FOD Buitenlandse Zaken, Buitenlandse Handel en Ontwikkelingssamenwerking

Dienst Bijzondere Evaluatie van de Belgische ontwkkelingssamenwerking

Studie van de Praktische Evalueerbaarheid van ge-(co)financierde interventies van de Belgische Samenwerking

Nathalie HOLVOET - Liesbeth INBERG - Bob PEETERS - Lisa POPELIER -Dirk VAN ESBROECK - Ellen VERHOFSTADT

Eindrapport

November 2015

Deze studie werd uitgevoerd door een samenwerkingsverband bestaande uit South Research en IOB (Universiteit Antwerpen), daarbij ondersteund door een begeleidingscomité.

November 2015

Grafische vormgeving: Dienst Communicatie FOD

Drukwerk: Drukkerij FOD

Evaluatie nr. S4/2014/03

Wettelijk depot: xxxxxx

Dit document is ook als pdf beschikbaar in het Frans op de CD-ROM in bijlage, en is verder verkrijgbaar bij de Dienst Bijzondere Evaluatie of via de website www.diplomatie.belgium.be/nl/Beleid/Ontwikkelingssamenwerking/Evaluatie

Dit rapport moet als volgt geciteerd worden:
Dienst Bijzondere Evaluatie / DBE (2015), *Studie van de Praktische Evalueerbaarheid van ge-(co)financierde interventies van de Belgische Samenwerking,* FOD Buitenlandse Zaken, Buitenlandse Handel en Ontwikkelingssamenwerking, Brussel.

Samenvatting

Achtergrond van deze studie

Deze studie situeert zich tegen de achtergrond van, enerzijds, het toenemend belang van evaluaties in de ontwikkelingssamenwerking, en, anderzijds, de vaststelling dat tot nu toe de kwaliteit, de bruikbaarheid en het effectief gebruik van evaluaties om verschillende redenen onder de verwachtingen zijn gebleven. De eerste fasen van de studie werden door DBE zelf uitgevoerd en vormden een belangrijke basis voor de tweede fase. Het bestek van de opdracht formuleerde als algemeen doel van de studie *"het leveren van een bijdrage tot het evalueerbaar maken van alle toekomstige acties op niet al te lange termijn"*.

De studie heeft in totaal 40 interventies in 4 landen (België, Benin, DRC en Rwanda) als voorwerp gehad. Het gaat om interventies in verschillende sectoren en uitgevoerd via een brede waaier van ontwikkelings-actoren, met (mede-)financiering van de Belgische federale overheid. Met de hulp van een studiekader (zie hierna) werden deze interventies gescoord op een aantal elementen die de evalueerbaarheid beïnvloeden. Deze scores maken een onderlinge vergelijking van de evalueerbaarheid van de 40 interventies mogelijk, maar laten niet toe om globale conclusies te trekken m.b.t. de graad van evalueerbaarheid van deze interventies of van interventies van de Belgische ontwikkelingssamenwerking in het algemeen.

Basisconcepten

De relatieve onbekendheid van de notie "evalueerbaarheid" en de aanvankelijke reserves, bij een aantal actoren, ten aanzien van deze studie hebben het studieteam ertoe aangezet om in eerste aanzet veel aandacht te besteden aan de opbouw van een draagvlak voor de studie en aan een eenduidige uitwerking van de noties van "evalueerbaarheid" en "evalueerbaarheidsbeoordeling". De studie heeft de OESO/DAC definitie gehanteerd die evalueerbaarheid definieert als *"de mate waarin een activiteit of project op een betrouwbare en geloofwaardige manier kan worden geëvalueerd"*. Evalueerbaarheidsbeoordeling wordt dikwijls in één adem genoemd met evalueerbaarheid en omschreven als *"een instrument dat toelaat te bepalen of een evaluatie aangewezen is in een bepaalde situatie"*.

Een belangrijk onderscheid dat verder wordt gemaakt is dat tussen theoretische evalueerbaarheid en praktische evalueerbaarheid. Theoretische evalueerbaarheid verwijst naar evalueerbaarheid in principe, zoals die kan worden afgeleid uit het ontwerp van de interventie, zonder de praktijk in aanmerking te nemen. Praktische evalueerbaarheid neemt die praktijk wèl in aanmerking en zal bijvoorbeeld nagaan of de gegevens over de voortgang van een interventie daadwerkelijk worden verzameld en of M&E systemen daadwerkelijk worden gebruikt, zodat die een bijdrage aan de verbetering van het beheer en de resultaten kunnen leveren.

Het studiekader dat werd ontwikkeld heeft met deze noties rekening gehouden en werd uitgewerkt in drie delen, in de studie 'dimensies' genoemd: het interventieontwerp, de praktijk m.b.t. uitvoering en beheer van de interventie en de rol van contextuele factoren. Deze drie dimensies werden verder uitgewerkt in 8 componenten en deze laatste in 62 items. Dit studiekader is een belangrijke resultaat van de studie en een nuttig instrument voor hen die verder met "evalueerbaarheid" aan de slag willen gaan.

Deze studie heeft evalueerbaarheid voornamelijk als een continuüm benaderd. Volgens deze benadering is geen enkele interventie perfect evalueerbaar in al zijn aspecten, terwijl het tegelijkertijd ook zeer onwaarschijnlijk is dat er interventies bestaan die helemaal niet kunnen geëvalueerd worden. Bovendien kan een evaluatie ook verschillende aspecten van een interventie behelzen: de vijf klassieke OESO/DAC of andere criteria, kan ze verschillende doelen nastreven (nadruk op accountability en/of op leren, …) en kan ze al dan niet gestuurd worden vanuit een evaluatiebeleid. In elk van deze situaties zal de evalueerbaarheid anders liggen.

Het belang van "evalueerbaarheid" en "evalueerbaarheidsbeoordeling"

De bepaling van de evalueerbaarheid van een interventie dient *niet* om de (ontwikkelings-)waarde van die interventie te bepalen. Het is perfect mogelijk dat erg waardevolle interventies moeilijk evalueerbaar zijn; er bestaat met andere woorden geen verband tussen (de mate van) evalueerbaarheid en de (ontwikkelings-) waarde van een interventie. Anderzijds is het belang van (de notie) evalueerbaarheid en (het instrument) evalueerbaarheidsbeoordeling onomstreden, ondanks hun relatief beperkte toepassing. Zo kan hun aanwending de kwaliteit van evaluaties verbeteren, en ervoor zorgen dat de evaluatieresultaten effectiever worden gebruikt. Een goede evalueerbaarheids-beoordeling kan uitgevoerd worden aan een kost die slechts een fractie vertegenwoordigt van het totale evaluatie-budget, maar verhoudingsgewijs een groot verschil kan maken. Verder – en dit wordt ook bevestigd door de resultaten van deze studie – is het toepassingsveld van een evalueerbaarheidsbeoordeling niet beperkt tot evaluatie als zodanig en kan zulke beoordeling in alle fasen van een interventie bijdragen tot een beter beheer.

Voornaamste resultaten en bevindingen

Kennis en gebruik van evalueerbaarheid en evalueerbaarheidsbeoordeling. De studie heeft vastgesteld dat het concept "evalueerbaarheid" en het instrument "evalueerbaarheidsbeoordeling" (of –toets) tot nu toe weinig gekend en gebruikt worden in de Belgische ontwikkelingssamenwerking. Elementen van een evalueerbaarheidstoets worden hier en daar wel in de praktijk gebracht (zonder ze als zodanig te benoemen), maar nergens is er sprake van een systematische toepassing. Deze vaststelling is betekenisvol omdat over de laatste decennia het belang van "evaluatie" gaandeweg is toegenomen en "evaluatie" tot een vast onderdeel van de beheerspraktijk in de ontwikkelingssamenwerking is uitgegroeid: actoren kunnen het zich eenvoudigweg niet meer veroorloven om *niet* te evalueren. Hoewel deze ontwikkeling in haar algemeenheid positief is, brengt ze ook het gevaar met zich mee dat evaluaties verworden tot rituele oefeningen, zonder authentieke betrokkenheid van de sleutelactoren. Een bewuster gebruik van de notie van "evalueerbaarheid' en het instrument "evalueerbaarheidstoets", met daaraan gekoppeld de mogelijkheid tot gefundeerde uitspraken m.b.t. de wenselijkheid van een evaluatie, kan in deze context een belangrijk instrument zijn tot een verbeterde en meer relevante en realistische invulling van de rol en functie van evaluatie(s) in de ontwikkelingssamenwerking.

Globale bevindingen. De 40 geanalyseerde interventies halen een gemiddelde score (op basis van de 62 items van het studiekader) die zich iets boven het middelpunt van de gehanteerde schaal situeert. Hoewel we aan de scores geen absolute betekenis kunnen geven, vormen die een goede indicatie van de voornaamste globale vaststelling van deze studie: interventies hebben in het algemeen een aantal sterke punten, maar ook nog een groot aantal werkpunten als ze hun evalueerbaarheid willen verbeteren. Verder valt op dat de distributie van de scores een normaalverdeling benadert, maar wel een grote spreiding kent, wat impliceert dat er op het vlak van kwaliteit van beheer nog grote verschillen bestaan onder de Belgische actoren en types interventies, ondanks een sturing van de financierende overheid.

Sterkere en zwakkere componenten en hun onderlinge beïnvloeding. Van de drie onderzochte dimensies (interventieontwerp, uitvoeringspraktijk, context) scoort de laatste merkelijk hoger dan de andere. Dit vormt een bijkomende illustratie van de marge voor verbetering van die dimensies waar actoren het meest vat op hebben (interventieplan en uitvoeringspraktijk). Meer specifiek scoren de onderdelen 'het voorgestelde M&E systeem', 'de kwaliteit van de basisinformatie m.b.t. de interventie-uitvoering' en 'de kwaliteit van de interventielogica en veranderingstheorie' het zwakste. De sterkste scores zijn te vinden bij de componenten die onder de context worden onderscheiden: 'de houding van de sleutelactoren t.a.v. (externe) evaluatie' en 'de invloed van de (politieke) en institutionele context'. Verder is gebleken dat onvolkomenheden in het interventieontwerp in vele gevallen doorwerken tijdens de uitvoering en op die wijze het niveau van evalueerbaarheid zowel rechtstreeks als onrechtstreeks beïnvloeden. Een goede ontwerpfase is immers veelal de voorbode van een kwalitatief goed interventiebeheer. Investeringen in een goede interventie-voorbereiding blijken zich later terug te betalen terwijl initiële zwaktes naderhand moeilijker blijken te corrigeren dan men op het eerste zicht zou verwachten; het initiële ontwerp blijkt met andere woorden in de praktijk dé referentie waarop de praktijk zich stoelt.

Verschillende gevolgen van zwakheden in het interventieontwerp voor monitoring en evaluatie. Er blijkt een verschil te bestaan in de gevolgen van een zwakke ontwerpfase voor monitoring enerzijds en evaluatie anderzijds. Blijkbaar is het mogelijk om op het – voornamelijk operationele – niveau van monitoring relatief eenvoudig correcties aan te brengen, zoals blijkt uit het feit dat de kwaliteit van het M&E systeem in de praktijk aanmerkelijk hoger ligt dan de kwaliteit van het voorgestelde M&E systeem, al blijven lacunes uit de beginfase doorwerken in de uitvoering. Lacunes in het interventieontwerp hebben echter zwaardere gevolgen voor de evaluatiefunctie omdat correcties daar moeilijker zijn aan te brengen en impliceren dat bepaalde belangrijke aspecten van een interventie (effectief bereiken van initiële doelgroepen, effecten van de interventie op verschillende sociale groepen, realisatie van assumpties en risico's) (bijna) niet geëvalueerd kunnen worden en dikwijls zelfs geheel buiten het blikveld van evaluaties blijven. In dat geval kan het samenspel van de hierboven vermelde factoren er ook toe leiden dat 'onafhankelijke' evaluaties de facto sterk van binnenuit worden gestuurd (of minstens bepaald) waardoor – gewild of ongewild – blinde vlekken niet worden onderkend en controversiële onderwerpen uit de weg worden gegaan of onderbelicht blijven. Er is met andere woorden een gevaar dat evaluaties zich slechts richten op de realiteit zoals die door de betrokken interventie bepaald of geïnterpreteerd is.

De globale evalueerbaarheidsscores per (OESO/DAC) evaluatiecriterium geven voor alle dimensies en componenten lagere scores voor duurzaamheid en vooral impact. Het feit dat hetzelfde score-patroon bijna steeds terugkomt, heeft uiteraard te maken met de verschillende moeilijkheidsgraad – in alle fasen van de interventiecyclus - van de evaluatie van deze criteria. Hoewel het moeilijk is om op dit vlak te veralgemenen, kan men stellen dat duurzaamheid en vooral impact moeilijker te evalueren zijn dan de drie andere criteria. Impact-evaluatie stelt hoge methodologische (en ook financiële) eisen terwijl de moeilijkheid bij de evaluatie van duurzaamheid meestal te maken heeft met de uitdaging om gefundeerde uitspraken te doen m.b.t. een situatie die zich pas in de toekomst zal voordoen. Verder werd vastgesteld dat ondanks de toegenomen aandacht voor duurzaamheid, deze onvoldoende wordt geïntegreerd in de beheerssystemen. Anderzijds scoren effectiviteit en vooral efficiëntie aanmerkelijk hoger, wat een indicatie vormt van de kwaliteit van het interventiebeheer, met name wat betreft monitoring en evaluatie. In dit kader is het ook belangrijk de invloed van de financierende overheid te onderkennen: DGD is in eerste instantie bekommerd om een correcte aanwending van de ter beschikking gestelde publieke fondsen en heeft daartoe belangrijke voorwaarden (via procedures, formats, …) opgelegd voor het beheer van de gefinancierde interventies, waarbij er veel aandacht wordt besteed aan aspecten die efficiëntie benadrukken, terwijl bijvoorbeeld impact duidelijk minder aandacht krijgt.

Daarnaast is de goede evalueerbaarheidsscore van efficiëntie ook zeker te verklaren door de belangrijke inspanningen die met name BTC en NGOs hebben ondernomen om hun M&E systemen uit te werken en te implementeren. Deze inspanningen zijn het gevolg van processen die reeds geruime tijd binnen deze organisaties spelen, maar ook een reactie op de geplande screening van de NGA's die in 2016 zal plaatsvinden en die onder meer de kwaliteit van de M&E systemen gaat bekijken. Bij de uitbouw van M&E systemen gaat het overigens in vele gevallen over vrij complexe initiatieven die zich gradueel ontwikkelen, waarbij de veranderingen op interventieniveau pas gaandeweg hun ingang vinden. Bij deze processen wordt een bottom-up benadering gevolgd, waarbij het implementatieniveau (inputs – activiteiten – outputs) eerst aan bod komt; dit is overigens ook het niveau waar het directe nut voor de betrokken organisaties het meest voelbaar is. Om verschillende redenen ligt het echter niet voor de hand dat M&E systemen ook gaandeweg "automatisch" de hogere niveaus van de doel-middelen ketting gaan integreren en dat met name de evaluatiefunctie even sterk wordt ontwikkeld als de monitoringfunctie. De stimuli en positieve randvoorwaarden die momenteel bestaan voor M&E op implementatieniveau (direct aantoonbaar nut, traditionele focus op het operationele, druk van DGD, relatief lage eisen voor opzet en uitvoering) bestaan immers minder of helemaal niet voor M&E m.b.t. outcomes en impacts. De verhoging van de evalueerbaarheid (in de eerste plaats m.b.t. de andere criteria dan efficiëntie) is bij ongewijzigd beleid en context in de toekomst dus helemaal niet zeker, ondanks de vooruitgang bij de ontwikkeling van M&E systemen gedurende de laatste jaren.

Resultaten van de comparatieve analyse. De evalueerbaarheidsscores *per land* verschillen erg weinig van elkaar, wat impliceert dat andere parameters waarschijnlijk van groter belang zijn. Slechts voor wat betreft de context zijn er om evidente redenen vrij grote verschillen, maar niettemin is de invloed van de context op de 'technische' evalueerbaarheid niet zo groot in de zin dat die niet erg belemmerend werkt. Een bepaalde institutionele context kan overigens zowel positief als negatief inwerken op de verschillende aspecten van evalueerbaarheid zoals de ervaringen in Rwanda hebben aangetoond.
De evalueerbaarheid van *interventies met een 'complexe' en 'minder complexe' veranderingstheorie* (TOC) verschilt niet wezenlijk van elkaar. Interventies met een complexe TOC scoren zelfs iets beter, mogelijk omdat actoren van deze interventies meer investeren in analyse en uitwerking van M&E systemen en –praktijk, en omdat zij – terecht of ten onrechte – er van uitgaan dat deze interventies moeilijker financierbaar zijn en het moeilijker is om hun resultaten aan te tonen.
De evalueerbaarheidsscores *per financieringskanaal* (type Belgische actor) geven grotere verschillen aan dan die voor de aard van de interventies en het land. Met name tussen 'bilateraal/NGOs/vakbonden' aan de ene kant en de 'andere actoren' aan de andere kant zijn de verschillen groot (betere scores voor de eerste groep), al zijn er ook in deze laatste groep voorbeelden van *good practices*. De voornaamste verklaring voor deze vaststelling ligt mogelijk in de (voor de 'andere actoren') lagere externe (DGD) eisen ten aanzien van interventieplan en uitvoeringspraktijk, die worden versterkt door het feit dat voor een deel van de tot deze groep behorende actoren 'ontwikkelingssamenwerking' geen hoofdtaak vormt.

Voornaamste aanbevelingen

De verschillende bij de Belgische ontwikkelingssamenwerking betrokken actoren hebben elkeen een belang en verantwoordelijkheid in het streven naar een betere evalueer-baarheid en nemen die best op vanuit een *gemeenschappelijk* gedefinieerd en gedragen kader en richtlijnen, waarmee elke groep vervolgens vanuit zijn eigen rol en specificiteit aan de slag gaat. Het studieteam beseft verder dat een aantal van de geformuleerde aanbevelingen veeleisend zijn, minstens in aanzet. Als zodanig kunnen zij pas goed worden toegepast als de beheerslast voor de betrokken actoren overeenkomstig kan worden verminderd, bij voorkeur door - overeenkomstig de visie van de Strategienota Ontwikkelingsresultaten - een andere invulling te geven aan de eisen m.b.t. de

uitwerking van interventievoorstellen, uitvoeringsrapportage, e.d.m. waarbij de focus verlegd wordt naar de ontwikkelings*resultaten* (outcomes, impact) ten koste van de operationele niveaus (middelen, activiteiten, outputs).

De **strategische aanbevelingen** betreffen:

(1) De systematische integratie, door alle actoren, van evalueerbaarheid en evalueerbaarheidsbeoordeling waarbij deze als een middel voor een meer performante ontwikkelingssamenwerking worden aanzien en niet als een hefboom voor bureaucratische controle of sturing (door donor en/of binnen organisaties). Evenmin mag/kan het de bedoeling zijn maximale evalueer-baarheid na te streven; het vergroten van evalueerbaarheid moet een constante zorg zijn maar die moet zich adequaat situeren binnen een specifieke context; steeds zal er een omslagpunt zijn waar de kost van het bereiken van een grotere evalueerbaarheid niet meer opweegt tegen de baten.

(2) De invoering van een coherente evalueerbaarheidsbeoordeling als belangrijk hulpmiddel bij een ex ante analyse van elke evaluatie om de potentiële baten van de evaluatie te analyseren en aan te tonen, en om op die wijze tot een gefundeerde beslissing te komen m.b.t. het al of niet laten uitvoeren van een evaluatie.

(3) Een verbetering van het voorbereidingstraject van interventies, waarbij meer moet worden ingezet op kwaliteit en minder op routine en 'meer van hetzelfde'. Aangezien het hier om een veeleisend proces gaat, is het belangrijk om een geleidelijke verbetering te beogen die op verschillende wijzen wordt ondersteund: via een aangepast kader (met incentives) van DGD, goede studies en evaluaties die de formulering kunnen ondersteunen (bv. aan het einde van vorige fasen) en via een vermindering van de administratieve eisen en regelgeving (gelieerd aan interventievoorstellen en –rapportage) die de ontwikkelingseffectiviteit niet ten goede komen.

(4) Een hernieuwde aandacht en opwaardering van de outcome en impact niveaus doorheen de gehele interventiecyclus (interventieontwerp, M&E, ….) via (onder meer) een duidelijke definiëring van deze basisnoties en hoe ze worden ingevuld in interventie-voorstellen en –rapporten, en de goede uitwerking van TOC met aanduiding van, enerzijds, directe (*intermediate*) outcomes waartoe interventies op basis van een duidelijke veranderingstheorie aantoonbaar kunnen bijdragen en, anderzijds, langere termijn effecten op maatschappelijk vlak.

(5) Een verdere ontwikkeling van de dikwijls reeds goed uitgewerkte M&E systemen en praktijken met de ambitie een goede evalueerbaarheid van effectiviteit, impact en duurzaamheid te bereiken. Zulke ontwikkeling gebeurt best op een geleidelijke wijze, waarbij middelen, instrumenten, capaciteit en ervaring overhands toenemen zodat gaandeweg meer complexe functies kunnen worden opgenomen en geïntegreerd.

(6) Verbonden met het vorige punt is het belangrijk dat een kader wordt gecreëerd waarbinnen deze (ambitieuze) veranderingen niet alleen worden gefaciliteerd, maar ook gestimuleerd en positief gewaardeerd. DGD speelt in dit verband een cruciale rol en zou in samenspraak met de andere sleutelactoren (a) de huidige regelgeving, instrumenten en procedures verder kunnen herzien en vereenvoudigen zodanig dat deze meer gericht zijn op de (beoogde) ontwikkelingseffecten, (b) incentives ontwikkelen om de M&E functie (en dan vooral de evaluatiefunctie) verder te ontwikkelen, waarbij de actoren beter in staat worden gesteld om kwalitatief sterke evaluaties te (laten) uitvoeren die ook duurzaamheid en impact analyseren, (c) een fonds oprichten ter financiering van studies en evaluaties op effect- en impactniveau waarvan het initiatief uitgaat van – bij voorkeur – het geheel van Belgische ontwikkelings-actoren. Dit fonds zou gezamenlijke oefeningen moeten financieren waarbij verschillende interventies van verschillende actoren zijn betrokken en studies en evaluaties uitvoeren die de middelen en capaciteiten van individuele actoren te boven gaan en/of voor hen minder interessant zijn.

(7) De geplande certificering van M&E systemen wordt best niet losgekoppeld van een bredere en meer geïntegreerde benadering, zoals die bv. door de voorziene screening van de NGAs minstens in aanzet is voorzien. Eerder dan een formele en gestandaardiseerde certificering lijkt het aangewezen om – in lijn met het de benadering in het verleden – goede incentives te ontwikkelen om de kwaliteit van het beheer van interventies en meer specifiek M&E systemen te verhogen, incentives die zich bij voorkeur enten op reeds aan de gang zijnde processen binnen de verschillende actoren, zoals (bijvoorbeeld) de uitwerking en toepassing (door de actoren zelf, of met externe begeleiding) van een diagnose-instrument dat de betrokken actoren (en DGD) toelaat om inzicht te krijgen in de sterke en zwakke punten van hun M&E systeem, en een *tailor-made* resultaatsgericht stappenplan te ontwikkelen ter verbetering van hun M&E systeem. Om deze belangrijke functies (blijvend) op te nemen, moet DGD over de nodige menselijke (en andere) middelen blijven beschikken.

De studie eindigt met enkele **operationele aanbevelingen** die voor een deel een concrete uitwerking van de strategische aanbevelingen betreffen:

(8) interventie-ontwerpen moeten meer aandacht schenken aan een betere omschrijving (differentiatie) van doelgroepen en de uitwerking van een goede baseline;
(9) het M&E beleid en zijn vertaling in de praktijk moeten worden vervolmaakt, met meer aandacht voor de ontwikkeling van een evaluatiefunctie en de articulatie van het eigen M&E beleid met dat van lokale actoren;
(10) de M&E functie op interventieniveau moet verder worden ontwikkeld, waarbij vooral de integratie van de verschillende M&E componenten in een coherent systeem en een betere betrokkenheid (vanuit het subsidiariteitsprincipe) van de betrokken actoren bij M&E belangrijk aandachtspunten moeten zijn;
(11) een meer bewuste benadering is nodig om het uiteindelijk gebruik van evaluaties te optimaliseren, onder meer door uiteindelijke gebruikers nauwer bij evaluaties te betrekken (in alle fasen) en door evaluaties te plannen en uit te voeren vanuit een portfoliobenadering die (op evaluatievlak) een gedifferentieerde benadering impliceert van de verschillende interventies.

Inhoudstafel

Lijst van de tabellen

Lijst met afkortingen

APEFE	Association pour la Promotion de l'Education et de la Formation à l'Etranger
BC	Begeleidingscomité
BTC	Belgische Technische Coöperatie
CUD	Coopération Universitaire au Développement
DAC	Development Assistance Committee
DBE	Dienst Bijzondere Evaluatie
DRC	Democratische Republiek Congo
DGD	Directie-generaal Ontwikkelingssamenwerking en Humanitaire Hulp
DFID	Department for International Development
FGD	Focus Groep Discussie
G&O	Gender en Ontwikkeling
IMEP	Independent Monitoring and Evaluation Project
IOB	Instituut voor Ontwikkelingsbeleid en –beheer
ITG	Instituut voor Tropische Geneeskunde
M&E	Monitoring en Evaluatie
MIS	Management Informatie Systeem
NGA	Niet-gouvernementele actor(en)
NGO	Niet-gouvernementele organisatie(s)
OESO	Organisatie voor Economische Samenwerking en Ontwikkeling
PCM	Project Cycle Management
SR	South Research
ToC	Theory of Change (veranderingstheorie)
TFD	Technisch en Financieel Dossier
UN	United Nations
VLIR-UOS	Vlaamse Interuniversitaire Raad, Universitaire Ontwikkelingssamenwerking
VN	Verenigde Naties
VVOB	Vlaamse Vereniging voor Ontwikkelingssamenwerking en Technische Bijstand

1. Inleiding

1.1 Achtergrond van deze studie

Verschillende overwegingen lagen aan de basis van het initiatief van de Dienst Bijzondere Evaluatie (DBE) om een studie te wijten aan 'Evalueerbaarheid'[1]. Evaluaties hebben gedurende het laatste decennium een steeds belangrijker plaats ingenomen in de beheerscyclus, zowel op interventieniveau als daarboven[2]. Ook is rond de rol van evaluaties grotere duidelijkheid gegroeid, waarbij in toenemende mate de nadruk wordt gelegd op de verschillende doelen die evaluaties kunnen nastreven: rekenschap afleggen over het gebruik van de ontvangen middelen, leren uit de ervaringen van het verleden om beter te doen in de toekomst en toekomstige beleidsvorming en –praktijk ondersteunen.

Gezien het toenemend belang van evaluaties in de ontwikkelingssamenwerking – evaluaties worden aanzien als cruciaal, worden dikwijls verplicht en zijn in vele gevallen in samenhang met monitoring uitgegroeid tot volwaardige systemen - is het erg belangrijk om aandacht te besteden aan de relevantie en kwaliteit van evaluaties en de opvolging van hun resultaten. Een eerste en belangrijke (maar dikwijls vergeten) stap daarbij bestaat er uit om na te gaan of aan alle voorwaarden is voldaan om een kwaliteitsvolle evaluatie uit te voeren, of met andere woorden de evaluatie beantwoordt aan een bestaande nood, of haar uitvoering haalbaar is in de gegeven context en met de beschikbare middelen en of de inzet van die middelen opweegt tegen de baten die de evaluatie verwacht wordt te genereren. Verder is het belangrijk rekening te houden met het feit dat er een erosie van de term en de praktijk van "evaluatie" heeft plaatsgevonden: ondanks het feit dat er meer evaluatie-ervaring is opgedaan, kan men zich bij vele van de als "evaluatie" aangeduide initiatieven afvragen of de minimum kwaliteitseisen wel degelijk vervuld zijn om van een daadwerkelijke evaluatie te kunnen spreken[3]. Deze erosie heeft deels te maken met een tekort aan expertise, maar waarschijnlijk nog veel meer met het feit dat evaluaties deel zijn gaan uitmaken van het "systeem" en daardoor dikwijls (te) routinematig worden geïnitieerd en uitgevoerd.

Naast deze internationale ontwikkelingen vormen ook de ervaringen op Belgische vlak een reden om een studie aan "evalueerbaarheid" te wijten. Ervaringen met externe evaluaties maar ook met monitoring suggereren immers dat in vele gevallen de initiële voorwaarden om goed te kunnen evalueren niet of slechts gedeeltelijk aanwezig zijn. Als zodanig kan men veronderstellen dat de ondernomen evaluatie-oefeningen slechts zeer ten dele hun ambities kunnen waarmaken. Anderzijds hebben vele Belgische ontwikkelingsorganisaties de laatste jaren geïnvesteerd in de ontwikkeling van hun M&E systemen, waardoor – in principe – de evalueerbaarheid zou moeten toegenomen zijn. Rond het uiteindelijke nut van deze systemen en dan met name van de resultaten van evaluaties zijn evenwel weinig gegevens beschikbaar. "Evalueerbaarheid" is in dit verband een interessante notie, omdat ze de besluitvorming m.b.t. de wenselijkheid van evaluatie (van een project, een programma, …) kan ondersteunen.

---

[1] Het begrip 'evalueerbaarheid' wordt in hoofdstuk 2 nader gedefinieerd.

[2] In deze studie gebruiken we de term interventie om het geheel aan te duiden van projecten, programma's, instrumenten, … van de Belgische ontwikkelingssamenwerking.

[3] Vanuit methodologisch standpunt wordt van evaluaties verwacht dat zij betrouwbare, valide en bruikbare resultaten opleveren. Zie ook hoofdstuk 2.

Vanuit bovenstaande vaststellingen heeft de notie van "evalueerbaarheid" de laatste tijd aan belangstelling gewonnen, terwijl er voordien minder aandacht aan werd besteed. DBE heeft daarom besloten een studie te wijden aan "evalueerbaarheid" waarbij de dienst zelf de eerste twee studiefasen voor zijn rekening heeft genomen. Zij betroffen de uitwerking van een analysekader van de theoretische evalueerbaarheid en een studie van de theoretische evalueerbaarheid van 43 interventies; het verslag van beide fasen werd bij de start van de studie ter beschikking gesteld (dat van fase 2 in draftversie).

1.2 Basiskarakteristieken van deze studie en haar context

Het is vooreerst belangrijk te beseffen dat de aard van deze opdracht op enkele belangrijke punten afwijkt van andere door de Dienst Bijzondere Evaluatie uitgeschreven opdrachten:

- deze opdracht is een *studie en geen evaluatie*; het komt er dus niet op aan om via deze opdracht te beantwoorden aan de klassieke doelen van een evaluatie, maar wel om te beantwoorden aan *enkele duidelijk omlijnde doelen* (zie 1.3 hierna) van de studie die vooral werden geformuleerd met het oog op de toekomst;
- zoals reeds vermeld, *werden de eerste twee fasen van deze studie reeds door de DBE zelf uitgevoerd*; ze betreffen de uitwerking van een analysekader van de theoretische evalueerbaarheid en een studie van de theoretische evalueerbaarheid van 43 interventies;
- zoals het bestek aangeeft, is de *DBE uitzonderlijk ook een stakeholder* bij deze studie; terwijl DBE in een "traditionele" DBE-opdracht voornamelijk is betrokken als leidend ambtenaar, is haar belang in deze studie groter in die zin dat de realisatie van de doelstellingen van deze studie DBE rechtstreeks aanbelangt;
- zoals ook uit de bestaande literatuur m.b.t. evalueerbaarheid blijkt, kan (en wordt) het begrip op verschillende wijzen worden gehanteerd. Daarom is het van belang om het concept van evalueerbaarheid goed af te lijnen. De definities in het bestek (en ook in de literatuur) betreffen zowel *evaluability* in de enge zin als *evaluability assessment* waarbij beide (d.w.z. de definitie en het instrument) niet steeds strikt van elkaar worden onderscheiden. De vermelde definities hebben goede handvaten aan voor de afbakening van de studie aangereikt, maar dienden tijdens het verloop van de studie verder te worden uitgeklaard (zie hoofdstuk 2). Daarbij was het *niet* de bedoeling om een uitputtende discussie rond het concept te voeren, dan wel om – pragmatisch - overeenstemming te bereiken over de te hanteren werkdefinitie voor het vervolg van de studie.

1.3 Doelen en reikwijdte van de studie

In een standaardaanpak voor een studie of evaluatie worden ex-ante duidelijk de doelen en het onderzoeksveld vooropgesteld die vervolgens tijdens de uitvoering (quasi) onveranderd blijven. Dit was in deze studie niet het geval in de zin dat de oorspronkelijke doelen en onderzoeksveld zoals in het studiebestek geformuleerd naderhand werden verruimd.

De **doelen** zoals die in het bestek van de opdracht (ref. S4/2014/01) werden geformuleerd, kunnen als volgt worden samengevat[4]:

[4] Zie punt B3 blz. 25 van genoemd document dat te vinden is in bijlage 1. Hierbij is het belangrijk te vermelden dat met name de tweede en derde specifieke doelstelling voor een groot deel ingegeven waren vanuit initiatieven die bij de lancering van de studie in de steigers stonden. Om verschillende redenen is de uitwerking van deze initiatieven echter niet gelopen zoals aanvankelijk gepland.

- de studie heeft een drieledige doelstelling:
    - samen met de betrokken partners de noodzakelijke, toereikende en realistische voorwaarden bepalen die het mogelijk maken om samenwerkingsinterventies op een objectieve manier te evalueren;
    - lessen trekken die nuttig zin voor de harmonisering en de certificatie van de evaluatiesystemen van de actoren waarin voorzien werd door de wetgever; en
    - nagaan in welke mate bovenvermelde evalueerbaarheidsvoorwaarden (al dan niet) overeenstemmen met de voorwaarden die noodzakelijk zijn voor de monitoring en het resultaatsgericht beheer van de interventies en met voornoemd wettelijke en reglementair kader;
- de algemene doelstelling is (een bijdrage te leveren tot) het evalueerbaar maken van alle toekomstige acties op een niet al te lange termijn.

Verder werd gesteld dat de resultaten van de studie in verschillende opzichten nuttig moesten zijn voor DBE, de diensten van DG-D en de Attachés, en de partnerorganisaties van de bilaterale en niet-gouvernementele samenwerking.

In de opstartfase van de studie bestond er rond bovenvermelde doelen nogal wat onduidelijkheid en, daaraan gekoppeld, een zeker wantrouwen bij een aantal actoren ten aanzien van de uiteindelijke bedoelingen en (bedoelde en onbedoelde) effecten van de studie, die o.m. in de eerste vergadering van het Begeleidingscomité werd geventileerd.

Het studieteam heeft daarom in een eerste fase erg veel aandacht besteed aan het creëren van een draagvlak voor de studie, onder meer door een betere uitwerking van het concept van evalueerbaarheid. Als een gevolg werd het voornaamste doel van de studie lichtjes bijgesteld (een bijdrage leveren aan een betere evalueerbaarheid van de acties van de Belgische samenwerking) en werd vooral de aandacht getrokken op het potentieel van de analyse van evalueerbaarheid voor de verbetering van het beheer van interventies, inclusief de bepaling van het nut van een evaluatie in een specifieke situatie en in het algemeen. Daarnaast werd aan de hierboven vermelde specifieke doelstellingen voor zover mogelijk en relevant de nodige aandacht besteed.

Tenslotte werd het voorstel aanvaard van het studieteam om de initiële (d.w.z. in het bestek van de studie) focus van de studie op *praktische* evalueerbaarheid te verbreden naar een focus op zowel theoretische als praktische evalueerbaarheid[5].

1.4 Structuur van dit syntheseverslag

De structuur van dit syntheseverslag is als volgt. Na dit inleidend hoofdstuk volgt een hoofdstuk dat aandacht besteedt aan de concepten van evalueerbaarheid en evalueerbaarheidsbeoordeling. Dit hoofdstuk wil ook het belang en praktisch nut van deze concepten aantonen en waarom het interessant kan zijn een evalueer-baarheidstoets uit te voeren. Dit hoofdstuk besteedt ook kort aandacht aan de onderzoeksmethodologie en gehanteerde studiekader. In hoofdstukken drie en vier worden de voornaamste bevindingen van de studie voorgesteld, terwijl hoofdstuk vijf de voornaamste conclusies en aanbevelingen bevat. Het verslag wordt vervolledigd met een aantal annexen: de referentietermen, beschrijving van de methodologie, het studiekader, de lijst met de 40 geanalyseerde interventies, de bibliografie, de lijst van de gecontacteerde personen en twee technische annexen (toelichting bij de statistische analyses en een uitgebreide scoringstabel).

[5] Deze begrippen worden in hoofdstuk 2 nader toegelicht.

2. Basisconcepten, studiekader en aanpak

2.1 Evalueerbaarheid en evalueerbaarheids-beoordeling

2.1.1 Definities

In de literatuur wordt de OESO / DAC (2002: p. 21) definitie van evalueerbaarheid het meest gebruikt: "Evalueerbaarheid is de mate waarin een activiteit of project op een betrouwbare en geloofwaardige manier kan worden geëvalueerd"[6]. Verder worden "evalueerbaarheid" en "evalueerbaarheidsbeoordeling" dikwijls in één adem genoemd. Dit laatste wordt dan omschreven als "een instrument dat toelaat te bepalen of een evaluatie aangewezen is in een bepaalde situatie"[7]. Deze koppeling geeft aan dat het belangrijk is om evalueerbaarheid niet alleen als concept te hanteren, maar om het te koppelen aan de evaluatiepraktijk en de noodzaak om na te gaan of een evaluatie gerechtvaardigd en haalbaar is, en in staat om nuttige informatie te verstrekken. Dit houdt in dat de bijdrage van een evaluatie aan de verbetering van het beheer en de resultaten van de actie wordt meegenomen in de beoordeling van de evalueerbaarheid.

Alvorens verder te gaan, is het belangrijk om kort de definitie van de OESO/DAC nader te ontleden omdat die definitie bedrieglijk is in haar eenvoud, waardoor belangrijke implicaties over het hoofd kunnen worden gezien. Centraal staan daarbij de noties 'betrouwbaar' en 'geloofwaardig' die frequent in de context van evaluaties worden gehanteerd, maar niet steeds worden geëxpliciteerd (overigens evenmin in de OESO/DAC definitie):

- een evaluatie is *betrouwbaar* wanneer haar onderzoeksresultaten stabiel en consistent zijn. Dit betekent dat als het evaluatieonderzoek zou worden overgedaan, het tot dezelfde bevindingen zou leiden. Elementen die de betrouwbaarheid beïnvloeden zijn onder meer de kwaliteit van de onderzoeksmethoden en -uitvoering, de reikwijdte van het onderzoek, de onafhankelijkheid van de informatie en de onderzoekers;
- een evaluatie is *geloofwaardig w*anneer de onderzoeksresultaten als valide en relevant worden beschouwd door de voornaamste betrokken partijen (stakeholders). Elementen die de geloofwaardigheid beïnvloeden zijn onder meer: aandacht en gevoeligheid voor context-specifieke dimensies (economische, culturele, sociale) van de interventie; brede dataverzameling en triangulatie tussen bronnen en een omstandige omschrijving van het dataverzamelings-proces; transparantie en onafhankelijkheid van het evaluatieproces; betrouwbaarheid van de meetinstrumenten; consistentie en coherentie in de bevindingen en tussen de bevindingen en de conclusies.

Een belangrijk onderscheid dat vaak wordt gemaakt is dat tussen '*theoretische evalueerbaarheid*' en '*praktische evalueerbaarheid*'. Theoretische evalueerbaarheid

[6] "Evaluability is the extent to which an activity or project can be evaluated in a reliable and credible fashion"; zie OECD/DAC (2002) Glossary of Key Terms in Evaluation and Results Based Management. Paris: OECD/DAC, p. 21.

[7] Zie bijvoorbeeld: Peter Dahler-Larsen (2013) "Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable", *Scandinavian Journal of Public Administration*, 16 (3): 29-46.

verwijst naar evalueerbaarheid in principe, zoals die kan worden afgeleid uit het ontwerp van de interventie, zonder de praktijk in aanmerking te nemen. Praktische evalueerbaarheid anderzijds zal bijvoorbeeld nagaan of de gegevens over de voortgang van een interventie *daadwerkelijk* worden verzameld en of M&E systemen daadwerkelijk worden gebruikt, zodat een bijdrage aan de verbetering van het beheer en de resultaten mogelijk wordt. Op deze manier is de link met (praktische) bruikbaarheid van de evaluatie snel gemaakt.

In deze studie hebben wij evalueerbaarheid voornamelijk als een **continuüm** benaderd. Geen enkele interventie is perfect evalueerbaar in al zijn aspecten, terwijl het tegelijkertijd ook zeer onwaarschijnlijk is dat er interventies bestaan die helemaal niet kunnen geëvalueerd worden. Bovendien kan een evaluatie ook verschillende aspecten van een interventie behelzen: de vijf klassieke DAC criteria (relevantie, efficiëntie, effectiviteit, impact en duurzaamheid) maar ook andere criteria. Het is echter niet noodzakelijk dat alle evaluaties aan al deze criteria tegelijk aandacht schenken. Bovendien kan evalueerbaarheid ook verband houden met het doel van de evaluatie. Terwijl evaluaties verschillende doelstellingen (accountability, leren, ondersteuning van het toekomstige beleid) kunnen hebben, is het niet nodig om deze tegelijk prominent op te nemen in alle evaluaties. Als bijvoorbeeld leren de centrale doelstelling is, zijn de eisen op het vlak van evalueerbaarheid anders dan wanneer verantwoording afleggen (accountability) centraal staat. Tenslotte is evalueerbaarheid ook afhankelijk van het evaluatiebeleid van een organisatie. In dat evaluatiebeleid kan er gekozen worden om de beschikbare middelen te gebruiken voor, bijvoorbeeld, het strategisch evalueren van innovatieve interventies, dan wel om alle interventies te evalueren; er kan verder gekozen worden om, bijvoorbeeld, vooral in te zetten op onafhankelijke evaluaties met een focus of verantwoording, dan wel op interne evaluaties met een sterk lerend karakter.

2.1.2 Waarom is de beoordeling van de evalueerbaarheid belangrijk?

Vooreerst is het belangrijk te stellen dat de bepaling van de evalueerbaarheid van een interventie **niet** dient om de (ontwikkelings-)waarde van die interventie te bepalen. Het is perfect mogelijk dat erg waardevolle interventies moeilijk evalueerbaar zijn; er bestaat met andere woorden geen verband tussen de mate van evalueerbaarheid en de (ontwikkelings-)waarde van een interventie. Evalueerbaarheidsbeoordeling heeft eerder te maken met de aard van de interventie en de context waarin ze wordt uitgevoerd. Zo zijn bijvoorbeeld de uitdagingen om interventies te evalueren die de empowerment van vrouwen nastreven doorgaans groter dan bij interventies die zich met drinkwatervoorziening bezig houden. Dit kan onder andere worden verklaard vanuit het gegeven dat "empowerment" iets helemaal anders kan betekenen voor de verschillende betrokken actoren, dat het om een iteratief proces gaat met verschillende dimensies en fasen, … terwijl het relatief eenvoudig is om het eens te raken over de betekenis van "drinkwatervoorziening". Daarnaast speelt ook de context een grote rol: het is moeilijk evaluaties uit te voeren in conflictzones of in interventies waar de betrokken partijen met getrokken messen tegenover elkaar staan.

Waar ligt dan wel het belang en nut van een beoordeling van de evalueerbaarheid? Hoewel de mate van evalueerbaarheid en de doenbaarheid van een evalueerbaar-heidsbeoordeling afhangen van verschillende factoren (zie hierboven), zijn er niettemin een aantal algemene antwoorden te geven op deze vraag:

- studies m.b.t. de kwaliteit van evaluaties wijzen uit dat deze in vele gevallen ontoereikend is of dat evaluaties weinig meerwaarde opleveren in vergelijking tot hun kostprijs; verder blijkt dat het effectief gebruik van evaluatieresultaten te wensen overlaat. Een goede beoordeling van de evalueerbaarheid kan voor deze problemen een gedeeltelijke oplossing bieden;
- als het de bedoeling is om een evaluatie uit te voeren, kan de beoordeling van de evalueerbaarheid doorgaans worden uitgevoerd aan een kost die slechts een fractie is van het totale evaluatiebudget; op die wijze kan men vermijden dat

middelen worden verspild, bijvoorbeeld aan evaluaties die niet gewenst of haalbaar zijn, of slecht ontworpen. Vooral bij de uitvoering van complexe evaluaties lijkt zulke analyse aangewezen, terwijl ze bijvoorbeeld voor goed geplande en uitgevoerde interventies met goed functionerende M&E systemen minder nodig is;

- verder zullen specifiek belang en nut van een evalueerbaarheidsbeoordeling afhangen van het ogenblik waarop die wordt uitgevoerd in de interventiecyclus (zie ook tabel 1 hierna):
    - *in de voorbereidingsfase van een interventie* zal een beoordeling van de evalueerbaarheid zich vooral richten op het interventieontwerp: is dit ontwerp coherent en volledig, zijn alle (ook impliciete) assumpties meegenomen, sluit het aan op de onderliggende analyse, …? Deze evalueerbaarheidsbeoordeling betreft de *theoretische* evalueerbaarheid zoals we die hierboven hebben gedefinieerd. Een goed (of verbeterd) interventieontwerp zal uiteraard de kwaliteit van de interventie uitvoering (inclusief het M&E systeem) en latere evaluaties ten goede komen;
    - *bij de opstart van een interventie* (of eventueel net daarvoor), zal een beoordeling van de evalueerbaarheid op basis van het interventieontwerp en het globale interventie-voorstel waardevolle elementen kunnen aanleveren voor de uitbouw van een M&E systeem;
    - *tijdens de uitvoering van een interventie* kan een beoordeling van de evalueerbaarheid aanduidingen geven rond de wenselijkheid, timing en haalbaarheid van een evaluatie door na te gaan in hoeverre de voorwaarden voor de adequate uitvoering van een evaluatie zijn vervuld (bv. wat is de houding van de betrokken partijen?); deze beoordeling kan ook nagaan of het interventieontwerp goed is aangepast aan externe ontwikkelingen of verworven inzichten en zo nodig suggesties aanleveren;
    - *bij de afronding van een interventie* zal de focus van de beoordeling vrij gelijkaardig zijn en aanduidingen verstrekken m.b.t. de wenselijkheid en haalbaarheid van een evaluatie (met dit verschil dat geen verbeteringen niet meer aangebracht kunnen worden); de focus zal vooral liggen op de uitdagingen voor een adequate evaluatie-uitvoering en hoe die op te lossen en welke de alternatieven zijn in dit verband;
    - *na de afloop van een interventie* ligt de focus van de beoordeling in het verlengde van die gehanteerd bij de afronding van een interventie; een belangrijk bijkomend aspect hier is de haalbaarheid om de betrokken stakeholders van de interventie te contacteren.

Verder is ook duidelijk dat het *systematisch* bekijken, door een ontwikkelingsorganisatie, van de evalueerbaarheid van interventies bij zal dragen tot een verbeterde interventiepraktijk en tot de ontwikkeling en uitvoering van een goed evaluatiebeleid, een goede afweging van de te evalueren interventies en van de verschillende evaluatie-doelen. Op deze wijze wordt het ook mogelijk een duidelijke scheiding te maken tussen "evalueerbaarheidsbeoordeling" en "wenselijkheid om te evalueren".

Tabel 1: Nut en belang van een beoordeling van de evalueerbaarheid[8]

Fase interventiecyclus	Doel van de beoordeling	Focus van de beoordeling	Beoogd resultaat
Formulering van de interventie	Verbetering van het interventieontwerp	Kwaliteit van het interventieontwerp	Verbeteringen en aanvullingen aan het interventieontwerp
Opstart van de interventie	Input leveren voor uitbouw M&E systeem	Beschikbaarheid en kwaliteit van de informatie en data-verzameling	Verbetering van het M&E systeem (inhoud en proces)
Uitvoering en afronding	• Beslissen of evaluatie nu of later plaatsvindt • Nagaan of inter-ventie goed is aangepast aan ontwikkelingen	• Beschikbaarheid, actualiteit en kwaliteit gegevens • Positie/oordeel betrokken partijen • Situatie in de ruimere omgeving	• Inzicht in moeilijk-heidsgraad van uitvoering evaluatie op dit ogenblik • Voorstel alternatieven qua timing en inhoud • Bijsturing interventie - ontwerp en M&E systeem
	Input leveren voor het ontwerp van een geplande evaluatie		• Inzicht in keuzes voor evaluatiedoelen, hoofdvragen, aanpak en benodigde expertise
Post-interventie	Nagaan of uitvoering evaluatie op betrouwbare en geloofwaardige wijze mogelijk is	• Beschikbaarheid en kwaliteit van de gegevens m.b.t. de interventie-effecten • Haalbaarheid van contactname met betrokken partijen	• Inzicht in moeilijk-heidsgraad van uitvoering evaluatie • Inzicht in mogelijkheden om (aspecten) van impact en duurzaamheid te evalueren • Inzicht in de wenselijkheid van de evaluatie

2.2 Het eigenlijke studiekader

Uitgaande van de studiedoelstellingen is getracht om een realistisch studiekader te ontwerpen, rekening houdend met de beschikbare middelen en de eis om ongeveer 40 interventies te bekijken. Het had met andere woorden geen zin om een overmatig uitgewerkt kader voor te stellen als dit nadien toch niet goed kan worden toegepast. Bijkomend wilden we ook vermijden dat een te gedetailleerd kader zou leiden tot een *'ticking the boxes'* aanpak, die onvermijdelijk ten koste zou gaan van de uiteindelijke kwaliteit en relevantie van de oefening. Concreet heeft dit geleid tot het scenario waarvoor per interventie in totaal vier dagen (inclusief het veldbezoek) beschikbaar waren voor de bepaling van de evalueerbaarheid, wat overeenkomt met het maximum aantal dagen dat voor de analyse van de evalueerbaarheid van een interventie ter beschikking kan worden gesteld[9].

Anderzijds konden we er niet onderuit te erkennen dat "evalueerbaarheid" meerdere dimensies en aspecten omvat, zoals de theoretische en praktische evalueerbaarheid. Een goed compromis tussen 'realisme' en 'volledigheid' werd o.i. bekomen door een hiërarchie aan te brengen in de verschillende onderdelen van het studiekader. Onder een aantal centrale noties/criteria, die we hierna *'componenten'* noemen, werden onderliggende factoren geclusterd die we in ons studiekader *'items'* hebben genoemd (men kan die aanpak vergelijken met die in evaluaties waar hoofdvragen worden geoperationaliseerd in beoordelingscriteria en deze laatste in indicatoren).

[8] Afgeleid van: G. Peersman, I. Guijt, T. Pasanen (2015) *Evaluability Assessment for Impact Evaluation, Guidance, checklists and decision support*. London: Methods Lab Publication, ODI (August 2015).

[9] We kunnen in dit verband opmerken dat een daadwerkelijke beoordeling van de evalueerbaarheid tot nu toe zelden plaatsvindt in de context van de Belgische samenwerking. Echter bevat het voortraject van vele evaluaties elementen die ook in een beoordeling van de evalueerbaarheid worden meegenomen.

Het studiekader laat verder toe alle componenten te scoren zodat aggregatie en vergelijking tussen interventies, onderdelen, actoren, landen … mogelijk zijn. De geaggregeerde score kan dan worden aanzien als een indicator van evalueerbaarheid die zich dan ergens in het continuüm tussen geen en totale evalueerbaarheid situeert. Bij de uitwerking van het studiekader hebben we ons vooral laten inspireren door de studie van Rick Davies voor DFID[10]. In vergelijking met Davies is ons studiekader evenwel meer gedetailleerd. Het kader bestaat uit **drie delen** (die we hierna als 'dimensies' gaan aanduiden):

- de analyse van het interventie-ontwerp (inclusief de onderliggende veranderingstheorie) waarmee vooral (maar niet uitsluitend) de theoretische evalueerbaarheid in rekening wordt gebracht. De analyse van het interventie-ontwerp (in strikte zin) werd aangewend om ook de kwaliteit van de geplande M&E systeem te beoordelen;
- de analyse van de praktijk m.b.t. de uitvoering en beheer van de interventie en de context (inclusief gebruik van menselijke en financiële middelen);
- de analyse van contextuele factoren; deze factoren kunnen een rol spelen zowel op het vlak van het interventie-ontwerp als dat van de systemen om gegevens te genereren, als bij de (toekomstige) uitvoering van de evaluatie zelf.

Het eerste deel behelst vooral de theoretische evalueerbaarheid (maar niet uitsluitend: het geplande M&E systeem is er geen onderdeel van), het tweede deel bekijkt de praktische evalueerbaarheid maar omvat deze niet helemaal, terwijl het derde deel elementen bevat die zowel de theoretische als praktische evalueerbaarheid kunnen beïnvloeden[11].

Tabel 2 op blz. 20 geeft een overzicht van het studiekader en van de items die voor elk van de vijf DAC evaluatiecriteria werden gescoord. Belangrijk is nog te vermelden dat we bij de uitwerking van het kader de DAC definitie van de vermelde evaluatiecriteria als uitgangspunt hebben genomen (zie kader).

In totaal omvat het studiekader dus 3 *dimensies* die verder zijn uitgewerkt in 9 *componenten* en 62 *items*, die alle werden gescoord[12]. De scoring gebeurde voor elke component en elk item aan de hand van een vijfpuntsschaal; voor de dimensies werd een gewogen gemiddelde berekend van de componenten[13]. Het volledige studiekader is opgenomen in annex 3.

[10] Davies (2013) *Planning evaluability assessments: A synthesis of literature with recommendations* (Report of a study commissioned by the Department for International Development). London: DFID (October 2013).

[11] Zoals hoger aangegeven werd door de DBE zelf een studie uitgevoerd van de theoretische evalueerbaarheid met als bedoeling als input te fungeren voor voorliggende studie met focus op praktische evalueerbaarheid. Echter gezien de steekproef van de DBE-studiefase niet volledig dezelfde was als deze van onderliggende studie werd snel duidelijk dat we genoodzaakt waren om ook in onze studie theoretische evalueerbaarheid mee te nemen. Enkel op deze wijze was het ook mogelijk om de verschillende aspecten van de praktische evalueerbaarheid, alsook de linken met theoretische evalueerbaarheid, goed te kunnen te bekijken.

[12] Later in het proces werd één van die negen componenten voor een deel geschrapt en voor een deel onder een andere component ondergebracht. De tabel hierna geeft het aangepaste studiekader weer.

[13] Zie verder en zie ook bijlage 2 voor meer uitleg.

Kader 1: De DAC evaluatiecriteria en hun interpretatie voor deze studie[14]

Relevantie: de mate waarin de doelstellingen van een ontwikkelingsinterventie consistent zijn met de eisen van de begunstigden, de noden van het land, de globale prioriteiten en het beleid van partners en donoren. *De studie heeft deze definitie aangehouden, maar in de praktijk vooral gekeken naar de twee eerste elementen.*

Effectiviteit: de mate waarin de doelstelling van een interventie werden bereikt (of verwacht wordt bereikt te zullen worden), waarbij rekening wordt gehouden met hun relatief belang. *Deze studie heeft deze definitie overgenomen, maar effectiviteit vooral beoordeeld op het niveau van de specifieke doelstelling (effect, outcome) van de interventies.*

Efficiëntie: de mate waarin de middelen en inputs (fondsen, expertise, tijd, etc.) van een interventie worden omgezet in resultaten[15]. *Deze studie heeft efficiëntie op gelijkaardige wijze geïnterpreteerd door te kijken naar de wijze waarop inputs worden geconverteerd in activiteiten en deze leiden tot outputs.*

Impact: de positieve en negatieve, voornaamste en secundaire lange-termijn effecten geproduceerd door een interventie, rechtstreeks of onrechtstreeks, bewust of onbewust. *Deze studie is van deze definitie uitgegaan, maar heeft op bepaalde ogenblikken ook de notie van impact gebruikt als de "direct toewijsbare effecten (causale relatie) van een interventie" [16].*

Duurzaamheid: de mate waarin de baten van een interventie blijven voortduren nadat substantiële ondersteuning is afgerond. De studie is van deze definitie uitgegaan.

[14] Zie OECD/DAC (2002) Glossary of key terms in evaluation and results based management. Paris: OECD/DAC.

[15] In andere documenten van de DAC kunnen 'resultaten' zowel outputs, outcomes als impacts zijn. In het compendium vervangt de DAC resultaten door het engere "outputs". We hebben hier gekozen voor de engere definitie omdat die het mogelijk maakt een duidelijk onderscheid aan te brengen tussen effectiviteit en efficiëntie dat voor deze studie belangrijk is. Het is belangrijk te beseffen dat dit niet de enige mogelijkheid was, en bovendien dat de bredere definitie wordt verkozen door de meeste experten in efficiëntie, zoals bijvoorbeeld Palenberg die er een studie over maakte voor BMZ die veel wordt geciteerd in de sector: Palenberg, M., Tools and Methods for Evaluating Efficiency in Development Interventions, BMZ Evaluation Division – Evaluation Working Papers, 2011, 131 p.

[16] In deze studie wordt de notie 'impact' op twee verschillende manieren gebruikt:

- als evaluatiecriterium in strikte zin, waarbij het wordt onderscheiden van (o.m.) 'effectiviteit' en refereert naar resultaten 'boven' het outcome-niveau; als zodanig verwijst 'impact' naar de algemene doelstelling(en) zoals geformuleerd in het logisch kader van een interventie;
- in het kader van een bespreking van (de praktijk van) 'impact-*evaluaties'* als een specifieke evaluatie-methodiek waarbij de nadruk ligt op effecten die "direct toewijsbaar" zijn aan een interventie (causaliteit) . Deze effecten kunnen zowel gesitueerd zijn op het outcome als de hogere niveaus van de ketting. Een impactevaluatie bestaat er dan in om de effecten die *toewijsbaar* zijn aan een interventie (causaliteit) te onderscheiden van de effecten die te wijten zijn aan andere factoren (externe factoren of de opzet en instrumenten van de evaluatie zelf). Deze invulling sluit aan bij die gehanteerd in de door DBE geïnitieerde impactevaluatie*: Impact meten: zoektocht naar de Graal, Ex-post impactevaluatie van vier projecten van de gouvernementele samenwerking.* In deze studie wordt impact gedefinieerd als: de effecten die *voortvloeien* uit het project op globaal niveau. Overigens betreffen de impactevaluaties die via DBE worden uitgevoerd vooral evaluaties van outcomes.

Onvermijdelijk kan dit dubbel gebruik van de notie 'impact' tot verwarring leiden; we hebben geprobeerd die in het vervolg van de studie tot een minimum te beperken.

Tabel 2: Samenvattende voorstelling van het studiekader

DAC Evaluatiecriteria Dimensies/Componenten	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid
	Aantal items dat werd gescoord				
1. Analyse van het interventie-design 1.1 Onderliggende analyse (7) 1.2 De interventielogica en veranderingstheorie (8) 1.3 Het voorgestelde M&E systeem (9) 1.4 De consistentie en adaptatie van de interventielogica en veranderingstheorie (3)	(7) (1) (5) (3)	(4) (6) (7) (3)	(0) (5) (9) (3)	(0) (5) (7) (3)	(0) (3) (7) (3)
2. Praktijk mbt. uitvoering en beheer van de interventie en van de context 2.1 Basisinformatie mbt. de interventie-uitvoering (11) 2.2 Het M&E systeem in de praktijk (12)	(5) (11)	(10) (12)	(11) (12)	(10) (12)	(8) (12)
3. De evaluatie-context 3.1 Houding van de sleutelactoren (9) 3.2 De bredere context (3) 3.3 Praktische elementen (2) (°)	(9) (3)	(9) (3)	(9) (3)	(9) (3)	(9) (3)
Geaggregeerde score	(44)	(54)	(52)	(49)	(45)
4. Suggesties mbt. het gebruikte M&E systeem en toekomstige evaluaties					
5. Feedback van de betrokkenen op onze analyse en suggesties					

(°) Dit onderdeel wordt slecht pro memorie vermeld omdat de parameters en scores voor dit onderdeel onvermijdelijk verbonden zijn met de organisatie van een *concrete* evaluatie.

2.3 Aanpak van de studie[17]

2.3.1 Steekproefbepaling

Overeenkomstig de vereisten van het bestek werden 40 interventies in vier landen (België, Benin, DRC en Rwanda) geanalyseerd; in elk land werden 10 interventies weerhouden[18]. Het ging daarbij telkens over interventies die qua uitvoering reeds ver waren gevorderd, eventueel via een voorgaande fase. De steekproefbepaling gebeurde niet aselect, maar beoogde een oververtegenwoordiging van kleinere sectoren en actoren en van atypische interventies (bv. studiebeurzen). Met het oog op een latere analyse werden verder bij de steekproef-bepaling naast het land twee bijkomende parameters gehanteerd: complexe versus minder complexe interventies[19] en het interventiekanaal (directe bilaterale samenwerking, indirecte samenwerking via NGOs en vakbonden, samenwerking via andere actoren).

De uiteindelijke steekproef bestond uit:

[17] Bijlage 2 bevat meer informatie m.b.t. de aanpak en methodologie van de studie.

[18] Een lijst met de 40 projecten is opgenomen in bijlage 4.

[19] Na een lang intern reflectieproces kreeg deze parameter een specifieke invulling: eerder dan een onderscheid in te voeren op basis van complexe en minder complexe *sectoren*, hebben we gekozen voor een opdeling tussen *interventies* met een complexe TOC en *interventies* met een minder complexe TOC. Indien bijvoorbeeld een bepaalde interventie in een 'minder complexe' sector op basis van de onderliggende TOC toch als 'complex' werd beoordeeld, kwam die in deze laatste categorie terecht. Het is ook belangrijk op te merken dat het gehanteerde onderscheid eveneens verschilt van dat tussen complexe en minder complexe interventies *als zodanig*, waarbij ook andere belangrijke factoren meespelen, zoals de omvang van de interventie, het aantal betrokken actoren, etc.

- 10 interventies in elk van de 4 landen;
- 24 (60%) complexe en 16 (40%) minder complexe interventies;
- 10 (25%) bilaterale interventies, 20 (50%) interventies uitgevoerd door NGOs of vakbonden, en 10 (25%) interventies uitgevoerd via andere actoren (APEFE/ VVOB, universiteiten, BIO, …).

2.3.2 Gegevensverzameling en –analyse

De gegevensverzameling gebeurde via een combinatie van secundaire en primaire dataverzameling. In de praktijk waren vooral volgende methodes van belang:

- studie en analyse van basisdocumenten (het interventievoorstel, de baseline (indien uitgevoerd), uitvoeringsrapporten en evaluaties, documenten m.b.t. M&E);
- interviews met de bij de uitvoering van de interventies betrokken sleutelactoren;
- focusgroep discussies (op actorniveau en interventieniveau).

Op basis van de verzamelde gegevens werd voor elke interventie het analysekader ingevuld d.m.v. een scoring van componenten en items. Om de subjectiviteit te beperken werd elke interventie (m.u.v. 5 interventies in België) door minstens twee evaluatoren gescoord en werd, bij verschillende scores, via discussie naar overeenstemming gezocht. De scores werden vervolgens samengebracht in tabellen waarop een aantal eenvoudige en meer complexe statistische analyses werden uitgevoerd (zie bijlagen 2 en 7), die de onderbouw hebben geleverd voor de bevindingen die in de twee volgende hoofdstukken worden voorgesteld.

2.3.3 Studiefasen: overzicht van de voornaamste stappen

De totale looptijd van de studie tot en met de afwerking van het eindverslag heeft ongeveer acht maanden in beslag genomen (van eind februari tot midden oktober 2015). De opstartperiode van de studie heeft relatief veel tijd gevraagd omwille van het vrij atypische onderwerp van de studie en de initiële weerstand bij een aantal actoren (zie hoger). Verhoudingsgewijs veel tijd werd besteed aan een bespreking van de basisconcepten en benadering van de studie, onder meer via de uitwerking van een verklarende nota. De studie van de basisdocumenten vond vooral plaats in de periode maart – mei, met een piloot-veldbezoek naar Rwanda in april-mei en bezoeken naar Benin en DRC in mei-juni. De projecten in België werden vooral in juni en juli geanalyseerd. De redactie van het draft syntheserapport vond voornamelijk in juli en augustus plaats, waarna het draft eindrapport eind september in het begeleidingscomité werd besproken.

Op cruciale scharniermomenten in het proces werden vergaderingen gehouden met een begeleidingscomité: bij de opstart van de studie, na het eerste veldbezoek en na de afwerking van het draft syntheseverslag.

3. Voornaamste globale bevindingen

Globaal overzicht

Onder dit hoofdstuk bekijken we de bevindingen vanuit een *globale* bril, d.w.z. voor wat betreft de 40 interventies *in hun geheel*. Deze globale analyse leidt tot globale conclusies die we naderhand (hoofdstuk 4) in sommige gevallen zullen moeten nuanceren naargelang het betrokken land, interventiekanaal of het niveau van complexiteit (van de TOC) van de interventie.

Onderstaande tabel (tabel 3) geeft een overzicht van de gemiddelde scores en de evalueerbaarheidsindex m.b.t. de acht componenten en vijf evaluatiecriteria die hiervoor (zie 2.2) reeds kort werden voorgesteld. Zoals reeds eerder aangegeven, werd elke component geoperationaliseerd in een aantal items (het aantal staat tussen haakjes vermeld) en zijn de aangegeven scores per component *niet* het gemiddelde van de item-scores per component, maar wel van de scores door de onderzoekers toegekend op het niveau van elke component en dit voor de 40 interventies.

Voor elke component (en elk item) werd een score gegeven die kon variëren van 1 tot 5, waarbij de waarde 1 werd toegekend voor een slecht presteren en een score 5 voor een goed presteren. Voor de eerste component (*De onderliggende analyse*) werd zo – bijvoorbeeld - een score 1 gegeven als zulke analyse niet werd uitgevoerd/niet werd geschreven, terwijl een score 5 werd gegeven indien deze analyse goed en volledig was en de doelen van de interventie ondersteunde. Een score 3 betekende dan dat de analyse goed was ontwikkeld, maar onvolledig (bijvoorbeeld geen analyse van de genderrelaties). In sommige gevallen (de gearceerde cellen) werd geen score toegekend omdat er geen (of slechts een erg indirecte) relatie bestond tussen de component en het evaluatiecriterium. De meer specifieke resultaten op component- en itemniveau worden in detail voorgesteld en besproken verderop in dit hoofdstuk.

In deze tabel en in meerdere gelijkaardige tabellen verderop in het rapport, maken we gebruik van kleuren om de verschillen in scores duidelijker aan te geven. Deze kleuren laten toe om *een eerste en erg globaal inzicht* te krijgen in de relatieve evalueerbaarheid van de interventies. Hoe hoger de score of index, hoe hoger de evalueerbaarheid. Van belang is hierbij om te beseffen dat de waarde van de index *niet in absolute termen* kan worden begrepen, maar alleen kan gebruikt worden in *relatieve* zin: om verschillende componenten en criteria *onderling* te vergelijken en om de 40 interventies *onderling* te vergelijken. Zo kunnen we bijvoorbeeld vaststellen dat van alle componenten de 'bredere context' het beste scoort, en '*het voorgestelde M&E systeem*' het minst goed. We kunnen eveneens vaststellen dat 'impact' het criterium is met (voor bijna elke component) de laagste score. Eveneens laat de tabel toe om snel de zwakste score te ontdekken: het voorgestelde M&E systeem op impact niveau.

Tabel 3: Evalueerbaarheidsindex per DAC criterium en component voor de 40 interventies[20]

Item	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
Dimensie 1 (interventieplan) (°)	**3,11**	**3,22**	**3,29**	**2,52**	**3,15**	**3,16**
1.1 De onderliggende analyse (7)	3,65	3,60				3,63
1.2 De interventielogica en veranderingstheorie (8)		3,00	3,50	2,10	3,25	2,96
1.3 Het voorgestelde M&E systeem (9)	2,45	2,83	2,85	1,95	2,46	2,51
1.4 Consistentie en adaptatie van de interventielogica en de veranderingstheorie	3,55	3,44	3,52	3,50	3,75	3,54
Dimensie 2 (uitvoeringspraktijk)	**2,99**	**3,25**	**3,51**	**2,33**	**2,73**	**2,96**
2.1 Basisinformatie m.b.t. de interventie uitvoering (11)	2,95	3,20	3,53	2,13	2,62	2,88
2.2 M&E systeem in de praktijk (12)	3,03	3,30	3,50	2,53	2,85	3,04
Dimensie 3 (context)	**3,89**	**3,98**	**3,98**	**3,84**	**3,89**	**3,91**
3.1 Houding van de sleutel actoren (9)	3,68	3,70	3,70	3,58	3,63	3,66
3.2 Bredere context 3)	4,10	4,25	4,25	4,10	4,15	4,17
Globale score evalueerbaarheid (°°)	**3,26**	**3,38**	**3,52**	**2,70**	**3,13**	**3,23**

(°) De scores m.b.t. de evalueerbaarheid op dimensie-niveau zijn ongewogen gemiddeldes van de scores op componentenniveau.
(°°) Voor de globale score evalueerbaarheid werd een gewogen gemiddelde berekend waarbij dimensie 1 en 2 even zwaar worden gewogen en dimensie 3 meetelt voor de helft van elk van de andere dimensies.

Index 4,01 – 5,00
Index 3,01 – 4,00
Index 2,01 – 3,00
Index 1,01 – 2,00

De tabel hierboven laat ons toe om reeds een aantal **globale bevindingen** te formuleren voor wat betreft de verschillende dimensies en componenten van het studiekader en de evaluatiecriteria:

- in eerste instantie valt op dat met uitzondering van dimensie 3 (de context) geen enkele dimensie, component of criterium een echt goede score behaalt: de maximale scores situeren zich rond 3,50 en vele scores zitten in de 'gele zone', d.w.z. onder het middelpunt van onze schaal. Dit duidt erop dat er globaal nog een substantiële marge voor verbetering is;
- op het vlak van de drie **dimensies**, kunnen we vaststellen dat dimensie 3 (de context) aanmerkelijk beter scoort dan de twee andere dimensies en dat van die twee dimensies de uitvoeringspraktijk (dimensie 2) zwakker scoort dan het interventieplan (dimensie 1)[21];
- op het vlak van de **componenten** valt weerom de goede score op van de twee componenten die deel uitmaken van de context; de zwakste componenten zijn het voorgestelde M&E systeem (dimensie 1) en de basisinformatie m.b.t. de interventie-uitvoering (dimensie 2);
- voor wat de **evaluatiecriteria** betreft, ligt de globale evalueerbaarheidsscore het laagste voor het impactcriterium; verder scoort ook duurzaamheid relatief zwak en scoren effectiviteit en vooral efficiëntie relatief hoog. Eigenlijk zorgen de bijna

[20] Niet alle 40 interventies konden voor elke component en elk DAC criterium gescoord worden. In tabel A1.2 van bijlage 7 wordt aangegeven wanneer en in welke mate minder interventies gescoord werden.

[21] Later (zie deel 3 van dit hoofdstuk) zullen we evenwel belangrijke opmerkingen plaatsen bij de goede score voor de componenten en items onder 'Context'.

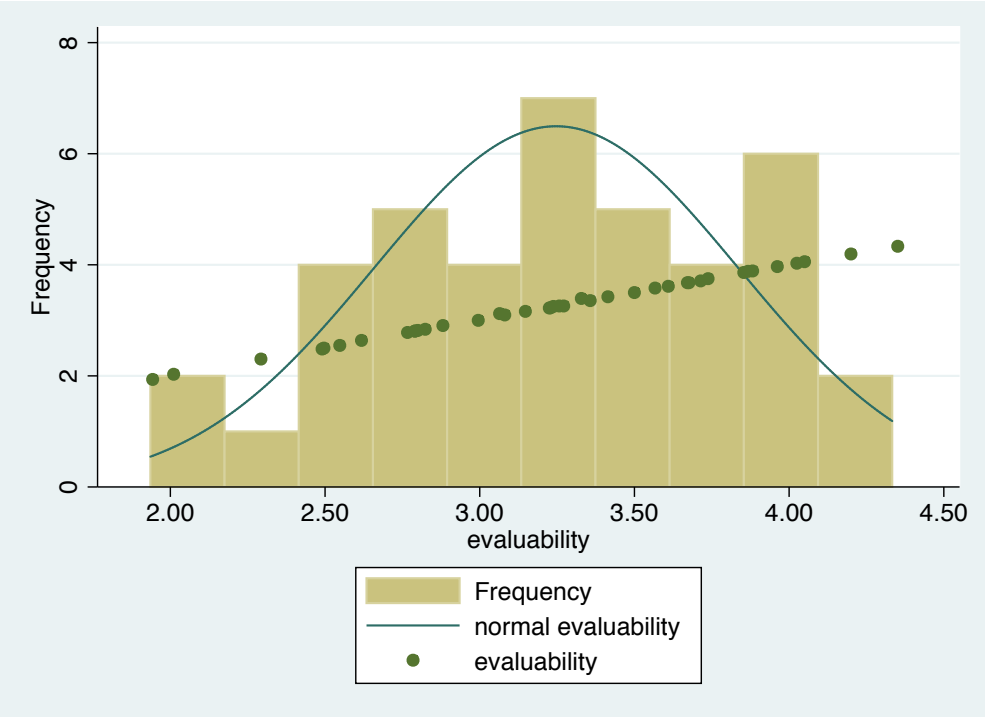
gelijke scores – voor alle evaluatiecriteria – van de derde dimensie (context) ervoor dat de relatieve verschillen tussen de vijf criteria, en dan vooral tussen impact enerzijds en de vier andere criteria anderzijds, globaal een beetje worden afgezwakt. Met andere woorden, als we met dimensie 3 (de context) geen rekening houden, zijn de verschillen tussen de evaluatiecriteria meer uitgesproken. Verder valt het volgende op:

- de evalueerbaarheid van *efficiëntie* blijkt relatief goed te scoren met voor bijna alle componenten een redelijke (groene) score; de enige zwakkere component is hier het (in het interventieplan) voorgestelde M&E systeem, maar dit wordt gecompenseerd door een vrij goede score m.b.t. het M&E systeem in de praktijk. De relatief goede score voor efficiëntie is een illustratie van wat we verder zullen omschrijven als een relatief goede M&E praktijk op operationeel niveau;
- wat de evalueerbaarheid van *effectiviteit* betreft, bevinden de meeste scores zich ook in de 'redelijke' (licht groene) zone, met uitzondering van 1.2 en daaraan gekoppeld 1.3 (interventielogica en veranderingstheorie; voorgesteld M&E systeem). Die relatieve zwakte duidt op een te beperkte aandacht voor met name het outcome niveau in het interventieplan;
- wat de *duurzaamheid* betreft, duiden de zwakke M&E systeem-scores (1.3 en 2.2) op een tekort aan aandacht voor duurzaamheid in het M&E systeem;
- wat het *impact*-criterium betreft, wijst de tabel aan dat de evalueerbaarheid van impact zeer matig tot zwak scoort voor alle componenten die met het interventiebeheer gelieerd kunnen worden (dimensies 1 en 2)[22]. Voor deze zwakke score zullen we hierna een aantal met elkaar verbonden oorzaken identificeren die in essentie te maken hebben met de beperkte aandacht die aan het impactniveau wordt geschonken door alle betrokken partijen. De lage evalueerbaarheidsscore voor impact ligt uiteraard in de lijn van wat kon verwacht worden gezien de methodologische uitdagingen van impactevaluatie maar ook, en belangrijker in het kader van deze studie, de relatief beperkte aandacht die door de actoren aan het impactniveau wordt geschonken.

Tenslotte hebben we globaal ook nagegaan hoe het gesteld is met de **distributie** van de globale evalueerbaarheidsscores van de 40 onderzochte interventies (zie figuur hierna). De figuur wijst op een (statistisch) vrij normale verdeling met de hoogste frequentie van interventies gesitueerd rond de gemiddelde score (3,23) van de evalueerbaarheidsindex. Opvallend is verder de relatief grote spreiding (minimale score van 1,94 en maximale score van 4,33) en het relatief grote aantal interventies (20% van de steekproef) met een hoge (4,00 of hoger) score.

[22] We maken hier abstractie van component 1.4 waar het aantal in ogenschouw genomen interventies (in totaal 10) te laag is om conclusies aan te kunnen verbinden.

Figuur 1: Distributie van de evalueerbaarheidsscores over de 40 interventies



3.1 Analyse van het interventieontwerp

Onder dit deel komen achtereenvolgens aan bod: de onderliggende analyse (7 items), de interventielogica en de veranderingstheorie (8 items), het voorgestelde M&E systeem (9 items) en de consistentie en aanpassing van de interventielogica en veranderingstheorie (3 items).

3.1.1 De onderliggende analyse

De kwaliteit van de onderliggende analyse heeft invloed op de kwaliteit van evaluaties die zich richten op relevantie en effectiviteit: zonder een goede analyse, inclusief een analyse van de (situatie van de) doelgroepen, kan een evaluatie moeilijk aantonen of een specifieke interventie relevant is voor de doelgroepen en daadwerkelijk het bedoelde effect heeft op de doelgroepen. Een (mogelijke) band tussen de onderliggende analyse en de evalueerbaarheid van andere DAC criteria (efficiëntie, impact en duurzaamheid) ligt evenwel minder voor de hand. Daarom werden bij deze component, anders dan voor de andere componenten, enkel scores gegeven voor de relevantie en effectiviteit.

Onder deze component worden zeven items besproken die hierna in tabel 4 met hun score zijn opgenomen.

Tabel 4: Voornaamste resultaten m.b.t. de onderliggende analyse

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heids Index
1.1. De onderliggende analyse	3,65	3,60				3,63
1.1.1 De doelgroepen zijn duidelijk afgebakend en beschreven	3,48	3,50				3,49
1.1.2. De rationale van de interventie en de (probleem)situatie van de doelgroep zijn duidelijk beschreven	3,85					3,85
1.1.3. De rol van de doelgroep(en) is duidelijk beschreven	3,60	3,55				3,58
1.1.4. De rol van de voornaamste betrokken actoren (exclusief doelgroep) is duidelijk	3,53	3,47				3,50
1.1.5. Genderanalyse is een integraal deel van de onderliggende analyse	2,40	2,40				2,40
1.1.6. De band tussen de onderliggende analyse en de interventiedoelen is duidelijk beschreven	4,00					4,00
1.1.7. De band tussen de rationale van de interventie en het sectorbeleid van het partnerland is duidelijk beschreven	4,10					4,10

***Belangrijke opmerking**: een meer uitgewerkte versie van de bovenstaande tabel (en van de hierna volgende gelijkaardige tabellen) is te vinden onder bijlage 8, waar elke score inhoudelijk wordt toegelicht. Deze inhoudelijke uitleg geeft een verantwoording van de score en is gebaseerd op een beoordelingsprotocol dat de onderzoekers hebben gehanteerd en er moest voor zorgen dat de onderzoekers op een consistente en meer objectieve wijze scoorden. Lezers die geïnteresseerd zijn in de feitelijke inhoud van elke score wordt dus aangeraden om bijlage 8 te raadplegen.*

In vergelijking met de zeven andere componenten scoort *de onderliggende analyse* globaal relatief goed; de andere componenten van deel 1 en deel 2 scoren lager. Alleen de twee componenten onder de evaluatie-context (deel 3) scoren hoger (zie Tabel 3). In vele gevallen is de aanpak het resultaat van een praktijk die over de jaren heen is gegroeid en vorm gegeven, waarbij er een taakverdeling is tussen de Belgische organisatie, (eventueel) haar lokale vertegenwoordiging en lokale actoren (zie box 1).

Box 1 : Onderliggende analyse en ontwikkeling van een M&E systeem bij Vredeseilanden Benin

Vredeseilanden heeft de strategische beslissing genomen om zich te specialiseren in de ondersteuning van enkele specifieke landbouwketens. De organisatie heeft een reeks specifieke en aan elke actor aangepaste procedures en instrumenten uitgewerkt om de coherentie van zijn terreinactie te waarborgen. Deze instrumenten laten toe dat de contextuele analyses in elke zone op coherente wijze worden uitgevoerd en bijdragen tot een proces van kapitalisatie.
Wat de onderliggende analyse betreft, heeft VECO verschillende documenten (zowel handleidingen als analyses) uitgewerkt die een grondige analyse van de context en de gekozen strategiëen (geconcretiseerd voor Benin via een regionaal planningsatelier) mogelijk maakt. Vervolgens worden in ateliers op landenniveau de indicatoren onderhandeld met de betrokken actoren. Het globale S&E beleid en systeem (inclusief de verschillende instrumenten) wordt evenwel op het hoofdkwartier in België uitgewerkt en naderhand op interventieniveau aangepast.

De bevindingen in tabel 4 laten zien dat op één uitzondering na *alle items* met betrekking tot de analyse relatief goed scoren voor zowel relevantie als effectiviteit, waarbij relevantie enigszins hoger scoort dan effectiviteit. Meer specifiek kunnen de verschillende scores als volgt worden geduid.

Doelgroepen worden in het algemeen vrij goed en eenduidig beschreven (items 1.1.1 tot en met 1.1.3). In vele gevallen worden doelgroepen evenwel niet duidelijk afgebakend, wat een evidente hypotheek legt op de latere evalueerbaarheid. De (probleem)situatie wordt veelal goed beschreven en eveneens hoe de interventie hiermee in verband staat, zonder dat er evenwel sprake is van uitgewerkte beschrijvingen. Vanuit de zorg voor een vermindering van de administratieve eisen, lijkt dit ook niet wenselijk, al vormt dit uiteraard wel een handicap vanuit het oogpunt van de evalueerbaarheid.

Met betrekking tot de beschrijving van de voornaamste betrokken actoren buiten de doelgroep (1.1.4) moet opgemerkt worden dat het verschil tussen doelgroepen en andere betrokken actoren niet altijd duidelijk is en in weinig voorstellen expliciet is terug te vinden, ook al omdat er niet daadwerkelijk wordt naar gevraagd. Vanuit het oogpunt van evalueerbaarheid zou het wenselijk zijn dat het geheel aan actoren beter beschreven wordt, evenals een meer gedetailleerde beschrijving van de intermediaire en uiteindelijke doelgroepen, inclusief hun rol. Tegelijkertijd moet rekening worden gehouden met het feit dat sommige interventies erg complex zijn, waarbij soms veel actoren betrokken zijn, wat de onderliggende analyse lastig maakt.

De integratie van een genderanalyse in de onderliggende analyse (1.1.5) is het zwakste scorende item binnen deze component. Deze zwakke scores zijn niet verrassend en bevestigen de bevindingen van de recent uitgevoerde evaluatie met betrekking tot de gendergevoeligheid van de Belgische ontwikkelingssamenwerking[23]. De afwezigheid van een genderanalyse kan de evaluatie van een interventie sterk ondermijnen, aangezien mogelijke redenen voor het falen van een interventie niet boven tafel (kunnen) komen. Meer specifiek, omdat gender relaties het gedrag van individuen sterk beïnvloeden, hebben deze relaties vaak een effect op de implementatie en resultaten van interventies. Het negeren van de invloed van gender heeft in veel gevallen negatieve consequenties voor de kwaliteit van evaluaties en de aangebrachte suggesties voor verbetering. Overigens valt het in dit verband op dat in vele interventies weinig aandacht besteed wordt aan sociale differentiatie binnen de doelgroep (niet alleen vanuit gender-standpunt) en dat doelgroepen te vaak als een homogene groep worden opgevoerd. Dit vormt een handicap bij de evaluatie van relevantie en effectiviteit omdat sociale differentiatie er kan voor zorgen dat een project meer of minder relevant is voor bepaalde (sub-)doelgroepen en dat een project via sociale mechanismen, die te weinig worden onderkend, ook qua effectiviteit verschillend kan scoren t.a.v. die verschillende groepen. Breder bekeken wordt op die wijze de tendens versterkt om niet, of te weinig, te kijken naar de effecten van interventies m.b.t. gelijkheid en gelijkwaardigheid (*equity*) terwijl deze principes toch centraal staan in de ontwikkelingssamenwerking[24].

De relatief hoge scores (binnen de 20% best scorende items) voor de beschrijving van de band tussen de onderliggende analyse en de interventie doelen (1.1.6) moeten enigszins genuanceerd worden. De indruk bestaat dat, zeker in het geval van een tweede of derde fase van een interventie of van programma's[25], de onderliggende analyse geschreven wordt in functie van doelstellingen die reeds voordien werden vastgelegd. Dit blijkt onder andere uit het feit dat verschillende alternatieve strategieën niet in overweging worden genomen, maar dat de analyse gericht wordt op één strategie die duidelijk is vanaf het begin. Refererend naar de klassieke PCM-cyclus, zouden we

[23] Zie Caubergs, L., Charlier, S., Holvoet, N., Inberg, L. and Van Esbroeck, D. (2014) *De lastige weg naar gelijkheid, Evaluatie Gender en Ontwikkeling in de Belgische Samenwerking: synthese rapport*. Brussel: Dienst Bijzondere Evaluatie van de Ontwikkelingssamenwerking, 133 blz.

[24] Terzijde kan worden opgemerkt dat gelijkheid/gelijkwaardigheid niet zijn ingebed in de vijf klassieke DAC criteria die tot op heden centraal staan in de meeste evaluaties.

[25] Met name bij NGAs bouwen programma's in vele gevallen verder op verworvenheden van het verleden. Dit is zeker niet negatief, maar houdt wel een aantal gevaren in zoals we hier aantonen.

kunnen stellen dat in vele gevallen er geen sprake is van een daadwerkelijke identificatie waarbij verschillende strategieën tegenover worden afgewogen en men dadelijk tot de formulering over gaat. Vanuit het standpunt van de evalueerbaarheid impliceert dit vooral bijkomende uitdagingen voor de evaluatie van de relevantie die zich in zulke situatie moet baseren op een analyse die niet breed en/of actueel is.

De band tussen de rationale van de interventie en het sectorbeleid van het partnerland (1.1.7) is in de meerderheid van de interventies duidelijk, wat zich vertaalt in de hoogste score binnen deze component, en een plaats binnen de 20% globaal best scorende items (zie ook tabel A3 2 van bijlage 7)[26]. Deze hoge score laat zien dat de verantwoordelijke actoren voor de interventies zich bewust zijn van het sectorbeleid en de interventies willen laten aansluiten bij dit beleid, zelfs als de toepassing (of toepasbaarheid) van dit beleid beperkt is. Waarschijnlijk speelt de toenemende ervaring binnen de Belgische ontwikkelingssamenwerking om te werken met meer globale meerjarige indicatieve programma's die streven naar meer coördinatie en aansluiting bij het (sector)beleid van het partnerland hier een rol. Ook de uitwisseling tussen de Belgische actoren en bestaande platformen in de partnerlanden kunnen een positieve invloed hebben.

3.1.2 De interventielogica en de veranderingstheorie

Onder deze component hebben we onderzocht in welke mate de interventielogica en de veranderingstheorie duidelijk omschreven zijn en tot op welk niveau (output, outcome, impact niveau)[27]. Zoals hierna doorheen de analyse van de verschillende items wordt aangegeven, is de kwaliteit van de interventielogica en de veranderingstheorie een belangrijke factor voor de evalueerbaarheid. De aandacht binnen ons analysekader voor het specifieke niveau tot waarop een interventielogica (alsook de cruciale schakels, assumpties, interne risico's, etc.) duidelijk en goed is uitgewerkt, heeft implicaties voor de evalueerbaarheid van de verschillende OECD/DAC criteria. M.a.w. een interventielogica die duidelijk en goed is uitgewerkt tot op het niveau van de outcomes maar niet tot op het impact-niveau bevordert de evalueerbaarheid van de effectiviteit maar impliceert een grote uitdaging voor de evalueerbaarheid van de impact. In totaal werden 8 items onder deze component geanalyseerd (zie tabel 5 op de volgende bladzijde).

De component *'interventielogica en veranderingstheorie'* heeft de op twee na laagste evalueerbaarheidsindex van de acht onderzochte componenten (zie Tabel 3). Zoals hierna verder wordt aangegeven, is het grote pijnpunt de uitwerking van de interventielogica en veranderingstheorie tot op het niveau van de impact. M.a.w. de aandacht bij de meeste interventies is voornamelijk toegespitst op de lagere niveaus en op de uitwerking van de implementatie-theorie terwijl de hogere niveaus, en meer specifiek het impactniveau (dat in onze analyse overeenkomt met het niveau van de algemene doelstellingen uit het logisch kader), vaak verwaarloosd worden. Desalniettemin zouden effecten op het impact niveau (met name de effecten die *rechtstreeks* voortvloeien uit de outcomes) toch al zichtbaar moeten zijn na het afsluiten van een interventie van drie jaar (duur van veel interventies overeenkomstig de financieringscyclus) en zeker na een vervolg-interventie van dezelfde duur.

[26] Deze vaststelling lijkt op het eerste gezicht niet zo indrukwekkend, maar is niettemin belangrijk omdat de meerderheid van de best scorende items tot dimensie 3 (de context) behoren); zie eveneens tabellen A3 1 en A3 2 van bijlage 7 voor meer details.

[27] Deze drie niveaus komen overeen met de outputs (tussentijdse resultaten), de specifieke doelstelling en de algemene doelstelling(en) van het logisch kader. Zie ook hoofdstuk 2.2 voor de invulling die in deze studie aan de vijf DAC criteria werd gegeven.

Tabel 5: Voornaamste resultaten m.b.t. het interventielogica en de veranderingstheorie

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaarheids-Index
1.2 De interventielogica en veranderingstheorie	2,80	3,00	3,50	2,10	3,25	2,96
1.2.1 Er is een duidelijk en correct onderscheid tussen outputs, outcomes en impact		3,65		2,65		3,15
1.2.2 De veranderings-theorie vanaf inputs tot outcomes en de uiteindelijke impact is duidelijk uitgewerkt		3,35		1,90		2,63
1.2.3 De veranderings-theorie is logisch en realistisch		4,15	4,40	3,20		3,92
1.2.4 Kritische en cruciale schakels in de ketting worden geïdentificeerd en kunnen worden getest		3,05	3,35	1,87		2,76
1.2.5 De veranderings-theorie omvat concrete maatregelen die de duurzaamheid van de baten van de interventie moeten waarborgen					3,50	3,50
1.2.6 Interne risico´s zijn duidelijk aangegeven en onderzocht/ ingeschat		3,00	3,10		2,80	2,97
1.2.7 Externe assumpties zijn duidelijk aangegeven en onderzocht/ ingeschat	2,80	3,10	3,20	2,15	2,90	2,83
1.2.8 De toewijzing van de voorziene middelen aan de outputs is duidelijk			3,64			3,64

De bevindingen in tabel 5 laten verder zien dat voor deze component het **effectiviteit** en **efficiëntie niveau** in het algemeen relatief goed zijn uitgewerkt. Relatief veel interventies scoren sterk op hun veranderingstheorie die logisch en realistisch is tot het output (70%) en outcome (60%) niveau; dit item (1.2.3) hoort overigens tot de 20% best scorende items. Daarbij worden kritische en cruciale schakels relatief vaak geïdentificeerd tot beide niveaus. Ook geven veel interventies interne risico´s en externe assumpties aan, hoewel deze lang niet altijd verder onderzocht worden. Hoewel de verschillen tussen bovengenoemde percentages voor output en outcome niveau niet zeer groot zijn, is het duidelijk dat het output niveau (efficiëntie), in het algemeen, beter is uitgewerkt dan het outcome niveau (effectiviteit).

Het feit dat een substantieel deel van de interventies cruciale elementen specifiëren op het niveau van outcomes en voornamelijk outputs is met name positief voor *leren op interventie-niveau*. Vooral voor de direct betrokken interventiestaf en de verantwoordelijke organisatie is het belangrijk inzicht te krijgen in die elementen die essentieel zijn bij de implementatie van een interventie en dus best overgenomen worden in toekomstige interventies, terwijl andere elementen die eerder bijkomstig blijken, weggelaten kunnen worden. Vooral in het geval van innovatieve interventies waar men onvoldoende inzicht heeft in wat werkt en wat niet, en waar nog onzekerheid bestaat over de meest aangewezen implementatie-modaliteiten, is de identificatie van de kritieke en cruciale elementen essentieel[28]. De identificatie en het testen van externe assumpties is dan weer voornamelijk van belang voor het afleggen van verantwoordelijkheid

[28] Dit sluit enigszins aan bij het idee van '*structured experiential learning*', zie Pritchett, L., Samji, S. and Hammer, H. (2013) "It's all about MeE: using structured experiential learning to crawl the design space", *Center for Global Development Working Paper 322*. Washington, D.C.: Centre for Gobal Development.

(*accountability*), aangezien het duidelijk maakt welke factoren van buitenaf invloed (kunnen) uitoefenen op de interventie.

Zoals hoger reeds aangegeven scoort de volgende stap in de veranderingstheorie, het **impact niveau**, duidelijk zwakker. Om precies te zijn, de zwakste elementen binnen de interventielogica en de veranderingstheorie betreffen in het bijzonder het impact criterium. Bij relatief veel interventies wordt de veranderingstheorie niet uitgewerkt tot op impactniveau en waar het impactniveau wordt gespecifieerd, bevindt het zich dikwijls op grote afstand van de outcomes waardoor een *'missing middle'* ontstaat en het onmogelijk wordt om de bijdrage van een individuele interventie tot die impact te bepalen[29]. Dit effect wordt versterkt doordat er vaak ook geen kritieke en cruciale schakels, noch externe assumpties worden gedefinieerd tussen het outcome en impactniveau (de score van dit item (1.2.4) behoort tot de 20% laagste van alle scores). Het is duidelijk dat dit de mogelijkheid tot en de kwaliteit van impactevaluaties ondermijnt, aangezien juist de analyse van deze elementen het mogelijk maakt te differentiëren tussen problemen op implementatieniveau, gebreken in de onderliggende veranderingstheorie en invloeden van andere externe factoren die buiten de controle van de interventie liggen. Bovendien kan vooral via de identificatie en het testen van kritische en cruciale schakels inzicht verkregen worden in de redenen waarom een interventie impact genereert of niet. Dit ondermijnt voor een stuk de leer/feedbackfunctie op het boven-interventieniveau die met name voor beleidsmakers (en andere actoren die niet direct betrokken zijn bij de implementatie van specifieke interventies) interessante inzichten m.b.t. outcome en impact moet opleveren. Tenslotte is het eveneens zo dat het uitwerken van een TOC tot op het impact-niveau ook inzicht kan bieden in die implementatie-modaliteiten die het meest effectief en efficiënt zijn en die de grootste kans bieden op het realiseren van impact (zonder dat er daarom daadwerkelijk ook steeds een evaluatie van de impact moet plaatsvinden).

De zwakke scores voor het impact criterium kunnen wellicht voor een stuk verklaard worden door het feit dat er - zeker bij kleine interventies of organisaties maar ook in het algemeen - weinig *incentives* zijn om het implementatieniveau te overstijgen: het uitwerken van een interventielogica op implementatieniveau heeft een veel groter direct nut voor de interventie/organisatie zelf en het is het niveau dat de interventie/ organisatie het meest onder controle heeft. Anderzijds zijn de linken met de hogere niveaus vaak onduidelijk en minder beheersbaar, is de identificatie van kritieke elementen en cruciale schakels op hogere niveaus moeilijker en noodzaakt die vaak een meer doorgedreven sector- en context-analyse. Bovendien werd er tot voor kort ook relatief weinig belang gehecht aan impact-theorie door de financierende instanties. Dit laatste blijkt onder meer uit het feit dat de formats van de projectvoorstellen toelaten dat aan dit niveau geen of weinig aandacht wordt besteed; evenmin wordt in het beoordelingsformat van DGD voor Zuid-acties aandacht besteed aan impact, mogelijks omdat financierende organisaties nog steeds voornamelijk in *"verantwoording"* zijn geïnteresseerd en daarnaast beseffen dat interventies slechts activiteiten en outputs daadwerkelijk onder controle (kunnen) hebben. Zoals verder in dit rapport zal worden aangetoond, heeft het ontbreken van een volledig uitgewerkte veranderingstheorie ook invloed op het M&E systeem (zowel op papier als in de praktijk), en zet het voornamelijk de evalueerbaarheid van impact zwaar onder druk.

Hoewel de manier waarop de interventies zijn uitgewerkt een negatief effect heeft op de evalueerbaarheid van de impact, zijn de scores voor **duurzaamheid** iets positiever. De meeste interventies formuleren concrete maatregelen en zelfs outputs die de duurzaamheid van de baten van de interventie moeten waarborgen. Dit weerspiegelt de toenemende zorg van organisaties om de resultaten van hun interventie te behouden na afloop van de interventies. Bij de inschatting van interne risico´s en externe assumpties wordt echter in veel mindere mate rekening gehouden met duurzaamheid: 28% van de

[29] Hiervoor bestaan oplossingen, zoals het invoegen van een extra niveau in de veranderingstheorie en het werken met intermediate and final outcomes om een tussenstap tussen outcomes en impact uit te werken. We komen hierop terug in de aanbevelingen.

interventies heeft interne risico´s gerelateerd aan duurzaamheid aangegeven en onderzocht, slechts 18% heeft dit gedaan voor externe assumpties.

Tenslotte is het zo dat alle interventies informatie verschaffen met betrekking tot de voorziene middelen en dat in slechts 10% van de interventies de toewijzing van de voorziene middelen/kosten aan de outputs helemaal niet duidelijk is. De goede score op dit item heeft een positief effect op en is eigenlijk cruciaal voor de evalueerbaarheid van de efficiëntie en is voor een stuk te danken aan de toenemende aandacht voor dit criterium bij de financiële formats opgelegd door DGD.

3.1.3 Het voorgestelde M&E systeem

Onder deze component hebben we onderzocht in welke mate het interventieontwerp informatie aanreikt m.b.t. het M&E systeem. De kwaliteit van het (voorgestelde) M&E systeem is om evidente redenen een belangrijke factor van evalueerbaarheid: hoe beter dit systeem is uitgewerkt, hoe hoger zijn potentie om informatie te genereren die de evalueerbaarheid ten goede zal komen.

Het analyseren van de negen items onder deze component heeft ons in staat gesteld een inzicht te krijgen in het voorgestelde M&E systeem. Daarbij is het belangrijk rekening te houden met het gegeven dat het inzicht in het *voorgestelde* M&E systeem nog niets zegt over de (kwaliteit van de) *daadwerkelijke* toepassing van dit systeem (deze daadwerkelijke toepassing komt in deel 2.2 aan bod).

In tabel 6 (zie volgende bladzijde) worden de voornaamste bevindingen kort in tabelvorm voorgesteld, zoals dit reeds voor de twee vorige componenten is gebeurd.

Het voorgestelde M&E systeem is de component met de laagste evalueerbaarheidsindex van de acht onderzochte componenten, en dit voor de vijf DAC evaluatiecriteria; zeven (!) van de negen items onder deze component behoren – minstens voor één criterium en dikwijls voor alle criteria - tot de groep van de 20% zwakst scorende items (zie ook tabel A3 3 van bijlage 7) . Vanuit het standpunt van de theoretische evalueerbaarheid hoeft dit *als zodanig* niet echt een probleem te zijn. Ex ante is vooral de kwaliteit van de onderliggende analyse en die van de interventielogica en veranderingstheorie belangrijk voor de evalueerbaarheid. De kwaliteit van het M&E systeem is vooral vanuit het standpunt van de praktische evalueerbaarheid belangrijk, maar dan gaat het veeleer om de M&E *praktijk* en niet zozeer over het M&E systeem op papier, zoals dit is voorgesteld in het interventievoorstel. Echter zal later blijken dat er voor vier criteria een sterke correlatie bestaat tussen de kwaliteit van het M&E ontwerp zoals voorgesteld in het interventievoorstel en de uiteindelijke M&E praktijk (zie bespreking van component 2.2 hierna). Vanuit deze optiek is de (kwaliteit van de) uitwerking van een M&E systeem in de beginfase een goede indicator voor de latere praktijk.

Tabel 6: Voornaamste resultaten m.b.t. het voorgestelde M&E systeem

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
1.3 Het voorgestelde M&E systeem	2,45	2,83	2,85	1,95	2,46	2,51
1.3.1 De voornaamste beoogde resultaten van de interventie zijn goed geoperationaliseerd		3,15	3,65	1,83	2,72	2,84
1.3.2 Waar nodig/relevant zijn de indicatoren sekse-specifiek en/of uitgesplitst naar andere relevante parameters		1,82	1,81	1,34	1,63	1,65
1.3.3 Het voorgestelde M&E systeem bevat een consistente vertaling van de onderliggende interventielogica en veranderingstheorie	3,00	3,45	3,65	2,30	2,69	3,02
1.3.4 De werkwijze om het bereiken van de interventie-resultaten en hun duurzaamheid op te volgen en te evalueren is duidelijk beschreven	2,70	3,15	3,28	2,13	2,62	2,78
1.3.5 De werkwijze om de assumpties op te volgen is duidelijk beschreven	1,85	1,85	1,90	1,55	1,82	1,79
1.3.6 De werkwijze om de interne risico's op te volgen is duidelijk beschreven	1,70	1,80	1,85	1,45	1,68	1,70
1.3.7 De personele en financiële middelen van het M&E systeem zijn duidelijk beschreven			2,40			2,40
1.3.8 Het voorgestelde MIS laat een toewijzing van uitgaven aan specifieke outputs en interventie-componenten toe			3,67			3,67
1.3.9 De wijze waarop het M&E systeem van de interventie zich afstemt/ verhoudt tot het nationale/ lokale M&E systeem is duidelijk beschreven	1,57	1,69	1,63	1,57	1,57	1,61

Er kunnen verschillende verklaringen worden aangegeven voor de zwakke score van deze component. Vooreerst zijn de eisen van de donor t.a.v. de beschrijving van het M&E systeem in het interventievoorstel niet sterk uitgewerkt; er wordt blijkbaar – terecht of ten onrechte – voor andere aspecten meer aandacht gevraagd. Interventiepromotoren investeren daarom waarschijnlijk relatief weinig in dit onderdeel. Deze hypothese wordt bevestigd door het feit dat de scores voor de eigenlijke M&E praktijk aanzienlijk (ongeveer 20%) hoger liggen (zie tabel 3). Ook hebben promotoren vaak niet de gewoonte om hun M&E systeem omstandig te omschrijven. Tenslotte speelt ook mee dat vanuit mentaal standpunt men pas echt aan M&E aandacht besteed van zodra de uitvoering is gestart; zoals we later zullen zien, zijn er een aantal interventies die hun M&E systeem pas bij de start daadwerkelijk beginnen te definiëren en uit te bouwen. Tenslotte speelt dat verschillende organisaties op het ogenblik van de studie (voorjaar 2015) pas vrij recent een M&E strategie hadden uitgewerkt (of daarmee nog bezig waren), waarbij de operationalisering op het terrein achter loopt.

Wanneer we naar de vijf DAC evaluatiecriteria bekijken, dan blijkt – niet verrassend – dat 'impact' het zwakste scoort, en effectiviteit en efficiëntie het beste, waarbij dient te worden opgemerkt dat de score voor deze beide criteria zich nog onder de midden-score van 3 bevindt. De verklaring voor de relatief hogere score voor effectiviteit en efficiëntie ligt uiteraard in het feit dat M&E systemen in eerste instantie gericht zijn op het te bereiken specifieke objectief en, vooral, op de te realiseren outputs, en – hiermee verbonden - ook aan het relatief grotere belang voor monitoring in vergelijking tot evaluatie in het M&E beleid van de meeste organisaties. Verder speelt ook de geringe aandacht voor impact in de interventielogica en veranderingstheorie (zie de bespreking

van component 1.2 hierboven) een belangrijke rol. Dat de scores niet echt goed zijn, heeft tenslotte ook te maken met het feit dat in een aantal gevallen blijkbaar de basisvaardigheden m.b.t. de formulering van goede indicatoren onvoldoende worden beheerst[30].

De stelling dat M&E systemen zich vooral richten op de outputs en onderliggende activiteiten en middelen wordt ook bevestigd door de relatief goede score, op het niveau van efficiëntie, voor wat betreft de operationalisering van de beoogde resultaten (1.3.1). Anderzijds is de score voor impact ook hier erg zwak, vooral omdat in vele voorstellen helemaal geen aandacht aan impact wordt besteed. Het voorgestelde M&E systeem vormt in die zin een voorafspiegeling van de M&E praktijk (zie hoofdstuk 2.2 hierna) waar we zullen vaststellen dat M&E systemen vrij goed zijn uitgebouwd op het operationele niveau (monitoring in de enge zin), maar dat hun kwaliteit en omvang vermindert naarmate men hoger gaat in de doel-middelen ketting en een bredere focus moet hanteren.

Ook wat betreft de vertaling - in het M&E systeem – van de onderliggende interventie-logica en veranderingstheorie (item 1.3.3), zijn de scores het beste op het vlak van efficiëntie en effectiviteit. M&E systemen blijken ook tot op zekere hoogte aandacht te hebben voor activiteiten en outputs die de duurzaamheid van de interventies moeten bevorderen (item 1.3.4); de betreffende scores zijn niet echt goed, maar wijzen niettemin op een expliciete aandacht voor duurzaamheid in vele interventies. Ook hier kunnen we een sterke parallel ontdekken met de bevindingen m.b.t. duurzaamheid bij de analyse van de vorige component (zie bespreking component 1.2 hierboven).

De (erg) zwakke scores m.b.t. het gebruik van sekse-specifieke indicatoren (en andere factoren van sociale differentiatie) (1.3.2) en de opvolging van externe assumpties (1.3.5) zijn tot op zekere hoogte te verklaren door het gebrek aan aandacht voor deze punten in de onderliggende analyse en de uitwerking van de interventielogica en de veranderingstheorie (zie bespreking van componenten 1.1 en 1.2 hierboven). Blijkbaar klopt ook hier de hypothese dat als deze elementen niet van in het begin af aan worden meegenomen (bijvoorbeeld via sekse-specifieke baseline gegevens en risico's), ze later (te) weinig aan bod komen. Anderzijds geldt ook het omgekeerde, zoals de *good practice* in Box 2 illustreert.

Box 2: Risico-analyse in het PASAB II project (Caritas)

In de projectdocumenten van PASAB II (*Projet d'Appui à la Sécurité Alimentaire au Bugesera*), een interventie van de NGO Caritas in Rwanda, worden per beoogd resultaat duidelijke risico's aangegeven die het behalen van de resultaten mogelijk beïnvloeden. Er wordt verder een risicoanalyse uitgevoerd waarbij voor elk van de risico's de mogelijke effecten worden aangegeven alsook acties en wijzigingen in het programma die zullen ondernomen worden in geval de risico's zich effectief realiseren.

Naast de opvolging van de resultaten in een format waar plaats voorzien is om kwalitatieve commentaren toe te voegen, worden de risico's ook opgevolgd via een specifieke format waarbij de voorziene risico's worden vergeleken met de actuele situatie op het terrein. Deze documenten bevatten nuttige informatie die ook kan gebruikt worden bij evaluaties omdat ze inzicht geven in achterliggende redenen voor het falen/ slagen van interventies.

De eveneens erg zwakke score voor wat betreft de werkwijze om interne risico's op te volgen (1.3.6) heeft o.i. te maken met het feit dat weinig organisaties geneigd zijn om deze risico's te beschrijven in hun interventievoorstellen (zie ook de bespreking van component 1.2 hierboven). Dit standpunt valt de begrijpen in de zin dat de organisaties weten dat hun voorstellen grondig zullen gescreend worden en daarom willen vermijden

[30] Met name wordt erg weinig rekening gehouden met de eis dat indicatoren 'specifiek' moeten zijn (de "S" van SMART): de indicator moet met name op 'specifieke' wijze refereren naar de via het objectief (output, outcome, impact) beoogd resultaat.

zich zwak op te stellen. Verder zijn er ook organisaties die stellen dat deze risico's behoren tot de interne keuken, geen rechtstreeks verband houden met de kwaliteit en uitvoering van de programma's en daarom niet (hoeven) gedeeld (te) worden. Deze argumenten lijken steek te houden als men de scores vergelijkt met de kwaliteit van de informatie m.b.t. praktijk van risicoanalyse die een betere score geeft (zie verder, item 2.1.9). Anderzijds is het vanuit het standpunt van evalueerbaarheid wenselijk en belangrijk dat informatie beschikbaar is over deze interne risico's, de wijze waarop men hiermee wil omgaan en hoe ze zullen worden opgevolgd.

De personele en financiële middelen die beschikbaar zijn voor het M&E systeem zijn in het algemeen niet of slechts erg beperkt beschreven (1.3.7). Slechts in een kwart van de voorstellen vinden we hierover substantiële informatie. Dikwijls worden bijvoorbeeld geplande evaluaties, zelfs wanneer die van enige omvang zijn, niet specifiek in het budget vermeld. Ook hier geldt waarschijnlijk dat men deze informatie niet cruciaal acht om op te nemen in een interventievoorstel. Vanuit evalueerbaarheidsstandpunt impliceert dit dat men weinig inzicht heeft in het relatieve belang dat ex ante aan M&E wordt gehecht.

De wijze waarop het M&E systeem van de interventie zich afstemt/verhoudt tot het nationale/ lokale M&E systeem is in de meeste gevallen niet duidelijk beschreven, zoals blijkt uit de erg lage score op dit item (1.3.9). Deze lage score is gedeeltelijk te verklaren vanuit het feit dat een aantal interventies erg klein zijn of met dingen bezig die op hoger vlak niet in M&E systemen worden geïntegreerd. Verder spelen ook andere factoren mee zoals het ontbreken, de twijfelachtige betrouwbaarheid of het niet functioneren van zulke systemen op nationaal en decentraal vlak (bijvoorbeeld in DRC) waardoor ze niet als adequaat worden beschouwd om te voldoen aan de opwaartse accountability tegenover de financierende overheid. Niettemin is het ook duidelijk dat te weinig interventies daadwerkelijk nagaan in welke mate hun M&E systeem, en dan met name de opvolging van sleutelindicatoren op outcome en impact niveau, zou kunnen aansluiten op wat op hogere en decentrale niveaus (en eventueel bij partner-organisaties) gebeurt[31].

De hoogste score onder deze component vinden we terug voor het item *"het voorgestelde MIS laat een toewijzing van uitgaven aan specifieke outputs en interventie-componenten toe"* (1.3.8). Vanuit het standpunt van evalueerbaarheid is dit een belangrijk verworvenheid die te maken heeft met de manier waarop donor budgetten wil opgemaakt zien en met het belang dat wordt gehecht aan een goed uitgewerkt budget in het voorstel. Door een format op te leggen dat verschillende categorieën van uitgaven met outputs verbindt, wordt belangrijke informatie voor een efficiëntie-analyse beschikbaar gesteld.

Last but not least is het belangrijk om op te merken dat er in weinig interventies sprake is van een daadwerkelijk M&E *systeem*. In de voorstellen is meestal wel informatie te vinden m.b.t. de componenten van zo'n systeem, maar deze zijn niet (of slechts gedeeltelijk) op elkaar afgestemd en samengebracht in een coherent geheel. Zo komt het bijvoorbeeld geregeld voor dat de risicoanalyse (interne risico's en externe assumpties) belangrijke factoren vermeldt die niet in het logisch kader zijn terug te vinden, en omgekeerd. Het ontbreken van een daadwerkelijk *systeem* trekt zich overigens door in de M&E praktijk.

3.1.4 De consistentie en aanpassing van de interventielogica en veranderingstheorie

Voor de evalueerbaarheid van interventies is het essentieel dat eventuele veranderingen in de interventielogica en in onderliggende veranderingstheorie *duidelijk* worden aangegeven en aangepast in het M&E systeem. Het is ook belangrijk dat duidelijk wordt aangegeven waarom en hoe die veranderingen werden doorgevoerd. Indien

[31] We komen op de gevolgen van dit manco terug onder 2.2 van dit deel.

veranderingen niet (duidelijk) worden aangegeven en aangepast in het M&E systeem, ontbreekt relevante informatie met betrekking tot de werkelijke situatie, wat kan leiden tot foutieve conclusies over de kwaliteit van interventies (gerelateerd aan alle OECD/DAC criteria) en de factoren die een effect hebben op de implementatie en effecten van interventies. Als het M&E systeem niet wordt aangepast, impliceert dit ook dat een deel van de vernieuwde interventie niet wordt opgevolgd.

In ongeveer een kwart van de interventies zijn bij de uitvoering veranderingen doorgevoerd in de interventielogica en de onderliggende veranderingstheorie[32]. De meerderheid van deze aanpassingen beperken zich tot het output en de onderliggende niveaus.

Tabel 7: Voornaamste resultaten m.b.t. consistentie en adaptatie van de interventielogica en de veranderingstheorie

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
1.4 Consistentie en adaptatie van de interventielogica en de veranderingstheorie	3,55	3,44	3,52	3,50	3,75	3,54
1.4.1 Eventuele in interventielogica en onderliggende veranderingstheorie zijn duidelijk aangegeven en beargumenteerd	4,27	4,00	3,76	4,20	4,17	4,03
1.4.2 Er is informatie beschikbaar over de visie en meningen van de voor-naamste betrokkenen partijen m.b.t. de eventuele veranderingen in inter-ventielogica en veranderingstheorie	2,82	2,63	2,43	2,80	2,67	2,63
1.4.3 Eventuele veranderingen in interventielogica en veranderings-theorie zijn op adequate manier ver-werkt in het M&E systeem	3,18	3,25	3,10	3,00	3,33	3,17

Deze component scoort relatief goed vergeleken met de andere componenten van dimensie 1 en dimensie 2 (zie tabel 7). De scores voor de drie items laten echter een grote variatie zien; relatief sterke scores voor het aangeven en beargumenteren van eventuele veranderingen in de interventielogica en onderliggende veranderingstheorie (die voor een deel tot de best scorende 20% behoren) en relatief zwakke scores voor de beschikbaarheid van informatie over de visie en meningen van de voornaamste betrokkenen m.b.t. de eventuele veranderingen.

Het feit dat veranderingen in de meeste interventies zijn aangegeven, is op zichzelf een positieve vaststelling die de kwaliteit van het beheer en tegelijkertijd de wil om veranderingen te documenteren en te argumenteren illustreert. Hoewel de scores voor het adequaat aanpassen van veranderingen in het M&E systeem redelijk zijn, is het niettemin verontrustend dat een groot deel van de interventies waar veranderingen werden doorgevoerd (en aangegeven) deze niet hebben doorgetrokken in het M&E systeem. Zwakke individuele scores voor dit item hangen vaak samen met zwakke scores voor het gehele M&E systeem: daar waar de systemen zwak zijn, worden veranderingen meestal niet doorgevoerd in het M&E systeem.

[32] Het beperkt aantal interventies dat hier kon geanalyseerd worden, noopt tot voorzichtigheid m.b.t. de bevindingen.

3.2 Analyse van de praktijk m.b.t. uitvoering en beheer van de interventie en de context

Onder dit deel komen twee sterk uitgewerkte componenten aan bod: de (beschikbaarheid van) basisinformatie (11 items) m.b.t. de interventie-uitvoering en het M&E systeem in de praktijk (12 items).

3.2.1 De beschikbaarheid van basisinformatie m.b.t. de interventie-uitvoering

Onder deze component hebben we bekeken of de basisinformatie m.b.t. de interventie-uitvoering aanwezig is. Dit betreft zowel informatie die aanwezig zou moeten zijn bij de aanvang van een interventie (zoals interventievoorstel, baseline) als informatie met betrekking tot de voortgang van de interventie. De aanwezigheid van basisinformatie m.b.t. de interventie-uitvoering is essentieel voor de evalueerbaarheid, aangezien er zonder deze informatie moeilijk een vergelijking kan worden gemaakt tussen de begin-, tussentijdse en eindsituatie, waardoor de voortgang in de implementatie en de effecten van een interventie moeilijk aan te geven en dus ook moeilijk te evalueren zijn.

Tabel 8: Voornaamste resultaten m.b.t. de beschikbaarheid van basisinformatie m.b.t. de interventie-uitvoering

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
2.1. Beschikbaarheid van basisinformatie m.b.t. de interventie-uitvoering	2,95	3,20	3,53	2,13	2,62	2,88
2.1.1. De basisdocumenten zijn beschikbaar	4,03	4,05	4,13	3,73	3,95	3,97
2.1.2. Er is baseline informatie m.b.t. de doelgroep beschikbaar		3,50	3,90	2,45	2,84	3,18
2.1.3. Er is baseline informatie voor de indicatoren m.b.t. de counterfactual beschikbaar		1,28	1,28	1,44	1,29	1,32
2.1.4. Er is baseline informatie voor-handen waarbij de relevante indicatoren sekse-specifiek en/of naar andere relevante parameters zijn uitgesplitst		1,76	1,86	1,51		1,71
2.1.5. De te verwachten info is beschikbaar m.b.t. de voortgang van de realisatie van de interventiedoelen		3,37	3,93	1,69		3,04
2.1.6. Er is informatie beschikbaar over de participatie van de oorspronkelijke doelgroep	3,26	3,31	3,41	2,84	3,11	3,19
2.1.7. Er is informatie ter beschikking m.b.t. de gegevensverzameling van de indicatoren	2,73	3,08	3,28	2,28	2,69	2,82
2.1.8. De gegevensverzameling laat in principe betrouwbare verzameling m.b.t. de indicatoren toe	3,00	3,12	3,47	2,52	3,00	3,04
2.1.9. De informatie m.b.t. de op-volging van de interne risico's is beschikbaar en de eventuele gevolgen voor de interventielogica en uitvoering zijn aangeduid		2,40	2,75	2,13	2,86	2,54
2.1.10. De informatie m.b.t. de opvolging van de assumpties is beschikbaar en de eventuele gevolgen voor de interventielogica en uitvoering zijn aangeduid	2,21	2,37	2,60	2,20	2,45	2,37
2.1.11. De interventie-uitgaven zijn goed gedocumenteerd/ geregistreerd en kunnen aan de outputs toegewezen worden			3,95			3,95

De component '*Beschikbare basisinformatie m.b.t. de uitvoering van de interventie'* heeft van de 8 onderzochte componenten de op één na laagste score op de evalueerbaarheidsindex (zie tabel 3). Dit is enigszins verrassend omdat men kan veronderstellen dat vanuit het oogpunt van M&E "informatie beschikbaar hebben" een allereerste stap is. Zoals blijkt uit de bespreking van de verschillende items is er een grote variatie in de scores, met vooral zeer lage scores voor de aanwezigheid van basisinformatie inzake een 'counterfactual' (wat vooral de evalueerbaarheid van impact bemoeilijkt) en voor de beschikbaarheid van informatie die gedesaggregeerd is volgens sekse of andere relevante parameters. Een meer genuanceerde vaststelling is dat er wel degelijk veel informatie verzameld wordt en beschikbaar is, maar dat deze informatie sterk geconcentreerd is op lagere niveaus van de ketting en weinig betrekking heeft op enerzijds elementen die *buiten de interventie* vallen (zoals bv. de counterfactual, de externe assumpties, etc.) en anderzijds op het data collectieproces zelf (soort van meta-informatie). Deze vaststelling hangt grotendeels samen met eerdere bevindingen bij de componenten m.b.t. de interventielogica (zie 1.2) en het voorgesteld M&E systeem (zie 1.3) waarbij reeds duidelijk werd dat de focus grotendeels ligt op het operationele en het implementatieniveau. Dat er geen traditie noch duidelijk geformuleerde vereisten bestaan m.b.t. het verstrekken van informatie rond de concrete opzet van M&E (wijze van dataverzameling en verwerking) kan eveneens meespelen.

Zoals tabel 8 laat zien, is het sterkste item binnen deze component "de beschikbaarheid van basisdocumenten" (2.1.1)[33]. Basisdocumenten betreffen onder andere interventie-voorstellen, het technische en financiële dossier, baseline rapporten en voortgang-rapportages. Tot het niveau van outcomes bevatten deze documenten relatief vaak informatie over de voortgang van de realisatie van de interventiedoelen (2.1.5), waarbij een vergelijking wordt gemaakt met de beginsituatie. Deze laatste informatie is van belang om een stap verder te kunnen gaan dan monitoring. Deze informatie is echter om evidente redenen nauwelijks aanwezig voor het impact niveau. De sterke score met betrekking tot de aanwezigheid van de basisdocumenten moet echter enigszins worden genuanceerd, omdat de informatie in sommige gevallen alleen aanwezig is in een taal die een deel van de actoren niet beheerst. Daarnaast zijn de beschikbare documenten niet altijd bekend bij alle actoren, wat een beperkt "ownership" van de inhoud van deze documenten bij deze actoren impliceert. Dit heeft invloed op de evalueerbaarheid van een interventie, omdat een evaluatie zich in deze gevallen baseert op informatie die niet bekend is bij alle actoren en het dus afhangt van de specifieke actoren die betrokken zijn bij de evaluatie of bepaalde informatie al dan niet beschikbaar is[34].

Twee items scoren voor de vijf DAC evaluatiecriteria buitengewoon zwak: de beschikbaarheid van baseline informatie m.b.t. de counterfactual (2.1.3) en de desaggregatie naar gender of andere parameters van relevante indicatoren in de baseline (2.1.4); beide items staan hoog op de lijst van de 20% zwakst scorende items (zie ook tabel A3 3 van bijlage 7). De afwezigheid van gedesaggregeerde informatie is niet verwonderlijk en terug te brengen tot de afwezigheid van een genderanalyse en gedesaggregeerde indicatoren in het vooropgestelde M&E systeem (zie hoger).

De counterfactual is een inschatting van de situatie die zich zou voordoen als de interventie niet zou hebben plaatsgevonden en is vooral belangrijk voor het methodologisch correct uitvoeren van impactevaluaties[35]. De slechte scores voor dit item worden mede veroorzaakt door het feit dat het niet altijd eenvoudig is een goede counterfactual te identificeren, dit enige methodologische kennis vereist en het om een praktijk gaat die

---

[33] … al moesten in een aantal gevallen de betrokken partijen meermaals worden aangemaand om deze informatie te verstrekken. Verder is het belangrijk te melden dat de kwaliteit van die documenten sterk verschilt; dit komt via de scores van de andere items tot uiting.

[34] Voor deze studie betekende dit eveneens dat scores gegeven op basis van de voorbereidende deskstudie met betrekking tot dit item voor heel wat interventies moesten worden aangepast tijdens de veldmissie. Indien een evaluator geen voorbereidend veldbezoek kan uitvoeren, heeft dit dus belangrijke implicaties voor de mate waarin hij/zij zich voldoende kan voorbereiden op de veldmissie en indirect op de evalueerbaarheid.

[35] 'Impact' wordt hier gebruikt in de betekenis van de aan de interventie direct toewijsbare effecten (causaliteit) (zie ook voetnoot 14) die dus zowel op het niveau van de 'outcomes' als de hogere liggende schakels in de ketting gesitueerd kunnen zijn.

tot op heden nog niet is ingeburgerd binnen de Belgische ontwikkelingssamenwerking. Het gebruik van een counterfactual blijkt eveneens vaak erg beperkend te worden geïnterpreteerd en verengd tot het gebruik van een RCT (*Random Controlled Trial*) type van controlegroep. Deze piste is inderdaad niet altijd mogelijk of zelfs wenselijk, maar er zijn tal van alternatieven beschikbaar die vaak minder omslachtig zijn maar blijkbaar eveneens minder gekend[36]. Voorbeelden zijn de selectie van een non-interventie groep via 'matching', de selectie van een vergelijkbare groep die betrokken is in een andere interventie (om inzicht te verkrijgen in differentiële impact), het gebruik van statistische controles, generieke controles en meer complexe types van voor-na vergelijkingen zoals panels, time series of schaduwcontroles[37]. Enkele interventies in de steekproef hebben een counterfactual gebruikt die niet complex en niet duur waren en enkele andere interventies zouden op relatief eenvoudige wijze een counterfactual kunnen identificeren.

In de meeste interventies is er baseline informatie m.b.t. de (participatie van de) doelgroep (2.1.2 en 2.1.6), maar deze informatie is meestal niet compleet en beperkt tot wie effectief bereikt wordt. Ook al omdat in het interventie-voorstel de informatie m.b.t. (de afbakening van) de doelgroep meestal erg algemeen is, blijft deze informatie ook vaak erg kwantitatief en geeft ze weinig kwalitatieve informatie over de situatie waarin de doelgroep zich bevindt. Dit is enigszins verwonderlijk omdat er vaak wel baseline studies uitgevoerd zijn, maar deze schetsen eerder een algemeen beeld van de regio of de sector, en geven geen diepgaande analyse van de specifieke karakteristieken van de doelgroep (en een counterfactual), noch van de factoren die mogelijk de implementatie en de resultaten van de interventie beïnvloeden. Het is net een dergelijke baselineanalyse die het mogelijk maakt om parameters (zoals onder meer gender) te identificeren die een invloed kunnen hebben op de implementatie en resultaten van een interventie en die dus best opgenomen worden in het M&E systeem gezien ze cruciale informatie leveren die de evaluatiekwaliteit en –exhaustiviteit zwaar beïnvloedt.

De aanwezigheid van zo ruim mogelijke gedesaggregeerde baseline informatie is essentieel voor de evalueerbaarheid van interventies, aangezien het een vergelijking tussen de begin- en eindsituatie mogelijk maakt. Specifieke informatie met betrekking tot de effectieve bereikte doelgroep in vergelijking met de oorspronkelijk gedefinieerde

---

[36] Zie Rossi P.H., Lipsey M.W. and H.E. Freeman (2004) *Evaluation: a systematic approach*, 7th edition. Thousand Oaks: Sage; Bamberger M., J. Rugh, M. Church and L. Fort (2004) "Shoestring evaluation: designing impact evaluations under budget, time and data constraints", *American Journal of Evaluation* 25 (1): 5-37.

[37] In het geval van *'matching'* wordt een controlegroep gezocht die gelijkaardig is aan de interventiegroep op die karakteristieken waarvan men vermoedt dat ze een invloed hebben op de resultaten van de interventie. Bij *statistische controles* volgt men een gelijkaardige logica maar in dit geval controleert men niet voor bepalende karakteristieken op het moment van de evaluatie-opzet en de dataverzameling maar op het ogenblik van de data-analyse (men zoekt m.a.w. bij de evaluatie-opzet geen gelijkaardige groep maar men gaat controleren voor de bepalende karakteristieken door ze als controle-variabelen aan een regressie toe te voegen, waardoor men de invloed van deze variabelen kan uitzuiveren). Nadeel bij voorgaande counterfactuals blijft dat ze dataverzameling vereisen bij een groep die niet betrokken is bij de interventie, iets wat niet altijd mogelijk of wenselijk is. In het geval van *generische controles* gebruikt men nationaal beschikbare data over de uitkomst die men wenst te bereiken met een specifieke interventie. Men gebruikt m.a.w. een volledige bevolking (of bevolking van een regio) als een controlegroep voor de interventiegroep. Generische controles zijn uiteraard de goedkoopste optie, maar niet altijd beschikbaar (meestal wel voor de sectoren van onderwijs, gezondheidszorg, etc.) en vaak niet verfijnd genoeg omdat de gemiddelde scores voor een grotere bevolking vaak niet als een goede proxy kunnen beschouwd worden voor de specifieke interventiegroep (die men net kiest omwille van afwijkende karakteristieken tegenover een gemiddelde bevolking). Een andere mogelijkheid zijn de meer gesofistikeerde vormen van voor/na vergelijking. In principe laat een voor/na vergelijking niet toe om een goede indicatie te krijgen van impact (impact vereist een met/zonder vergelijking). Echter door meer metingen te doen van de voor-en na situatie kan men de voor/na vergelijking versterken: in het geval van *panel* volgt men een aantal huishoudens voor, doorheen en na de interventie, in het geval van time series gebruikt men minimaal 30 observaties voor de interventie en extrapoleert men op basis hiervan om een idee te krijgen van de situatie na de interventie (=zonder interventie). Deze informatie vergelijkt men met de werkelijke situatie na de interventie (= met de interventie) en dit geeft een goed idee van de werkelijke impact. Tenslotte kan men kiezen voor schaduwcontroles: in dit geval zal de evaluator pogen om op basis van eigen ervaring, gesprekken met lokale sectorexperts, gesprekken met begunstigden, etc. proberen om een zo goed mogelijke inschatting te maken van de situatie die zou bestaan zonder de interventie. Dit is de minst wetenschappelijke methode maar hoe sterker de informatiebronnen en de kennis waarop men zich baseert, hoe meer deze 'counterfactual' de werkelijke 'zonder-situatie' kan benaderen.

doelgroep is dan weer essentieel om na te gaan of de interventie haar oorspronkelijke doelgroep bereikt. Deze info is een eerste stap in de analyse van de manier waarop de interventie al dan niet haar doelgroep bereikt en laat toe om te evalueren of de interventie zodanig is ontworpen dat de vooropgestelde doelgroep ook effectief bereikt kan worden. Verder leidt het ontbreken van informatie met betrekking tot drop-outs gemakkelijk tot een overschatting van de effecten van interventies.

De scores voor de items met betrekking tot de opvolging van interne risico´s en assumpties (2.1.9 en 2.1.10) zijn matig tot zwak voor alle criteria (ze behoren beide tot de 20% zwakst scorende items), waarbij over het algemeen de opvolging van interne risico´s enigszins beter scoort dan de opvolging van externe assumpties. Daarnaast scoort het output niveau enigszins beter dan het outcome niveau, en het outcome niveau beter dan het impactniveau. Deze vaststelling is niet verwonderlijk en hangt samen met de eerder vermelde focus op het implementatieniveau en op die elementen die direct verbonden zijn met de interventie.

Voor de meeste interventies is er tot het outcome niveau informatie aanwezig met betrekking tot de gegevensverzameling van de indicatoren (2.1.7), maar slechts voor enkele interventies is de informatie volledig en bevat deze onder andere details over wat wordt verzameld, door wie, met welke frequentie en dekking en met welke methodes van dataverzameling (zie box 3 voor een voorbeeld van een goede praktijk). Informatie met betrekking tot de gegevensverzameling is verder vaak erg gefragmenteerd aanwezig, zowel verspreid over enkele documenten als verspreid over verschillende hoofdstukken in hetzelfde document. Dergelijke informatie over het dataverzamelings-proces is belangrijk voor de evalueerbaarheid van interventies omdat de bestaande informatie vaak een belangrijke eerste gegevensbron is voor een evaluatie, waardoor het bijgevolg ook essentieel is om inzicht te krijgen in de kwaliteit van de informatie. Met betrekking tot de betrouwbaarheid van de gegevensverzameling (2.1.8) zijn de belangrijkste condities voor betrouwbare dataverzameling (steekproef groot genoeg, triangulatie, goede frequentie, onafhankelijke gegevensverzamelaar) veelal voldoende voor effectiviteit en efficiëntie evaluaties, maar in (veel) mindere mate voor impact, duurzaamheid en relevantie evaluaties[38].

Box 3: Duidelijke basisinformatie m.b.t. de werking van het M&E systeem bij PROTOS (programma in Rwanda)

De NGO Protos heeft een document 'scenario voor opvolging indicatoren' ontwikkeld waarbij per indicator duidelijk wordt aangegeven welke personen en organisaties betrokken zijn bij de datacollectie. Tevens wordt ook aangegeven wie de gegevens zal analyseren. Bovendien beschrijven deze scenario's via welke methode, wanneer en met welke periodiciteit de gegevens voor de indicator in kwestie moeten worden verzameld. Er wordt eveneens vermeld welke opleidingen eventueel voorzien worden met betrekking tot een specifieke methode of tool. Dergelijke opvolgingsscenario's kunnen vooral nuttig zijn om een globaal overzicht te behouden wanneer verschillende partners instaan voor specifieke onderdelen van de datacollectie en analyse, maar ze kunnen ook belangrijk zijn om een externe evaluator een overzicht te geven van verantwoordelijkheden en procedures van dataverzameling en analyse.

3.2.2 Het M&E systeem in de praktijk

Onder deze component is gekeken naar hoe het M&E systeem in de praktijk werkt (of heeft gewerkt) en werden 12 items gedefinieerd die deels overeenkomen met de items onder 1.3 (*Het voorgestelde M&E systeem*). Waarschijnlijk is de kwaliteit van het voorgestelde M&E systeem gecombineerd met de kwaliteit van de toepassing van dit systeem de component die de evalueerbaarheid van een interventie het sterkst beïnvloedt. Zwakheden bij andere componenten kunnen tot op grote hoogte worden opgevangen via een goed uitgewerkt en functionerend M&E systeem. Omgekeerd zal de

[38] Dit item kon slechts voor 24 interventies worden gescoord.

praktische evalueerbaarheid van een interventie laag scoren als de M&E functie slecht werkt, zelfs als men beschikt over een interventie-ontwerp van goede kwaliteit.

Tabel 9: Het M&E systeem in de praktijk

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
2.2 Het M&E systeem in de praktijk	3,03	3,30	3,50	2,53	2,85	3,04
2.2.1 De visie over en de rol van monitoring en evaluaties (M&E en onafhankelijke evaluaties) zijn duidelijk		3,23	3,31	2,90	2,73	3,05
2.2.2 Er zijn voldoende tijd, middelen en personeel voorzien voor de adequate werking van het M&E systeem	3,67	3,87	4,15	3,21	3,63	3,71
2.2.3 De verantwoordelijkheden en procedures m.b.t. het verzamelen en analyseren van M&E gegevens zijn duidelijk gedefinieerd	3,50	3,85	4,08	2,64	3,41	3,50
2.2.4 De verantwoorde-lijkheden en procedures m.b.t. besluitvorming op basis van de analyse van de M&E gegevens zijn duidelijk gedefinieerd	3,62	3,75	4,03	2,95	3,56	3,58
2.2.4bis De voornaamste betrokken partijen waren het eens met het voorgestelde M&E systeem (inclusief hun rol hierbinnen)[39]	4,18	4,20	4,33	4,23	4,18	4,22
2.2.5 Het personeel dat instaat voor M&E is competent en onafhankelijk	2,95	3,11	3,32	2,73	3,00	3,02
2.2.6 Het M&E systeem van de interventie is afgestemd op het nationale/lokale M&E systeem	2,59	2,88	3,12	2,52	2,64	2,75
2.2.7 Er is een interne drive voor strategische sturing en leren	3,72	3,80	3,90	3,70	3,72	3,77
2.2.8 De resultaten van M&E worden gebruikt voor learning	3,05	3,08	3,30	2,93	3,00	3,07
2.2.9 De resultaten van M&E worden gebruikt voor accountability	3,23	3,40	3,63	2,95	3,08	3,26
2.2.10 Er hebben evaluaties en/of studies plaatsgevonden die van goede kwaliteit zijn en nuttige informatie aanleveren [40]	2,67	2,78	2,75	2,22	2,58	2,61
2.2.11 Het M&E systeem wordt regelmatig op zijn kwaliteit bekeken en eventueel aangepast	2,74	2,90	3,00	2,80	2,85	2,86

Voor *alle* criteria scoort de *daadwerkelijke* M&E praktijk substantieel (meer dan 20%) hoger dan het in het interventieplan *voorgestelde* M&E systeem (zie tabel 3 en punt 1.3 van deel 1 van dit hoofdstuk). Dit bevestigt de eerder geformuleerde hypothese dat aan de presentatie van het M&E systeem in het interventievoorstel relatief weinig aandacht wordt besteed om verschillende redenen (zie 1.3). Het grote verschil qua score is echter eveneens een illustratie van het belang dat aan M&E wordt gehecht in de interventie-*uitvoering* en heeft positieve implicaties op de evalueerbaarheid. De M&E en dan vooral monitoring is dus veelal beter ingebed in de praktijk dan vanuit het interventievoorstel kan besloten worden. Dit is uiteraard een goede zaak, niet alleen vanuit het oogpunt van de evalueerbaarheid, maar vooral omdat (goede) M&E een belangrijk onderdeel vormt van het interventiebeheer en substantieel kan bijdragen tot betere interventies.

Anderzijds bestaat er een sterke correlatie tussen de scores m.b.t. de kwaliteit van het M&E ontwerp (zoals voorgesteld in het projectvoorstel) en de kwaliteit van de M&E

[39] Dit item was aanvankelijk onder dimensie 1 opgenomen, om tijdens de analyse naar component 2.2 te worden verplaatst. Om praktische redenen werd de oude nummering evenwel aangehouden.

[40] De gegevens voor deze item hebben slechts betrekking op 27 van de 40 onderzochte interventies. In de 13 andere interventies konden we niet verwachten dat er reeds een evaluatie of studie had plaatsgevonden.

praktijk, en dit voor alle evaluatiecriteria met uitzondering van impact[41]. De correlatie is het meest uitgesproken voor effectiviteit en duurzaamheid, wat suggereert dat het van belang is om met name deze criteria goed vanaf het begin in het M&E systeem te integreren. Ondanks het feit dat er een sterke correlatie bestaat, zijn er bij de 40 onderzochte interventies wel voorbeelden gevonden van een zwakke kwaliteit van het M&E systeem in het interventievoorstel, die naderhand werd gecorrigeerd door bijvoorbeeld in de eerste semester van de interventie-uitvoering sterk in te zetten op de uitbouw van een M&E systeem. Het voordeel van zulke benadering is dat *ownership* van het systeem beter kan worden gewaarborgd en het eenvoudiger is om andere actoren bij de uitwerking en toepassing van het systeem te betrekken. Verder impliceert een sterke correlatie ook niet noodzakelijk causaliteit (van papier naar praktijk): zoals hierna beschreven is het mogelijk dat de M&E praktijk ook het M&E-voorstel op papier beïnvloedt, wat bijvoorbeeld zeker tot de mogelijkheden hoort bij interventies die een opvolging zijn van voorgaande interventies.

Anderzijds scoort deze component slechts gemiddeld in vergelijking met de zeven andere componenten die we hebben onderzocht (tabel 3). Dit geeft aan dat er nog ruimte voor verbetering is. Dit lijkt ons ook normaal omdat de gesprekken met de interventie-uitvoerders ons hebben geleerd dan in vele gevallen de bestaande M&E systemen vrij recent werden ingevoerd of vorm gegeven. In vele interventies gaat het daarbij om een proces dat nog niet is afgerond, maar waarbij de intentie bestaat om zowel de structuur als de uitvoering van het systeem te vervolmaken. Deze vervolmaking heeft in eerste instantie te maken met het tot een daadwerkelijk systeem smeden van de verschillende componenten die dikwijls reeds (in aanzet) aanwezig zijn, maar nog niet op elkaar afgestemd werden. Verder viel het ook op dat in vele interventies *good practices* m.b.t. M&E blijken te bestaan maar waarover op geen enkele wijze wordt gerapporteerd en die dikwijls ook los staan van andere M&E praktijken. Vanuit de optiek van evalueerbaarheid mag dus worden aangenomen dat deze in de toekomst nog zal verhogen via een verdere ontwikkeling van de M&E praktijk.

Verder lijkt men er al vrij goed in te slagen om te monitoren op het operationele vlak: de aanwending van de middelen, de opvolging van de activiteiten en de voortgang op het niveau van de outputs zijn in vele gevallen een intrinsiek onderdeel geworden van de uitvoeringspraktijk. Evenwel blijkt de evaluatieve component van het M&E systeem, en dan met name het niveau van impact en duurzaamheid, minder aandacht te krijgen. De mogelijke oorzaken hiervan worden hierna besproken.

Als we de scores van de vijf DAC criteria vergelijken, komt een patroon naar voor dat gelijkaardig is aan dat van de andere componenten, met name een laagste score op het vlak van impact en een hoogste voor efficiëntie. Verder valt op dat de scores dezelfde rangorde vertonen als voor de beschrijving van het M&E systeem in het interventie-voorstel; de beschrijving van het M&E systeem is met al zijn onvolkomenheden dus een voorafspiegeling van de latere praktijk. De lage scores voor impact maar ook voor duurzaamheid impliceren dat de uitdagingen die reeds intrinsiek (ex ante) aanwezig zijn om deze beide criteria in evaluaties mee te nemen, nog worden vergroot. Aan de andere kant scoren effectiviteit en vooral efficiëntie vrij goed, wat positief is vanuit het standpunt van evalueerbaarheid, temeer daar deze beide criteria in veel evaluaties centraal worden gesteld.

Uit onze analyse blijkt dat de visie m.b.t. (rol van) monitoring en evaluatie op interventieniveau relatief zwak ontwikkeld is, wat vrij verrassend is in het licht van de in ontwikkeling zijnde M&E praktijk (2.2.1), al zijn er een aantal interessante uitzonderingen waarbij veranderingen in beleid aanleiding geven tot de uitwerking van een M&E aanpak en beleid en worden ondersteund met voldoende personele middelen (zie box 4) . Niet verrassend daarbij zijn voor impact en duurzaamheid de scores lager, wèl dat het verschil aanzienlijk is; blijkbaar vallen deze twee criteria dikwijls buiten het

[41] Dit laatste heeft waarschijnlijk te maken met het feit dat impact weinig expliciete aandacht krijgt in M&E systemen (zowel in het voorstel als in de praktijk) wat ook de kans op afstemming tussen voorstel en praktijk verkleint.

blikveld bij de ontwikkeling van een M&E beleid. Als we wat dieper graven, vinden we nog enkele andere redenen. Zo is het goed mogelijk dat het beleid tot op zekere hoogte de praktijk volgt (en niet omgekeerd) en dat die praktijk vooral op operationeel vlak reeds goed vorm gegeven is. Als zodanig is het begrijpelijk dat monitoring in beleidsdocumenten doorgaans beter is uitgewerkt dan evaluatie. Verder is het ook duidelijk dat visie- en beleidsvorming rond monitoring en evaluatie dikwijls vanuit de (Belgische) verantwoordelijke organisatie wordt aangestuurd, maar dat het tijd vraagt om dit beleid naar interventies te vertalen o.a. omdat een 'nieuw' of ander beleid in regel slechts na samenspraak met partners kan worden toegepast. In vele gevallen is dit proces nog niet afgerond, wat begrijpelijk is gezien de relatief recent toegenomen aandacht voor M&E. Een gevolg van dit alles lijkt dat het globaal vrij goed zit met de evalueerbaarheid op het vlak van effectiviteit en efficiëntie (en om andere reden ook van relevantie), maar dat dit minder het geval is voor impact en duurzaamheid.

Box 4: Hoe gewijzigd beleid tot een verbeterde M&E aanpak kan leiden

Gedurende de laatste jaren heeft de strategie van Artsen Zonder Vakantie een grondige wijziging ondergaan. Terwijl vroeger het accent lag op zendingen van Belgische medici die ter plekke medische prestaties gingen leveren, wordt de huidige werking opgebouwd vanuit de behoeften van de partnerorganisaties. Deze fundamentele omslag is ook doorgetrokken in het M&E instrumentarium. De organisatie heeft een brede waaier aan M&E instrumenten ontwikkeld die moeten toelaten de effecten van het programma beter op te volgen. Centraal staat daarbij een instrument om de realisaties op het vlak van capaciteitsversterking te meten waarbij wordt uitgegaan van een door de partner zelf gedefinieerde/gewenste veranderingen en onder meer '*progress markers*' worden gedefinieerd en opgevolgd. Verder worden ook de kwaliteit en effecten van de zendingen van de Belgische medici systematisch gemeten via onder meer een gezamenlijk (door medicus en partner) opgesteld zendingsrapport (volgens een gestandaardiseerd format) en een door elke partij opgesteld specifiek rapport; deze documenten worden nadien samengebracht in een synthesenota. Op basis van deze synthese wordt aan alle betrokkenen feedback gegeven.

In tabel 9 valt ook de erg hoge score op m.b.t. de mate van overeenstemming, onder de voornaamste actoren, rond hun rol in het M&E systeem (2.2.4bis). Deze score moeten we evenwel nuanceren omdat M&E veelal de zaak is van alleen de projectploeg, met erg weinig (verwachte of gevraagde) inbreng van de andere sleutelactoren. Slechts in meer complexe interventies of organisaties blijken er zich problemen voor te doen, bijvoorbeeld wanneer hogere niveaus excessieve (in de ogen van lagere echelons) eisen opleggen m.b.t. dataverzameling, verwerking en rapportage.

Het onderzoek wijst ook uit dat middelen (tijd, personeel, fondsen) voor M&E doorgaans voldoende aanwezig zijn – al scoort het impactniveau hier om begrijpelijke redenen zwakker (2.2.2). Dit is vanuit het oogpunt van evalueerbaarheid een positieve vaststelling die evenwel enige nuancering behoeft. Nadere discussies met projectteams gaven immers aan dat er wel voldoende tijd was om de 'basale' M&E functies zoals het verzamelen en verwerken van gegevens op te nemen, maar dat dikwijls de tijd ontbrak om die gegevens diepgaand te bespreken en te analyseren. In de meeste interventies wordt M&E immers niet specifiek toegewezen aan iemand (of een dienst) maar wordt de M&E functie opgenomen door meerdere teamleden, doorgaans elk verantwoordelijk voor de verzameling van de gegevens m.b.t. 'hun' output (of onderdeel daarvan) en dus tot het operationele niveau beperkt (inputs – activiteiten – outputs). De focus op het operationele niveau impliceert dus dat projectteams vooral met het 'hier en nu' bezig zijn, en er in de processen weinig momenten worden ingebouwd waarbij zaken van op afstand worden bekeken; dit kan ook een reden zijn voor de relatief zwakke score m.b.t. het (regelmatig) herbekijken van de werking van M&E systemen (2.2.11). Een andere reden waarom middelen als voldoende worden beschouwd, is de relatief beperkte aandacht die wordt besteed aan evaluatie (in de volle betekenis van het woord); dit zal nader in deel 3 van dit hoofdstuk worden uitgewerkt.

Verantwoordelijkheden en procedures voor de verzameling en analyse van M&E gegevens en besluitvorming op dit vlak zijn vrij goed vastgelegd, met uitzondering van het impactniveau omdat dit dikwijls niet wordt meegenomen (2.2.3 en 2.2.4). Monitoring (en evaluatie) op hogere niveaus maakt dikwijls eenvoudigweg geen deel uit van het M&E takenpakket; zulke analyse wordt aanzien als een collectieve verantwoordelijkheid, waardoor niemand zich specifiek voelt aangesproken[42]. Hoewel er een gezonde *peer pressure* bestaat, blijkt eenieder "slechts" voor de monitoring van "zijn" (of haar) deel van de interventie-uitvoering in te staan. Opvolging van *globale* interventie-effecten en zeker van impact staan daarboven en vereisen bovendien een meer specifieke en veeleisende aanpak op het vlak van gegevensverzameling en –analyse waarvoor meer specifieke methodologische kennis en ruimere middelen vereist zijn. Daarvoor wordt in de uitbouw van vele M&E systemen niet echt een plaats ingeruimd waardoor er ook geen verantwoordelijkheden op dit niveau worden aangeduid. Bovendien is het zo dat de informatie die verkregen wordt door impactanalyse vaak het specifieke interventieniveau overstijgt waardoor het binnen interventies/organisaties als minder nuttig en direct bruikbaar wordt beschouwd en derhalve geen prioriteit. Dergelijke informatie is evenwel bijzonder nuttig op een hoger (beleids)niveau maar vertoont a.h.w. een publiek goed-karakter wat de gebrekkige investering in impact- en duurzaamheidsevaluaties door specifieke interventies mede verklaart.

Terzijde willen we opmerken dat de wenselijkheid en mogelijkheid om andere actoren bij M&E te betrekken, slechts zelden grondig wordt bekeken. Op die wijze gaan een aantal opportuniteiten verloren op het vlak van bijvoorbeeld leren, maar ook wat betreft de verduurzaming van M&E systemen en de mogelijkheid om de last en verantwoordelijk-heid voor M&E beter te spreiden over de verschillende actoren.

Competentie en onafhankelijkheid van het personeel dat instaat voor M&E scoren minder hoog dan de meeste andere items onder deze component (2.2.5). Een eerste verklaring ligt in het feit dat binnen kleine interventies het moeilijk is om daadwerkelijk onafhankelijke M&E te organiseren, bijvoorbeeld door die toe te vertrouwen aan een aparte dienst of persoon; pas bij grotere interventies of programma's is dit mogelijk (zie box 5 op de volgende bladzijde). Wat de competentie betreft, zijn er weinig indicaties van specifieke deskundigheid of investeringen in specifieke vorming. De meeste stafleden die met M&E bezig zijn, worden dus on the job opgeleid. Globaal leidt dit evenwel tot weinig problemen omdat (1) er veel uitwisseling en een goede leercultuur bestaat binnen de teams (zie hierna), (2) de M&E taak zich veelal tot het operationele niveau beperkt, waar zich minder methodologische uitdagingen stellen, (3) vele interventies voor belangrijke activiteiten (bv. vormingen) een goed standaardinstrumentarium hebben ontwikkeld. Echter is het duidelijk dat M&E (en dan vooral evaluatie) op hogere niveaus hogere eisen stelt waaraan ongetraind interventiepersoneel niet steeds kan voldoen. Hierbij kunnen we opmerken dat die methodologische uitdagingen op deze niveaus onvoldoende worden onderkend en dat de facto aan het outcome en impact niveau binnen de monitoring weinig aandacht wordt besteed. Tenslotte is ook opgevallen dat in vele gevallen de kwaliteit en continuïteit van M&E systemen te lijden hebben onder personeelsverloop of interne verschuivingen. Vooral wanneer het M&E systeem nog weinig is geformaliseerd, leidt dit tot problemen, wat een bijkomende illustratie vormt van het belang van een meer uitgewerkte beschrijving van een M&E systeem.

[42] In sommige gevallen moeten lokale actoren ook aan monitoringvereisten voldoen die opgelegd zijn door verschillende actoren en vaak slecht op elkaar afgestemd.

Box 5: De uitwerking van een interventie overstijgende M&E aanpak (BTC Katanga)

Het TFD van het EDUKAT project (uitgevoerd door BTC in de DRC) bevat duidelijke indicaties m.b.t. de opzet en organisatie van een M&E systeem, die gebaseerd zijn op het algemeen beleid en de handleidingen die BTC op dit vlak heeft ontwikkeld. Binnen dit kader heeft EDUKAT het vrij uitzonderlijke initiatief genomen om een werkplan uit te werken (*Plan de travail baseline*) die probeert om de verschillende M&E activiteiten in een coherent geheel samen te brengen. Dit werkplan werd uitgewerkt na een herwerking van het initiële logische kader (dat aanzien werd als nogal ingewikkeld en niet helemaal coherent) en definieert de stappen voor de uitwerking van een baseline, inclusief een opvolgingsmatrix (afgeleid van het herwerkte logische kader), een plan voor het beheer van de risico's en de effectieve toepassing van de opvolgingsmechanismen.

De uitvoering van het plan is de verantwoordelijkheid van een programmabeheerder die instaat voor het transversale beheer van M&E in de verschillende bilaterale projecten die BTC in Katanga uitvoert. De keuze voor een specifieke verantwoordelijke is gemaakt vanuit de wens om leren en continuïteit van het M&E systeem te waarborgen, en dit binnen een kader van een langdurig (12 jaar) engagement. Op die wijze moet het ook mogelijk zijn om een proces van institutionele versterking te ondersteunen via een betrokkenheid van lokale structuren bij de gegevensverzameling en –analyse en de opvolging van de M&E resultaten.

De resultaten van onze analyse voor wat betreft het gebruik van M&E voor leren zijn vrij goed (2.2.7), met uitzondering van die voor impact. Vanuit de sterke oriëntatie van de M&E op het operationele niveau, is het evident dat de score goed is op het vlak van efficiëntie. Het gebruik van de M&E resultaten voor accountability scoort iets beter dan het gebruik voor learning doelen (2.2.9). Waarschijnlijk kan dit o.a. verklaard worden door het feit dat voor accountability de verzamelde gegevens als zodanig in grote mate volstaan, terwijl voor learning verdergaande analyse, reflectie en internalisering nodig zijn waarvoor, zoals hoger aangegeven, de tijd in vele gevallen ontbreekt. Aan de andere kant wordt 'accountability' veelal vrij eng geïnterpreteerd en is die de facto beperkt tot accountability naar de donor toe en (in een aantal gevallen) indirect naar de bredere achterban, bv. via websites. Erg zelden is er sprake van bewuste accountability naar sleutelactoren zoals doelgroepen. Hieraan gerelateerd is het ook zo dat er relatief weinig feedback wordt gegeven (over de dataverzameling en de bevindingen op basis van de geanalyseerde data) naar lokale sleutelactoren (vaak ook de doelgroep) die betrokken zijn in processen van dataverzameling. Dergelijke feedback is essentieel om te vermijden dat lokale dataverzameling een ritueel proces wordt dat de betrouwbaarheid van de data (en dus ook de evalueerbaarheid) negatief beïnvloedt en op langere termijn ook de duurzaamheid van het M&E systeem. Deze noodzaak aan terugkoppeling geldt niet alleen voor de relatie tussen lokale interventieverantwoordelijken en de sleutelactoren (inclusief de doelgroepen), maar eveneens die tussen Belgische organisatie en lokale interventieverantwoordelijken, en die van DGD tegenover de indirecte actoren en BTC.

In vele interventies is er echter wél een sterke drive voor strategische sturing en leren die zich ook vertaalt in de uitvoeringpraktijk (2.2.7). Dit schept een kader waarin eventuele onvolkomenheden van het M&E systeem (zoals op het vlak van competenties en onafhankelijkheid en de beperkte aandacht voor de hogere niveaus in de doel-middelen ketting) toch voor een deel worden opgevangen. Ook zorgt dit voor een interventiecultuur die open staat ten aanzien van kritiek en bevraging, een belangrijk 'transversaal' voordeel voor de evalueerbaarheid.

Reeds eerder werd aangegeven dat de evaluatie-praktijk in de interventies (te) weinig aandacht krijgt. Dit wordt ook geïllustreerd door de zwakke score (het item met de laagste score onder deze component, en dit voor alle criteria; dit item staat ook voor 3 criteria in de lijst van de zwakst scorende items) m.b.t. het track record van de onderzochte interventies op dit vlak (2.2.10)[43]. Wij zien hiervoor verschillende verklaringen die nader zullen besproken worden onder deel 3. Hier kan worden volstaan

[43] … waarbij we moeten opmerken dat we dit item slechts bij 27 interventies konden scoren.

met een verwijzing naar de sterke focus op monitoring in de vele interventies, waarbij men er dikwijls terecht of ten onrechte van uit gaat dat (externe) evaluatie geen meerwaarde te bieden heeft. Anderzijds zijn er ook voorbeelden van een goede praktijk, waarbij de opzet en uitvoering van evaluaties reeds ex ante tot op grote hoogte is bepaald en over een langere periode toegepast, zodat die een aanzienlijke meerwaarde kunnen bieden (zie box 6).

Box 6 : Uitwerking van een duidelijk evaluatiebeleid en –praktijk

Als NGA werkzaam in het domein van de ontwikkelingseducatie, heeft ITECO een duidelijk beleid ontwikkeld op het vlak van planning, monitoring en evaluatie. Wat de evaluatie van zijn vormingsacties betreft, heeft ITECO een standaardbenadering uitgewerkt die drie complementaire dimensies inhoudt die verschillende schakels van de onderliggende veranderingstheorie betreffen : de evaluatie van hetgeen tijdens de vorming is geleerd, de toepassing (door de deelnemers) van het geleerde en de effecten van de vorming (via het in praktijk brengen van de nieuwe verworvenheden) op niveau van de organisatie en de omgeving van de deelnemers. De resultaten van deze analyse laten ITECO toe om het pedagogisch domein te overstijgen en zich een beeld te vormen van haar bijdragen tot de gewenste transformatie.
Op het niveau van het programma dat door de Belgische overheid wordt gecofinancierd, heeft ITECO de steun van een specialist in educatie en vorming om de M&E instrumenten van de organisatie te verfijnen en goede indicatoren te definiëren. De uitkomst van dit proces was, onder meer, een logisch kader van erg goede kwaliteit, met een duidelijke baseline en jaarlijkse targets, die een goede opvolging van de gerealiseerde voortgang moeten waarborgen en een ideale basis vormen voor evaluaties.

Een laatste punt betreft de afstemming van project M&E systemen op nationale en decentrale systemen (2.2.6). De scores hiervoor zijn, waarschijnlijk niet verrassend, vrij laag, behalve voor interventies in Rwanda (zie hoofdstuk 4.1). In veel interventies, voornamelijk deze buiten het bilaterale kanaal, gaat het hier om een overweging waar men niet echt mee bezig is. Dit ligt door de band anders in bilaterale interventies, maar daar spelen dikwijls praktische moeilijkheden, zoals de zwakte of afwezigheid van M&E systemen op nationaal of decentraal niveau.

Geen enkel project lijkt overigens de ambitie te hebben om de bestaande M&E systemen te verbeteren en te verduurzamen. Door dit van de interventies te willen verwachten, leggen we misschien de lat te hoog. Niettemin zou elk project expliciet moeten nagaan of zijn M&E systeem (of enkele van de sleutel-indicatoren) niet met nationale, lokale of partnersystemen kan worden gealigneerd. Het is immers duidelijk dat het ontbreken van een link met nationale of decentrale systemen belangrijke implicaties heeft m.b.t evalueerbaarheid, met name als men ex-post wil evalueren en daarbij bijvoorbeeld de duurzaamheid van door een project gerealiseerde baten wil nagaan.

3.3 De evaluatiecontext

Om de invloed van de context op de evalueerbaarheid van de interventies na te gaan, werd de assumptie aangenomen dat voor elke interventie een onafhankelijke externe evaluatie zou moeten worden uitgevoerd. Verder werd ervoor gekozen om praktische context-elementen zoals de situatie op het vlak van veiligheid, infrastructuur, etc. niet mee te nemen in de analyse, omdat die erg specifiek zijn voor elke interventie en moeilijk te hanteren in het kader van deze algemene studie. Het weze evenwel duidelijk dat dit soort elementen in de realiteit van groot belang zijn bij het bepalen van de evalueerbaarheid; daarom werd er ook voor gekozen ze expliciet te vermelden in het analysekader.

3.3.1 De houding van de sleutelactoren

De houding van sleutelactoren (zoals de nauw bij de uitvoering van de interventie betrokken actoren, maar ook lokale overheidsinstanties, donor(en) en andere in de regio werkzame organisaties) is van groot belang voor de evalueerbaarheid van interventies. Een negatieve houding van actoren maakt het werk van de evaluatoren in alle omstandigheden erg moeilijk, zelfs als de andere voorwaarden, zoals een goed interventieplan of een goed functionerend M&E systeem, in grote mate vervuld zijn.

Tabel 10: Bevindingen m.b.t. de houding van de sleutelactoren

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
3.1. Houding van de sleutelactoren	3,68	3,70	3,70	3,58	3,63	3,66
3.1.1 De voornaamste gebruikers van de evaluatie en hun ver-wachtingen/belang t.a.v. de evalu-atie zijn duidelijk gedefinieerd [44]	2,36	2,36	2,36	2,27	2,27	2,33
3.1.2 De sleutelactoren zijn vragende partij voor (of minstens geïnteresseerd in) een evaluatie	3,45	3,45	3,45	3,40	3,45	3,44
3.1.3 De verwachtingen van de sleutelactoren t.a.v. de evaluatie (proces en resultaten) zijn onderling compatibel	4,11	4,11	4,19	4,11	4,16	4,14
3.1.4 De verwachtingen van de sleutelactoren t.a.v. de evaluatie zijn realistisch (tov beschikbare middelen)	4,32	4,23	4,49	4,11	4,32	4,29
3.1.5 De voornaamste gebruikers werden betrokken/zullen betrokken worden bij het evaluatieproces	3,53	3,52	3,58	3,55	3,55	3,55
3.1.6 De voornaamste stakeholders (incl. doelgroep) werden betrokken/ zullen betrokken worden bij het evaluatieproces[45]	3,62	3,62	3,62	3,56	3,56	3,59
3.1.7 De onderlinge relaties tussen de sleutel-actoren zijn "gezond"	4,26	4,24	4,24	4,22	4,22	4,24
3.1.8 Het is mogelijk om alle sleutelactoren te contacteren zonder risico op wederzijdse beïnvloeding	4,63	4,64	4,64	4,63	4,63	4,64
3.1.9 Er is een positieve houding t.a.v. onafhankelijke evaluatie bij alle sleutelactoren	4,18	4,18	4,18	4,18	4,21	4,18

Zoals blijkt uit tabel 10, overheersen positieve scores. Het gaat hier dan ook om een component die globaal hoog scoort, met vijf items in de lijst van de 20% best scorende items (zie ook tabel A3 2 van bijlage 7)[46]. Verder valt op dat de onderlinge verschillen tussen de vijf evaluatiecriteria minimaal zijn (waarbij impact en duurzaamheid niettemin ook hier weer het zwakst scoren), wat te verklaren is vanuit de overweging dat de attitude van de sleutelactoren veelal 'onafhankelijke evaluatie' *in haar geheel* betreft. Het is duidelijk dat een goede globale score met betrekking tot de houding van de sleutelactoren een positieve bevinding is in het licht van de evalueerbaarheid. Zulke positieve houding speelt niet alleen tijdens het evaluatieproces, maar is veelal een indicatie van een openheid en bereidheid tot bevraging en kritiek, die ook belangrijke factoren zijn voor een goede interventie-uitvoering.

[44] Dit item kon slechts bij 22 interventies zinvol geanalyseerd worden; de interventies die niet konden gescoord worden waren veelal te recent opgestart om een externe evaluatie te kunnen uitvoeren.

[45] Dit item kon slechts bij 26 interventies zinvol worden geanalyseerd.

[46] Verderop in het rapport zullen we evenwel vaststellen dat de resultaten per land kunnen verschillen; zie deel 4 (comparatieve analyse) hierna.

'De mate waarin de voornaamste gebruikers van de evaluatie en hun verwachtingen en belangen t.a.v. de evaluatie duidelijk zijn gedefinieerd' (3.1.1) is het item met de laagste score. We zien hiervoor verschillende verklaringen. Een eerste verklaring is dat, zoals reeds hoger aangegeven, het beleid en de praktijk m.b.t. (onafhankelijke) evaluatie op interventieniveau relatief zwak is uitgebouwd (in tegenstelling tot monitoring). Eigenlijk is men op interventieniveau weinig (expliciet) met evaluatie bezig en daardoor wordt er ook weinig nagedacht over de verschillende parameters met betrekking tot evaluatie: de mogelijke doelstellingen van de evaluatie (accountability – leren –beleidsondersteuning) worden weinig tegenover elkaar afgewogen en dus ook de daaruit volgende keuzes m.b.t tot de voornaamste evaluatiecriteria en de mogelijke (toekomstige) gebruikers van evaluatieresultaten (die niet noodzakelijk samenvallen met de sleutelactoren bij de uitvoering van de interventie[47]). Slechts in de eigenlijke aanloop naar de evaluatie krijgen deze punten enige aandacht, maar blijven ze veelal onvoldoende uitgewerkt, zoals onder meer blijkt uit de analyse van de referentietermen voor evaluaties. Het gebrek aan een duidelijke inspanning (en duidelijke keuzes) m.b.t. dit item heeft een invloed op de evalueerbaarheid: wanneer de verwachtingen/belangen van de voornaamste gebruikers niet goed zijn geïdentificeerd en/of er geen duidelijke keuzes werden gemaakt m.b.t de evaluatie-doelen, leidt dit onvermijdelijk tot een lagere evalueerbaarheid en - vooral - een lagere kwaliteit van de evaluatieproducten en -effecten. De vele evaluatierapporten die zelf aangeven dat ze bepaalde aspecten onvoldoende in de diepte hebben kunnen analyseren zijn hiervan een duidelijke illustratie.

De score met betrekking tot het item of de sleutelactoren vragende partij zijn voor een evaluatie of daarin minstens geïnteresseerd zijn (3.1.2), is relatief goed maar situeert zich toch onder de gemiddelde score van de items onder deze component. Niettemin is in meer dan 85% van de interventies minstens een meerderheid van de actoren geïnteresseerd of een vragende partij. De bereidheid tot en/of interesse in een externe bevraging kan onder meer verklaard worden door de sterke resultaatgerichtheid binnen bepaalde organisaties (bv. BTC) of landen (bv. Rwanda) maar ook door de open en democratische organisatiecultuur (met name merkbaar binnen NGAs) . Anderzijds impliceert een negatieve houding niet automatisch een afkeer van externe evaluaties. Interventies die deze houding aannemen, doen dit dikwijls vanuit de onderstelling dat externe evaluaties geen meerwaarde bieden (ten aanzien van interne evaluaties en monitoring resultaten); in enkele gevallen speelden ook negatieve ervaringen met eerdere externe evaluaties een rol.

De hoge scores met betrekking tot de onderlinge verenigbaarheid van de verwachtingen van de sleutel-actoren t.a.v. evaluatieproces en –resultaten en m.b.t. de realiteitswaarde van die verwachtingen (3.1.3 en 3.1.4) verdienen enige nuancering die aantoont dat hun positieve invloed op de evalueerbaarheid minder groot is dan we op het eerste gezicht zouden denken. Vooreerst kan worden opgemerkt dat de verwachtingen van de sleutelactoren realistisch zijn omdat het ambitieniveau m.b.t. externe evaluaties meestal niet erg hoog ligt. Er zijn, bijvoorbeeld, weinig interventies die daadwerkelijk willen investeren in een grondige evaluatie die impact of duurzaamheid centraal stellen. Anderzijds, als er veel van de evaluatoren wordt gevraagd (in vergelijking tot de beschikbare middelen), dan wordt daarbij ook impliciet aanvaard dat er een *trade off* is t.a.v. de diepgang van de analyse. Wat de onderlinge compatibiliteit van de verwachtingen betreft, zijn er weinig problemen omdat de definiëring van de voornaamste gebruikers en hun verwachtingen niet systematisch gebeurt, zoals hierboven reeds aangegeven. Op die wijze blijven mogelijke verschillen in verwachtingen en belangen onderbelicht of worden ze toegedekt door brede evaluatieobjectieven en -thema's te formuleren.

[47] De groep van (mogelijke) gebruikers van evaluaties kan verschillen van die van de sleutelactoren. De laatste groep betreft de actoren die een rechtstreeks belang in het project hebben en nauw bij de uitvoering zijn betrokken. De 'gebruikers' van de evaluatie omvatten deze sleutelactoren maar kunnen ook andere partijen bevatten zoals de donoren, leden van een netwerk waar het project deel van vormt, nationale planningsdiensten, …

De betrokkenheid van de *gebruikers* bij het evaluatieproces (3.1.5) scoort globaal goed, maar niet zeer goed. Van belang is dat slechts in uitzonderlijke gevallen externe evaluaties gemonopoliseerd worden door één enkele actor. Het verleden heeft ons geleerd dat er in zulke gevallen veelal weinig gebeurt met de evaluatieresultaten. Echter stelt een adequate betrokkenheid van de gebruikers hoge eisen ten aanzien van de evaluatie-uitvoering en de evalueerbaarheid. Daarom is het belangrijk die betrokkenheid (voortraject, uitvoering, na-traject) adequaat te doseren, d.w.z. oog te hebben voor de doenbaarheid van de evaluatie, om op die wijze het proces niet nodeloos te verzwaren. Deze oefening wordt in vele interventies nog niet gemaakt (ook al omdat de gebruikers en hun verwachtingen en de doelen van de evaluatie niet expliciet worden gedefinieerd): de meeste interventies kiezen vrij intuïtief voor een zekere vorm van betrokkenheid met name bij voor- en na-traject, zonder na te gaan of dit de beste keuze vormt. De betrokkenheid van *sleutelactoren* (3.1.6) scoort lichtjes beter dan die van de voornaamste gebruikers, wat te verklaren valt vanuit de overweging dat het evidenter is om deze actoren bij de evaluatie te betrekken. Een belangrijke vaststelling in dit verband is dat slechts iets meer dan één derde van de interventies de doelgroepen expliciet bij evaluaties betrekt. Enkele *good practices* tonen aan dat een gepaste betrokkenheid van gebruikers en sleutel-actoren gemakkelijker tot stand kan komen binnen een bredere benadering die deze betrokkenheid over een langere periode centraal stelt, niet alleen in het kader van evaluaties, maar via een systemische aanpak die start bij de project-voorbereiding en wordt verdergezet tijdens de uitvoering (zie box 7).

Box 7 : Brede aanpak bij evaluaties zorgt voor hoge betrokkenheid van sleutelactoren

Het PARZS project (*Projet d'Appui au Renforcement des Zones et Départements Sanitaires du Mono-Couffo et de l'Atacora-Donga*) dat door BRC in de gezondheidssector in Benin wordt uitgevoerd heeft sterk ingezet op de toepassing van de principes van *outcome mapping* (OM) in zijn plannings-, monitoring- en evaluatiesysteem. De definiëring, door de verschillende actoren en instanties betrokken bij het project, van gedragsveranderingen en *progress markers* liggen aan de basis van de M&E systemen. In het PARZS M&E systeem zijn verschillende momenten van auto-evaluatie en controle voorzien, die samen een goede opvolging waarborgen. Peer evaluaties en het OM systeem zijn aangepast aan de lokale context en houden rekening met de capaciteiten (op het vlak van monitoring) en verwachtingen van de betrokken partijen. Tijdens de projectuitvoering heeft dit systeem gaandeweg vorm gekregen en wordt het gebruikt door de verschillende betrokken partijen die verantwoordelijkheid gekregen hebben. Verschillende principes en elementen van het systeem zijn ook overgenomen in het M&E systeem van de gezondheidssector dat is opgezet vanuit de principes van RBF (*Results Based Financing*).

In de meeste interventies zijn de relaties tussen de sleutelactoren gezond (3.1.7). Waar dit niet het geval is, heeft dit veelal te maken met incidenten uit het verleden of met tegenstellingen tussen actoren die reeds lang bestaan en soms de specifieke context overstijgen. In de meeste interventies zijn de relaties tussen de actoren gezond kritisch, wat impliceert dat men niet overdreven vriendelijk is voor elkaar. Het is evenwel mogelijk dat een meer 'doorgedreven' definiëring van belangen van sleutelactoren en evaluatiedoelen en –inhoud de relaties meer op scherp zou kunnen stellen, al hoeft dit niet noodzakelijk negatief te zijn.

Het feit dat de relaties tussen de sleutelactoren in de regel gezond zijn, vormt zeker een verklaring voor de hoge score (hoogste score onder deze component) voor het item *'het is mogelijk om alle sleutelactoren te contacteren zonder risico op wederzijdse beïnvloeding'* (3.1.8). Daarnaast is het ook duidelijk dat in de meeste interventies een goed begrip aanwezig is m.b.t. de rol van (externe) evaluatie in het ontwikkelingsproces en de wijze waarop men zich ten aanzien van zulk proces dient te positioneren. Dit blijkt verder uit de hoge score op het item dat naar de houding van de sleutelactoren t.a.v. onafhankelijke evaluatie heeft gepeild (3.1.9). In meer dan 80% van de interventies hebben (bijna) alle sleutelactoren een positieve houding ten aanzien van onafhankelijke evaluaties, wat ons een belangrijk gegeven lijkt vanuit het standpunt van evalueerbaarheid.

3.3.2 De bredere context

Deze laatste component van ons analysekader kijkt naar de bredere context, m.a.w. de context buiten de directe omgeving van de interventies. Het is duidelijk dat deze context de evalueerbaarheid kan beïnvloeden, zowel in positieve als negatieve zin.

De bredere context is de component die het beste scoort van alle componenten (de drie items van deze component behoren tot de 20% best scorende (zie ook tabel A3 2 van bijlage 7), ondanks het feit dat in de studie landen werden opgenomen waar de institutionele en politieke omgeving bepaalde uitdagingen kan stellen t.a.v. onafhankelijke evaluaties (tabel 3). Tussen de verschillende criteria zijn er weinig verschillen; zoals kon worden verwacht, scoort impact iets lager. Dit is ook het geval voor relevantie, wat kan te maken hebben met het feit dat bij de evaluatie van relevantie moeilijkheden in de bredere context zwaarder doorwegen dan bij de andere criteria die zich veeleer op de interventie '*an sich*' concentreren.

Tabel 11: Overzicht van de bevindingen m.b.t. de bredere context

	Relevantie	Effectiviteit	Efficiëntie	Impact	Duurzaam-heid	Evalueerbaar-heidsindex
3.2 De bredere context	4,10	4,25	4,25	4,10	4,15	4,17
3.2.1 De bredere institutionele en politieke context staat positief t.a.v. onafhankelijke evaluatie	4,40	4,45	4,45	4,40	4,45	4,43
3.2.2 De socio-culturele context op het niveau van de doelgroepen maakt correct verzamelen van informatie mogelijk	4,35	4,45	4,45	4,35	4,40	4,40
3.2.3 Lokale deskundigheid met het vereiste profiel voor de evaluatie is beschikbaar	4,25	4,30	4,25	4,20	4,25	4,25

Hoe dan ook zijn de hoge scores een bijkomende illustratie van het feit dat onafhankelijke evaluaties globaal zonder noemenswaardige moeilijkheden kunnen worden uitgevoerd. Echter hebben gesprekken met de betrokkenen geleerd dat er zich in bijna elke evaluatie moeilijkheden voordoen, maar dat daar vrij gemakkelijk oplossingen voor gevonden worden die een goede uitvoering van de evaluatie en dus ook een goede evalueerbaarheid waarborgen.

Bovenstaande vaststelling dient evenwel in perspectief te worden geplaatst. Het is vooreerst niet uit te sluiten dat contextfactoren die potentieel de evalueerbaarheid onder druk zetten ook onze eigen studie beïnvloed hebben, door bv. sociaal gewenste antwoorden die tot positieve scores leiden. Meer in het algemeen hebben de onderzoekers te weinig tijd gehad om de invloed van de context grondig te analyseren (bv. om goede gesprekken te voeren met een brede waaier aan (potentiële) gebruikers van de evaluatie). Eigenlijk is een diepgaande studie op landenniveau nodig om de invloed van de context daadwerkelijk in te kunnen schatten. De focus die deze studie heeft gehanteerd m.b.t. de evaluatiecontext was gezien de beperkte tijd en middelen – noodgedwongen - vooral evaluatie-technisch en niet evaluatie-politiek. Bij zulke 'enge' invulling van de context kan het geen verwondering baren dat vrij gemakkelijk oplossingen worden gevonden voor contextuele moeilijkheden die zich aandienen.

Verder lijkt ons ook een meer diepgaande reflectie nodig rond de vraag of men niet te optimistisch is ten aanzien van de mogelijkheden om *onafhankelijk* te evalueren? In dit verband is het vooreerst belangrijk om enkele bevindingen met betrekking tot de vorige component mee te nemen. Zo hebben we vastgesteld (zie 3.1) dat het definiëren van toekomstige gebruikers van de evaluatie en van hun verwachtingen en belangen vrij zwak scoort. Omdat aan dit aspect te weinig aandacht wordt besteed, is de kans vrij groot dat evaluatie-oefeningen vooral 'binnen het systeem' worden opgezet en uitgewerkt. Immers, ook bij onafhankelijke evaluaties betreft die onafhankelijkheid vooral de *uitvoering* – door onafhankelijk experten – van de evaluatie, maar houden die experten zich veel minder (of helemaal niet) bezig met de bepaling van de

evaluatiedoelen en -vragen waarop de evaluatie in eerste instantie een antwoord moet geven onder meer omdat ze niet worden betrokken bij de uitwerking van de referentietermen. Op die wijze is het mogelijk dat ook 'onafhankelijke' evaluaties vrij verregaand gestuurd worden en met name controversiële onderwerpen (die bijvoorbeeld bepaalde partijen in verlegenheid kunnen brengen) uit de weg gaan of onderbelichten. Als evaluaties delicate ontwerpen laten liggen, is het evident dat er weinig of minder problemen zullen zijn op het vlak van de institutionele en politieke context; we kunnen hier van een vorm van 'zelfcensuur' spreken die ook in andere situaties veel voorkomt. Anderzijds kunnen bekwame en handige evaluatoren proberen om ook aandacht te besteden aan '*out of the box'* aspecten. Hoe dan ook is het duidelijk dat in zulke context met name het leerpotentieel van een evaluatie maar ook haar globale relevantie onder druk staan. Deze vaststelling nodigt ons ook uit om een nuance toe te voegen aan de definitie van evalueerbaarheid in de zin dat 'evalueerbaarheid' ook impliceert dat alles wat in principe *moet* geëvalueerd worden, ook daadwerkelijkheid *kan* geëvalueerd worden.

De score m.b.t het item '*De socio-culturele context op het niveau van de doelgroepen maakt correct verzamelen van informatie mogelijk'* (3.2.2) is eveneens hoog. Dit wil niet zeggen dat er op het niveau van de doelgroepen geen uitdagingen liggen in dit verband, maar wel dat deze (voldoende?) worden onderkend en men over oplossingen beschikt (of denkt te beschikken) t.a.v. deze uitdagingen. Eveneens is er m.b.t. specifieke 'moeilijke' thema's zoals gender-gerelateerd geweld over de laatste jaren, bijvoorbeeld in Oost-Congo, erg veel (evaluatie-)ervaring opgedaan waardoor de moeilijkheden kleiner zijn dan een aantal jaren geleden. Anderzijds geldt op dit vlak een identiek gevaar dan dat in de vorige paragraaf beschreven, m.n. dat bepaalde delicate onderwerpen uit de weg worden gegaan of niet daadwerkelijk behandeld. Verder heeft het onderzoeksteam de indruk dat de methodologische uitdagingen van onderzoek op doelgroepniveau nog al eens onderschat worden[48]. Het feit dat men goede relaties met doelgroepen onderhoudt is immers geen reden om de methodologische en andere eisen t.a.v. goed onderzoek minder strikt toe te passen. Verder worden interventies alsmaar complexer, met een veelheid aan actoren, en vragen ze daarom een grote evaluatie-expertise. Tenslotte is er in vele gevallen sprake van evaluatie- of onderzoeks-*fatigue* die sterk vertekende resultaten kan genereren.

Slechts in uitzonderlijke gevallen hebben interventies moeite om evaluatoren met het vereiste profiel (voldoende technische en methodologische expertise, onafhankelijkheid, beschikbaarheid aan marktconforme tarieven) te contracteren (3.2.3). Deze positieve score is zeker een illustratie van de toenemende beschikbaarheid van (met name) lokale expertise, wat vanuit het oogpunt van evalueerbaarheid erg belangrijk is. Anderzijds is ook hier een nuancering op haar plaats. Doordat de 40 interventies van onze steekproef voor het merendeel 'klassieke' interventies betreffen, zijn er weinig problemen om goede experten te vinden. Ervaringen met de evaluatie van minder klassieke vormen van samenwerking, zoals budgethulp, wijzen uit dat dit niet steeds het geval is[49]. Eveneens kan meespelen dat het (dikwijls) ontbreken van scherpe doelstellingen en evaluatievragen impliceert dat ook de eisen t.a.v. de benodigde expertise niet scherp zijn gesteld.

3.3.3 Praktische elementen

Voor de volledigheid willen we in herinnering brengen dat er nog een derde component bestaat die van belang is bij de analyse van de evaluatie-context en de bepaling van de evalueerbaarheid. Het gaat hier om een aantal praktische elementen met (veelal) rechtstreekse invloed op de evalueerbaarheid, zoals de veiligheidssituatie, de weersomstandigheden, de staat van de lokale infrastructuur, de geografische omvang van het interventiegebied en de bereikbaarheid, etc. Verder kan ook de timing van de

[48] We hebben dit – belangrijk – aspect onvoldoende kunnen onderzoeken om hierover verregaande uitspraken te doen.

[49] In het bestek werd reeds geopteerd om budgethulp niet mee te nemen in deze studie.

evaluatie de evalueerbaarheid sterk beïnvloeden: met name het veldwerk in evaluaties valt bij voorkeur niet samen met verkiezingen, religieuze feesten, de vakantieperiode, periodes van grote drukte binnen de interventie, etc. Ook kunnen bepaalde gebeurtenissen (bv. conflicten binnen organisaties of interventies) of gelijkaardige initiatieven (bv. van andere donoren) voor verwikkelingen zorgen. Door het erg interventie-specifieke karakter van deze elementen, hebben we deze in onze analyse niet mee willen nemen.

4 Comparatieve analyse

Dit hoofdstuk bouwt verder op de analyse in hoofdstuk 3 uitgevoerd en probeert na te gaan in hoeverre er verschillen zijn in de evalueerbaarheid op basis van bepaalde parameters die reeds eerder bij de bepaling van de steekproef werden gehanteerd[50]. Meer bepaald gaat het daarbij om (1) de vier **landen**, elk met 10 interventies vertegenwoordigd in de steekproef, (2) **de graad van complexiteit** van de interventies, waarbij we een onderscheid hebben gemaakt tussen interventies met een 'complexe' en 'minder complexe' TOC, waarbij 40% van de interventies als 'minder complex' werden gecatalogeerd en 60% als 'complex'[51], en (3) het type **actor**, waarbij 25% van de interventies onderdeel uitmaakt van de bilaterale samenwerking, 50% van de samenwerking via NGOs en vakbonden, en 25% via andere actoren (APEFE/VVOB, universitaire samenwerking, BIO, ITG, BOS+, VVSG). Deze analyse steunt in grote mate op een statistische verwerking van de scores die in bijlage 7 verder wordt toegelicht.

4.1 Vergelijking van de evalueerbaarheid op landen-niveau[52]

De tabel hierna presenteert de scores op *componenten*-niveau en voor wat betreft de drie grote dimensies van het studiekader. Daarbij is het belangrijk te vermelden dat gelijkaardige scores op componenten niveau verschillen tussen de landen op *item*-niveau kunnen verbergen. Deze potentiële verschillen worden geanalyseerd door de gemiddelde scores per DAC-criterium op componenten- en itemniveau tussen de vier landen te vergelijken.

Uit de tabel blijkt dat er *globaal* niet echt een duidelijk patroon naar voor komt voor wat betreft verschillen tussen de evalueerbaarheidsscores van de vier landen[53]. Er is enkel een duidelijk verschil daar waar we dat het meest mogen verwachten, met name onder de derde dimensie die de invloed van de context op de evalueerbaarheid heeft geanalyseerd: hier blijken op het eerste zicht DRC en dan vooral Rwanda zwakker te scoren dan de twee andere landen. De relatief kleine verschillen tussen de scores voor de andere dimensies en componenten suggereren dat *andere parameters* dan het land van groter belang kunnen zijn bij het verklaren van de verschillen tussen de scores (zie bespreking onder 4.2 en 4.3 hierna).

---

[50] Voor meer informatie i.v.m. de steekproefbepaling, zie hoofdstuk 2.3.1, bijlage 4 (lijst van alle opgenomen interventies) en bijlage 2 waar meer informatie wordt gegeven m.b.t. de onderzoeksmethodologie.

[51] Initieel wilde de studie een vergelijking maken tussen interventies in "harde" en "zachte" sectoren vanuit de veronderstelling dat interventies in harde sectoren wellicht meer evalueerbaar zijn dan in zachte, maar dit bleek niet werkbaar. Na een vrij lange interne reflectie bleek vooral de complexiteit van de TOC van de interventie belangrijk en werd overgegaan tot een onderscheid tussen 'complexe' en 'minder complexe' interventies, met name interventies met een complexe en minder complexe TOC. Zie punt 2.3.1 en punt 2 van bijlage 2 (Beschrijving van de aanpak en onderzoeksmethodologie) voor meer details.

[52] Dit onderdeel is onder meer gebaseerd op de vier landennota's die werden opgesteld en met de betrokken organisaties gedeeld; deze nota's hebben evenwel geen officieel karakter.

Tabel 12: Overzicht van de evalueerbaarheidsscores per land

	België	Benin	DRC	Rwanda	Evalueerbaar-heidsindex (°)
Dimensie 1 (interventieplan)	**3,14**	**2,94**	**3,24**	**3,32**	**3,16**
1.1 De onderliggende analyse	3,25	3,80	3,15	4,30	3,65
1.2 De interventielogica en veranderingstheorie	2,70	2,63	3,58	2,95	2,96
1.3 Het voorgestelde M&E systeem	2,55	2,18	2,34	2,98	2,51
1.4 Consistentie en adaptatie van de interventielogica en de veranderingstheorie	4,08	3,17	3,90	3,03	3,54
Dimensie 2 (uitvoeringspraktijk)	**2,98**	**2,67**	**3,05**	**3,15**	**2,96**
2.1 Basisinformatie m.b.t. de interventie uitvoering	2,88	2,68	2,90	3,08	2,88
2.2 M&E systeem in de praktijk	3,08	2,66	3,20	3,22	3,04
Dimensie 3 (context)	**4,17**	**4,18**	**3,85**	**3,45**	**3,91**
3.1 Houding van de sleutel actoren	3,78	3,76	3,50	3,58	3,66
3.2 Bredere context	4,56	4,60	4,20	3,32	4,17
Globale score evalueerbaarheid (°°)	**3,28**	**3,08**	**3,29**	**3,28**	**3,23**

(°) De score m.b.t. de globale evalueerbaarheidsindex is een score berekend op basis van indexen (geen gemiddeldes). Anderzijds zijn de scores berekend voor de landen gemiddelde scores (geen indexen). Als zodanig is het mogelijk dat de globale evalueerbaarheidsscore lichtjes afwijkt van het gemiddelde van de vier landenscores. De verschillen zijn evenwel erg klein en veranderen niets aan de inhoud van de analyse.
(°°) Voor de globale score evalueerbaarheid werd een gewogen gemiddelde berekend waarbij dimensie 1 en 2 even zwaar worden gewogen en dimensie 3 meetelt voor de helft van elk van de andere dimensies.

Uit de meer doorgedreven statistische analyse op component en vooral itemniveau komen wel een aantal significante verschillen naar voor tussen landen (die hierna worden besproken) waarbij evenwel geen eenduidig patroon te ontdekken valt (zie Tabel A2 3 in bijlage 7 voor meer details). Een nadere bestudering van bovenstaande tabel, een vergelijking van de gemiddelde scores die per DAC-criterium op componenten- en itemniveau werden behaald[54], gecombineerd met de resultaten van de statistische analyse laten ons toe de verschillen tussen landen beter te duiden.

De resultaten van de verdere analyse liggen in lijn met de observaties bij tabel 12; het duidelijkste verschil vinden we op het niveau van de bredere context (3.2) waar de lagere score voor de interventies in Rwanda opvalt voor alle vijf DAC-criteria[55]. Op item niveau vinden we enkel statistisch significante verschillen met betrekking tot de institutionele en politieke context (3.2.1)[56]. De scores geven aan dat deze context de evalueerbaarheid van de interventies in Rwanda op een negatieve manier beïnvloedt omdat de uitvoering van onafhankelijke evaluaties sterk kan belemmerd worden. Echter, de socio-culturele context op het niveau van de doelgroepen (3.2.2) en de beschikbaarheid van lokale deskundigheid (3.2.3) zijn geen factoren die, wanneer een vergelijking tussen landen wordt gemaakt, de evalueerbaarheid van de interventies op een verschillende manier beïnvloeden. Ook vallen de relatief goede scores op voor DRC. Deze scores zijn niet significant verschillend van de scores voor Benin en België, m.a.w. ondanks de politieke en institutionele moeilijkheden in het land zijn er op basis van deze studie geen aanwijzingen dat die contextuele factoren de evalueerbaarheid van interventies in het land (meer) negatief beïnvloeden[57].

[54] De gemiddelde scores per land kunnen teruggevonden worden in tabel A2 1 van Bijlage 7. De resultaten en een korte beschrijving van de bijhorende statistische testen werden opgenomen in dezelfde bijlage.

[55] … maar niettemin blijft ook de Rwanda score globaal hoog (zeker in vergelijking met de andere dimensies). Het is hier echter wel van belang te wijzen op de kanttekeningen die geplaatst werden in vorig hoofdstuk bij de manier waarop 'context' in de studie werd benaderd (zie de bespreking onder deel 3).

[56] Zie ook bijlage 7 voor meer informatie m.b.t. de statistische analyses die werden uitgevoerd.

[57] Overigens is het belangrijk op te merken dat de verschillen tussen de landen ook kunnen te maken hebben met verschillen op het vlak van het interventiekanaal (bilateraal versus andere) en de sector/het type interventie (dienstverlening versus – bijvoorbeeld – lobby en advocacy). Er is getracht om per land een vrij gelijkaardige steekproef samen te stellen, maar dit is slechts gedeeltelijk gelukt omwille van het beperkt aantal

Voor de andere significante verschillen is het moeilijker een eenduidige interpretatie te vinden gezien de deze verschillen zich niet altijd tussen dezelfde landen bevinden. Wat echter wel opvalt is dat bij de significante verschillen, Rwanda vaak als één van de beste scoort. Het lijkt er dus op dat de context in Rwanda de evalueerbaarheid van interventies zowel op een positieve (1.1, 1.2, 1.3, 2.1 en 2.2) als een negatieve (3.2 en 1.4) manier beïnvloedt. We kunnen dit verklaren door de erg duidelijke beleidscontext (zowel op papier als in de praktijk) waardoor Rwanda zich onderscheidt van de andere landen. De performantie-cultuur in het land zorgt er onder meer voor dat voor elke interventie een duidelijke rationale en probleemsituatie kan/moet worden geformuleerd, dat de band tussen analyse en interventiedoelen duidelijk en eenduidig is/moet zijn, en dat interventies meer dan elders oog (moeten) hebben voor de relatie tussen interventie en het beleid van het land. Deze doorgedreven performantie-cultuur zorgt er ook voor dat M&E systemen, vooral wanneer deze gericht zijn op opwaartse accountability, sterker ontwikkeld zijn. De noodzaak aan afstemming op lokale systemen beïnvloedt evenzeer de duurzaamheid van interventies (zie hoge scores voor Rwanda op duurzaamheids-criterium onder componenten 1.2 en 1.3).

Zoals eerder aangegeven moeten we, gezien het beperkt aantal observaties voor deze component, voorzichtig zijn met de interpretatie van de component betreffende de consistentie en adaptatie van de interventielogica en de veranderingstheorie (1.4)[58]. Wat betreft de evalueerbaarheid van efficiëntie, impact en duurzaamheid vinden we geen significante verschillen tussen de landen. Er is tussen de landen ook geen significant verschil in het aangeven en beargumenteren van de eventuele veranderingen (1.4.1). De scores zijn enkel verschillend voor de relevantie- en effectiviteitscriteria bij de items met betrekking tot de verwerking van de veranderingen in het M&E systeem (1.4.3) en de beschikbaarheid van informatie over de visie en meningen van de voornaamste betrokkenen m.b.t. de eventuele veranderingen (1.4.2). Vooral de interventies in Rwanda scoren lager op deze sub-items. Een doorgedreven performantie-cultuur kan de scores bij het beantwoorden van deze component op verschillende manieren beïnvloed hebben. De sterke nadruk op *'quick wins'* kan er bv. voor zorgen dat er minder openheid is bij interventies om eventuele moeilijkheden, falen en daaraan gerelateerde veranderingen in interventies te rapporteren. Dit hangt sterk samen met de hierboven beschreven neiging tot zelfcensuur. Bovendien is het zo dat het gezien de snelheid waarmee hervormingen (uitgewerkt door de overheid) effectief worden geïmplementeerd in Rwanda, het voor interventies onmogelijk is om alle veranderingen die hierdoor veroorzaakt worden te rapporteren. Dit zorgt er echter voor dat er een kloof ontstaat tussen wat op papier bestaat enerzijds en de werkelijkheid anderzijds, wat de evalueerbaarheid negatief beïnvloedt.

Voor de relatief hoge score in DRC voor wat betreft de kwaliteit van de interventielogica en veranderingstheorie hebben we niet dadelijk een verklaring. Een mogelijk verklarende factor is het feit van 7 van de 10 interventies in DRC een tweede of derde fase zijn van eerdere interventies, wat tot een geleidelijke verbetering van de interventielogica en veranderingstheorie kan hebben bijgedragen.

interventies per land en de vele parameters waarmee rekening moest worden gehouden; verder is de steekproef te klein om eventuele interactie-effecten te checken. In dit verband is de situatie m.b.t. de interventies in België vrij apart. Vele interventies in België verschillen sterk van die in het Zuiden omdat zij in eerste instantie een aanbod uitwerken voor hun doelgroepen die eerder 'consumenten' zijn dan actieve betrokken actoren; als een gevolg zijn ze ook minder bereikbaar en doorgaans minder geïnteresseerd in evaluaties. Andere specifieke problemen in sommige van deze interventies zijn de veranderende doelgroepen en de moeilijkheid om doelgroepen die in eerdere fases werden bereikt naderhand nog te traceren.

[58] Dit punt geldt ook bij de comparatieve analyses volgens andere parameters die hierna volgen, en zal daar niet worden herhaald.

4.2 Vergelijking van de evalueerbaarheid op basis van de complexiteit van de interventies

De tabel hierna presenteert de scores op *componenten*niveau en voor wat betreft de drie grote dimensies van het studiekader. Zoals dit het geval was voor de analyse op landenniveau, is het belangrijk te vermelden dat gelijkaardige scores op componentenniveau verschillen op *item*niveau kunnen verbergen. Deze potentiële verschillen worden geanalyseerd door de gemiddelde scores per DAC-criterium op componenten- en itemniveau tussen complexe en minder complexe interventies te vergelijken.

Tabel 13: Overzicht van de evalueerbaarheidsscores voor interventies met een complexe en minder complexe TOC

Item	Minder complexe interventies	Complexe interventies	Evalueerbaar-heidsindex (°)
Dimensie 1 (Interventieplan)	**2,87**	**3,30**	**3,16**
1.1 De onderliggende analyse	3,31	3,83	3,65
1.2 De interventielogica en veranderingstheorie	3,03	2,92	2,96
1.3 Het voorgestelde M&E systeem	2,11	2,78	2,51
1.4 Consistentie en adaptatie van de interventielogica en de veranderingstheorie	3,03	3,67	3,54
Dimensie 2 (Uitvoeringspraktijk)	**2,82**	**3,07**	**2,96**
2.1 Basisinformatie m.b.t. de interventie uitvoering	2,68	3,03	2,88
2.2 M&E systeem in de praktijk	2,95	3,10	3,04
Dimensie 3 (Context)	**3,82**	**3,98**	**3,91**
3.1 Houding van de sleutel actoren	3,44	3,80	3,66
3.2 Bredere context	4,20	4,15	4,17
Globale score evalueerbaarheid (°°)	**3,04**	**3,34**	**3,23**

(°) De score m.b.t. de globale evalueerbaarheidsindex is een score berekend op basis van indexen (geen gemiddeldes). Anderzijds zijn de scores berekend voor complexe en minder complexe interventies *gemiddelden* (geen indexen). Als zodanig is het mogelijk dat de globale evalueerbaarheidsscore lichtjes afwijkt van het gewogen gemiddelde van de scores van mindere complexe (40%) en complexe (60%) interventies. De verschillen zijn evenwel erg klein en veranderen niets aan de inhoud van de analyse.
(°°) Voor de globale score evalueerbaarheid werd een gewogen gemiddelde berekend waarbij dimensie 1 en 2 even zwaar worden gewogen en dimensie 3 meetelt voor de helft van elk van de andere dimensies.

Uit de tabel hierboven blijkt dat 'complexe interventies' (d.w.z. interventies met een complexe TOC) door de band (voor 6 van de 8 componenten) iets beter scoren dan minder complexe interventies. Dit is een verrassende conclusie omdat we initieel vertrokken zijn van de onderstelling dat complexe interventies moeilijker evalueerbaar zijn dan minder complexe. Deze hypothese schijnt eigenlijk alleen op te gaan voor component 1.2 waar we vaststellen dat de score voor *'de interventielogica en de veranderingstheorie'* iets hoger ligt bij minder complexe interventies, maar het verschil is zo miniem dat we hier eigenlijk geen conclusies uit mogen trekken. Als we al een verklaring kunnen geven voor de vaststelling op het niveau van deze component, dan zou het te maken kunnen hebben met het feit dat het bij complexe interventies minder evident is om kritieke elementen te identificeren omdat er minder empirische evidentie bestaat waarop men zich kan baseren.

Een algemene verklaring voor de relatief goede score van complexe interventies zou kunnen zijn dat men in deze interventies juist omwille van de TOC complexiteit meer aandacht gaat besteden aan, bijvoorbeeld, de onderliggende analyse en het M&E systeem. Dit zou impliceren dat de betrokken actoren meer aandacht besteden aan 'leren' omdat men beseft dat er nog maar weinig empirische evidentie beschikbaar is en men dus meer aandacht moet besteden aan analyse en het M&E systeem om bijstelling beter mogelijk te maken. Anderzijds is men zich, vanuit *accountabilty* standpunt, bewust van het feit dat de donor misschien meer sceptisch staat tegenover 'moeilijke' complexe interventies waarvan de effectiviteit moeilijker te bewijzen is.

Om een beter inzicht te kregen in de verschillen m.b.t. de evalueerbaarheid tussen interventies met een complexe en interventies met een minder complexe TOC, hebben we een vergelijking gemaakt tussen de gemiddelde scores die per DAC-criterium op componenten- en itemniveau werden behaald[59]. Bij de vergelijking tussen interventies met een complexe en minder complexe TOC, vinden we enkel significant verschillende scores voor 2 componenten, namelijk voor het voorgestelde M&E systeem (1.3) en de houding van de sleutelactoren ten aanzien van onafhankelijke evaluaties (3.1). De verschillen bevinden zich op het niveau van relevantie, effectiviteit en efficiëntie wat betreft het eerste item en enkel op het niveau van effectiviteit en efficiëntie wat betreft het tweede item.

Bij een nadere analyse van deze beide componenten valt in de eerste plaats op dat interventies met een complexe TOC significant hoger scoren. Wanneer we de vergelijking maken tussen de onderliggende items dan zien we dat deze trend behouden blijft, met uitzondering van het item dat handelt over de identificatie van de gebruikers van een evaluatie en hun verwachtingen (3.1.1). Voor interventies met een complexe TOC kan het identificeren van de gebruikers van een evaluatie en hun verwachtingen moeilijker zijn omwille van, bijvoorbeeld, het grote aantal verschillende actoren met verschillende verwachtingen binnen dergelijke interventies, omwille van het feit dat er vaak nog minder ervaring is met een complexe(re) veranderingstheorie waardoor er meer onzekere factoren kunnen zijn en de identificatie van de gebruikers en hun verwachtingen moeilijker is, …

Dat interventies met een minder complexe TOC voor de vermelde DAC-criteria significant lager scoren op het voorgestelde M&E systeem (1.3) en daarbinnen lager op het item betreffende de vertaling van de onderliggende interventielogica in het voorgestelde M&E system lijkt moeilijk te verklaren… Misschien speelt mee dat zulke interventies vooral worden uitgevoerd door minder ervaren organisaties, waardoor ook technische componenten of items zwakker scoren[60].

De hogere score voor interventies met een meer complexe TOC wat de onderlinge compatibiliteit van de verwachtingen betreft (3.1.3), kan, zoals reeds aangehaald in deel 3.1 van vorig hoofdstuk, mede een verklaring vinden in volgende redenering: er zijn weinig of minder problemen omdat, zoals door de score voor 3.1.1 aangegeven, de definiëring van de voornaamste gebruikers en hun verwachtingen niet systematisch gebeurt. Op die wijze blijven mogelijke verschillen in verwachtingen en belangen van deze gebruikers onderbelicht.

De hogere score voor interventies met een complexe TOC met betrekking tot het item of de sleutelactoren vragende partij zijn voor een evaluatie of daarin minstens geïnteresseerd (3.1.2) kan verschillende oorzaken hebben (zoals ook reeds aangehaald in deel 3.1 van vorig hoofdstuk) : interventies die lager scoren op dit item hebben geen expliciet negatieve houding t.a.v. externe evaluaties omdat de lagere score kan duiden op de veronderstelling dat externe evaluaties geen extra meerwaarde zullen bieden ten aanzien van reeds bestaande interne evaluaties en de voorhanden zijnde monitoring resultaten. Als interventies met een minder complexe TOC gemakkelijker monitoring resultaten kunnen opvolgen, bijvoorbeeld, dan is het mogelijk dat in deze interventies de meerwaarde van onafhankelijk evaluaties lager wordt ingeschat.

In de meeste interventies zijn de relaties tussen de sleutelactoren gezond. Het kan zijn dat interventies met een minder complexe TOC lager scoren op dit item (3.1.7) omdat, zoals we hierboven reeds besproken hebben, de interventies met een complexe TOC een minder duidelijke definiëring hebben gemaakt van de gebruikers en hun belang en verwachtingen m.b.t. evaluatiedoelen en –inhoud. Zoals we aangegeven hebben onder punt 3.1 van vorig hoofdstuk, is het mogelijk dat een meer 'bewuste' definiëring van belangen en verwachtingen m.b.t. evaluatiedoelen en –inhoud de relaties meer op

[59] De gemiddelde scores kunnen teruggevonden worden in tabel A2 7 van bijlage 7. Een korte beschrijving van de bijhorende statistische testen werden opgenomen in Annex 7.

[60] De steekproef is evenwel te klein om hier voldoende betrouwbare uitspraken over te doen.

scherp zou kunnen stellen. Anderzijds, als we het resultaat *'de relaties tussen de sleutelactoren zijn gezond'* combineren met de hoge score (hoogste score onder deze component) voor het item *'het is mogelijk om alle sleutelactoren te contacteren zonder risico op wederzijdse beïnvloeding'* zou dit er ook kunnen op wijzen er bij de interventies met een meer complexe TOC gezien de onzekerheid meer interesse is in verschillende opinies (om van daaruit meer te kunnen leren).

Op het niveau van de evalueerbaarheid van impact en duurzaamheid zijn er geen significante verschillen tussen interventies met een complexe TOC en interventies met een minder complexe TOC. Een verklaring hiervoor kan gevonden worden in wat reeds over deze niveaus in de bovenstaande analyses werd beschreven, d.i. een mindere uitwerking van deze niveaus in vergelijking met de andere drie DAC criteria. Verder merken we graag op dat er, voor wat betreft de componenten die de onderliggende analyse (1.1) en de onderliggende interventielogica en de veranderingstheorie (1.2) omvatten, voor geen enkele van de relevante DAC-criteria significante verschillen gevonden worden tussen interventies met een complexe TOC en interventies met een minder complexe TOC. Hoewel interventies met een complexe TOC eerder geassocieerd worden met meer moeilijkheden inzake uitbouw van de interventielogica en TOC, bijvoorbeeld, blijkt dit in de praktijk niet te leiden tot analyses en/of interventielogica die de evalueerbaarheid van deze interventies negatief beïnvloeden. Voor alle interventies zijn er relatief goede scores voor deze componenten. Ook bleek uit terreinbezoeken dat bij interventies met een meer complexe TOC, ondanks de moeilijkheden, veel energie wordt besteed om onderliggende analyses en veranderingstheorieën uit te werken en concreet te vertalen in een consistente interventielogica.

4.3 Vergelijking van de evalueerbaarheid op actor-niveau

De tabel hierna presenteert de scores op *componenten*niveau en voor wat betreft de drie grote dimensies van het studiekader. Zoals dit het geval was voor de analyse op landen- en complexiteitsniveau, is het belangrijk te vermelden dat gelijkaardige scores op componentenniveau verschillen op *item*niveau kunnen verbergen. Deze potentiële verschillen worden geanalyseerd door de gemiddelde scores per DAC-criterium op componenten- en itemniveau tussen de drie types actoren te vergelijken.

Tabel 14: Overzicht van de evalueerbaarheidsscores per actor

Item	BTC	NGOs en vakbonden	Andere	Evalueerbaar-heidsindex (°)
Dimensie 1 (interventieplan)	**3,52**	**3,20**	**2,59**	**3,16**
1.1 De onderliggende analyse	4,10	3,75	2,90	3,65
1.2 De interventielogica en veranderingstheorie	3,00	2,98	2,90	2,96
1.3 Het voorgestelde M&E systeem	2,82	2,70	1,82	2,51
1.4 Consistentie en adaptatie van de interventielogica en de veranderingstheorie	4,17	3,38	2,75	3,54
Dimensie 2 (uitvoeringspraktijk)	**2,97**	**3,09**	**2,70**	**2,96**
2.1 Basisinformatie m.b.t. de interventie uitvoering	2,86	3,03	2,62	2,88
2.2 M&E systeem in de praktijk	2,95	3,10	2,78	3,04
Dimensie 3 (context)	**4,07**	**3,93**	**3,72**	**3,91**
3.1 Houding van de sleutel actoren	3,74	3,72	3,44	3,66
3.2 Bredere context	4,40	4,14	4,00	4,17
Globale score evalueerbaarheid (°°)	**3,41**	**3,30**	**2,86**	**3,23**

(°) De score m.b.t. de globale evalueerbaarheidsindex is een score berekend op basis van indexen (geen gemiddeldes). Anderzijds zijn de scores berekend voor de actoren gemiddelden (geen indexen). Als zodanig is het mogelijk dat de globale evalueerbaarheidsscore lichtjes afwijkt van het gewogen gemiddelde van de interventies, interventies van 'BTC', van 'NGOs en vakbonden' en van 'Andere' maken respectievelijk 25%, 50% en 25% van het bestand uitmaken.
(°°) Voor de globale score evalueerbaarheid werd een gewogen gemiddelde berekend waarbij dimensie 1 en 2 even zwaar worden gewogen en dimensie 3 meetelt voor de helft van elk van de andere dimensies.

Uit de tabel hierboven blijkt dat, in vergelijking met de twee vorige parameters, er tussen de verschillende actoren of 'kanalen' relatief grotere verschillen qua evalueerbaarheid optreden. Deze doen zich uiteraard niet zozeer voor op het vlak van de derde dimensie (invloed van de context), maar wel bij de twee andere dimensies en dan vooral bij dimensie 1 (het interventieplan). Uit de gegevens van de tabel blijkt dat de evalueerbaarheid van BTC en NGO/vakbondsinterventies een stuk hoger ligt dan die van de andere actoren, wat niet impliceert dat deze verschillen noodzakelijk opgaan voor individuele interventies binnen elke actorgroep; dit geldt a fortiori voor de groep 'Andere actoren' die een brede waaier aan organisaties en praktijken bevat.

Een nadere bestudering van bovenstaande tabel en een vergelijking van de gemiddelde scores die per DAC-criterium op componenten- en itemniveau werden behaald[61] laten ons toe de verschillen tussen actoren beter te duiden. De resultaten van de statistische testen (zie bijlage 7) tonen aan dat er tussen de drie verschillende actoren enkel significante verschillen zijn op het niveau van het ontwerp van de interventie (deel 1). Anderzijds hebben de context en de praktijk van uitvoering van de interventie voor deze drie interventiekanalen geen significant verschillende invloed op de evalueerbaarheid van de interventies.

Verder blijkt uit de statistische analyse dat de verschillen betreffende het ontwerp van de interventie zich situeren op het niveau van de onderliggende analyse (1.1) en het voorgestelde M&E systeem (1.3).

Het verschil in de onderliggende analyse (1.1) uit zich voornamelijk in verschillen in de items die handelen over de afbakening en beschrijving van doelgroepen (1.1.1), de beschrijving van de rol van doelgroepen en van de rol van de actoren (1.1.3 en 1.1.4) en het al dan niet aanwezig zijn van een goede genderanalyse (1.1.5). Wat betreft de eerste drie vermelde items vallen vooral de hoge scores van BTC interventies en de lage scores van de interventies in het "Andere" interventiekanaal op. Mogelijke verklaringen hiervoor kunnen gevonden worden in het feit dat er bij de BTC aan alle interventies gevraagd wordt dezelfde, goed uitgewerkte en gedetailleerde formats te gebruiken bij het opstellen van interventievoorstellen en het globale technische en financiële dossier. In deze formats wordt ook voldoende belang gehecht aan het beschrijven van de doelgroep en de rol die doelgroepen en/of verschillende actoren zullen opnemen in de interventie. Verder gaat zowel bij BTC (via het MoreResults traject) als de NGOs (via initiatieven van de federaties en veranderingstrajecten bij vele NGOs) veel aandacht naar verbetering van het projectbeheer. Dit staat in contrast met het interventiekanaal "Andere" waar in bepaalde gevallen zelfs geen projectvoorstel en/of onderliggende analyse wordt vereist (om fondsen aan te vragen) en waar de gebruikte formats vaak minder volledig en meer divers blijken te zijn. Daarbij komt dat voor sommige van deze "Andere" actoren (zoals bijvoorbeeld de steden en gemeenten) ontwikkelingssamenwerking slechts een nevenactiviteit vormt waarin ze niet echt zijn gespecialiseerd.

Wat betreft de genderanalyse (1.1.5), is er een hogere score terug te vinden voor het interventiekanaal 'NGO's en vakbonden'. Een mogelijke verklaring hiervoor kan zijn dat deze actoren historisch een grotere focus hebben op het bereiken van sociaal zwakkere doelgroepen waar gender relaties veel makkelijker kunnen gelinkt worden aan onderliggende verklarende mechanismen, waardoor de noodzaak voor integratie van genderanalyse en vertaling naar het M&E systeem meer voor de hand ligt. De relatief zwakke score van de BTC-interventies kan zijn oorzaak vinden in het feit dat, wanneer de onderzocht interventies werden gepland en uitgevoerd, er slechts sporadisch aandacht was voor gender. De aandacht voor gender in de MoreResults handleiding verschilt niet wezenlijk van die in eerdere richtlijnen.

Er worden tussen de drie verschillende interventiekanalen verder geen significant verschillende scores gevonden voor de items betreffende de rationale van de interventie

[61] De gemiddelde scores kunnen teruggevonden worden in tabel A2 4 bijlage 7. De resultaten en een korte beschrijving van de bijhorende statistische testen werden opgenomen in dezelfde bijlage.

(1.1.2), de band tussen de onderliggende analyse en de interventiedoelen (1.1.6) en de positie van de interventie t.a.v het lokale sectorbeleid (1.1.7). Dit is een bijkomende indicatie om te stellen dat goede formats een grote rol spelen bij de uitwerking van het plan maar dat er achter de interventies van de drie verschillende interventiekanalen wel degelijk een rationale schuilt waarbij logische interventiedoelen worden vooropgesteld en waarbij rekening gehouden wordt met de beleidscontext waarbinnen de interventie kadert. Er worden verder ook geen significante verschillen gevonden tussen de drie interventiekanalen wat betreft de interventielogica en de veranderingstheorie (1.2). Deze bevinding wijst er mede op dat per interventiekanaal geen verschillen zijn in de technische competentie om interventies te kaderen binnen een veranderingstheorie en om een interventielogica uit te werken.

Wat betreft de verschillen die voor de drie verschillende interventiekanalen worden gevonden op het niveau van het voorgestelde M&E systeem (1.3) zien we drie verschillende "trends". Vooreerst hebben de interventies van het interventiekanaal "Andere" een significant lagere score op het componentniveau van het voorgestelde M&E systeem die ook terugkomt in de items met betrekking tot het operationaliseren van de interventieresultaten (1.3.1), de werkwijze om de interventieresultaten op te volgen en te evalueren (1.3.4), de beschrijving van de financiële en personele middelen van het M&E systeem (1.3.7), het al dan niet aanwezig zijn of gebruik van een MIS (1.3.8) en het M&E systeem als consistente vertaling van de interventielogica (1.3.3). Zoals reeds eerder vermeld, gelden er voor dit type van interventiekanalen andere voorwaarden en vereisten wat betreft rapportage en inhoudelijke format van rapporten. Binnen dit type interventiekanaal is het niet altijd een vereiste om geoperationaliseerde doelstellingen voorop te stellen en wordt er ook niet verwacht dat deze op een planmatige manier worden opgevolgd. Ook is het niet verwonderlijk dat praktische zaken zoals de middelen ter beschikking van het M&E systeem minder goed worden belicht gezien er bij de interventies in dit type interventiekanaal veel minder sprake is van een reeds bestaande M&E cultuur en praktijk. Deze algemene bevinding sluit uiteraard niet uit dat bepaalde individuele interventies wel goede M&E praktijken hebben voorgesteld en uitgewerkt. Daarbij kan worden opgemerkt dat de scores m.b.t. M&E systeem in de praktijk veel dichter bij elkaar liggen dan bij het voorgestelde M&E systeem.
Ten tweede merken we op dat de interventies in het interventiekanaal "NGO en vakbonden" significant hoger scoren voor wat betreft de duidelijke beschrijving van de werkwijze om assumpties op te volgen (1.3.5). Een verklaring zou kunnen zijn dat voor de interventies van NGO's en vakbonden het overheidsbeleid behoort tot de externe context en dat men zich in deze interventies bewust is dat het overheidsbeleid deel uitmaakt van de externe assumpties die de interventies kunnen beïnvloeden. Anderzijds behoort het overheidsbeleid voor de bilaterale interventies veel minder tot de externe assumpties (maar eerder tot interne risico's) en zijn andere externe assumpties die de interventies potentieel kunnen beïnvloeden veel moeilijker te identificeren. Interventies in het interventiekanaal "Andere" zijn dan weer minder gebonden aan een overheidsbeleid en kunnen moeilijkheden ondervinden om andere externe assumpties te identificeren.
Ten derde scoren de interventies binnen het interventiekanaal "BTC" hoger wat betreft een duidelijke beschrijving van de wijze waarop het M&E systeem is afgestemd op lokale/nationale M&E systemen (1.3.9). Dit verschil kunnen we gemakkelijk verklaren vanuit de uitvoeringsmodaliteiten waarbinnen BTC interventies uitgevoerd worden en waarbij een afstemming op lokale/nationale politiek en de daarbij horende opvolgingssystemen een meer expliciet aandachtspunt is.

5 Conclusies en aanbevelingen

5.1 Voornaamste conclusies

De hierna voorgestelde conclusies en geleerde lessen worden in twee delen voorgesteld. Het eerste deel is vooral een samenvatting van de voornaamste bevindingen, terwijl het tweede deel vooral analytisch van aard is.

5.1.1 Synthese van de voornaamste resultaten en bevindingen

Betekenis en belang van "evalueerbaarheid" in de ontwikkelingspraktijk

Evalueerbaarheid wordt gedefinieerd als de mate waarin een activiteit of project op een betrouwbare en geloofwaardige wijze kan worden geëvalueerd. Het bepalen van de "haalbaarheid" van een evaluatie wordt daarbij verbreed tot het nagaan van de *wenselijkheid* van een evaluatie. Dit impliceert dat evalueerbaarheid zich vooral op drie dimensies concentreert: (de kwaliteit van) het interventieontwerp, (de kwaliteit van) de interventie-uitvoering (inclusief het M&E systeem) en de rol van de bredere interventieomgeving.

Het eerste en meest evidente belang van evalueerbaarheid vertrekt van de overweging dat het uitvoeren van een evalueerbaarheidstoets (of –beoordeling) slechts een fractie van de middelen vereist die nodig zijn voor de uitvoering van de evaluatie als zodanig, terwijl de potentiële meerwaarde van deze check aanzienlijk is. Via een beoordeling van de evalueerbaarheid wordt met andere woorden duidelijk of een evaluatie (op een bepaald tijdstip, in een bepaalde context) wenselijk en haalbaar is, en zo ja onder welke voorwaarden.

Een adequaat gebruik van "evalueerbaarheid" en de uitvoering van een evalueerbaarheidsbeoordeling dienen evenwel niet noodzakelijk met de opzet van evaluaties te worden verbonden maar hebben *een ruimer nut* dat het beheer van ontwikkelingsinterventies in zijn geheel ten goede kan komen. Zo kan een analyse van de evalueerbaarheid het interventieontwerp verbeteren, een belangrijke inbreng leveren voor het ontwerp en de uitwerking van een M&E systeem en belangrijke inputs leveren m.b.t. wenselijkheid, timing, aanpak en doelen van een evaluatie en op die wijze bijdragen tot de bruikbaarheid en effectief gebruik van de evaluatieresultaten en dus het uiteindelijk nut van de evaluatie.

Uit het bovenstaande blijkt het belang van zowel het concept evalueerbaarheid als van het instrument "evalueerbaarheidstoets": hun potentiële toepassing gaat een stuk verder dan (het organiseren van) evaluaties en betreft het gehele interventiebeheer. Echter heeft de studie vastgesteld dat het concept "evalueerbaarheid" en het instrument "evalueerbaarheidstoets" tot nu toe weinig gekend en gebruikt worden in de Belgische ontwikkelingssamenwerking. Elementen van een evalueerbaarheidstoets worden hier en daar wel in de praktijk gebracht (zonder ze als zodanig te benoemen), maar nergens is er sprake van een systematische toepassing.

Deze vaststelling is betekenisvol omdat over de laatste decennia "evaluatie" gaandeweg belangrijker is geworden en een vast onderdeel is geworden van de beheerspraktijk op het niveau van interventies, programma's en de ontwikkelingssamenwerking in haar

geheel: de relevantie en het nut van evaluaties worden niet in vraag gesteld en actoren kunnen het zich – eenvoudigweg – niet meer veroorloven om *niet* te evalueren. Evalueren is met andere woorden een verplichting geworden. Hoewel deze ontwikkeling in haar algemeenheid positief is, brengt ze ook het gevaar met zich mee dat evaluaties verworden tot rituele oefeningen, zonder authentieke betrokkenheid van de sleutelactoren. Een bewust gebruik van de notie van 'evalueerbaarheid' en het instrument 'evalueerbaarheidstoets', met daaraan gekoppeld de mogelijkheid tot gefundeerde uitspraken m.b.t. de wenselijkheid van een evaluatie, kan in deze context een belangrijk instrument zijn tot een verbeterde en meer relevante en realistische invulling van de rol en functie van evaluatie(s) in de ontwikkelingssamenwerking.

Evalueerbaarheid globaal bekeken

De 40 geanalyseerde interventies halen een globale score (op basis van de 62 items van het studiekader) die zich iets boven de midden-score situeert (zie tabel 3). Hoewel we aan de scores geen absolute betekenis kunnen geven, vormt de globale score een goede indicatie van de voornaamste vaststelling van deze studies: **interventies hebben in het algemeen een aantal sterke punten, maar ook nog een groot aantal werkpunten als ze hun evalueerbaarheid willen verbeteren**. Verder valt op dat de (statistische) distributie van de scores een normaalverdeling benadert, maar wel een grote spreiding kent: de laagst scorende interventies komen qua score nog niet aan de helft van de hoogst scorende. Omdat "evalueerbaarheid" sterk is verbonden met de interventie-beheerspraktijk, impliceert deze vaststelling dat er qua beheer nog grote verschillen bestaan onder de Belgische actoren en types interventies, ondanks een sturing van de financierende overheid.

Van de drie onderzochte dimensies (interventieplan, uitvoeringspraktijk, context) scoort de laatste merkelijk hoger dan de andere. Dit vormt een bijkomende illustratie van de marge voor verbetering van die dimensies waar actoren het meest vat op hebben (interventieplan en uitvoeringspraktijk). Verder suggereert de score dat, tenminste voor de 4 onderzochte landen, de uitdagingen op contextniveau de evalueerbaarheid niet zwaar hypothekeren. Bij nader inzien moeten we deze interpretatie evenwel nuanceren omdat de eventuele beperkingen op contextniveau niet in de diepte geanalyseerd konden worden.

Als we kijken naar de globale scores per evaluatiecriterium vallen de lagere scores voor duurzaamheid en vooral impact op: ondanks de toegenomen aandacht voor duurzaamheid, blijft deze onvoldoende geïntegreerd in de beheerssystemen; wat impact betreft, zorgt een samenspel van factoren voor een moeilijke evalueerbaarheid. Anderzijds scoren effectiviteit en vooral efficiëntie aanmerkelijk hoger dan de midden-waarde van de ontwikkelde schaal, wat een indicatie vormt - doorheen de verschillende dimensies en componenten - van kwaliteit in het interventiebeheer, met name wat betreft monitoring en evaluatie.

Het interventieontwerp

De globaal goede score m.b.t. **de onderliggende analyse** wijst op een zekere traditie en vaardigheid m.b.t. uitwerking van analyses die het gevolg blijkt te zijn van zowel interne processen als eisen van de donoren voor wie zulke analyse een belangrijk onderdeel vormt bij de beoordeling van subsidieaanvragen. Verder kunnen de initiatiefnemers in vele gevallen bogen op ervaringen uit het verleden (vervolg-interventies). Een belangrijke positieve vaststelling is de ruime aandacht voor de beleidscontext in vele analyses, die voor een stuk voortkomt uit de betere netwerking en platformwerking tussen de Belgische actoren.

Anderzijds zijn er op het vlak van de analyse ook enkele belangrijke knelpunten: de zwakke aandacht, op het vlak van de doelgroepen, voor gender en (factoren van) sociale differentiatie in het algemeen; de beperkte aandacht voor sleutelactoren (andere dan doelgroepen) en de vernauwing van de analyses die dikwijls eerder dienen om

strategische en beleidskeuzes te verantwoorden dan dat ze er de basis van vormen. Deze lacunes beïnvloeden de evalueerbaarheid negatief omdat ze de beoordeling van de relevantie van de gemaakte beleidskeuzes bemoeilijken en het o.m. moeilijk maken om te oordelen wie effectief wordt bereikt (in vergelijking met de planning), of er op doelgroep niveau verdringingseffecten hebben gespeeld.

De component **interventielogica en veranderingstheorie** scoort relatief zwak, wat te wijten is aan de geringe aandacht voor hogere niveaus in de doel-middelen ketting. In de praktijk wordt vooral aandacht besteed aan het implementatieniveau (bij screening wordt daar vooral naar gekeken) en blijken er weinig incentives en dus belangstelling te bestaan om dat niveau te overstijgen: niet alleen heeft het implementatie-niveau een groter en direct nut voor de betrokken organisaties, evenzo is de financierende overheid de facto vooral in verantwoording (m.b.t. de correcte aanwending van middelen) geïnteresseerd en besteden bijvoorbeeld de ontwikkelde formats weinig aandacht aan het impact niveau. Een bijkomende moeilijkheid is dat impacts, als ze dan al geformuleerd worden, zich op erg grote afstand in de doel-middelen ketting situeren, waardoor er een '*missing middle'* ontstaat in de veranderingstheorie. Omwille van de bestaande intrinsieke uitdagingen, het ontbreken van externe incentives en de beperkte kennis en inzicht in bestaande (en realistische) methodes en opportuniteiten van impactevaluatie, bestaan er reeds van in de voorbereidingsfase belangrijke moeilijkheden die de evalueerbaarheid van impact en de daadwerkelijke uitvoering van impactevaluaties (d.w.z. voor wat betreft de outcomes en de directe effecten van deze outcomes) in een later stadium hypothekeren.

Aan de andere kant zorgt de goede focus op 'uitvoering' voor een goede score voor het efficiëntie criterium die zich ook in de daadwerkelijke praktijk doortrekt (bijvoorbeeld in de goede koppeling van het gebruik van middelen aan de outputs).

De kwaliteit van **het M&E voorstel in het interventievoorstel** scoort zwak. Hiervoor zijn er verschillende oorzaken: ontbreken van incentives om ex ante veel te investeren in de beschrijving van het M&E systeem, het feit dat slechts recent de aandacht voor M&E fors is toegenomen (vele actoren hebben hard gewerkt aan de uitwerking van een M&E beleid) maar deze niet geheel is doorgedrongen tot op interventieniveau, en de keuze van sommige actoren om slechts bij de start van een interventie te investeren in de uitwerking van een M&E systeem. Vanuit evalueerbaarheidsstandpunt is de beschikbaarheid, ex ante, van een M&E ontwerp, slechts belangrijk voor zoverre dit naderhand een goede M&E praktijk impliceert. In de praktijk blijkt goede ex ante aandacht inderdaad dikwijls een voorbode van een goede praktijk, maar zijn er ook interventies waar een zwakke ex ante score naderhand gecorrigeerd wordt door een goede praktijk. Verder blijken (op papier) goed uitgewerkte M&E voorstellen ook belangrijk in het licht van het grote personeelsverloop en de vele taakveranderingen in vele interventies. Tenslotte is duidelijk dat initiële pijnpunten (bijvoorbeeld zwakke aandacht voor impact en gender; geen specifieke beschrijving en toewijzing van middelen voor monitoring en vooral voor evaluatie) later in vele gevallen terugkomen. Dit geldt – a fortiori – voor de vaststelling dat er slechts zelden sprake is van een daadwerkelijk M&E *systeem*; in vele gevallen worden componenten van zo een systeem voorgesteld, maar is er nog veel werk om een goede interne coherentie te waarborgen.

Veranderingen in interventielogica en veranderingstheorie tijdens de uitvoering worden meestal goed gesignaleerd en geargumenteerd, een positieve vaststelling vanuit het standpunt van evalueerbaarheid. Echter worden deze veranderingen in vele gevallen niet consistent doorgetrokken in het M&E systeem, wat een indicatie vormt van de nog zwakke institutionele inbedding van dit systeem.

De uitvoeringspraktijk

De **basisinformatie met betrekking tot de voortgang van de interventie-uitvoering** is relatief goed beschikbaar maar (onder meer omwille van de taal, soms ook de complexiteit) niet steeds toegankelijk (en gekend) voor alle sleutelactoren. Verder is deze informatie een afspiegeling van de informatie die beschikbaar is m.b.t.

het interventieplan: ze concentreert zich vooral op het implementatieniveau, blijft oppervlakkig voor wat betreft aspecten van sociale differentiatie (inclusief gender) en besteedt weinig aandacht aan hogere niveaus in de doel-middelenketting. Aansluitend doet de informatie een "*inward looking*" ingesteldheid in het interventiebeheer vermoeden, met weinig aandacht voor externe assumpties, de mogelijkheid tot ontwikkelen van een counterfactual en de aanpak en kwaliteit van het proces van gegevensverzameling. De vrij zwakke score wat betreft dit item moet echter worden genuanceerd omdat de veldbezoeken in een belangrijk aantal gevallen *good practices* aan het licht brachten waarvan in de basisdocumenten geen spoor was terug te vinden.

Het **M&E systeem in de praktijk** is waarschijnlijk de component die de evalueerbaarheid van een interventie het sterkst beïnvloedt. Zwakheden elders kunnen immers tot op zekere hoogte worden opgevangen door een coherent en goed functionerend M&E systeem; omgekeerd zal met name de praktische evalueerbaarheid van een interventie bemoeilijkt worden als het M&E systeem slecht functioneert. In het licht van deze overweging is het feit dat de *daadwerkelijke* M&E praktijk aanmerkelijk beter scoort dan het *voorgestelde* M&E systeem een positieve vaststelling. Bij nader inzien blijkt de M&E praktijk in vele gevallen 'werk in uitvoering' dat nog teveel tot het operationele niveau beperkt blijft om redenen die reeds gedeeltelijk hierboven werden aangegeven en terug te voeren zijn tot onvolkomenheden in het interventieontwerp en de goede maar al bij al relatief recente aandacht voor M&E die maakt dat beleidsbeslissingen op dit vlak nog niet zijn doorgetrokken naar het lokale niveau. Verder blijken een aantal mechanismen te spelen die de kenmerken van de M&E praktijk beïnvloeden: de nauwe relatie tussen operationele monitoring en de kerntaken van programmapersoneel (in tegenstelling tot resultaatsgerichte monitoring die dikwijls veeleisender is en "verderaf" staat), het ontbreken van indicatoren op impact en (soms) outcome-niveau zodat deze niveaus in de M&E praktijk onderbelicht blijven, het ontbreken van (aandacht voor) specifieke competentieontwikkeling op het vlak van M&E, het feit dat M&E voor velen een bijkomende verantwoordelijkheid is en het tekort aan tijd, middelen en procedures om de M&E resultaten goed te gebruiken voor grondige analyse, besluitvorming en bijsturing. Deze zwakke punten worden evenwel deels gecompenseerd door een op reflectie en leren gerichte organisatiecultuur. Verder blijkt dat M&E beleid en –praktijk tot nu toe vooral "monitoring" centraal stellen ten koste van evaluatie, wat verschillende oorzaken heeft: het gegeven dat het M&E beleid nog gestalte moet krijgen in de praktijk, waarbij het logisch is dat de aandacht initieel naar monitoring gaat, de grotere methodologische vereisten bij evaluatie in vergelijking met monitoring, en de twijfel (met name bij organisaties met een performant monitoringsysteem en in complexe interventies) t.a.v. de mogelijke meerwaarde van externe evaluaties waardoor er voor zulke evaluaties weinig middelen worden voorzien (zie ook hierna).

Een laatste belangrijke vaststelling heeft te maken met de sterke *'inward looking'* houding die de M&E praktijk van vele Belgische ontwikkelingsactoren karakteriseert en zich vooral vertaalt in een beperkte rol in M&E beleid en –uitvoering voor de interventie-actoren (andere dan de verantwoordelijke organisatie en het interventieteam), het reduceren van accountability tot donor-accountability en de beperkte aandacht voor de afstemming van M&E met lokale en nationale systemen (inclusief die van de partner) en voor de mogelijke rol die Belgische actoren in het opzetten of verbeteren van deze systemen kunnen/moeten spelen.

Samenvattend bevat de huidige uitvoeringspraktijk zowel mogelijkheden als beperkingen ten aanzien van evalueerbaarheid, waarbij het belangrijk lijkt te beseffen dat deze studie een momentopname levert van een praktijk die aan sterke verandering onderhevig is.

De invloed van de evaluatiecontext

De **houding van sleutelactoren** ten aanzien van (onafhankelijke) evaluaties is een belangrijke factor van evalueerbaarheid. Het onderzoek bracht aan het licht dat deze houding in weinig gevallen een rechtstreekse belemmering vormt. Sleutelactoren stellen zich in de regel positief en geïnteresseerd op t.a.v. externe evaluaties en weten ook

welke rol en houding (bv. respect voor de autonomie van de evaluator; geen beïnvloeding) van hen wordt verwacht. Echter geldt ook hier dat de realiteit wat complexer is dan op het eerste gezicht lijkt: in de praktijk blijkt immers dat slechts een zeer beperkt aantal actoren bij het evaluatieproces worden betrokken en de belangen en verwachtingen van een brede waaier aan actoren bij evaluaties niet worden meegenomen (o.m. als gevolg van een onduidelijk M&E beleid, een te zwakke evaluatiepraktijk en een scepticisme t.a.v. de meerwaarde van externe evaluaties). Door de belangen en verwachtingen van deze actoren niet mee te nemen, worden fundamentele keuzes m.b.t. evaluatie veelal bepaald door de '*usual suspects'* die op die wijze de evaluatie kunnen sturen alvorens die feitelijk van start is gegaan.

De resultaten m.b.t. de invloed van de **bredere context** leveren een gelijkaardig beeld op. Op het eerste zicht blijken onafhankelijke evaluaties zonder noemenswaardige moeilijkheden uitgevoerd te kunnen worden en kunnen de uitdagingen die zich aandienen adequaat worden opgelost. Echter is het mogelijk dat contextfactoren die de evalueerbaarheid onder druk (kunnen) zetten ook deze studie hebben beïnvloed (bijvoorbeeld via het verstrekken van sociaal gewenste antwoorden), temeer omdat ook deze studie (zoals vele evaluaties) onder aanzienlijke tijdsdruk werd uitgevoerd. Met andere woorden is het mogelijk dat het ontbreken van grote problemen kan worden verklaard vanuit een focus op de 'technische' dimensies van evalueerbaarheid ten koste van de politieke dimensies die om verschillende redenen niet expliciet werden meegenomen.

De invloed van de specifieke landen-context op de evalueerbaarheid

De evalueerbaarheidsscores per land verschillen erg weinig van elkaar, wat impliceert dat andere parameters waarschijnlijk van groter belang zijn. Slechts voor wat betreft de context zijn er om evidente redenen redelijk grote verschillen, maar niettemin is de invloed van de context op de 'technische' evalueerbaarheid niet zo groot in de zin dat die niet zwaar belemmerend werkt. Een bepaalde institutionele context kan overigens zowel positief als negatief inwerken op verschillende aspecten van evalueerbaarheid. In het geval van Rwanda bijvoorbeeld heeft de uitgesproken performantie-cultuur een positieve invloed op de vraag naar monitoring, het effectief uitwerken van monitoringssystemen (die vooral gericht zijn op opwaartse accountability), terwijl dezelfde focus op performantie en *quick-wins* het rapporteren van mislukkingen en veranderingen kan belemmeren wat tezelfdertijd een negatieve invloed uitoefent op de leer/feedbackfunctie en de evalueerbaarheid onder druk zet.

De invloed van de graad van complexiteit van de interventies op de evalueerbaarheid

De evalueerbaarheid van interventies met een 'complexe' en 'minder complexe' TOC verschilt niet wezenlijk van elkaar. Interventies met een complexe TOC scoren zelfs iets beter, mogelijk omdat actoren van deze interventies meer investeren in analyse en uitwerking van M&E systemen en –praktijk, omdat zij – terecht of ten onrechte – er van uitgaan dat deze interventies moeilijker financierbaar zijn en het moeilijker is om hun resultaten aan te tonen. Verder kunnen de kleine verschillen tussen beide types interventies ook te maken hebben met het feit dat bepaalde parameters van complexiteit (zoals het aantal betrokken actoren) bij de samenstelling van de twee groepen niet werden meegenomen.

De invloed van het financieringskanaal op de evalueerbaarheid

De evalueerbaarheidsscores per financieringskanaal (type Belgische actor) geven grotere verschillen aan dan voor de aard van de interventies en het land. Met name tussen 'bilateraal/NGOs/vakbonden' aan de ene kant en de 'andere actoren' aan de andere kant zijn de verschillen groot (betere scores voor de eerste groep), al zijn er ook in deze laatste groep voorbeelden van *good practices*. De voornaamste verklaring voor deze vaststelling ligt mogelijk in de (voor de 'andere actoren') lagere externe eisen ten

aanzien van interventieplan en uitvoeringspraktijk, die worden versterkt door het feit dat voor een deel van de tot deze groep behorende actoren 'ontwikkelingssamenwerking' geen hoofdtaak vormt. Anderzijds hebben BTC en NGOs/vakbonden ook als 'sector' sterk ingezet op verbeterde M&E (vooral monitoring).

5.1.2 Analyse

Als we de scores m.b.t. de vijf evaluatiecriteria onderling vergelijken, valt op dat bijna voor alle items (en dus ook alle componenten en dimensies) het criterium efficiëntie het hoogste scoort, gevolgd door effectiviteit, duurzaamheid en impact. De score voor relevantie situeert zich veelal in het midden. Het feit dat hetzelfde score-patroon bijna steeds terugkomt, heeft uiteraard te maken met de verschillende moeilijkheidsgraad – in alle fasen van de interventiecyclus - van de evaluatie van deze criteria. Hoewel het moeilijk is om op dit vlak te veralgemenen, kan men stellen dat duurzaamheid en vooral impact moeilijker te evalueren zijn dan de drie andere criteria. Impactevaluatie stelt hoge methodologische eisen terwijl de moeilijkheid bij de evaluatie van duurzaamheid meestal te maken heeft met de uitdaging om gefundeerde uitspraken te doen m.b.t. een situatie die zich pas in de toekomst zal voordoen.

Er zijn evenwel ook andere redenen die het verschil tussen de scores van de vijf criteria kunnen verklaren. In de eerste plaats zijn de effecten van de sturing vanuit de donoren vooral voelbaar op efficiëntieniveau. Donoren (en dan met name DGD) zijn in eerste instantie bekommerd om een correcte aanwending van de ter beschikking gestelde publieke fondsen en hebben daartoe belangrijke voorwaarden opgelegd voor het beheer van de interventies die zij financieren. Procedures en formats expliciteren in belangrijke mate de prioriteiten van de donor al kunnen ook de betrokken actoren, via hun dialoog met de overheid, een aantal accenten leggen. Overeenkomstig blijken de gehanteerde formats (met name die voor interventievoorstellen en voortgangsrapportage) een belangrijke rol te spelen bij het bepalen van de evalueerbaarheid doordat ze bijvoorbeeld sterk aspecten verbonden aan efficiëntie benadrukken, maar weinig aandacht besteden aan impact. Overigens zijn deze formats ook tot op zekere hoogte geïnspireerd door het verlangen een resultaatsgericht beheer vorm te geven, bijvoorbeeld door de eis om bij het budget voorziene uitgaven aan outputs te koppelen. Daarnaast is de goede evalueerbaarheidsscore van efficiëntie ook zeker te verklaren door de belangrijke inspanningen die met name BTC en NGOs hebben ondernomen om hun M&E systemen uit te werken en te implementeren. Deze inspanningen zijn het gevolg van processen die reeds geruime tijd binnen deze organisaties spelen, maar die ook een reactie vormen op de geplande screening van de NGA's die in 2016 zal plaatsvinden en die ondermeer de kwaliteit van de M&E systemen gaat bekijken[62]. Bij de uitbouw van M&E systemen gaat het in vele gevallen over vrij complexe initiatieven die zich gradueel ontwikkelen, waarbij de veranderingen op interventieniveau pas gaandeweg hun ingang vinden. Bij deze processen wordt ook, vrij logisch, een bottom-up benadering gevolgd, waarbij het implementatieniveau (inputs – activiteiten – outputs) eerst aan bod komt; dit is overigens ook het niveau waar het directe nut voor de betrokken organisaties het meest voelbaar is en het niveau waarin DGD het sterkst is geïnteresseerd.

Om verschillende redenen ligt het echter niet voor de hand dat M&E systemen gaandeweg ook "automatisch" de hogere niveaus van de doel-middelen ketting gaan integreren en dat met name de evaluatiefunctie even sterk wordt ontwikkeld als de monitoring. De stimuli en positieve randvoorwaarden die momenteel bestaan voor M&E op implementatieniveau (direct aantoonbaar nut, traditionele focus op het operationele, druk van DGD, relatief lage eisen voor opzet en uitvoering) bestaan immers minder of

[62] Naast deze screening is ook een *certificering* voorzien van de M&E systemen van de actoren van de Belgische samenwerking. Deze certificering is een uitvloeisel van de DGD management respons op de meta-evaluatie van de programma's van NGOs die eveneens in opdracht van DBE werd uitgevoerd (Meta-evaluatie van de programma's van niet-gouvernementele actoren, Juli 2013). Het verslag bevat een aanbeveling m.b.t. de opzet en uitvoering van deze certificering (zie hoofdstuk 5.2).

helemaal niet voor M&E m.b.t. outcomes en impacts[63]. Een belangrijke conclusie die hieruit voortvloeit is dat ondanks de vooruitgang bij de ontwikkeling van M&E systemen over de laatste jaren, de verhoging van de evalueerbaarheid (in de eerste plaats m.b.t. de andere criteria dan efficiëntie) bij ongewijzigd beleid en context helemaal niet zeker is.

Een andere belangrijke vaststelling van deze studie is dat onvolkomenheden in het interventieplan in vele gevallen doorwerken gedurende de uitvoering en op die wijze het niveau van evalueerbaarheid zowel rechtstreeks als onrechtstreeks beïnvloeden. Dit geldt onder meer voor:

- de integratie van gender en andere factoren van sociale differentiatie: als er initieel (tijdens de analyse van de context en doelgroepen) geen aandacht is voor gender, is de kans groot dat dit ook het geval is tijdens de uitwerking van de baseline en naderhand van het M&E systeem en zal tijdens de uitvoering gender onderbelicht blijven en niet als aandachtspunt opgenomen worden in monitoringsystemen en evaluaties;
- een zwakke ex ante aandacht voor het impactniveau (bijvoorbeeld door de afwezigheid van hogere niveaus of een '*missing middle'* in de TOC) die zich doorzet in een onvolledige interventielogica (zoals geëxpliciteerd in het logisch kader) en vandaar ook in het voorgestelde M&E systeem en naderhand in de M&E uitvoeringspraktijk waarbij vele interventies de facto geen enkele aandacht voor het impactniveau hebben en impactanalyse evenmin in hun evaluaties meenemen. Als zodanig kunnen interventies of bredere programma's gedurende verschillende financieringscycli doorgaan zonder dat er reflectie plaatsvindt rond de mogelijke effecten van de actie, laat staan dat deze in kaart worden gebracht;
- het feit dat de karakteristieken, belangen, … van actoren en andere betrokken groepen ex ante onvoldoende in kaart worden gebracht, impliceert dat zij niet of onvoldoende bij de uitwerking van M&E systemen worden betrokken, niet of onvoldoende worden meegenomen in de baseline en dus ook niet in het M&E systeem in de praktijk, waarin ze evenmin een rol spelen. Verderop in de interventiecyclus impliceert dit ook dat de kans dat belangen en verwachtingen van deze actoren worden meegenomen bij de voorbereiding en uitvoering van evaluaties eerder gering is, met kwalijke gevolgen voor de evalueerbaarheid maar ook voor het uiteindelijk nut van de evaluaties;
- indien de wijze waarop het M&E systeem van de interventie zich afstemt/ verhoudt tot het nationale/lokale M&E systeem niet in het voorstel is beschreven, baart het geen verwondering dat het M&E systeem van de interventie in de praktijk ook niet wordt afgestemd op dat nationale of lokale systemen.

Zoals onder meer wordt geïllustreerd door de voorbeelden hierboven, heeft de studie aangegeven dat de kwaliteit van het interventieontwerp - of beter: van de ontwerpfase - een factor is die de evalueerbaarheid rechtstreeks en onrechtstreeks sterk beïnvloedt. Een goede ontwerpfase is immers veelal de 'voorbode' van een kwalitatief goed interventiebeheer. Investeringen in een goede interventie-voorbereiding blijken zich later terug te betalen terwijl initiële zwaktes blijken naderhand moeilijker te corrigeren dan men op het eerste zicht zou verwachten; het initiële ontwerp blijkt met andere woorden in de praktijk dé referentie waarop de praktijk zich stoelt. Hoe deze vaststelling verzoenen met het huidige terechte streven naar administratieve vereenvoudiging zal onder onze aanbevelingen aan bod komen.

Verder blijkt er wel een gradueel verschil te bestaan in de gevolgen van een zwakke ontwerpfase voor monitoring enerzijds en evaluatie anderzijds. Onze analyse heeft aangetoond dat, om een aantal redenen, de M&E praktijk vanuit evalueerbaarheids-standpunt beduidend beter scoort dan het M&E ontwerp. Blijkbaar is het dus mogelijk

[63] De recent uitgewerkte 'Strategienota Ontwikkelingsresultaten, legt evenwel sterk de nadruk op het outcome niveau en stelt bijvoorbeeld (blz. 3) dat outputs *niet* als ontwikkelingsresultaten worden beschouwd. Als deze strategienota goed wordt opgevolgd en consequent vertaald in veranderingen op het vlak van eisen en procedures, zou ze in dit verband mee voor een omslag kunnen zorgen.

om op dit – voornamelijk operationele – niveau relatief eenvoudig correcties aan te brengen, al blijven lacunes uit de beginfase doorwerken (cfr. supra). Deze lacunes hebben echter zwaardere gevolgen voor de evaluatiefunctie omdat correcties moeilijker zijn aan te brengen en lacunes in interventieontwerp en -uitvoering impliceren dat bepaalde belangrijke aspecten van een interventie (effectief bereiken van initiële doelgroepen, effecten van de interventie op verschillende sociale groepen, realisatie van assumpties en risico's) (bijna) niet geëvalueerd kunnen worden en dikwijls zelfs geheel buiten het blikveld van evaluaties blijven. In dat geval kan het samenspel van de hierboven vermelde factoren ertoe leiden dat 'onafhankelijke' evaluaties de facto sterk van binnenuit worden gestuurd (of minstens bepaald) waardoor – gewild of ongewild – blinde vlekken niet worden onderkend en controversiële onderwerpen uit de weg worden gegaan of onderbelicht blijven. Er is met andere woorden een gevaar dat evaluaties zich slechts richten op de realiteit zoals die door de betrokken interventie bepaald of geïnterpreteerd is. Het feit dat het globale nut en relevantie van evaluaties (als verplicht onderdeel van een goede ontwikkelingspraktijk) en de wenselijkheid van evaluaties zelden in vraag worden gesteld, kan dit gevaar nog versterken. In zulke situaties zullen zelfs ervaren evaluatoren moeite hebben om de evaluatie-foci waar nodig in vraag te stellen; in vele gevallen zullen ze daartoe zelfs niet de kans krijgen of bereid zijn.

Samenvattend blijken de vele initiatieven om het interventiebeheer te verbeteren en op die wijze bij te dragen tot een grotere evalueerbaarheid en (waarschijnlijk) eveneens tot een grotere ontwikkelingsperformantie belangrijke resultaten te hebben geboekt. Er blijven echter belangrijke lacunes bestaan. Zo is het bijna geheel ontbreken van inzicht in de impact en de duurzaamheid van de effecten van interventies die grotendeels met publiek geld zijn gefinancierd vanuit maatschappelijk standpunt problematisch. Hetzelfde kan worden gesteld m.b.t. de onvolledig ontwikkelde evaluatiefunctie waardoor onafhankelijke evaluaties binnen een de facto beperkend kader moeten plaatsvinden en daardoor hun doelen op het vlak van leren en accountability slechts gedeeltelijk kunnen bereiken.

5.2 Aanbevelingen

Voorafgaande opmerking

In het bestek van de studie, onder punt B3, wordt gesteld dat de studie nuttig moet zijn voor drie verschillende groepen: de DBE, de diensten van DG-D en de Attachés en de partners van de bilaterale en niet-gouvernementele samenwerking. In de visie van het studieteam heeft elk van deze groepen een belang en verantwoordelijkheid in het streven naar een betere evalueerbaarheid en nemen ze die best op vanuit een *gemeenschappelijk* gedefinieerd en gedragen kader en richtlijnen, waarbinnen elke groep dan vanuit zijn eigen rol en specificiteit mee aan de slag gaat. Evalueerbaarheid en evalueerbaarheidsbeoordeling vormen op die wijze een bijkomende aandachtspunt in de overlegtraditie zoals die reeds bestaat tussen de verschillende actoren. Als zodanig zijn de strategische aanbevelingen voor eenieder van belang. Wat de operationele aanbevelingen betreft, zijn deze voornamelijk voor de partners van de samenwerking van belang.

Het studieteam beseft dat een aantal van de geformuleerde aanbevelingen veeleisend zijn, minstens in aanzet, misschien niet op termijn. Als zodanig kunnen zij pas worden toegepast als de beheerslast voor de betrokken actoren op een aantal vlakken overeenkomstig kan worden verminderd. De pogingen om te komen tot administratieve vereenvoudiging zijn in dit kader een noodzakelijke maar onvoldoende voorwaarde. Eveneens is het van belang tot een *andere invulling* te komen m.b.t. de uitwerking van interventievoorstellen, uitvoeringsrapportage, e.d.m. waarbij de focus gaat naar de ontwikkelingsresultaten (outcomes, impact) in plaats van de operationele niveaus (middelen, activiteiten, outputs). Deze overweging loopt door verschillende van de hierna uitgewerkte aanbevelingen.

5.2.1 Strategische aanbevelingen

1. *Deze studie heeft geprobeerd de notie en praktijk van de beoordeling van de evalueerbaarheid te concretiseren binnen de context van de Belgische ontwikkelingssamenwerking. Hieruit is gebleken dat het concept en vooral de daadwerkelijke beoordeling van de evalueerbaarheid tot nu toe vrij onbekend zijn in de samenwerking. Verder heeft de studie aangetoond dat de evalueerbaarheidsbeoordeling voor verschillende doeleinden kan worden aangewend die verder gaan dan deze in het bestek aangegeven. Evalueerbaarheid beoordelen hoeft niet noodzakelijk te worden gekoppeld aan het organiseren van een evaluatie. Zo kan een goede toepassing van een evalueerbaarheidstoets bijvoorbeeld belangrijke leereffecten genereren en op die wijze bijdragen tot een betere praktijk. Meer in het algemeen kan het bekijken van een interventie vanuit het oogpunt van evalueerbaarheid op vele wijzen bijdrage tot een beter beleid en praktijk in de ontwikkelingssamenwerking.*
   Deze studie beveelt daarom aan dat alle actoren, op hun niveau, evalueerbaarheid en de beoordeling van evalueerbaarheid op een meer systematische wijze integreren in hun werking en beide vooral zien als een middel voor een meer performante ontwikkelingssamenwerking. Het gebruik van evalueerbaarheid mag niet verworden tot een hefboom voor bureaucratische controle of sturing (door donor, binnen organisaties) maar moet onderdeel vormen van veranderings- en leerprocessen die in vele organisaties reeds op gang zijn gekomen. Evenmin mag/kan het de bedoeling zijn maximale evalueerbaarheid na te streven; het vergroten van evalueerbaarheid moet een constante zorg zijn maar moet zich adequaat situeren binnen een specifieke context; steeds zal er een omslagpunt zijn waar de kost van het bereiken van een grotere evalueerbaarheid niet meer opweegt tegen de baten.

2. *Monitoring en evaluatie ontwikkelen zich (of zijn reeds ontwikkeld) tot goed institutioneel uitgewerkte systemen met een duidelijk beleid en (dikwijls) gestandaardiseerde methoden. Deze evolutie is positief en een illustratie van het geloof in de waarde van evaluatie (en monitoring) als onderdeel van een goede ontwikkelingspraktijk. In deze context bestaat evenwel het gevaar dat de waarde en baten van met name evaluatie (ex ante) niet meer in vraag worden gesteld en ex post niet meer worden geanalyseerd, wat onder meer leidt tot rituele evaluaties of evaluaties met een beperkt nut of bruikbaarheid zonder dat hier veel vragen bij worden gesteld.*
   Deze studie beveelt daarom aan dat bij elke evaluatie ex ante een expliciete actie wordt ondernomen om de potentiële baten van een evaluatie te analyseren en aan te tonen, eerder dan impliciet aan te nemen dat die bij evaluaties steeds aanwezig zijn, om op die wijze tot een gefundeerde beslissing te komen m.b.t. het al of niet laten uitvoeren van een evaluatie. De invoering van een coherente evalueerbaarheidsbeoordeling is een belangrijk hulpmiddel in dit proces; zij kan worden uitgevoerd door de betrokken actoren, eventueel ondersteund door de aangezochte evaluator van de interventie.

3. *De studie heeft belangrijke onvolkomenheden aan het licht gebracht in de voorbereidingsfase van vele interventies: doelgroepen worden slechts oppervlakkig beschreven en afgebakend (als homogene groepen, zonder rekening te houden met elementen van sociale differentiatie), belangrijke interventie-actoren worden in het geheel niet of slechts erg summier beschreven en vele analyses zijn uitgewerkt om reeds eerder genomen strategische keuzes te verantwoorden (terwijl deze keuzes juist op die analyses gebaseerd zouden moeten zijn). Deze onvolkomenheden hebben belangrijke gevolgen voor de kwaliteit van het interventieontwerp en voor het vervolg van de interventiecyclus en ook voor de evalueerbaarheid.*
   Deze studie beveelt een verbetering van het voorbereidingstraject aan, waarbij meer moet worden ingezet op kwaliteit en minder op routine en 'meer van hetzelfde'. Aangezien het hier om een veeleisend proces gaat, is het belangrijk om een

veranderingstraject uit te zetten met het oog op geleidelijke verbetering dat op verschillende wijzen wordt ondersteund: via een aangepast kader (met incentives) van DGD, via goede studies en evaluaties die de formulering kunnen ondersteunen (bv. aan het einde van vorige fasen) en die inzoomen op sleutelvragen en domeinen waarrond de kennis nog onvoldoende is, en via een vermindering van administratieve eisen en regelgeving (gelieerd aan interventievoorstellen en –rapportage) die de ontwikkelingseffectiviteit niet ten goede komen. Deze administratieve vereenvoudiging conformeert zich bij voorkeur met de principes van de recent uitgebrachte 'Strategienota Ontwikkelingsresultaten' die stelt dat het resultaatbeleid van DGD op de outcomes in plaats van op de inputs en outputs zal worden gericht.

4. *Uit de analyse van de 40 interventies is gebleken dat er op operationeel vlak (inputs, activiteiten, outputs, …) veel vooruitgang is gemaakt op het vlak van M&E, die de evalueerbaarheid ten goede komt. Echter gaat er om verschillende redenen weinig aandacht naar het impact niveau terwijl het outcome niveau (samenvallend met de specifieke doelstelling van de interventies) op verschillende wijzen wordt geïnterpreteerd en niet steeds daadwerkelijke veranderingen op doelgroep-niveau beoogt. Dit is problematisch: ontwikkelingssamenwerking gaat immers in essentie om maatschappelijke verandering, maar dit type verandering kan in vele interventies niet worden aangetoond of wordt zelfs niet expliciet nagestreefd.*
   De studie pleit daarom (en overeenkomstig de Strategienota Ontwikkelings-resultaten) voor een hernieuwde aandacht en opwaardering van de outcome en impact niveaus doorheen de gehele interventiecyclus (interventieontwerp, M&E, ….) Dit impliceert vooreerst een duidelijke definiëring van deze basisnoties en van hoe ze moeten worden toegepast. Meer specifiek beveelt de studie aan om:
   - er in interventievoorstellen op toe te zien dat outcomes (het niveau van het specifiek objectief in het logisch kader) worden gedefinieerd als daadwerkelijke *veranderingen* (in beleid, gedragingen, praktijk) op het niveau van doelgroepen en lokale instituties; het gaat om intermediaire resultaten die (via een causaal verband) bijdragen tot de gewenste impact;
   - in interventievoorstellen er eveneens voor te zorgen dat de TOC tussen de (huidige) effecten en impacts goed is uitgewerkt door (bijvoorbeeld) aandacht te besteden aan, enerzijds, directe (*intermediate*) outcomes waartoe interventies op basis van een duidelijke veranderingstheorie aantoonbaar kunnen bijdragen en, anderzijds, lange termijn effecten op maatschappelijk vlak; zulke TOC kan dan een belangrijk instrument worden voor een evaluatie die nagaat in hoeverre vastgestelde veranderingen (outcomes en impacts) aan een interventie kunnen toegewezen worden (dan wel het gevolg zijn van andere invloeden);
   - (in lijn met internationale ontwikkelingen) efficiëntie ruimer te definiëren, zodat outcome en impact ook worden onderzocht op de verhouding tussen kosten en baten.

   Deze 'opwaardering' van het outcome en impact niveau moet uiteraard worden doorgetrokken in de interventiecyclus-praktijk en het instrumentarium dat momenteel wordt aangewend in de verschillende fasen (interventievoorstellen, rapportage, M&E systeem, …).

5. *De studie heeft geleerd dat vele interventies, veelal als gevolg van het beleid van de uitvoerende organisaties, een toenemend belang hechten aan de uitbouw van M&E beleid en -systemen en onderliggende verantwoordingsmechanismen voor het gebruik van de ontvangen middelen. Deze evolutie is positief maar de ontwikkeling en werking van deze systemen vraagt relatief veel middelen en tijd van dikwijls reeds overbevraagde stafleden. Om die reden (maar ook omwille van de beperkte aandacht voor het outcome en impactniveau – zie aanbeveling 4) blijkt dat de M&E praktijk veelal op het implementatie-niveau blijft steken: de aandacht gaat naar de relatief gemakkelijk op te volgen inputs en outputs ten koste van het opvolgen van (meer complexe) veranderingsprocessen. Hiermee samenhangend besteden M&E systemen overmatig aandacht en middelen aan monitoring ten koste van evaluatie.*

Deze ontwikkeling komt de evalueerbaarheid van efficiëntie ten goede, maar bemoeilijkt deze van effectiviteit en vooral die van impact en duurzaamheid.
De studie beveelt aan dat de uitvoerende partners van de Belgische samenwerking in gezamenlijk overleg een "voortschrijdend" M&E beleid, strategie en praktijk ontwikkelen die kunnen/moeten starten met de uitwerking van systemen en praktijken op implementatie-niveau (zoals die nu reeds in vele organisaties bestaan) maar zich daartoe niet mogen beperken. Het is integendeel noodzakelijk dat deze systemen en praktijken zich gaandeweg verder ontwikkelen en met name een goede evalueerbaarheid van effectiviteit, impact en duurzaamheid ambiëren. Zulke ontwikkeling (en de daarmee gepaard gaande toename van de evalueerbaarheid) gebeurt dus best op een geleidelijke wijze, waarbij middelen, instrumenten, capaciteit en ervaring overhands toenemen zodat gaandeweg meer complexe functies kunnen worden opgenomen en geïntegreerd. Goede operationele monitoring legt dan een solide basis voor evaluaties die zich kunnen richten op specifieke vragen en aspecten die door de monitoring niet worden opgenomen. Binnen de huidige context zal het van belang zijn deze aanbeveling goed af te stemmen met de geplande screening van NGAs en de harmonisering en certificering van de interne opvolgings- en evaluatiesystemen van de Belgische ontwikkelingsactoren (zie ook aanbeveling 7 hierna).

6. *De hierboven voorgestelde aanbevelingen zijn in verschillende opzichten vrij veeleisend en moeilijk realiseerbaar als we de beperkingen in rekening brengen waarmee de meeste Belgische ontwikkelingsactoren worden geconfronteerd. Deze beperkingen hebben zowel te maken met een kwantitatief tekort aan middelen als met het tekort aan (kwalitatieve) ervaring en expertise.*
   Het is daarom belangrijk dat er een kader wordt gecreëerd waarbinnen deze ambitieuze veranderingen niet alleen worden gefaciliteerd, maar ook gestimuleerd en gewaardeerd. DGD speelt in dit verband een cruciale rol en zou in samenspraak met de andere sleutelactoren:
   - de huidige regelgeving, instrumenten en procedures verder kunnen herzien en vereenvoudigen zodanig dat deze meer gericht zijn op de (beoogde) ontwikkelings*effecten* en een verlichting van de administratieve en financiële rapportageverplichtingen;
   - incentives ontwikkelen om de M&E functie (en dan vooral de evaluatiefunctie) verder te ontwikkelen, waarbij de actoren beter in staat worden gesteld om kwalitatief sterke evaluaties te (laten) uitvoeren die ook duurzaamheid en impact analyseren; een belangrijke initiële incentive in dit verband kan zijn om (met name in nieuwe interventies) toe te laten dat baselines worden ontwikkeld bij de start van de interventie in plaats van tijdens de voorbereidingsfase;
   - de oprichting van een fonds ter financiering van studies en evaluaties op effect- en impactniveau waarvan het initiatief uitgaat van – bij voorkeur – het geheel van Belgische ontwikkelingsactoren. Dit fonds zou *gezamenlijke* oefeningen moeten financieren waarbij verschillende interventies van verschillende actoren zijn betrokken en studies en evaluaties uitvoeren die de middelen en capaciteiten van individuele actoren te boven gaan en/of minder direct interessant zijn voor heen. DBE kan daarbij (mee) instaan voor de kwaliteitsbewaking zonder evenwel voor het beheer van deze evaluaties verantwoordelijk te zijn. De werking van dit fonds zou zich ook moeten laten inspireren door de noodzaak aan meer *"public good"* evaluaties van de Belgische samenwerking.

7. *De geplande certificering van de M&E systemen van de Belgische ontwikkelings-actoren is een belangrijk gegeven in het kader van de bevindingen en aanbevelingen van deze studie, waarbij we in herinnering brengen dat één van de doelen van deze studie (zoals geformuleerd in het bestek) is om … "lessen te trekken die nuttig zijn voor de harmonisering en de certificatie van de evaluatiesystemen van de actoren".. Evenzo is het belangrijk te wijzen op de geplande (2016) screening van de NGAs (met het oog op hun toegang tot de volgende financieringscyclus) waarbij het M&E*

systeem één van de screeningselementen vormt. De DBE is door de wetgever belast met de uitvoering van de geplande certificering, maar is niet betrokken bij de screening van de NGAs.

Vanuit de bevindingen van deze studie wordt in dit verband het volgende aanbevolen:

- deze studie heeft aangetoond dat de kwaliteit van M&E systemen (in ontwerp en uitvoering) sterk wordt bepaald door een samenspel van andere elementen van het interventie-beheer en de institutionele context en dat, omgekeerd, het M&E systeem op zijn beurt de kwaliteit van die andere componenten van het beheer beïnvloedt. Als zodanig lijkt het niet aangewezen om – zoals nu is voorzien - de certificering van deze systemen los te koppelen van een bredere en meer geïntegreerde benadering (zoals die o.m. in de NGA-screening minstens in aanzet beter is gewaarborgd);
- deze studie heeft ook aangetoond dat DGD een belangrijke rol gespeeld heeft (en in de toekomst kan spelen) in de kwalitatieve verbetering van het beheer van de Belgische samenwerking (en indirect in het verhogen van de performantie van die samenwerking). De bevindingen van deze studie (die overigens schijnen gelijk te lopen met die van een nog in uitvoering zijnde impactevaluatie van NGA acties) zouden ertoe moeten leiden dat DGD over de (menselijke en andere) middelen blijft beschikken om die belangrijke rol verder in te vullen. Of een formele certificering de beste benadering is om die rol in te vullen, valt nog af te wachten. De studie heeft alvast vastgesteld dat erg veel actoren reeds met de uitwerking van M&E systemen aan de slag zijn gegaan, o.m. met het oog op de geplande screening waarvan de modaliteiten via een dialoog van DGD met de betrokken actoren zijn bepaald. Eveneens is vastgesteld dat de voortgang bij de uitwerking van M&E systemen binnen organisaties sterk kan variëren (tussen landen, tussen verschillende interventies binnen hetzelfde land) en dat de lokale context (bv. in fragiele staten) daarbij erg bepalend kan zijn. Het gaat hier om in essentie *kwalitatieve processen* waarbij de context een belangrijke parameter is en de geleidelijkheid en de opbouw van *ownership* centraal (moeten) staan. Er bestaat een grote kans dat een certificeringsaanpak die – onvermijdelijk – vrij formalistisch en gestandaardiseerd moet worden uitgevoerd en vooral een "foto" gaat opleveren van de situatie op een bepaald ogenblik (zonder band met het lopende proces) verstorend gaat werken t.a.v. de processen die momenteel lopen en uiteindelijk contraproductieve effecten kan genereren vanuit het oogpunt van evalueerbaarheid. Dit wil niet zeggen dat DGD de verdere uitwerking van M&E systemen geheel aan de actoren moet overlaten. DGD moet, als het dat wenst, toegang kunnen hebben tot de resultaten van de aan gang zijnde M&E processen. Het zou daartoe bijvoorbeeld middelen kunnen ter beschikking stellen voor de uitwerking en toepassing (door de actoren zelf, of met externe begeleiding) van een diagnose-instrument dat de betrokken actoren (en DGD) toelaat om inzicht te krijgen in de sterke en zwakke punten van hun M&E systeem, dat hen toelaat om een *tailor-made* resultaatsgericht stappenplan te ontwikkelen ter verbetering van hun M&E systeem en -praktijk waarbij aangegeven wordt hoe en op welke termijn specifieke componenten van het M&E systeem en de praktijk zullen bijgesteld worden, en waarbij de actoren wordt gevraagd om over de voortgang in het stappenplan te rapporteren aan DGD.
- aansluitend op het vorige punt en ingeval men er voor kiest de certificering overeenkomstig het huidige plan uit te voeren, moet vermeden worden dat die certificering uitgroeit tot een pure meta-evaluatie-aanpak van M&E in de sector. Verder is het belangrijk te beseffen dat, zelfs als men overgaat tot certificering van M&E systemen, het in de toekomst hoe dan ook belangrijk zal blijven om voldoende talrijke en omvangrijke *externe* evaluaties van de kwaliteit van het werk van de ontwikkelingsactoren te blijven uitvoeren (in aanvulling op de door de actoren zelf geïnitieerde evaluaties, zie ook de operationele aanbevelingen hierna). Hiervoor zijn echter andere instrumenten

beschikbaar (of moeten die worden beschikbaar gesteld - zie onder meer aanbevelingen 6 en 11).

5.2.2 Operationele aanbevelingen

Uit de analyse van de factoren die de evalueerbaarheid beïnvloeden (zie hoofdstuk 3 en 4) kunnen vrij gemakkelijk een aantal aanbevelingen m.b.t. *'good practices'* worden afgeleid die niet alleen de evalueerbaarheid maar de ontwikkelingsperformantie in het algemeen ten goede komen. Daarom beperken wij ons hierna tot enkele belangrijke operationele aanbevelingen die veelal implicaties hebben in de verschillende fasen van de interventiecyclus en/of in onze ogen van groot operationeel belang zijn. Enkele van die aanbevelingen overlappen gedeeltelijk met strategische aanbevelingen die hierboven werden voorgesteld.

8. **Verbeteringen aan het interventie-ontwerp**
   Zoals reeds eerder aangegeven (zie aanbeveling 3) worden deze verbeteringen bij voorkeur opgevat als een geleidelijk proces. Belangrijke initiële stappen zijn daarbij (voor zover nodig):
   - een betere beschrijving (differentiatie) en afbakening van directe doelgroepen;
   - de ontwikkeling (of verbetering) van een goed uitgewerkte baseline waarbij (1) de resultaten van het vorige punt worden meegenomen en (2) er een definiëring en graduele vervolmaking (naar "boven" toe) van de veranderingstheorie plaatsvindt en, overeenkomstig, (3) de formulering van de beoogde effecten (specifieke doelstelling) en impact (inclusief overeenkomstige indicatoren) en cruciale assumpties wordt verfijnd en/of vervolledigd, (4) rond de weerhouden indicatoren ook baseline gegevens worden verzameld.

9. **Vervolmaking van het M&E beleid en zijn vertaling in de praktijk**
   Een toenemend aantal organisaties heeft een M&E beleid uitgewerkt. In vele gevallen vertoont dit beleid en vooral zijn toepassing in de praktijk echter een aantal onvolkomenheden die sterk kunnen verschillen van organisatie tot organisatie. De aanbevelingen hierna betreffen tekorten die vrij veel voorkomen maar niet noodzakelijk voor alle betrokkenen opgaan.
   Bij de ontwikkeling en implementatie van het M&E beleid is het belangrijk om:
   - de evaluatiefunctie (in strikte zin) sterker te ontwikkelen (inclusief de relatie tot monitoring);
   - meer aandacht te hebben voor de uitvoering van het beleid op landen- en interventieniveau, waarbij de lokale gegevenheden worden meegenomen (betrekken lokale partners bij de totstandkoming van het lokale beleid en de lokale praktijk; articulatie tussen eigen M&E beleid en dat van de lokale actoren);
   - een duidelijke aanpak uit te werken waarbij de M&E inspanningen en resultaten worden gebruikt voor daadwerkelijke analyse, leren en verandering.

10. **Ontwikkeling van de M&E functie op interventieniveau**
    Naast de noodzaak aan verhoogde aandacht voor de 'hogere niveaus' in de doel-middelen ketting en voor evaluatie (zie aanbevelingen 4 en 5) is het belangrijk dat:
    - de opzet en functionering van M&E systemen goed wordt gepland ; dit impliceert niet noodzakelijk dat deze systemen reeds tijdens de formulering van de interventie worden uitgewerkt, maar wel dat hun opzet, uitbouw en functionering goed wordt gedefinieerd en ingepland waarbij ook aandacht wordt besteed aan de benodigde menselijke en financiële middelen, aan capaciteitsversterking, etc.;
    - de verschillende M&E componenten daadwerkelijk worden samengebracht in een coherent *systeem* dat het hele M&E veld omvat, overlappingen vermijdt

tussen de verschillende inhoudelijke componenten en een goede definiëring van de benodigde middelen (kwantitatief en kwalitatief) en de taakverdeling (rol verschillende actoren) impliceert;
- de rol van de verschillende actoren in M&E systemen wordt geoptimaliseerd, waarbij wordt vermeden dat M&E het monopolie wordt van de partners verantwoordelijk voor de uitvoering en veeleer vanuit het subsidiariteits-principe wordt nagegaan wie wat waar en hoe aan de M&E kan bijdragen;
- voldoende aandacht wordt besteed aan de terugkoppeling van de monitoring-resultaten naar alle betrokkenen, om op die wijze het bewustzijn in het belang van (goede) gegevensverzameling blijvend te voeden en bij te dragen tot de duurzaamheid van het systeem;
- nagegaan wordt hoe het M&E systeem van de interventie zich verhoudt tot lokale en nationale M&E systemen en, waar nodig en relevant, welke rol de interventie kan spelen om die lokale en nationale systemen op te zetten, te verbeteren en te verankeren.

11. **Evaluaties voorbereiden en uitvoeren vanuit een constante aandacht voor hun uiteindelijk gebruik**
Deze aanbeveling bouwt verder op (o.m.) aanbeveling 2 en gaat uit van de premisse dat de kwaliteit van een evaluatie in eerste instantie afhangt van haar effectief gebruik die vooral door haar nut en bruikbaarheid wordt bepaald. Een grondige reflectie m.b.t. de toekomstige baten van de evaluatie (aanbeveling 2) is daartoe een nodige maar onvoldoende voorwaarde.
Teneinde het uiteindelijk gebruik van evaluaties te optimaliseren is het belangrijk om:
    - overeenkomstig aanbeveling 2 in elke evaluatie een evalueerbaarheids-beoordeling te voorzien, die onder meer de potentiële baten van een geplande evaluatie analyseert en aantoont;
    - de planning en uitvoering van een evaluatie te kaderen in een "portfolio benadering" waarbinnen niet alle interventies in al hun aspecten dienen te worden geëvalueerd maar in functie van de beschikbare middelen, eisen naar accountability, omgevingsfactoren, etc. strategische keuzes worden gemaakt. Het kan bv. ook interessant zijn de interventies in verschillende groepen in te delen naargelang de reeds bestaande inzichten over de werking en impact van de interventie:
        - interventies waarvan voorgaande evaluaties en studies reeds veelvuldig hebben aangetoond dat ze impact hebben (vaak inter-venties met een minder complexe TOC), hoeven niet steeds opnieuw geëvalueerd te worden op impact. Voor deze groep van interventies hoeft er dus geen counterfactual gespecifieerd te worden en volstaat het de effectiviteit en efficiëntie te evalueren;
        - interventies waarvan men weet dat ze goed geïmplementeerd zijn en die men wil herhalen of substantieel uitbreiden, maar waarvan men onvoldoende weet of ze impact genereren worden best wel aan een grondige impactevaluatie onderworpen. Voor deze interventies is het noodzakelijk dat ze ook effectief op hun impact kunnen geëvalueerd worden en dat er van bij het ontwerp van de interventie hiermee ook rekening wordt gehouden. Dit impliceert dat de verschillende aspecten van evalueerbaarheid van impact die doorheen de studie bekeken werden ook effectief aandacht krijgen (onder meer TOC tot op niveau van impact, identificatie van cruciale elementen tot op niveau van impact, vertaling van TOC in M&E systeem tot op niveau van impact, specificatie van counterfactual, goede baseline, etc.). Vooral de aanwezigheid van data voor een counterfactual[64] en betreffende

[64] Dit noodzaakt niet noodzakelijk het uitvoeren van een RCT (random controlled trial). In hoofdstuk 2 werden verschillende mogelijkheden voor de identificatie van een counterfactual beschreven.

externe assumpties is belangrijk vooraleer tot een impactevaluatie over te gaan[65];
    - interventies waarvan men onvoldoende weet of de gekozen modaliteit van implementatie de beste is, kunnen best experimenteren met verschillende modaliteiten tijdens de implementatie en inzetten op evaluatie die de werking en effectiviteit van deze verschillende modaliteiten vergelijkt (zie in dit verband ook het idee van *structured experiential learning*, voetnoot 28).
- de uiteindelijke gebruikers van de evaluatie (bij de interventie betrokken actoren en anderen) reeds van bij de voorbereiding bij de evaluatie te betrekken (door bijvoorbeeld hun belangen en verwachtingen mee te nemen);
- referentietermen op te stellen met goed omlijnde en doordachte hoofdvragen en die eveneens evaluatoren toelaten (of zelfs aansporen) om buiten het strikte kader van de evaluatiedoelen en –hoofdvragen onderzoek uit te voeren.

[65] Terwijl de afwezigheid van een expliciete TOC tot op het niveau van impact tijdens een impactevaluatie nog kan geremedieerd worden (omdat verantwoordelijken van interventies vaak een impliciete TOC hebben zonder die op papier te zetten), is de afwezigheid van een M&E systeem in de praktijk dat valide data verzamelt voor de interventie-groep/situatie en een counterfactual moeilijker op te vangen bij een impactevaluatie.

Bijlagen

(in apart document)

1. Bestek
2. Beschrijving van de onderzoeksmethodologie en aanpak
3. Studiekader
4. Lijst van de 40 geanalyseerde interventies
5. Lijst van de belangrijkste geraadpleegde documenten
6. Lijst van de belangrijkste gecontacteerde personen
7. Toelichting bij de statistische analyses en analyse van de sterkste en zwakte items
8. Gedetailleerd score overzicht