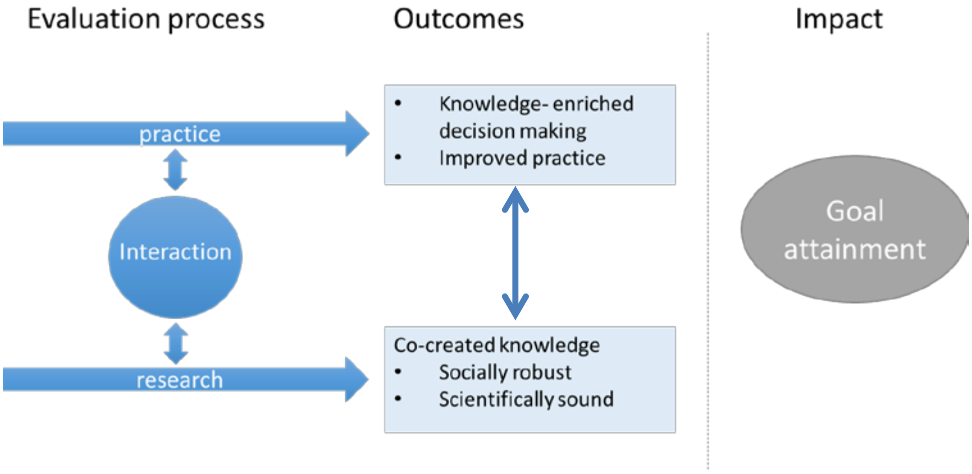
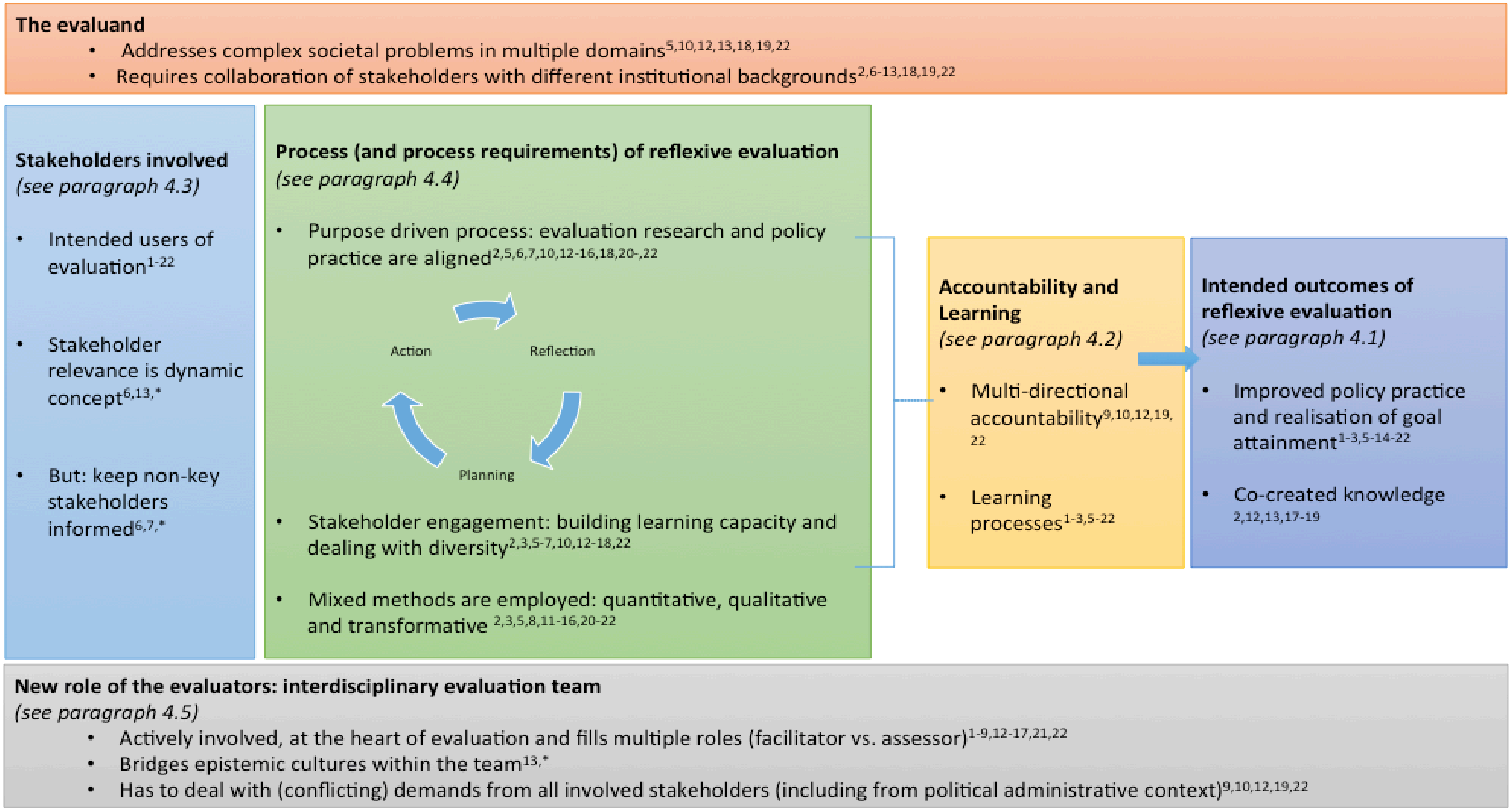


*Figure 3. Expected outcomes of a reflexive evaluation*

Figure 3 illustrates these two main outcomes we expect to see as a result of a reflexive evaluation process. Whereas in more traditional forms of evaluation the research process to generate knowledge for the evaluation and the practice under investigation are part of separate realms with limited interaction between them, in a reflexive evaluation, overlap (to a smaller or larger extent) is created between the realms of research and practice. This has implications for the types of outcomes that we can expect to see within a reflexive evaluation. On the one hand, we expect (policy) practices to becomes more informed and enriched by the knowledge generated through the evaluation process. Likewise, we expect the research process to become more practice-informed, resulting in knowledge that is more socially robust whilst remaining scientifically sound.4.1.1 Improved policy practiceA reflexive evaluation has been called 'utilisation-focused'; this means that it should lead to knowledge-enriched decisions by stakeholders in the area under investigation. Through knowledge-enriched decisions, the performance of programmes, interventions, or policies is enhanced. Policy practice may become more efficient and effective through reflecting on the evaluation process and outcomes, and at the same time, the evaluation process stimulates the policy practice to become more responsive to changes in the policy context. Ultimately, the idea behind reflexive evaluation is that enriched decision making and improved policy practice leads to the attainment of (medium and longer term) goals and ambitions at different levels in the policy area. This has implications for our thinking about the process of evaluation (see section 4.4) and for the knowledge that is produced during the process. Next we will discuss the characteristics of the kind of knowledge that we can expect a reflexive evaluation to generate, i.e. knowledge that will enrich decision-making and improve policy practice.



**The evaluand**

- Addresses complex societal problems in multiple domains[5,10,12,13,18,19,22]
- Requires collaboration of stakeholders with different institutional backgrounds[2,6-13,18,19,22]

**Stakeholders involved**
*(see paragraph 4.3)*

- Intended users of evaluation[1-22]
- Stakeholder relevance is dynamic concept[6,13,*]
- But: keep non-key stakeholders informed[6,7,*]

**Process (and process requirements) of reflexive evaluation**
*(see paragraph 4.4)*

- Purpose driven process: evaluation research and policy practice are aligned[2,5,6,7,10,12-16,18,20-,22]

- Stakeholder engagement: building learning capacity and dealing with diversity[2,3,5-7,10,12-18,22]
- Mixed methods are employed: quantitative, qualitative and transformative [2,3,5,8,11-16,20-22]

**Accountability and Learning**
*(see paragraph 4.2)*

- Multi-directional accountability[9,10,12,19,22]
- Learning processes[1-3,5-22]

**Intended outcomes of reflexive evaluation**
*(see paragraph 4.1)*

- Improved policy practice and realisation of goal attainment[1-3,5-14-22]
- Co-created knowledge [2,12,13,17-19]

**New role of the evaluators: interdisciplinary evaluation team**
*(see paragraph 4.5)*

- Actively involved, at the heart of evaluation and fills multiple roles (facilitator vs. assessor)[1-9,12-17,21,22]
- Bridges epistemic cultures within the team[13,*]
- Has to deal with (conflicting) demands from all involved stakeholders (including from political administrative context)[9,10,12,19,22]

1. Abma & Stake (2001), 2. Edelenbos & van Buuren (2005), 3. Flowers (2010), 4. Friedman, Rothman & Withers (2006), 5. Gamble (2008), 6. Guba & Lincoln (1989), 7. Huebner & Betts (1999), 8. Kazi (2003), 9. Kuindersma et al. (2006), 10. Lehtonen (2014) 11. Mark, Henry & Julnes (1998), 12. Van der Meer & Edelenbos (2006), 13. Van Mierlo et al. (2010), 14. Patton (1984), 15. Patton (1994), 16. Pawson & Tilley (1997), 17. Preskill & Torres (2000), 18. Regeer et al. (2009), 19. Regeer et al. (2016), 20. Stake ( 1976), 21. Stake (1991), 22. Teisman et al. (2002), * plus additional papers that were added to our initial literature selection

Figure 4. Conceptual framework of reflexive evaluation

### *4.1.2 Co-created knowledge*

The participative and interactive character of the reflexive evaluation brings together different perspectives and thereby knowledge from different stakeholders, by which an in-depth understanding of the programme, its underlying mechanisms, boundaries and involved stakeholders is developed. The knowledge generated is highly contextualised, as the process of its production occurs in intense interaction between the stakeholders, in the context of its application, and is thereby referred to as co-created knowledge. Co-created knowledge is therefore described as being 'socially robust' (Nowotny 1999, 2000); it is context-appropriate, broadly supported and sustainable. It is similar to the concepts of 'negotiated knowledge' (Bruijn & Heuvelhof 2008), and 'joint fact-finding' (Buuren et al. 2004; referring to Ehrmann & Stinson 1999); as a result, a shared understanding of knowledge is realised. As it is co-created in deliberation and negotiation, it simultaneously allows for multiple understandings for different parties.

Through this joint process of interpreting and creating knowledge, the knowledge is more likely to be perceived as independent and authoritative in different fields of research and practice (Van Buuren et al. 2004). Knowledge that is perceived as independent and authoritative is regarded as essential for the knowledge to be used in policy change – which, in this case, is the purpose of co-created knowledge. This is highly relevant; if stakeholders do not recognise their own point of view in, for instance, final reports on findings, they will be more inclined to dismiss the findings as 'false' or not trustworthy. Especially when asymmetrical power relations are of influence it is thus essential to ensure the views of those with power are sufficiently met in the evaluation research, without of course dismissing the perspectives of those with less power.

#### *The science-policy interface and uncertainty communication*

Another important aspect of knowledge co-creation is that perceived validity as well as the meaning of knowledge may differ between parties, especially between the science and policy fields. Research on the science-policy interface has shown that there are multiple and multi-faceted barriers between scientific knowledge production and the use of this knowledge in the policy context (e.g. In 't Veld 2000; Jasanoff 2004; Van der Sluijs 2010). An important misconception that was revealed by detailed analyses of processes at the science-policy interface is the idea that scientific knowledge is transferred from science to policy implementation in a linear and neutral fashion. Rather, the use of scientific knowledge in policy processes is seen as a process of active interpretation and sense making and thereby reconstruction of meaning. Actors from science and from policy make different sense out of the same data because of their respective frames of reference. It is suggested that early and regular interaction between the involved scientists and policy actors aids the mutual understanding of frames of reference, a constructive dialogue about research findings, and the development of boundaries objects to accommodate different local meanings.

A related concern is the difference in perception between scientists and policy actors regarding the extent to which the produced knowledge can be considered certain. The need for 'certain' knowledge is high in a policy context; after all, a policy maker needs to make concrete decisions on the basis of the acquired knowledge. Conversely, the possibility to produce 'certain' knowledge is rather low in a scientific context. In fact, certainty is often produced outside of the context in which knowledge is created (Bucchi & Trench 2008; Collins & Evans 2002). That is, even if uncertainties are communicated in original evaluation reports, they often disappear in the process of wider dissemination of the evaluation results (e.g. in newspaper headings). If policy decisions are based on these decontextualised findings, this may have far-reaching consequences. When knowledge is co-created, the level of certainty of the acquired knowledge will likely be known by the involved parties; both scientists and policy actors. However, parties that operate at a larger distance from this co-creational process – such as members of the House of Representatives or the Provincial Council – may also use the findings in their decision-making process.. In complex settings, where outcomes are the result of multiple interlinked processes, it is suggested to make the process of knowledge production transparent, to explicate the reasons for uncertainty. By combining outcome-evaluations with process-evaluations, the relationships between what happened are studied, for instance by establishing so-called Context-Mechanism-Outcome (CMO) configurations, derived from the field of realist evaluation research (Pawson & Tilley 1997). Other scholars add that uncertainties in research findings need to be communicated more clearly, to prevent the drawing of unsound conclusions (e.g. Kloprogge et al. 2007, regarding uncertainty communication). Both approaches suggest including an account of the process (e.g. policy strategies and their context, or scientific modeling) that led to certain results in the communication of evaluation outcomes.

### *What type of knowledge?*

Furthermore, as co-created knowledge is developed by the integration of the perspectives of all involved stakeholders, it is thereby likely to result in more scientifically as well as socially informed policy practice. Hence, the evaluation delivers findings that may support socially desirable and thereby sustainable policy decisions (Klaassen et al., 2014). In order to determine what kinds of knowledge questions could be beneficial to generate the type of knowledge needed to improve practice and realise goals, the typology introduced by Pohl and Hadorn (2008) can be used. They suggest that knowledge for changing (policy) practices can be ideally formed based on questions regarding: 1) the current situation 2) the aspired situation, and 3) the transformation from the current to the aspired situation. Thus, next to assessing the current situation regarding multiple components of the policy arrangement and the socio-economic-ecological situation in the problem area (type 1 questions), a reflexive evaluation could benefit from generating knowledge on the aspired situation according to the different stakeholders involved (type 2 questions) as well as from generating knowledge on the different pathways towards the aspired situation (type 3 questions).

**Key message: intended outcomes of a reflexive evaluation**

The ultimate impact of a reflexive evaluation is goal attainment achieved by the outcomes of:

1. **Improved policy practice** and knowledge-enriched decision-making

*as a product of*

2. **Co-created knowledge** deliberated by processes of joint-fact finding, by which it adheres to the perspectives of all involved stakeholders to generate socially-robust solutions

## 4.2 Accountability and Learning

A reflexive evaluation is said to fulfil, at least, two functions; first, it shows whether, and to what extent, the aspired goal attainment has indeed been reached, and second, it results in learning on the part of the stakeholders involved. Accountability and learning within the process of a reflexive evaluation are intermediate outcomes in the process of achieving the major outcomes of co-created knowledge and improved policy practice, as discussed in the previous section. In this section we will explain in more detail the concepts of accountability and learning, the relationships between these concepts (by introducing the idea of *multi-directional accountability*) and the use of static and dynamic evaluations to unite both accountability and learning within a reflexive evaluation.

### *4.2.1 Accountability and learning*

In any policy context, an evaluation functions to adhere to the call for accountability, and should test *'efficiency, output and outcomes of policies against their (initial) goals'* (Van der Meer & Edelenbos 2006, p. 202). Herein, *'the means by which individuals and organisations report to a recognised authority and are held responsible for their actions'* are central. Accountability towards commissioners, funders and authorities is also described as upwards, functional or financial accountability (Edwards & Hulme 1996, p. 967). While it is only natural for a funder to inquire after spent funds, this emphasis on evaluation for accountability in determining predefined goal attainment (for example, in an impact assessment) *'is also perceived to have considerable unfavourable effects'* (Regeer et al. 2016). For instance, a programme often relies on positive assessments in order for funding to be continued and, as a result, providing information to satisfy such top-down demands may gain priority over obtaining information to inform decision-making to support goal attainment (Ebrahim 2005). Additionally, it may be an incentive for what Lehtonen (2014) has termed 'strategic misinterpretation' of findings to prevent being held accountable for negative outcomes, and subsequently ensure continuation of the programme or project. Moreover, in evaluation for accountability purposes, approaches that are discrete and proven are generally favoured over approaches that are more innovative and uncertain, and adhere to a more emergent design (Regeer et al. 2016). Such proven approaches to evaluation assume the programme or project under evaluation is stable, with activities, goals and intended outcomes that may be univocally described (Regeer et al. 2009; 2016). However, in literature we find that many contemporary programmes and projects have more long term and intangible objectives that may be redefined during the course of the programme or project. Redefining the objectives may ultimately result in a fundamental change of the initial approach and direction of the initiative (Lehtonen 2014; Regeer et al. 2016).

As also described in our introduction, what furthermore adds to the complexity of realising accountability through evaluation, is the increased multi-actor and multi-level governance character of contemporary policy processes (Hajer 2003; Fischer 2006). Sometimes a division of responsibilities makes it possible to determine accountability per actor, but when the results are realised in co-operation between different stakeholders it is important to realise that the outcomes of the policy implementation are the result of interaction and argumentation (Fischer & Forester 1993; Guba & Lincoln 1989; Van der Meer & Edelenbos 2006).

Naturally, such developments have altered the demands on evaluation. In literature, primarily a call for evaluation for learning purposes is seen (Van der Meer & Edelenbos 2006; Van Mierlo et al. 2010). Such evaluations accommodate the involvement of multiple perspectives surrounding a programme or project, and allow for complex situations with emerging properties. Furthermore, learning becomes increasingly important in a multi-actor setting due to increased uncertainty and ambiguity surrounding a policy programme – there is thus more to learn (Van der Meer & Edelenbos 2006). However, it is generally believed that evaluation approaches that focus on learning diverge too far from assessing goal attainment or impact and thereby lose sight of the accountability purpose of evaluation. Guijt (2010) suggests that what keeps accountability and learning separated is a deeply rooted belief that accountability is not learning. Several authors, nevertheless, contest this notion and have suggested ways to overcome the proposed opposition between the two concepts, for instance: i) redefining accountability (e.g. Ebrahim 2005; Perrin 2002; Regeer et al. 2016) or adapting its meaning to different policy contexts (e.g. Zapico-Goñi 2007), ii) articulating the complementary values of learning and accountability to be used as design criteria for the evaluation approach (e.g. Van Mierlo et al. 2010), or iii) reformulating accountability evaluation questions from a learning perspective by regarding accountability as one of the learning purposes (Guijt 2010).

### 4.2.2 Multidirectional accountability

For uniting accountability and learning in one evaluation, authors have further explored the relations between these concepts. What has been found is when examining the concept of accountability more comprehensively is that both concepts are not as unrelated to each other as generally believed. So far, the accountability we have discussed is upwards or *vertical*: towards a funder or commissioner, also referred as financial accountability (Richmond et al. 2003; Ebrahim 2005). However, any organisation, programme, or project is not just accountable to its funders, but also to the actors and stakeholders that are affected by or otherwise involved in the implementation of a policy (e.g. policymakers, citizens, and partner organisations) (Ebrahim 2005). This is referred to as *horizontal* accountability (which subsumes the concept of downwards accountability, also found in literature, see Ebrahim et al. 2005; Regeer et al. 2016). As a reflexive evaluation engages not just the end-users of the evaluation findings (i.e. the policy actors) but also those parties that are affected or involved, it is instrumental in enhancing horizontal accountability. Finally, it may be stated that programmes or projects are also accountable to themselves and their own intended goals – their mission, which Ebrahim (2005) has termed *internal accountability*. Here, especially in changing and unpredictable environments, evaluation may be used as an instrument to ensure attainment of a programme’s internal mission, in addition to externally formulated objectives (Regeer et al. 2016).

Incorporating the theory of vertical, horizontal, and internal accountability can lead us to reconceptualise accountability as *multidirectional*, providing a more nuanced outlook on the role of accountability within learning in a reflexive evaluation. Both horizontal and internal accountability may be greatly enhanced using a reflexive evaluation where there is emphasis on a participative and responsive approach. By providing insight in the process of co-operation among multiple stakeholders in policy implementation and the underlying interpretations and interactions of the stakeholders involved, causal relations that underlie policy outcomes are revealed (Van der Meer & Edelenbos, 2006). Knowledge on the workings of the programme, its boundaries and mechanisms contributes to the learning processes of the participating stakeholders, ultimately contributing to an improved practice. This has also been termed the results-based use of evaluation (Kirkhart 2008). Simultaneously, interactive reflection on one’s own – and each other's – practice, perspective or frame can contribute to mutual understanding and alignment of strategies in order to improve policy implementation and goal attainment. This process-based use of the evaluation – through its collaborative and reflective character – is a more implicit process that also contributes to a better understanding of the programme by the participants, and thereby also to improved policy practice

(Kirkhart 2008). Insight is thereby obtained in stakeholders' own perspectives or frames and shared, by which potential conflicts may be resolved. The discrepancy between the intentions of a project or programme and its actual practice may thereby be decreased, which Ebrahim refers to as internal accountability to the mission of a programme or project (2005). Learning is thus an outcome of multiple processes during the reflexive evaluation and thereby occurs at several phases.

### *4.2.3 Uniting accountability and learning*

Presented as intermediate outcomes of the reflexive evaluation in Section 4.2.1 are accountability and learning, concepts that are proposed as irreconcilable in evaluation research generally (Fischer & Forester 1993; Lehtonen 2014; Van der Meer & Edelenbos 2006). Reasons for uniting these concepts have been elaborately described previously in this review; in this section we focus on the practical requirements in the evaluation arrangement design in order to successfully adhere to both accountability and learning. In the literature we located three practical guidelines that support realising an accountability track and learning processes. These are i) incorporation of both static and dynamic sub-evaluations, ii) ensuring sub-evaluations are both externally and internally conducted and, iii) the use of mixed and transformative research methods. Here we highlight static and dynamic evaluations and external and internal evaluations; we will discuss the use of mixed methods in a later section on the process requirements of a reflexive evaluation.

#### *Static and dynamic sub-evaluations*

Practically, the design of the reflexive evaluation may adhere to accountability and learning by intentionally incorporating sub-evaluations that are either more *static* or more *dynamic*. Static sub-evaluations assess impact in relation to goal attainment to realise upwards accountability towards the commissioner and funders of the evaluation, and prove that progress has been made. Dynamic sub-evaluations emphasise learning and co-operation, and allow for emerging goals and objectives on which participants continuously reflect (Teisman et al. 2002). By intentionally, consciously, and transparently designing the arrangement to meet both evaluation demands, many stakeholder groups are satisfied. This is not to say that a static evaluation, for instance, does not involve any learning processes; intermediate findings may of course be collectively interpreted and debated by the involved stakeholders.

#### *Internally and externally conducted evaluations*

Similarly, to adhere to both accountability and the learning processes, the sub-evaluations must possess components of both internal and external evaluations. Internal evaluations are conducted by the evaluated themselves, while external evaluations are conducted by an independent party. Internally conducted evaluations run the risk of being overly 'soft' on the findings and may also result in externalising 'bad' outcomes (Teisman et al. 2002). Simultaneously, external evaluations may not be optimally aligned with the evaluated, and negative outcomes may be attributed to the agency conducting the evaluation (Van der Meer & Edelenbos 2006). In both situations, learning is compromised. However, when aspects of both types are combined, accountability and learning may simultaneously be addressed. For instance, an internally conducted evaluation in which an additional externally conducted assessment is incorporated, contributes to the learning processes of the participants, but also to the perceived validity and accountability of the outside world (Teisman et al.

2002). Similarly, an external evaluation guided by a project team where the evaluated are represented by one or more team members is likely to find more endorsement from the evaluated, contributing to their learning processes while preserving independence and validity.

**Key message 1:**
***Multidirectional accountability* and *learning* are intermediate outcomes of the reflexive evaluation that work towards the co-creation of knowledge and improved policy practice**

1. ***Multi-directional accountability*** is directed to the commissioners of the evaluation (upwards accountability), and to other stakeholders affected by the programme or project (horizontal accountability) and based on co-created knowledge
2. ***Learning*** may be referred to as internal accountability, is stimulated through evaluation results or evaluation process, and is based on co-created knowledge

**Key message 2:**
**Accountability and learning are key elements of a learning focused evaluations, and can be united by:**

1. **Employing both *static* and *dynamic* evaluations**
   static evaluations adhere to (upwards) accountability, while dynamic evaluations address the learning processes of the participants (and thereby horizontal and internal accountability)

2. **Elements of *external* and *internal* evaluations are combined**
   this allows for evaluations that are sufficiently aligned with policy practice and simultaneously are independent enough to have weight and credibility

## 4.3 Stakeholders involved in a reflexive evaluation

A reflexive evaluation aims to involve stakeholders of the policy programme in the evaluation for several reasons. From an instrumental substantive perspective, involving a broad range of stakeholders provides experiential knowledge and perspectives of local practitioners that will lead to a more robust knowledge base on which a project (and evaluation) can be developed (Patton 1984; Huebner & Betts 1999; Edelenbos & Van Buuren 2005; Reed et al. 2009). Paying attention to the variety of stakeholders and their perspectives during the process of the evaluation encourages *'enough understanding, appreciation, information sharing, legitimacy or commitment to produce a credible evaluation that will eventually be used'* (Bryson et al. 2011, p. 3).

### *4.3.1 Stakeholder selection*

During any policy programme, however, aspects such as time and resource constraints place tensions on the range of stakeholders to engage and the intensity with which they are involved. Due to these constraints it is unlikely that a reflexive evaluation can effectively address all stakeholder's perspectives equally well (Bryson et al. 2011). The project team that guides the evaluation, thus has to consider a trade-off when selecting stakeholders: between involving a range of stakeholders that accurately represents the entire scope of the policy programme and ensures sufficient socially robust outcomes, and the limited time and resources available for realisation of the programme evaluation. Scholars have, in response, argued to narrow the range of stakeholders, often arguing to involve 'the relevant stakeholders' of a programme (Stake 1991; Patton 1984; Edelenbos & Van Buuren 2005).

For participative evaluations where time and other resources place constraints on the process of the evaluation, many authors agree that the stakeholders to involve should mostly be the primarily intended end-users of the evaluation findings (Patton 2008). However, to ensure sufficient and continuous support for the programme and its evaluation, stakeholders that are *'less key'* (Bryson et al. 2011, p. 3) should not be ignored.

In order to guide this process of stakeholder selection, different authors have suggested typologies to identify classes of stakeholders. The intended users are those actors that '*have responsibility to apply evaluation findings and implement recommendations'* (Patton 2000: p. 426). Similarly, Guba and Lincoln speak of involving *'those persons involved in producing, using and implementing the evaluand'*, such as the developers of the programme, its funders, those involved in implementing the programme locally, but also the commissioner of the evaluation; together these are termed *agents* (1989, p. 40). The second and third class of stakeholders that need to be addressed in a reflexive evaluation arrangement, according to Guba and Lincoln, are the *beneficiaries* and *victims*; those individuals or groups that may either benefit or are negatively affected by the programme, respectively.

A different and often used framework for identifying relevant stakeholders is that of Eden and Ackerman (1998) (see Figure 5). It considers levels of influence and interest of stakeholders with regard to the phenomenon concerned, in this case the reflexive evaluation. From this framework it follows again that intended-users are *key players* – in policy evaluation these are policymakers and implementers of the policy at various levels of governance. Their level of influence is high (the evaluation probably largely depends on the provision of information by intended end-users) as well as their level of interest (their work is being evaluated, which may yield essential information that is needed to make policy adjustments but may also lead to a public judgement of their performance). Commissioners and funders of the programme likewise are intended users and have a high interest in and influence on the evaluation endeavour. So called *context setters* have little interest in the evaluation and its findings but are, however, highly influential in the development of the programme and therefore may be a risk-group – they may sabotage the evaluation and hence, they should be managed and monitored (Reed et al., 2009). For instance, policy programmes may be heavily influenced by the political administrative context they function in. It is then relevant to consider stakeholders to engage to ensure the evaluation research is embedded within this context, by which political administrative support is ensured (Boonstra & Kuindersma, 2008). Other context setters in policy evaluations could be public engagement groups or businesses that provide resources in implementing policies. *Subjects* have a low influence, but a high interest. They lack impact-capacity but may become more influential when they form alliances with other stakeholder groups and therefore deserve consideration, especially if the evaluation findings may affect them (Bryson et al. 2011; Reed et al. 2009). Lastly, the *crowd* are those people who have both little influence and interest in the outcomes of a project and may just need to be informed about the evaluation and its findings (Bryson et al. 2011). Considering this matrix, in a reflexive evaluation the stakeholder groups that the evaluation team should engage with are thus primarily the key players (the end-users of the evaluation findings), followed by the context setters and to a lesser extent, the subjects.

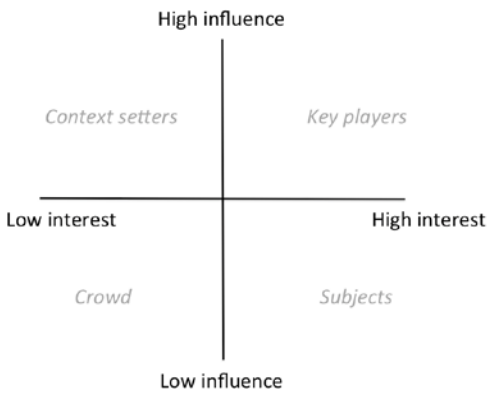


*Figure 5. Interest/influence matrix after Eden and Ackermann (1998).*

Nevertheless, levels of influence or interest (Eden & Ackerman 1998) or stake (Guba & Lincoln 1989) remain hard to determine, especially when this is done by one person or one small group that has a limited perspective on the entire range of the programme and its potential consequences. Therefore, Guba and Lincoln advocate that stake is a concept that should be negotiated between identified stakeholders, in a setting where those who wish to, may present their case (1989). Similarly, the Reflexive Monitoring in Action approach (Van Mierlo et al. 2010) and the stakeholder profiling approach (De Cock Buning, Regeer, & Bunders 2008) recommend the organisation of collaborative sessions to negotiate who should be involved in the evaluation.

Another important distinction that may ease the task of identifying the relevant stakeholders for the evaluation is between programme-stakeholders and evaluation-stakeholders. Bryson et al. (2011) argue that for an effective evaluation, the stakeholders included should primarily have a stake in the evaluation research, rather than in the programme itself. Evaluation stakeholders are those actors that are '*intimately tied to the purposes of the evaluation'* (Bryson et al. 2011, p. 3). During a reflexive evaluation where the evaluation is integrated within the programme, there is inevitably overlap between programme and evaluation stakeholders. Acknowledging this distinction, however, helps to determine who is relevant and when.

### 4.3.2 Stakeholder selection in dynamic context

Despite many available methods and tools for stakeholder analysis (e.g. Bryson et al. 2011; Broerse et al. 2008), what further adds to the complexity of determining relevant stakeholders during a reflexive evaluation are the emergent objectives and boundaries of a policy programme. Stakeholder-roles develop and emerge over time; the relevance of stakeholders is not static (Guba and Lincoln 1989) and their interest and influence levels are dynamic and dependent on the project's context (Reed et al., 2009). Consequentially, when an evaluation arrangement comprises multiple sub-projects, the relevant stakeholders may also be different per sub-project.

To deal with these issues, Bryson et al. (2011) suggest stakeholder analyses should occur multiple times during the process of an evaluation. *'Attending to and engaging with evaluation stakeholders typically must occur every step along the way'* (Bryson et al. 2011, p. 3): during the evaluation planning, the evaluation design, the data collection, the data analysis and lastly, the decision-making or implementation of the findings. In each phase, different stakeholders may be relevant as each phase comes with its own specific demands. For instance, during the data collection phase, stakeholders with abundant knowledge in the field may be of greater value than during the decision-making phase. By recurrently determining the relevant stakeholders, there is also ensured that the stakeholders' involved in any point in time will be more engaged with the process, as their views and perspectives are more relevant and valued. It indirectly contributes to streamlining the process of the evaluation. Stakeholders' relevance should thus be a recurrent concept to be negotiated during the evaluation, also within the different sub-projects (Bryson et al. 2011).

**Key message: stakeholder selection**

1. **Several methodologies are available for determining and selecting the relevant stakeholders**
   in literature it is advised to distinguish between actors that have a stake in the evaluation and actors that have a stake in the programme under evaluation

2. **Key stakeholders are the intended users of the evaluation findings and have an active role in shaping the evaluation process**
   scholars advise to also engage stakeholders that are less key in order to ensure continuous support, e.g. through informing them on the evaluation progress

3. **Stakeholder relevance is a dynamic concept and may therefore differ per point in time and per sub-evaluation project**
   therefore regular stakeholder analyses may ensure the relevant groups are involved at any given time

## 4.4 The process (and process requirements) needed to realise the intended outcomes

From the literature we have distinguished a number of process features that are believed to support the successful implementation of the evaluation. These features regard multiple levels in the evaluation; the entire evaluation arrangement, but also the single evaluations that take place within the sub-projects. The process includes the integration of evaluation research within policy practice, the multi-actor setting of the evaluation, and the use of mixed and transformative methods.

### *4.4.1 Purpose driven process: evaluation research and policy practice are aligned*

A reflexive evaluation follows a deliberative and collaborative process of impact assessment and collective learning simultaneously. Thereby, the reflexive evaluation strives to contribute to the likelihood of goal attainment – it is purpose driven. To realise this, first, the evaluation objectives are determined by the participants. Second, through frequent cycles of interpretation and adaptation intermediate outcomes are used to fine-tune the alignment between evaluation and policy practice. Third, the sub-evaluations within the evaluation arrangement need to be adequately linked to ensure the policy programme and its evaluation function as an effective whole.

*The stakeholders negotiate the evaluation objectives and boundaries*

To increase the effectiveness of the evaluation research, it is suggested by many authors that the objectives of the evaluation are deliberated and determined by the participants, rather than solely pre-determined by a commissioner (Edelenbos & Van Buuren 2005; Guba & Lincoln 1989; Patton 1984; Preskill & Torres 2000). Ensuring that the evaluation caters to the needs of the participants increases the impact of the evaluation. As Edelenbos and Van Buuren (2005: p. 606) state: *'a learning evaluation is a form in which users (the evaluated) and executors of evaluation (evaluators) shape the evaluations in close interaction and consultation'.* Importantly, the participants shape the evaluation according to their pressing concerns, learning objectives and goals in practice; the evaluation is a reflection of their needs (Guba & Lincoln 1989; Lehtonen 2014; Patton 1984; Regeer et al. 2009). In the process of design the participants and the evaluators together determine the research questions (Flowers 2010; Huebner and Betts 1999; Patton 1984), indicators and boundaries (Huebner and Betts 1999) that guide the evaluation and its sub-evaluations. Simultaneously, top-down determined evaluation objectives (for instance, enforced by the commissioners) are also incorporated in the evaluation framework. The validity and legitimacy of the evaluation in the eye of all stakeholders (including the commissioners) and thus the overall impact of the evaluation findings is increased (Van Buuren et al. 2004; Van der Meer & Edelenbos 2006).

A risk that needs to be considered here is the possibility that participants may not all see eye-to-eye regarding the evaluation objectives and its boundaries. Consensus is not an objective of the reflexive evaluation - indeed, the evaluation aims for acknowledging a plurality of perspectives where goals and objectives are openly deliberated. This is more elaborately discussed in paragraph *4.3.2 Stakeholder engagement: learning capacity.* Similarly, a risk is that the political administrative context in which the programme or project and its evaluation function, may require evaluation objectives and boundaries that conflict with the needs of the participants. The, often, large distance between the institutional background organisations of the commissioners and the stakeholders of the policy programme or project may result in insufficient understanding of each other's practices and thereby conflicting demands. It is the aim of the reflexive evaluation to bring both groups in dialogue to interactively reflect on their practice, by which the evaluation objectives are deliberated, as mentioned previously. The evaluation framework should eventually satisfy the demands of both parties. Factors that influence this, such as trust, are also discussed in paragraph 4.3.2. The project team that guides the evaluation may further help to ensure support and understanding for the programme and its evaluation on higher operational and strategic levels by investing in a good relation with the commissioner party. In paragraph *4.4 The interdisciplinary evaluation team*, the role of the project team and how these team members may further deal with the political administrative context is discussed.

Formulating and shaping research questions in the context of a reflexive evaluation can be aided by using the distinction proposed by Pohl and Hadorn (2008) between research question generating knowledge about 1) what is 2) what should be and 3) how we come from were we are to were we should be (which is referred to as transformation knowledge), as discussed previously in section *4.1.2 Co-created knowledge*. Regeer et al. (2009) suggest that the generation of transformation knowledge is aided by formulating learning questions on a learning agenda and following subsequent learning agenda's as well as solutions over time (by which it becomes a Dynamic Learning Agenda) (also see Van Veen et al. 2014). A challenge that the project team may encounter during this phase is deciding between conflicting perspectives and learning questions of participants. It is unlikely the project team can address all questions and information needs in the evaluation design, and the team may therefore have to help negotiate and prioritise, should the participants have conflicting perspectives that remain hard to overcome.

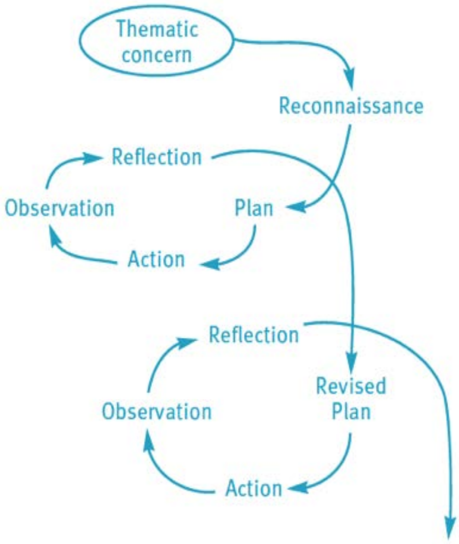


*Figure 6. Action-research spiral* (Kemmis & McTaggart 1982)

### *Frequent cycles of reflection and adaptation*

Secondly, many authors conceptualise the process of a reflexive evaluation as a continuous iterative process of reflection and adaptation wherein research and practice are aligned together (Broerse et al. 2013; Lehtonen 2014; Mark et al. 1998; Van Mierlo et al. 2010; Patton 1984; Guba & Lincoln 1989; Preskill & Torres 2000). The objectives of the evaluation are constantly debated and fine-tuned according to the intermediate outcomes of the evaluation. This continuous cycle of reflection and adaptation requires constant interaction between the evaluators and participants of the reflexive evaluation (Edelenbos & Van Buuren 2005; Gamble 2008; Preskill and Torres 2000), which encompasses frequent linking back of findings of the evaluation (Edelenbos & Van Buuren 2005) to respond to the information need of the participants (Stake 1976) and active monitoring in practice to give input into the evaluation (Lehtonen 2014). In this way the intended end-users of the evaluation are involved in the entire evaluation process (Patton 1984) and the results can directly be linked to practice to learn about their application, thereby sharpening the evaluation outcomes.

The action research spiral (Figure 6) (Kemmis & McTaggart 1982) helps to describe how learning is integrated in all elements of the evaluation process. It starts with learning from the variety of perspectives the different participants may have to come to a common understanding of the diversity of the evaluation subject (Guba & Lincoln 1989). Furthermore, during the cycle from action to reflection continuous learning is stimulated in practice to guide and refine intervention strategies and to link the results to the wider context of application (Regeer et al. 2009). Practically, realising interactive reflection requires sufficient time and space planned for the evaluators and the participants to meet and for learning to occur (Edelenbos & Van Buuren 2005; Preskill and Torres 2000). Most importantly, there must be sufficient time for the participants to partake in a process of interactive interpretation of (intermediate) findings of the sub-evaluations (Kuindersma et al. 2006; Van Mierlo et al. 2010; Regeer et al. 2009). Such moments of interactive reflection provide new input for the proceedings of the reflexive evaluation.

As a consequence of the more elaborate approach of a reflexive evaluation and building in such learning cycles, there is a risk that the process of improving policy may be slowed down. For the evaluation to remain effective and to maintain its momentum, continuous attention is required for its design; time and timing of sub-evaluations and moments of interaction need to be adequately managed (Teisman et al. 2002).

### *Sub-evaluations and their findings are linked*

Thirdly, to further support the integration of the evaluation research within practice, it is also crucial to ensure that the sub-evaluations within the evaluation arrangement are adequately linked. This additionally ensures that the single evaluations strengthen one another's processes and findings (Van der Meer & Edelenbos 2006). For instance, sub-evaluations that take place at different operational levels may incorporate one another's findings, which simultaneously prevents double work and increases the efficiency of the evaluation. Also, representatives of different operational levels may be included in the different sub-evaluation teams to support both adequate linkage between sub-evaluations and overall integration of the evaluation within policy practice (Teisman et al. 2002). Furthermore, accurate timing of the (sub-)evaluation(s) is of significance for the evaluation to have an effect on policy practice (Teisman et al. 2002). For this, insight is required in the administrative and political processes at all levels (Boonstra & Kuindersma 2008). Structuring the reflexive evaluation as an evaluation arrangement implies that the project team is involved in all levels of the evaluation arrangement; it requires a helicopter view to ensure all tracks are running smoothly and are well aligned. Regarding the high number of stakeholders and the multiple sub-projects within the arrangement, this means that the design must be well thought out and that there is high quality communication across the sub-projects.

**Key message: Purpose driven process: evaluation research and policy practice are aligned**

1. **The stakeholders negotiate the evaluation objectives and boundaries**
   by which the evaluation research accommodates the needs of all parties

2. **There are frequent cycles of reflection and adaptation**
   by which evaluation research and policy practice are continuously realigned; it requires sufficient time and space for frequent interaction among the participants and evaluators

3. **The sub-evaluations and their findings are linked**
   to ensure the evaluation arrangement functions as a whole; adequate design and high quality communication across all levels are prerequisites

### *4.4.2 Stakeholder engagement - learning capacity*

The involved stakeholders will have a more diverse role during a reflexive evaluation compared to classical evaluation approaches. While during more classical evaluation approaches, stakeholders are often primarily passive informers who supply data, during a reflexive evaluation their role is overall more active (Kuindersma et al. 2006).  The level of their involvement may differ per sub-evaluation, but also per evaluation phase, depending on what is deemed appropriate at that point in time. For instance, stakeholders may be involved in determining the evaluation research questions, the appropriate methods for research, collecting data and the analysis thereof. Their role may differ from passive, to consulting, to actively participating in the decision-making process surrounding the evaluation.

The process of the reflexive evaluation asks from the participants a divergence from their usual approach, as they are required to acknowledge and understand each other's point of view and to subsequently collaborate, and to be willing to learn from the evaluation findings. The extent to which the reflexive evaluation is successful in stimulating learning processes in the participants, is influenced by an individual's and their organisation's learning capacity (Edelenbos & Van Buuren 2005). Learning occurs on an individual level, which subsequently may promote organisational learning (Argyris & Schön 1978). In the literature we found several aspects that together influence an individual's learning capacity, which will be discussed next. The evaluation project team may encourage the participant's learning capacity through several ways in order to improve the evaluation process and its outcomes, which are also discussed.

#### *Dealing with diversity*

To start, a reflexive evaluation aims to bring together stakeholders with different – sometimes opposing – perspectives and values, from different institutional backgrounds and from different operational levels. The reflexive evaluation acknowledges this pluralism of values and it is therefore necessary that participants are at least open and flexible towards alternative constructions (Guba & Lincoln 1989; Preskill & Torres 2000; Van Mierlo et al. 2010). Schön and Rein add that *'the participants must be able to put themselves in the shoes of other actors* […] *and they must reflect on their own action frames.'* (1994, p. 187). A frame is a structure that provides *'a perspective from which […] a situation can be made sense of and acted on'* (Rein and Schön 1993, p. 146) and may be a cause of conflict if frames of involved stakeholders are opposing. Though consensus is not an objective (which would be naïve), recognition of each other's underlying values regarding the policy programme and the evaluation will allow for more effective collaboration as it allows for better understanding of each other's points of view (Friedman, Rothman & Withers 2006), or underlying frames (Schön and Rein 1994). We do not wish to suggest this is an easy process; the stakeholders may differ in their problem definitions as well as the appropriate solutions and may be highly persistent to other notions (Sabatier, 1988), which as a result stands in the way of learning processes

(Van Buuren et al. 2004). In the literature, however, we find several approaches the project team may take on, that aspire an open dialogue among the stakeholders to overcome their conflicts.

For instance, Van Mierlo et al. (2010) propose the system analysis, which is a tool to obtain insight into a project's barriers and its driving forces. Through collectively analysing barriers and opportunities of a programme with the participants, they exchange their perspectives and views on the project, as well as on its barriers and opportunities. It encourages participants to see matters from each other's point of view and to find common ground, which supports successful collaboration and finding joint solutions (Van Mierlo et al. 2010). Similarly, an interactive frame analysis may support a reflective process that stimulates openness to other frames and frame-reflective discourse, during which participants explore conflicts as well as potential resolutions (Metze & Van Zuydam 2013; Rein & Schön 1993).

### *Mutual trust*

Additional to gaining insight in other perspectives, the evaluation arrangement inspires active collaboration between the participants themselves, and with the project team. For fruitful collaboration and learning a level of mutual trust is required. A lack of trust may be a reason for participants not to share their knowledge and insights (Regeer et al. 2016), which would decrease the quality of the evaluation. Trust here is defined as *refraining from opportunistic behaviour*, following Van Buuren et al. (2004). Asymmetrical power structures are important causes for distrust (Lehtonen 2014) and therefore deserve consideration in evaluation arrangements where multiple operational levels are involved. Though such power dimensions are usually present, it is hard to predict where and when these may start to play a role and affect the course of the evaluation (Gamble 2008). Trust promotes the flow of information, which is crucial for initiating mutual learning processes. For trust to be established, '*intensive and enduring interactions and therefore time'* is required (Van Buuren et al. 2004, p. 17). The project team may encourage mutual trust by the approaches discussed previously; the system and frame interactive analysis (Van Mierlo et al. 2010; Rein & Schön 1993). Additionally, to specifically address issues such as distrust, the project team may adopt Benjamin and Greene's (2009) network characterisation during which aspects such as power imbalances are collectively made explicit. By explicating the imbalances and surfacing the tension, distrust may be openly discussed and potentially resolved before it becomes a destructive obstacle (Gamble 2008).

### *Willingness to learn*

Next to being open to other perspectives, and mutual trust, the participants are asked to learn from the (intermediate) evaluation findings and adapt their policy practice accordingly. It is paramount that they are open and willing to learn, which asks a level of flexibility and capability of adapting to an evolving situation (Patton 1984; Huebner and Betts 1999; Preskill and Torres 2000; Edelenbos & Van Buuren 2005; Van Mierlo et al. 2010; Lehtonen 2014). It is likely that when starting out with a reflexive evaluation, not all participants are equally willing to learn and adapt their approaches, which may compromise the impact of the evaluation findings. Authors suggest that the participants' willingness to learn increases when they regard the evaluation as beneficiary to their personal or institutional cause (assuming they also view the evaluation as trustworthy, discussed previously) (Edelenbos & Van Buuren 2005; Patton 2000). The project team may support this by ensuring that the findings are considered useful through aligning the evaluation objectives to the needs and interests of the participants, and thus by making sure the evaluation is sufficiently utilisation-focused (Patton 1984).

Trust once more plays a crucial role here; the extent to which the participants trust the evaluation findings – the extent to which they find the findings sufficiently independent and authoritative – affects whether the findings will inform their decisions: whether they are willing to learn from it (Van Buuren et al. 2004; Teisman et al. 2002). To maintain their perceived independent and authoritative status, the evaluators need to keep sufficient distance from the participants, which is difficult during a reflexive evaluation that requires high levels of interaction between the evaluators and participants. Edelenbos and Van Buuren (2005) suggest that the researchers divide the different roles among the project team; some team members more involved and in interaction with the participants, and some members more in the background conducting external assessments. Also, findings may be perceived as independent and authoritative if stakeholders with different frames recognise their own perspectives in them. Van Buuren et al. (2004) suggest that in order for this to be true, the actors should be involved in the research process; in determining the research questions, the methods and jointly interpreting the findings.

Finally, scholars have argued that the willingness of participants to learn may be furthermore be supported by an initial sense of urgency and need for change among the participants (Patton 1984; 2000; Edelenbos & Van Buuren 2005; Gamble 2008). This helps to energise the policy programme and to focus the efforts of individual stakeholders. Without a sense of urgency, the evaluation does not have sufficient authority to support and improve policy practice and participants of an evaluation will be less inclined to be actively involved and similarly will unlikely be open to other perspectives, learning and adapting (Van Mierlo et al. 2010). However, a shared sense of urgency among all stakeholders is not necessarily present. Van Mierlo et al. (2010) propose an actor analysis and causal analysis during the start-up phase of an evaluation as it offers tools to ensure participants develop a shared understanding of the issue at hand and an approach to solving it. An actor analysis provides insight in what roles actors play within the programme and the subsequent causal analysis helps to clarify factors that sabotage the programme in any way, in a reflective manner. An insufficient feeling of urgency or lack of involvement among participants may be addressed by this approach.

### *Dynamic representatives of stakeholder-groups*

Lastly, there is another aspect that influences the learning capacity of the participant parties. When these organisations are large and comprise high numbers of employees, it is not unlikely that during each moment of interaction between participants and evaluators, different individuals are present. This has effects on the overall learning capacity of the organisation; if a different individual attends each time, the process is recurrently newly initiated. However, it may also be argued that if it is always the same individual, he or she is individually responsible for promoting organisational learning. Finding a balance between 'new' and 'old' participants per organisation may be more efficient in stimulating organisational learning.

**Key message: stakeholder engagement - learning capacity**

1. **Stakeholders fulfil different roles during a reflexive evaluation**
   ranging from passive to consulting to active, depending on the sub-evaluation and evaluation phase

2. **The effect of the reflexive evaluation is partly determined by the stakeholders learning capacity, which is influenced by the following conditions:**

   i. **Stakeholders are open towards different – sometimes opposing – perspectives and values**
   which may further be encouraged by the evaluator through organising a system analysis or interactive frame analysis

   ii. **There is mutual trust between the stakeholders (commissioners as well as participants), and also between stakeholders and evaluators**
   which is a prerequisite for learning to occur and may be encouraged by similar approaches: system and interactive frame analysis, and a collective network characterisation

   iii. **The stakeholders are willing to learn and adapt their practice accordingly**
   for the evaluation to have effect; usability but also perceived trustworthiness of evaluation findings are important factors

   iv. **Representatives of organisations may be different individuals during each moment of interaction, which affects the overall learning capacity of their organisation**
   therefore it is advised to balance involvement of participants from an organisation that are 'new' and that have been involved in the evaluation process before

### *4.4.3 Mixed methods are employed*

To come to optimally usable answers in the context of the complex and emergent nature of a reflexive evaluation, no set of research methods exists that can be decided on *a priori*. Employing a diversity of research methods increases perceived validity by the participants, as well as the usability of the findings (Teisman et al. 2002). For instance, next to measuring and monitoring indicators through quantitative research, in-depth interviews and focus-group discussions may be held to gain understanding in the context of the measurement. Some actors regard quantitative data as more insightful to learn from, while others believe qualitative findings are most informative (Teisman et al. 2002). By mixing both methods and combining the findings, both 'type' of participants may obtain novel insights through communal sense-making. Important is that the different evaluation methods used are adequately linked so that these may strengthen one another's findings, through triangulation (Teisman et al. 2002). Little is said in the selected literature about how to adequately link findings from qualitative and quantitative research methods; additional literature on mixed methods needs to be consulted  (e.g. Creswell 2009). If done well, a reflexive evaluation may yield high quality data that provides in-depth insight in how the programme functions in its context, and in the causal pathways that determine the outcomes (Edelenbos & Van Buuren 2005; Gamble 2008; Kazi & Spurling 2002; Mark et al. 1998; Patton 1984).

### *Data volume*

Due to the size of the programme and the evaluation arrangement, and the necessity to have in-depth information on the programme and its functions by employing mixed methods, it is likely that the evaluation results in large volumes of data of which the participants are required to make sense of (Gamble, 2008). The vast amount of information may be overwhelming, and when not adequately addressed, may cause important aspects to be overlooked. By ensuring data is regularly and timely interpreted, the data bulk remains digestible (Gamble 2008). Using visual aids such as graphs, theories of change (Hoogerwerf 1984; Weiss 1997), or mind maps (Guijt 2010) may be instrumental for making sense of the massive bulk of information.

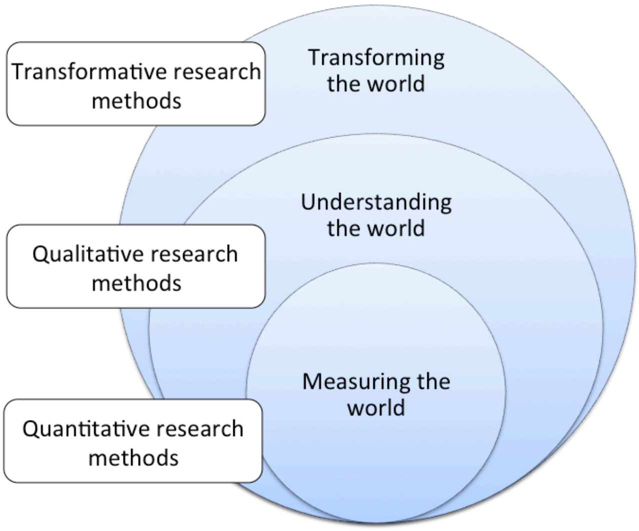

Transformative research methods
Transforming the world
Qualitative research methods
Understanding the world
Quantitative research methods
Measuring the world

*Figure 7. Relation between three types of research methods, based on the work of Cummings & Regeer et al. 2013*

### *Transformative research methods*

In addition to quantitative and qualitative research methods, transformative research methods that are participative in nature are used in reflexive evaluations to stimulate (social) change. Through interactive processes of deliberation and dialogue, new understandings are developed and aligned with local contexts and institutional contexts of participants (Regeer et al. 2009). Participants are involved in giving meaning to the results to ensure usable contextually sound outcomes (Preskill and Torres 2000). Knowledge is thus co-produced and integrated in a process involving academic and non-academic partners, also referred to as transdisciplinary research (Thompson Klein et al. 2001; Regeer and Bunders 2007). In this way, research methods can be applied to understand and at the same time facilitate social change (Mertens, 2003), building on the early work of Guba and Lincoln on ‘naturalistic inquiry’ (1982, p. 133) and ‘constructivism’ (1994, p. 105). A transformative research approach can be applied in two ways. First, it implies employing the usual methods (e.g. interviews, focus group discussions) differently, building in elements for learning and change. Through adopting regular methods in a transformative manner they can contribute to the alignment of evaluation and policy practice, resulting in clear starting points to initiate change. And second, a transformative research approach entails employing methods specifically designed to encourage learning, or to design such new methodological tools if required (examples are Dynamic Learning Agenda, System Analysis, Frame Analysis etc., see Van Mierlo et al. 2010; de Wildt-Liesveld et al. 2015). The relation between these three research methods – quantitative, qualitative and transformative – is shown in Figure 7.

**Key message: mixed methods**

1. **Qualitative and quantitative methods are used**
   i. the different evaluation methods assess impact and encourage learning processes, and help

to engage different types of participants

ii. by linking the different methods, individual sub-evaluations strengthen one another

2. **Due to multiple methods in the diversity of sub-projects of the evaluation arrangement a massive data bulk should be anticipated on**

   the project team may ensure the participants interpret the data regularly and timely to ensure no crucial data is overlooked

3. **Transformative methods are used to encouraged (social) change**

   i. the project team may employ the regular (quantitative and qualitative) methods differently by building in elements that support learning and change

   **ii.** additionally, methods specifically designed for learning may be used

## 4.5 The interdisciplinary evaluation team

Large-scale evaluation projects that require interventions at multiple levels are generally driven by a project evaluation team, which carries the responsibility for realising the evaluation objectives. Similar to the demands a reflexive evaluation places on the participants, the evaluation team too is required to have a different outlook on the evaluation. During a reflexive evaluation, the role of the evaluators differs greatly from classical evaluations (Patton 1984; Preskill & Torres 2000; Regeer et al. 2009; Stake 1976; Van der Meer & Edelenbos 2006). Patton (2000, p. 425) notes: '*The evaluator facilitates judgment and decision making by intended users rather than acting as a distant, independent judge.'* During a reflexive evaluation the evaluators are actively involved in the entire process of the evaluation: the team members are assessors, but additionally may monitor the activities, negotiate different perspectives and advise the participants during decision-making (Huebner and Betts 1999; Preskill and Torres 2000). Most of all, the evaluators are facilitators who ensure meaningful discourse among all those involved in the evaluation and encourage learning to occur through collaborative meaning giving and through encouraging reflective thinking (Gamble 2008; Patton 1994; Stake 1976). The evaluators affect all steps in the reflexive evaluation, and must be responsive to the needs of the participants to ensure optimal alignment of the evaluation and policy practice (Regeer et al. 2009; Van Mierlo et al. 2010).

The evaluators are thus at the heart of the programme and are change managers who share responsibility in ensuring a desired outcome (Stake 1991; Patton 1994; Huebner and Betts 1999; Preskill and Torres 2000; Edelenbos and Van Buuren 2005; Gamble 2008; Regeer et al. 2009; Van Mierlo et al. 2010). This new role causes tensions, as it requires from the evaluation team to work together much more intimately with the participants than during classical evaluations. As a result, their independence, authority and credibility as perceived by participants and the outside world may be at stake. Simultaneously, becoming too distant may be a reason for distrust and may reduce the evaluators' legitimacy according to the participants, which in turn will reduce the impact of the evaluation (Edelenbos & Van Buuren 2005; Teisman et al. 2002). The evaluation team is therefore faced with the challenge of balancing their level of involvement to ensure maximum impact of the evaluation findings (Van Mierlo et al. 2010). Edelenbos & Van Buuren (2005) suggest that in order to deal with this challenge, the different roles the evaluation team is required to fulfil (e.g. assessor, mediator, facilitator) should be distributed over different team members. This way, members that function as facilitators can build trust and rapport with the participants, while other members may focus their efforts more on assessing policy strategies. By dividing these roles, legitimacy and credibility of the evaluation towards the outside world, as well as towards the participants, is increased. To further balance the level of involvement during the evaluation, to maintain independence and credibility, while simultaneously working together intensively with the participants, the team members may be regarded as boundary people. Boundary people mediate between different worlds, which they thereby aim to align (Regeer et al. 2016). They thus answer to the different demands of participants, without being biased by any specific group. Regeer et al. (2009) and Van Mierlo et al. (2010) suggest that in order to support those conducting novel approaches to evaluation, dedicated members of the team may act as monitors, by tracking a Dynamic Learning Agenda at the level of the evaluation team too. Challenges encountered by the evaluation team are put on the agenda, reflected upon and possible solutions are devised and tried out (see also section 4.4.1).

In literature, it is suggested that an evaluation team will be most effective when it is interdisciplinary, and knowledge from different fields is brought to the table. Through collaboration, knowledge co-creation occurs to enrich the process of the reflexive evaluation. Mutual trust, equal power for all team members and high quality communication are prerequisites for the evaluators to effectively work together in such a complex process (Regeer & Bunders 2009; Edelenbos & Van Buuren 2005). Additionally, for such an interdisciplinary collaboration to be successful, the differences between epistemic cultures of the evaluation team members need to be bridged (Regeer & Bunders 2003). Intensive reflective interaction between the team members can support exploring the potential conflicts caused by different epistemic cultures, as well as their potential solutions. By reflective interaction, boundary concepts may emerge that further encourage project team members to effectively collaborate (Metze & Van Zuydam 2013). Boundary concepts are multi-interpretable and hence allow for new interpretations and, subsequently, practice. Such concepts thus facilitate a '*collaborative construction for an alternative direction'* (Metze & Van Zuydam 2013, p. 5). By acknowledging the different epistemic cultures and allowing for reflective interaction, the interdisciplinary evaluation team will likely be able to effectively guide the reflexive evaluation and contribute to obtaining the programme's objectives.

Another difficulty the evaluation team may encounter is caused by the diverse demands from all involved stakeholders. Being at the heart of the programme and its evaluation, implies the evaluation team communicates with all parties and has to – to some extent – satisfy all actors to ensure their commitment. For instance, conflicts on evaluation scope or content may arise which the evaluation team may be required to solve. The team needs to deal with those demands without compromising the quality of the evaluation. Though many authors acknowledge this negotiating or mediating role of the evaluators (Edelenbos & Van Buuren 2005; Preskill & Torres 2000; Patton 2000), clear guidelines on balancing the diverse demands of participants while ensuring their commitment and investment are not available. Additionally, it is likely that political administrative demands cause the commissioner to press more for establishing the programme's impact (or failure thereof) (Teisman et al., 2002), while the participants may express a larger interest in learning for improving their practice (Patton 2000). As the latter likely requires more time, this may be a cause for friction between the two evaluation functions.

Also, top-down decisions made in a political administrative context may strongly affect the evaluation research and its boundaries, over which the project evaluation team has little control. The evaluation team may deal with this by ensuring a good relation and high quality communication with the commissioners of the evaluation to mobilise political administrative support. For instance, by having frequent interaction with a representative of the commissioner party that is more intimately involved in the evaluation process, this representative may promote the evaluation research and its process requirements on sufficiently high levels (Boonstra & Kuindersma 2008). Similarly, the team may be responsive to the potentially sensitive political administrative context by engaging the commissioners in the process of the evaluation. However, there is a risk of becoming too involved in the political system by which remaining unbiased becomes more difficult; for this, the project evaluation team must remain vigilant. Finally, it is important that there is transparency with respect to decisions that are influenced by the administrative context – transparency between the project evaluation team and the commissioners, but also towards the outside world and all the involved stakeholders (Boonstra & Kuindersma 2008). It is this transparency and open communication that makes the reflexive evaluation robust in the diversity of contexts brought along by the involved stakeholders and that will benefit the outcomes to be socially relevant, scientifically sound and useful in practice.

**Key message: the interdisciplinary evaluation team**

1. **The interdisciplinary evaluation team has to fulfil different roles**
   evaluators have an active role and function as e.g. assessors, facilitators, mediators and negotiators

2. **The team has to balance its level of involvement with the participants**
   in order to maintain independence and credibility; it is recommended to subdivide the different roles over team members

3. **Within the interdisciplinary team, epistemic cultures need to be bridged**
   aspects such as mutual trust, equal power and high quality communication are prerequisites, and intensive reflective interaction may support bridging epistemic cultures (e.g. by the emergence of boundary concepts)

4. **The evaluation team has to deal with (conflicting) demands from all involved stakeholders**
   without compromising the quality of the reflexive evaluation arrangement – clear guidelines on how to ensure this are not available

5. **The political administrative context may influence the evaluation**
   the project team may deal with this by realising a good relation with the commissioners where communication is transparent for all stakeholders - nevertheless, the project team must remain vigilant the pressure of the demands of the commissioners do not bias the evaluation scope

# 5. Quality control of a reflexive evaluation

In the literature review presented in chapter 4, we have selected and synthesised what was found regarding the five themes in the included publications, but moreover, we have identified a number of challenges that surface regularly in the literature for which concrete action perspectives remain unclear. These challenges may be a result of current lack of knowledge, but we would argue that they are also inherently part of any reflexive evaluation in complex (policy) settings (see Regeer and Bunders 2007). How to translate identified issues into concrete action perspectives for evaluators will, to a certain extent, always be context- and investigator-dependent. This implies that in any initiative to conduct a learning-focused evaluation, tailor-made strategies to deal with these emerging challenges will need to be developed. We suggest that tracking the challenges encountered – as well as the strategies devised in response to these challenges – should be part of conducting a reflexive evaluation. Indeed, learning does not only take place on the part of the involved stakeholders regarding the programme or issues under evaluation, but also on the part of the involved evaluators regarding the evaluation approach itself.

The Athena Institute has been asked by PBL to reflect on the quality of the reflexive evaluation of the Natuurpact. Quality control by assessing attainment of process and outcome criteria of the evaluation approach itself must be the ultimate paradox of a reflexive evaluation, which starts from the position that outcome criteria can not *a priori* be determined, are dynamic in nature, and need to be formulated in a process of deliberation together with the involved end-users of the results of the evaluation. However, the fact that the process of a reflexive evaluation is dynamic, iterative and emergent, does not imply that it can be done in a sloppy and arbitrary way. On the contrary – a learning-focused evaluation approach evaluation needs to be methodologically sound for the outcomes of the evaluation to have sufficient validity and legitimacy and needs to be designed and facilitated in a systematic, meticulous and transparent way. Therefore, the process criteria, as developed in the framework, are continuously monitored by the designated project team members (notably the Athena Institute). The Athena institute will publish a report based on these monitoring activities in December 2016, in which a reflection is provided on the evaluation team's approach to implementing the criteria, leading to an overview of the added value of the reflexive evaluation for the stakeholders involved and suggestions are provided for improvement in the process of conducting a learning focused evaluation based on the Dynamic Learning Agenda of the evaluation team.

## 5.1 Monitoring the intended outcome of the Natuurpact evaluation

Several deliverables are explicated in the evaluation plan. These consist of intermediate outcomes, reports of different deliberative sessions and a final report of the evaluation 2016, with underlying scientific reporting. The intermediate reports will be used in the deliberative process with actors involved to co-create knowledge with the aim of shared sense-making and joint fact-finding to co-create new knowledge. In the review of the Natuurpact evaluation we will assess to what extent the input from this deliberative process is visible in the development of a final evaluation report. Herein we will focus on the extent to which the outcomes are based on joint-fact finding and co-creation of knowledge. We will especially focus on the multi-directional accountability herein; in other words questioning whether the evaluation report is directed to its commissioners as well as its intended users. This reflection will be done through a document analysis of the intermediate outcomes, reports of deliberative sessions and the final report, including underlying documents. Furthermore, we will interview a sample of the project team, researchers, commissioners and end-users of the evaluation for their view on the outcomes. We will reflect on the potential for improved policy practice and increased likelihood of goal-attainment, as the ultimate outcomes of the evaluation.

## 5.2 Monitoring the stakeholder involvement in the Natuurpact evaluation

Several moments of stakeholder involvement are planned in the evaluation plan. Furthermore, the plan is emergent in its design, leaving room for additional stakeholder interaction when needed. The stakeholder involvement in the Natuurpact evaluation will be evaluated on the process of selecting the relevant stakeholders for each interaction in a dynamic manner. We will review how decisions are made for each moment of interaction on who to involve when. Hereby we want to find out to what extent the intended users have an active role in shaping the evaluation process and how the political administrative context of the evaluation influenced this process. This will be done through monitoring decision making on the involvement of stakeholders in the project team and in sub-projects of the evaluation. Furthermore, we will analyse the actual invitation lists and people that attend, with the aim of looking for discrepancies in attendance for different groups to analyse the actual representation of the stakeholders involved. Herein, we make a distinction between the following actors: project team members, researchers, policy-makers provinces, other provincial staff, representatives of societal organisations, relevant corporations, and commissioners.

## 5.3 Monitoring the process of the learning-focused evaluation of the Natuurpact

An evaluation arrangement has been designed that includes a combination of approaches that together are expected to yield answers to the research questions described in the evaluation plan, which are based on the input of the participants during the evaluation framework development. However, the needs of the participants are likely to develop over time. Therefore, the reflexive evaluation has a flexible design that allows for adaptations according to the needs of the participants, but also to unexpected developments in the political administrative context. Hereby it is expected that the evaluation integrates policy and research and thereby becomes purpose driven. This will be reviewed by monitoring how the research and policy practice are continuously realigned in different sub-research projects and how these sub-projects are in their turn interlinked to ensure that the evaluation arrangement functions as a whole. Furthermore, we will interview a sample of the project team, researchers, commissioners and end-users of the evaluation for their view on the purpose driven aspects of the evaluation.

In two collaborative learning sessions all stakeholders come together to give meaning to intermediate results. Furthermore, several reflection workshops are planned in the context of sub-research projects with relevant stakeholders. In these sessions we will monitor stakeholder engagement and evaluate their opinion on their involvement through evaluations after the sessions. Furthermore, we will interview a sample of the project team, researchers, commissioners and end-users of the evaluation for their view on trust issues and willingness to learn during the sessions.

All the sub-research projects are intertwined and often address multiple research questions. The reflexive evaluation adopts smaller and larger action-learning loops in its arrangement to enable a continuous iterative process of reflection and adaptation by both research and practice. The evaluation as a whole is expected to give insight in the three-year cycle on the status of goal attainment (accountability) and reflection on policy adaptation (learning) with relevant key stakeholders. In the review we will reflect on the extent to which the evaluation contains static and dynamic components, adhering to upwards, horizontal and internal accountability. This will be evaluated through analysing the design of different sub-projects and reflecting on upwards and downwards accountability with the researchers in the project team.

Throughout the entire learning focused evaluation process, comprising multiple sub-research projects in an evaluation arrangement, a mixed methods approach is employed. It means that multiple data collection methods are used to gain insight into a research question, such as desk studies, interviews, group discussions and modelling. Data collected through this variety of research methods is subsequently juxtaposted, to verify and enrich findings. In the reflexive evaluation of Natuurpact, multiple researchers are involved in the collection and interpretation of data throughout the process. We will in relation to the mixed method design reflect on whether: the design of interview and focus group scripts is done collaboratively and well in advance to allow time for adjustments to different needs and to increase focus on the desired outputs; data analysis is done by at least two researchers and collaboratively discussed by the project team; discussions in the project team also involve self-reflecting on potential biases and predispositions influencing the studies; and this self-awareness of the researchers is also fostered by brainstorming and exchanging ideas with people who are less involved in the research process. This will be reviewed during the process of the evaluation in the meetings with several research projects and in retrospect through interviewing a sample of the researchers involved.

The Natuurpact evaluation does not take place in isolation but in a political administrative context, partly represented by its commissioners and the steering board, but also by the political influences at play at provincial level, and expectations that exist in parliament and the provincial councils. In a reflexive evaluation, the project team members are expected to be sensitive to the political administrative context of the evaluation, and responsive by engaging the commissioners during the learning sessions and by being open and transparent in the evaluation process regarding mutual expectation. Close involvement of the commissioner in a reflexive evaluation is advised to ensure upwards accountability and create support for horizontal and internal accountability, however it also implies a challenge with regard to the political power the commissioner has with regard to the Natuurpact evaluation. The tensions that arise in this process will be monitored and reflected upon in the review. Since the selected literature does not give clear guidance on how to deal with issues arising from the political administrative context we hope to develop insights through our analysis in how this aspect influences the learning-focused evaluation and what strategies are developed in response.

## 5.4 Monitoring the interdisciplinary team guiding the learning-focused evaluation of the Natuurpact

Whilst being engaged in organising a learning process among a diversity of stakeholders involved in the shaping and execution of nature policy in The Netherlands as *facilitators*, the evaluation team of a reflexive evaluation at the same time acts as a team of *researchers.* Naturally, the researchers involved need to comply with the norms and rules for quality control of their respective epistemic cultures to ensure methodological soundness. This counts for the involved policy scientists, ecologists, cost-effectiveness researchers, and transformative researchers. For all these researchers the data collection methods as well as interpretation of the data need to comply with the norms and rules of methodological soundness. In our reflection we will review to what extent all involved scientists could comply with the norms and rules from their epistemic culture and where consensus needed to be found, and comfort zones were left. Multiple moment for reflection are build in at project team meetings and special team building days that give input for this reflection. Herein, we will not focus on how the team adheres to their epistemic norms and rules, but more on their learning processes on dealing with them and understanding the norms of others. We will specifically focus on the growth process the team experiences in bridging epistemic cultures for effective collaboration. Thus, the review will focus on the growth process in the interdisciplinary collaboration throughout the evaluation by analysing the discussions in the project meetings and team building activities.

In the evaluation, the interdisciplinary team will interact with the commissioners and end-users of the evaluation. This all takes place in the realm of a political administrative context. The review will focus on how the project team balances its involvement with the participants and deals with conflicting demands herein. We will review how the interdisciplinary team maintains its independence and credibility by subdividing different roles over the team members and how they deal with the demands of stakeholders without compromising the quality. Herein, building rapport and transparency are important aspects. To assess this, multiple moments of reflection are used in the project team meetings and a learning history will be made by the researchers, zooming in on the most significant moments in the evaluation process.

## 5.4 Extra attention for tough issues in the learning-focused evaluation

To conclude, in a reflexive evaluation, an iterative and reflexive approach is applied to develop an evaluation framework and (preliminary) outcomes, which are refined during the course of the research through continuous interaction with a diversity of stakeholders. Rather than purely theory-testing or theory-building, a reflexive evaluation can be characterised as continuously iterating between theory-testing and theory-building, in a process of mutual exploration and reflection.

In the literature review on learning-focused evaluations it became clear that some aspect of it are not yet fully developed and remain intangible in the selected literature. Through the application in the learning-focused evaluation of the Natuurpact, we aim to develop more insight in these issues by following them up with the Dynamic Learning Agenda (DLA). The DLA tool supports learning processes by explicating tough issues together with the stakeholders involved. We would like to adopt this tool in the project team meetings to ensure regular reflections on tough issues that may otherwise remain hidden and hamper the process of the learning focused evaluation. The DLA is flexible to the needs of the participants, in the sense that questions maybe solved and/or not longer relevant and therefore dropped while at the same time questions can be added.

Out of our literature search and the operationalization of the evaluation plan we came to the following learning agenda that we would like to follow up in the learning-focused evaluation of the Natuurpact and will reflect on in our review in 2016:

- What strategies are employed to deal with possible intransigence in the political administrative context of the Natuurpact evaluation?

- What challenges emerge as a result of the multi-project *evaluation arrangement* and how are these challenges resolved?
- To what extent does the evaluation include knowledge generation on: a) the current situation, b) the aspired situation, and c) the transformative process (or system knowledge, target knowledge, and transformative knowledge)?
- In what way has the policy theory fostered learning by considering its dynamic nature and connecting it to action theory of policy actors?
- What strategies have contributed to the fostering of mutual understanding and bridging of epistemic cultures within and beyond the project team?
- What strategies were used to strengthen the mixing of different research methods and avoid weakening of its parts?
- What strategies were used to unite accountability and learning purposes of the evaluation of Natuurpact?

# 6. Conclusion

The vision on nature of the Dutch government states that 'nature belongs amidst society and not only in protected areas. This is beneficial for the economy and biodiversity' (Rijksoverheid, 2014). By elucidating the relation between nature, society and economy, like the Natuurpact ambitions, the nature vision embraces the complex, multi-facetted and multi-actor character of nature policy in the Netherlands. In this challenging context PBL decided to take a novel evaluation approach that reflects the complexities of contemporary policy practice in which the learning and accountability function of evaluation are united; the reflexive evaluation. With this review we aimed to provide a scientific justification for this novel approach in the context of the Natuurpact programme. Secondly, this document serves as a basis for reviewing the application of the reflexive evaluation in practice to identify points of improvement and to demonstrate the added value in relation to more traditional evaluation approaches. We introduced quality control measures to support an effective process design and implementation thereof in order to achieve optimal evaluation outcomes and lessons for further use of this evaluation method.

Chapter 2 presented the evaluation design based on the evaluation plan of the Natuurpact (2014-2027) and key elements of the reflexive evaluation were related to this design. First, the participative development of the evaluation framework was discussed as a pre-condition for the reflexive evaluation of nature policy in the Netherlands. Second, the evaluation plan was discussed as well as its emergent and flexible character to deal with the complex and multi-stakeholder environment of the reflexive evaluation. Herein, the determined research questions, research process, and intended outcomes were related to the key characteristics of a reflexive evaluation. Furthermore, the evaluation team in charge of the Natuurpact evaluation is discussed in relation to its interdisciplinarity as well as intended activities to bridge epistemic cultures and deal with the demands of the diverse stakeholders and the political administrative context.

After the methodology for the literature in chapter 3, chapter 4 provided an overview of the key elements of new evaluation approaches as described in the scientific literature. Herein, the importance of carefully selecting and engaging different stakeholders in a dynamic context was stressed. In a purpose driven process, wherein stakeholders have an active role, research related to accountability and learning is executed to evaluate the policy programme. The outcomes of the reflexive evaluation are thereby socially robust and can be used by relevant stakeholders for both accountability and learning purposes. To implement a reflexive evaluation, an interdisciplinary research team is required with actively involved members that bridge epistemic cultures and deal with the demands of the diverse stakeholders and the political administrative context.

In Chapter 5 finally multiple perspectives were explored regarding warranting methodological soundness of the approach. First, epistemological quality norms and rules are taken into account in the sub-evaluations. Second, the knowledge production process of a reflexive evaluation, e.g. in learning sessions, warrants a reconsideration of concepts such as validity and reliability of scientific knowledge. And third, specific process and outcome criteria that have been developed for reflexive evaluations are monitored continuously during the evaluation of the Natuurpact. By reflecting on the implementation of the reflexive evaluation of the Natuurpact, insight is provided in how the approach may be improved.

### *Encountered limitations of our literature review*

To increase the feasibility of our literature study, naturally we were required to make a number of demarcations that provided essential boundaries to make our search more effective. As a result, our choices inevitably have also caused other aspects related to reflexive evaluation to receive less emphasis, while they are of significance for the Natuurpact evaluation. To start, our literature search resulted primarily in evaluation approaches that focused on establishing individual learning processes – of the policy professional, the societal partner, etc. – by collective reflection on perspectives and practices. Literature that concerns 'system learning' (signifying permanent system change, where new structures, networks and organisational embedding are realised, Arkesteijn et al. 2015) is relevant for the Natuurpact as, partly due to the decentralisation and the empowerment of new actors, it may be argued that the reflexive evaluation of Natuurpact requires a system perspective. More literature on system learning may further enrich our knowledge and approach. Similarly, policy learning concerns learning processes on multiple levels within the policy field (e.g. Kemp & Weehuizen 2005) and may therefore also be of value to our understanding of the aspired learning processes of a reflexive evaluation.

Additionally, during the writing of this review we encountered a number of elements of reflexive evaluation that to our opinion were not sufficiently explained or discussed in our selected literature. When possible, we have added literature to our selection, but in some cases time limitations have caused these elements to remain insufficiently explored. For instance, several authors argue that during reflexive evaluation, a mixed methods approach is appropriate to ensure high quality data. They add that these different (qualitative and quantitative) research approaches should be adequately linked – but how this linkage is best realised, is rarely touched upon. In our view the Natuurpact evaluation would benefit from a more in-depth study of literature on mixed methods (e.g. Creswell 2009) to obtain clear how-to information to ensure the different methods are indeed adequately linked.

Lastly, there is a field of research that we briefly touched upon when discussing the intended outcomes of reflexive evaluation: the science-policy interface. It is a highly relevant field as it provides insight in how knowledge may be transferred from science to policy, and what boundaries or barriers inhibit effective communication of knowledge, which may effect how the policy field uses this information. Related is uncertainty communication of scientific findings and how this process may effect policy decisions and resulting policy outcomes. A more systemic inventory of research on the science-policy interface may benefit the approach to reflexive evaluation and thereby positively affect how the evaluation findings are used by policy professionals. Future research will add such knowledge to our theoretical framework.

In conclusion, this review made clear what the key characteristics are of a reflexive evaluation that can be operationalised in the Natuurpact evaluation. The embedding in scientific literature enhances the credibility of the methodology. Attention for quality control will augment this credibility in the course of the evaluation by supporting an efficient and effective implementation of the reflexive evaluation. By documenting the emergent process of the evaluation carefully in every step, an optimal implementation can be realised and lessons learned will be documented for further development of this new approach by PBL. The process, outcomes and lessons learned will together show the potential added value of the reflexive evaluation for policy practice and evaluation research in complex and multi-stakeholder contexts.

### *Acknowledgements*

*This review has benefited from the critical scrutiny of many conscientious and constructive reviews. I would like to thank the following persons for their special efforts: Jurian Edelenbos, Bas Arts and Hens Runhaar for their critical review on the draft of "Characteristics of reflexive evaluation - a literature review conducted in the context of the Natuurpact (2014-2027) evaluation". Rob Folkert, Froukje Boonstra, Arianne de Blaeij and Dirk-Jan van der Hoek for their constructive feedback and support as the project team of the Natuurpact (2014-2027) evaluation during the writing process. Hein Haak, Sido Mylius, Peter Janssen, Eva Kunseler and Femke Verwest for their constructive feedback from PBL on different draft versions in the writing process.*

## References

Abma, T. A. (1996). *Responsief Evalueren*. Erasmus University Rotterdam.

Abma, T. A., & Stake, R. E. (2001). Stake's responsive evaluation: Core ideas and evolution. *New Directions for Evaluation*, *2001*(92), 7–22.

Argyris, C., & Schon, D. (1978). Organizational learning: A theory of action perspective. *Reading, Mass.: Addision Wesley*.

Arkesteijn, M., van Mierlo, B., & Leeuwis, C. (2015). The need for reflexive evaluation approaches in development cooperation. *Evaluation*, *21*(1), 99–115.

Boonstra, F. G., & Kuindersma, W. (2008). *Leren van de evaluatie reconstructie zandgebieden. Methode, proces en politiek-bestuurlijke inbedding*. Wageningen.

Broerse, J. E. W., de Cock Buning, T., Roelofsen, A., & Bunders, J. F. G. (2008). Evaluating Interactive Policy Making on Biotechnology: The Case of the Dutch Ministry of Health, Welfare and Sport. *Bulletin of Science, Technology & Society*, *29*(6), 447–463.

Bruijn, de H., & Heuvelhof, ten E. (2008). *Management in Networks - On Multi-actor Decision Making*. Taylor & Amp; Francis Ltd.

Bryson, J. M., Patton, M. Q., & Bowman, R. A. (2011). Working with evaluation stakeholders: A rationale, step-wise approach and toolkit. *Evaluation and Program Planning*, *34*(1), 1–12.

Bucchi, M., & Trench, B. (Eds.). (2008). *Handbook of Public Communication of Science and Technology*. Londen and New York: Routledge Taylor & Francis Group.

Buuren, M., Edelenbos, J., & Klijn, E. (2004). Managing knowledge in policy networks. Organising joint fact-finding in the Scheldt Estuary. *Conference on Democratic Network Governance,* 1–27. Retrieved from http://repub.eur.nl/resource/pub_10165/

Cherryholmes, C. H. (1992). Notes on Pragmatism and Scientific Realism, *21*(6), 13–17.

Cock Buning, de T., Regeer, B. J., & Bunders, J. F. G. (2008). *Biotechnology and Food: Towards a societal agenda in 10 steps*. *Biotechnology and Food Safety*. Den Haag. doi:10.1016/B978-0-409-90260-0.50012-9

Collins, H. M., & Evans, R. (2002). The Third Wave of Science Studies: Studies of Expertise and Experience. *Social Studies of Science*, *32*(2), 235–296. doi:10.1177/0306312702032002003

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (3rd ed.). Thousand Oaks: SAGE Publications Inc.

Cummings, S., Regeer, B. J., Ho, W., & Zweekhorst, M. (2013). Proposing a fifth generation of knowledge management for development : investigating convergence between knowledge management for development and transdisciplinary research. *Knowledge Management for International Development Journal*, *9*(2), 10–36.

Douglas, M., & Wildavsky, A. (1982). Risk and Culture: An Essay on the Selection of Technical and Environmental Dangers. *Journal of American Studies*, *18*(1), 145–146.

Ebrahim, A. (2005). *Accountability Myopia: Losing Sight of Organizational Learning*. *Nonprofit and Voluntary Sector Quarterly* (Vol. 34). doi:10.1177/0899764004269430

Edelenbos, J., & van Buuren, A. (2005). The learning evaluation: a theoretical and empirical exploration. *Evaluation Review*, *29*(6), 591–612. doi:10.1177/0193841X05276126

Eden, C., & Ackerman, F. (1998). *Making Strategy: The Journey of Strategic Management*. California: SAGE Publications Inc.

Ehrmann, J. R., & Stinson, B. L. (1999). Joint fact-finding and the use of technical experts. In *The consensus building handbook* (pp. 375–99). Thousand Oaks, CA: Sage.

Fischer, F. (2006). Participatory Governance as Deliberative Empowerment: The Cultural Politics of Discursive Space. *The American Review of Public Administration*, *36*(1), 19–40. doi:10.1177/0275074005282582

Fischer, F., & Forester, J. (1993). *The Argumentative Turn in Policy Analysis and Planning*. Duke University Press.

Flowers, A. B. (2010). Blazing an evaluation pathway: Lessons learned from applying utilization-focused evaluation to a conservation education program. *Evaluation and Program Planning*, *33*(2), 165–171. doi:10.1016/j.evalprogplan.2009.07.006

Gamble, J. A. A. (2008). *A Developmental Evaluation Primer*. *The J.W. McConnel Family Foundation*. Canada.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth Generation Evaluation* (1st ed.). Newbury Park, CA: SAGE Publications Inc.

Guijt, I. (2010). Accountability and Learning: Exploding the Myth of Incompatibility between Accountability and Learning. In J. Ubels, N.-A. Acquaye-Baddoo, & A. Folwer (Eds.), *Capacity Development in Practice* (pp. 277–291). Earthscan. Retrieved from http://www.snvworld.org/en/publications/capacity-development-in-practice-complete-publication

Hajer, M. (2003). Policy without Polity ? Policy Analysis and the Institutional Void. *Policy Sciences*, *36*, 175–195.

Heron, J., & Reason, P. (1997). A Participatory Inquiry Paradigm. *Qualitative Inquiry*, *3*(3), 274–294. doi:10.1177/107780049700300302

Hoogerwerf. (1984). Het ontwerpen van overheidsbeleid: een handleiding met toelichting. *Bestuurswetenschappen*, *38*, 4–23.

Hoppe, R., & Hisschemöller, M. (1996). No Coping with Intractable Controversies: The Case for Problem Structuring in Policy Design and Analysis. *Knowledge and Policy: The International Journal of Knowledge Transfer and Utilization*, *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.

Kahane, A. (2007). *Solving Tough Problems - An Open Way of Talking, Listening, and Creating New Realities* (2nd ed.). Berrett-Koehler Publishers.

Kazi, M. a F., & Spurling, L. J. (2002). Realist Evaluation for Evidence-Based Practice. *Evaluation*, (October 2000), 1999–2002.

Kemmis, S., & McTaggart, R. (1982). *The Action Research Reader*. Victoria: Deakin University Press.

Kemp, R., & Weehuizen, R. (2005). *Policy Learning: What does it mean and how can we study it? Publin Report No. D15*. Oslo. Retrieved from http://www.sba.oakland.edu/faculty/mathieson/mis524/resources/readings/innovation/innovation_in_the_public_sector.pdf

Kirkhart, K. E. (2000). Reconceptualizing Evaluation Use: An Integrated Theory of Influence. *New Directions for Evaluation*, (88), 5.

Klaassen, P., Kupper, F., Rijnen, M., Vermeulen, S., & Broerse, J. (2014). *RRI Tools: Policy brief on the state of the art on RRI and a working definition of RRI*. Amsterdam. Retrieved from www.rri-tools.eu

Kloprogge, P., Sluijs, J. P. Van Der, & Wardekker, A. (2007). *Uncertainty Communication. Issues and good practice*.

Kuindersma, W., & Boonstra, F. G. (2005). Methoden van beleidsevaluatie onder de loep. Een zoektocht naar nieuwe vormen van beleidsevaluatie voor het Milieu-en Planbureau.

Kuindersma, W., Boonstra, F. G., Boer, S. De, Gerritsen, a L., Pleijte, M., & Selnes, T. a. (2006). Evalueren in interactie. De mogelijkheden van lerende evaluaties voor het Milieu-en Natuurplanbureau.

Lehtonen, M. (2014). Evaluating megaprojects: from the “iron triangle” to network mapping. *Evaluation*, 278–295. doi:10.1177/1356389014539868

Mark, M. M., Henry, G. T., & Julnes, G. (1998). A realist theory of evaluation practice. *New Directions for Evaluation*, *1998*(78), 3–32. doi:10.1002/ev.1098

Mayne, J. (2003). Results-Based Governance: Collaborating for Outcomes. In A. Gray, B. Jenkins, F. Leeuw, & J. Mayne (Eds.), *Collaboration in Public Services:The Challenge for Evaluation* (pp. 105–130). London: Transaction.

Meer, van der F.-B., & Edelenbos, J. (2006). Evaluation in Multi-Actor Policy Processes: Accountability, Learning and Co-operation. *Evaluation*, *12*(2), 201–218. doi:10.1177/1356389006066972

Mertens, D. M. (2003). Mixed Methods and the Politics of Human Research: the Transfromative-Emancipatory Perspective. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social & Behavioral research*. Thousand Oaks: Sage Publications, Inc.

Metze, T. A. P., & Van Zuydam, S. (2013). Pigs in the City: Reflective Deliberations on the Boundary Concept of Agroparks in The Netherlands. *Journal of Environmental Policy & Planning*, (August), 1–18. doi:10.1080/1523908X.2013.819780

Mierlo, van B., Regeer, B. J., van Amstel, M., Arkesteijn, M., Beekman, V., Bunders, J., … Leeuwis, C. (2010). *Reflexive monitoring in action: A guide for monitoring system innovation projects*. *A guide for monitoring system innovation projects*. Wageningen/Amsterdam. Retrieved from http://www.falw.vu.nl/en/Images/Reflexive monitoring in Action B van Mierlo and B Regeer 2010_tcm24-399363.pdf

Nowotny, H. (2000). Re-thinking Science : From Reliable Knowledge to Socially Robust Knowledge. In *Jahrbuch 2000 des Collegium Helveticum* (pp. 1–19). Zurich.

Patton, M. Q. (1984). An alternative evaluation approach for the problem-solving training program: A utilization-focused evaluation process. *Evaluation and Program Planning*, *7*(2), 189–192. doi:10.1016/0149-7189(84)90045-4

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: SAGE Publications.

Patton, M. Q. (1994). Developmental Evaluation. *American Journal of Evaluation*, *15*(3), 311–319. doi:10.1177/109821409401500312

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.

Pawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. In *Evaluation for the*

*21st century: A handbook* (pp. 405–418). Thousand Oaks, CA, US: Sage Publications.

Perrin, B. (2002). Towards a New View of Accountability. In *Symposium on Promoting Organisation Learning via Evaluation: the New Accountability?* (Vol. 2002). Seville.

Pohl, C., & Hirsch Hadorn, G. (2008). Methodological challenges of transdisciplinary research. *Nature Sciences Sociétés*, *16*, 111–121. doi:10.1051/nss

Preskill, H., & Torres, R. (2000). The Learning Dimension of Evaluation Use. *New Directions for Evaluation*, *2000*(88), 25–37. doi:10.1002/ev.1189

Reed, M. S., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., … Stringer, L. C. (2009). Who's in and why? A typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management*, *90*(5), 1933–1949. doi:10.1016/j.jenvman.2009.01.001

Regeer, B. J., & Bunders, J. F. G. (2003). The epistemology of transdisciplinary research: from knowledge integration to communities of practice. *Interdisciplinary Environmental Review*, *5*(2), 98–118.

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., Hoes, A.-C., van Amstel-van Saane, M., Caron-Flinterman, F. F., & Bunders, J. F. G. (2009). Six Guiding Principles for Evaluating Mode-2 Strategies for Sustainable Development. *American Journal of Evaluation*, *30*(4), 515–537. doi:10.1177/1098214009344618

Rein, M., & Schön, D. (1993). Reframing policy discourse. In F. Fischer & J. Forester (Eds.), *The Argumentative Turn in Policy Analysis* (pp. 145–166). Durham, NC: Duke University Press.

Richmond, B. J., Mook, L., & Quarter, J. (2003). Social accounting for nonprofits: Two models. *Nonprofit Management and Leadership*, *13*(4), 308–324.

Rijksoverheid. (2014). *Rijksnatuurvisie 2014: Natuurlijk verder*. Den Haag.

Sabatier, P. A. (1988). An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy Sciences*, *21*(2-4), 129–168. doi:10.1097/01.cmr.0000205019.23612.a1

Sluijs, van der J. P. (2010). Uncertainty and complexity: the need for new ways of interfacing climate science and climate policy. *From Climate Change to Social Change: Perspectives on Science–Policy Interactions. Edited by Driessen P, Leroy P, Van Vierssen W:. Utrecht: International Books*, (September), 31–49. Retrieved from http://cxdd.broceliande.kerbabel.fr/files/cxdd/Modesl_of_Science_and_Policy_essay_0.pdf

Stake, R. E. (1991). Excerpts from: "Program Evaluation, Particularly Responsive Evaluation." *American Journal of Evaluation*, *12*(1), 63–76. doi:10.1177/109821409101200109

Teisman, G., van der Meer, F.-B., Erik-hans, K., Edelenbos, J., Klaassen, H. L., & Reudink, M. A. (2002). *Evalueren om te leren. Naar een evaluatiearrangement voor de vijfde nota RO*. Rotterdam.

Veen, van S. C., de Wildt-liesveld, R., Bunders, J. F. G., & Regeer, B. J. (2014). Supporting reflective practices in social change processes with the dynamic learning agenda : an example of learning about the process towards disability inclusive development Saskia C . van Veen *,. *Int. J. Learning and Change*, *7*(3/4), 211–233.

Veld, in 't R. J. (2000). *Willens en wetens: de rollen van kennis over milieu en natuur in beleidsprocessen*. Den Haag: Raad voor Ruimtelijk, Milieu-en Natuuronderzoek.

Watson, R. T. (2005). Turning science into policy: challenges and experiences from the science-policy interface. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 471–477. doi:10.1098/rstb.2004.1601

Weiss, C. H. (1997). Theory based evaluation: Past, present, and future. *New Directions for Evaluation*, *1997*(76), 41–55.

Wildt-Liesveld, de R., Regeer, B. J., Bunders, J. F. G., de Wildt-Liesveld, R., Bunders, J. F. G., & Regeer, B. J. (2015). Governance strategies to enhance the adaptive capacity of niche experiments. *Environmental Innovation and Societal Transitions*, *16*, 154–172. doi:10.1016/j.eist.2015.04.001

Zapico-Goñi, E. (2007). Matching Public Management, Accountability and Evaluation in Uncertain Contexts: A Practical Suggestion. *Evaluation*, *13*(4), 421–438. doi:10.1177/1356389007082130

## Appendix 1: articles and reports included in the literature review

| | Authors | Year | Title | Theory/evaluation school | Context |
|---|---|---|---|---|---|
| 1 | Abma, T. & Stake, R. | 2001 | Stake's responsive evaluation: core ideas and evolution | Responsive evaluation | Evolution of responsive evaluation |
| 2 | Edelenbos, J. & Van Buuren, A. | 2005 | The Learning Evaluation: A Theoretical and Empirical Exploration | The Learning Evaluation | Case: Stimulation Program on Citizen and Environment. Aimed to increase the role of citizens in the development of environmental policy |
| 3 | Flowers, A.B. | 2010 | Blazing an evaluation pathway: Lessons learned from applying utilisation-focused evaluation to a conservation education program | Utilisation-Focused Evaluation | Case: Evaluation of an educational program of nature conservation |
| 4 | Friedman, V.J., Rothman, J. & Withers, B. | 2006 | The Power of Why: Engaging the Goal Paradox in Program Evaluation | Action Evaluation/Goal setting in multi-actor evaluation | Case: evaluation of education program to bring together different streams of Judaism |
| 5 | Gamble, J.A.A. (primer, grey literature) | 2008 | A Developmental Evaluation Primer | Developmental Evaluation | Project: SSI initiative which examined the capacity for social innovation to address social problems |
| 6 | Guba, E.G. & Lincoln, Y.S. | 1989 | Fourth Generation Evaluation | Fourth Generation Evaluation | |
| 7 | Huebner, A.J. & Betts, S.C. | 1999 | Application to Positive Youth Development | Fourth Generation Evaluation | Project: program that aimed to determine indicators for positive youth development |
| 8 | Kazi, M.A.F. & Spurling, L. | 2002 | Realist Evaluation for Evidence-Based Practice | Realist Evaluation | Project: Integration of single-subject designs in family centres |
| 9 | Kuindersma, W., Boonstra, F.G., de Boer, S., Gerritsen, A.L., Pleijte, M. & Selnes, T.A. | 2006 | Evalueren in interactie | Learning evaluation | Environmental and nature policy |
| 10 | Lethonen, M. | 2014 | Evaluating megaprojects: from | Learning-focused | 'Megaprojects'; large complex |

| | | | | | |
|---|---|---|---|---|---|
| | | | the 'iron triangle' to network mapping | evaluation | industrial and infrastructure projects |
| 11 | Mark, M.M., Henry, G.T. & Julnes, G. | 1998 | A realist theory of evaluation practice | Emergent Realist Evaluation | Evaluation of educational programmes |
| 12 | Meer, van der, F. & Edelebos, J. | 2006 | Evaluation in Multi-Actor Processes: Accountability, Learning and Co-operation | Evaluation Arrangement | Dutch spatial planning |
| 13 | Mierlo, van, B.C., Regeer, B.J., van Amstel, M., Arkesteijn, M.C.M., Beekman, V., Bunders J.F.G., de Cock Buning, T., Elzen, B., Hoes, A.C. & Leeuwis, C.<br>(Guide to RMA; grey literature) | 2010 | Reflexive Monitoring in Action: a guide for monitoring system innovation projects | Reflexive Monitoring in Action | System innovation projects |
| 14 | Patton, M.Q. | 1984 | An alternative approach for the problem-solving training program: an utilisation-focused evaluation process | Utilisation-Focused Evaluation | Evaluation of a problem-solving educational program |
| 15 | Patton, M.Q. | 1994 | Developmental Evaluation | Developmental Evaluation | Project: development of a community leadership program<br>Project: development of a program to support diversity in schools |
| 16 | Pawson, R. & Tilley, N. | 1997 | An introduction to scientific realist evaluation | Realist evaluation | Social programmes embedded in social systems |
| 17 | Preskill, H. & Torres, R.T. | 2000 | The Learning Dimension of Evaluation Use | Transformative learning in organisational contexts | Companies, organisations |
| 18 | Regeer, B.J., Hoes, A.C., van Amster-van Saane, M., Caron-Flinterman, F.F. and Bunders, J.F.G. | 2009 | Six Guiding Principles for Evaluating Mode-2 Strategies for Sustainable Development | Interactive Learning & Action/Evaluation of Mode-2 Strategies | Mode-2 Strategies for sustainable development |
| 19 | Regeer, B.J., De Wildt-Liesveld, R., Van Mierlo, B., Bunders, J.F.G. | 2016 | Exploring ways to reconcile accountability and learning in Triple P niche experiments | Aligning accountability and learning | Triple P projects, niche experiments |

| | | | | | |
|---|---|---|---|---|---|
| 20 | Stake, R.E. | 1976 | A Responsive Evaluation of Two Programmes | Responsive Evaluation | Educational programmes |
| 21 | Stake, R.E. | 1991 | Excerpts from ‘Program Evaluation, Particularly Responsive Evaluation | Responsive Evaluation | Educational programmes |
| 22 | Teisman, G., Van der Meer, F., Klein, E., Klaassen, H. L. & Reudink, M.A. | 2002 | Evalueren om te leren | Evaluation arrangement | Dutch spatial planning |

Athena Instituut
De Boelelaan 1085
1081 HV Amsterdam
T 020 598 7031
I www.falw.vu.nl/nl/onderzoek/athena-institute