EVALUATION

Meta-evaluation of Project and Programme Evaluations in 2015–2017

Evaluation on Finland's Development Policy and Cooperation

2018/3

EVALUATION

META-EVALUATION OF PROJECT AND PROGRAMME EVALUATIONS IN 2015-2017

Dr. Stefan Silvestrini (Team Leader)
Dr. Susanne Johanna Väth (Deputy Team Leader)
Dr. Cornelia Römling
Michael Lieckefett
Petra Mikkolainen

Consortium composed of:

Lead Company

In collaboration with:

2018/3

This evaluation was commissioned by the Ministry for Foreign Affairs of Finland to Particip GmbH. This inception report is the product of the authors, and responsibility for the accuracy of the data included in this inception report rests with the authors. The findings, interpretations, and conclusions presented in this report do not necessarily reflect the views of the Ministry for Foreign Affairs of Finland.

This report can be downloaded through the home page of the Ministry for Foreign Affairs
http://formin.finland.fi/developmentpolicy/evaluations

Contact: EVA-11@formin.fi

ISBN 978-952-281-563-7 (pdf)
ISSN 2342-8341

Cover design and layout: Innocorp Oy/Milla Toro

CONTENTS

BOXES

FIGURES

TABLES

ACRONYMS AND ABBREVIATIONS

ADB	Asian Development Bank
AfDB	African Development Bank
AU	African Union
BEAM	Business with Impact
CGF	Green Climate Fund
CSO	Civil Society Organisations
DAC	Development Assistance Committee
DFID	Department for International Development
ECOSOC	United Nations Economic and Social Council
EMS	Evaluation Management Services
EQ	Evaluation question
EU	European Union
EVA-11	Development Evaluation Unit
FAO	Food and Agriculture Organization of the United Nations
FE	Final evaluation
Finnfund	Finnish development finance company
Finnpartnership	Finnish Business Partnership Programme
GEF	Global Environment Facility
GEFIEO	Independent Evaluation Office of the Global Environment Facility
GEWE	Gender equality and women's empowerment
GNI	Gross National Income
HEI ICI	Higher Education Institutions Institutional Cooperation Instrument
HIV / AIDS	Human immunodeficiency virus infection and acquired immune deficiency syndrome
HRBA	Human Rights Based Approach
IDB	Inter-American Development Bank
IL	Intervention Logic
ILO	International Labour Organization

IOOI	Input Output Outcome Impact
IR	Inception Report
ITT	Invitation to tender
LF	Logframe
MFA	Ministry for Foreign Affairs (Finland)
MTE	Mid-term evaluation
MTR	Mid-term review
NDF	Nordic Development Fund
NORAD	Norwegian Agency for Development Cooperation
ODA	Official Development Assistance
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary least squares
PBS	Programme-based support
PT	Programme Theory
QA	Quality Assurance
RBM	Results-Based Management
ROM	Results Oriented Monitoring
ToC	Theory of Change
ToR	Terms of Reference
UN	United Nations
UNCTAD	United Nations Conference on Trade and Development
UNDP	United Nations Development Programme
UNEP	United Nations Environment Programme
UNFCCC	United Nations Framework Contract on Climate Change
UNFPA	United Nations Population Fund
UNICEF	United Nations Children's Fund
WBG	World Bank Group

TIIVISTELMÄ

Suomen ulkoministeriö (UM) tilaa säännöllisesti metaevaluointeja. Tässä toimeksiannossa toteutettiin metodologinen laatuarviointi ja summatiivinen sisältöarviointi 51 hajautetusta evaluointiraportista, jotka on tehty syyskuun 2015 ja elokuun 2017 välisenä aikana kahden-, monen- ja monen-kahdenvälisistä interventioista sekä 45 tehtävänkuvauksesta. Kaksivaiheinen monimetodinen analyysi noudatti osallistavaa lähestymistapaa. Se perustui kattavien standardoitujen arviointivälineiden käyttöön, temaattiseen koodaukseen, tilastojen yhteenvetämiseen ja sisällön laadun analyysiin.

Koska interventioiden kokonaispopulaatiosta ei ole tarpeeksi tietoja, emme voi arvioida, missä määrin tämä raporttien otos edustaa koko Suomen kehitysyhteistyön tätä osaa. Siksi suosittelemme, että kaikki interventiot kartoitetaan, jotta stratifioitu otanta voidaan toteuttaa tulevaisuudessa.

Löydöstemme mukaan 60 % arvioiduista tehtävänkuvauksista niiden laatu on tyydyttävä. Kuitenkin havaitsimme myös useita heikkouksia, jotka tuovat ilmi useita kapasiteettiin liittyviä puutteita UM:ssä. Koska tyypillisesti korkeampi tehtävänkuvausten laatu liittyy korkeampaan raportin laatuun, suosittelemme erityisesti, että evaluointimanuaalia parannetaan, jotta voidaan parantaa arviointimetodologioiden ja -käytäntöjen tuntemusta ja harkita olemassa olevien rakenteiden parantamista.

Rapottien laatuun liittyen havaitsimme, että löydökset perustuvat usein heikkoon metodologiaan, mutta vaikuttavat silti suhteellisen luotettavilta. Noin kaksi kolmasosaa raporteista sisältää joitakin laadullisia puutteita, ja noin kolmasosaa sisältää merkittäviä laadullisia puutteita. Siksi on suositeltavaa parantaa laadunvarmistusta ja varmistaa metodologinen asiantuntijuus arvioijia rekrytoitaessa.

Arviointiraporttien mukaan kokonaislaatu arvioidaan laadultaan kohtalaiseksi tai paremmaksi 70 %:ssa interventioista. Relevanssia pidetään vahvuutena ja kestävyyttä suurimpana haasteena.

Avainsanat: *meta-evaluaatio, systemaatiinen arviointi, monimetodinen lähestymistapa, Suomen kehitysyhteistyö, maailmanlaajuinen*

REFERAT

Finlands utrikesministerie beställer regelbundet meta-utvärderingar. Inom ramen för detta uppdrag gjordes en metodisk kvalitetsgranskning och övergripande innehållsbedömning av 51 rapporter från decentraliserade utvärderingar av bi-, multi- och multi-bilaterala insatser som genomfördes mellan september 2015 och augusti 2017, samt 45 uppdragsbeskrivningar. Analysen genomfördes i två steg med hjälp av flera olika metoder och på ett sätt som främjade deltagande. Standardiserade bedömningsverktyg, tematisk kodning, sammanfattande statistik och kvalitativ innehållsanalys tillämpades.

Eftersom tillräcklig information saknas om hela volymen av bi-, multi- och multi-bilaterala insatser har vi inte kunnat avgöra om urvalet av utvärderingsrapporter är representativt för denna del av Finlands utvecklingssamarbete. Vi rekommenderar därför att samtliga bi-, multi- och multi-bilaterala insatser inventeras och kategoriseras för att möjliggöra ett stratifierat urval i framtiden.

Analysen visar att 60 % av uppdragsbeskrivningar i urvalet är av godkänd kvalitet. Trots det har vi hittat många brister som pekar på ett behov av kapacitetsutveckling inom utrikesministeriet. Eftersom det i allmänhet finns ett samband mellan kvaliteten på uppdragsbeskrivningar och kvaliteten på utvärderingsrapporter rekommenderar vi starkt att utvärderingshandboken uppdateras i syfte att öka kunskapen om utvärderingsmetoder och praxis, samt att ministeriet överväger att förbättra befintliga strukturer.

Vad gäller utvärderingsrapporternas kvalitet noterar vi att slutsatser ofta dras på basis av undermåliga metoder, men samtidigt verkar vara någorlunda tillförlitliga. Två tredjedelar av rapporterna har vissa kvalitetsbrister medan en tredjedel har betydande sådana brister. Vi rekommenderar därför att kvalitetssäkringen förbättras och att högre krav ställs på metodkompetens när utvärderare upphandlas.

Enligt utvärderingsrapporterna är den övergripande kvaliteten hos bi-, multi- och multi-bilaterala insatserna godkänd eller bättre för 70 % av insatserna. Styrkan ligger i insatsernas relevans. Den största utmaningen är att förbättra bärkraften.

Nyckelord: *metautvärdering, metodisk genomgång, multi-metod, finskt utvecklingssamarbete, global*

ABSTRACT

The Ministry of Foreign Affairs of Finland (MFA) regularly commissions meta-evaluations. In this assignment, 51 decentralised evaluation reports of bi-, multi-, and multi-bilateral interventions conducted between September 2015 and August 2017 and 45 ToRs were subject to methodological quality and summative content assessment. The two-stage multi-method analysis followed a participatory approach. It built on comprehensive standardised assessment tools and thematic coding. Summary statistics and qualitative content analysis were applied.

Given the lack of information on the whole population, we cannot assess the representativeness of this sample of reports for this part of Finnish development cooperation. We therefore recommend to run an inventory of interventions to enable stratified sampling in future.

We find that overall quality is satisfactory for 60% of the assessed ToR. Nevertheless, numerous weaknesses were identified which reveal capacity gaps within MFA. Given that on average higher ToR quality is associated with higher report quality, we highly recommend to improve the evaluation manual, to enhance knowledge of evaluation methodologies and practices and to consider improving existing structures.

Regarding the quality of the reports, we observe that findings are often based on weak methodologies but appear to be somewhat reliable. Two thirds of the reports feature some, one third substantial quality flaws. Thus, we recommend to enhance quality assurance and to ensure methodological expertise when recruiting evaluators.

According to the evaluation reports, the overall quality is assessed as of moderate quality or better for 70% of the interventions. Relevance is considered as strength and sustainability as greatest challenge.

Keywords: *meta-evaluation, systematic review, mixed-methods approach, Finnish development cooperation, worldwide*

YHTEENVETO

Suomen ulkoministeriö (UM) tilaa säännöllisesti metaevaluaatioita. Ne toteutetetaan kehitysevaluoinnin yksikön (EVA-11) kautta, joka on toiminnallisesti itsenäinen yksikkö ja raportoi suoraan alivaltiosihteerille.

Metaevaluaatioiden toteuttamisen taustalla olevat **perusteet** ovat kahdenlaisia: niitä pidetään arvokkaana välineenä

i) ”lisäämään vastuullisuutta ja avoimuutta kumppanimaita, suurta yleisöä, kansanedustajia, korkeakouluja, tiedotusvälineitä ja UM:n ulkopuolisia kehitysyhteistyön ammattilaisia kohtaan” (kts. tehtävänkuvaus) ja

ii) analysoimaan UM:n evaluaatiotoiminnan kokonaislaatua yhdistämällä tuloksia sekä kokemuksia ja mitä on opittu monista Suomen rahoittamista kehitysyhteistyön interventioista.

Tämän metaevaluaation kohteena oli 51 hajautettua evaluaatioraporttia ja niiden 45 vastaavaa tehtävänkuvausta, jotka toteutettiin syyskuun 2015 ja elokuun 2017 välisenä aikana. Asiakirjat sisälsivät 23 keskipitkän aikavälin ja 28 lopullista evaluaatiota koskien yksittäisiä kahdenvälisiä, monenvälisiä tai monen-kahdenvälisiä projekteja ja ohjelmia, joita olivat tilanneet eri temaattiset yksiköt, UM:n alueelliset osastot, suurlähetystöt ja monenväliset kumppanit.

Tämän tehtävän **pääasiallinen tarkoitus** on seuraava: se pyrkii antamaan

i) tarkkoja johtopäätöksiä ja suosituksia, joiden avulla UM voi parantaa hajautettujen arviointien laatua ja arviointien hallintokäytäntöjä sekä edistää arviointikapasiteetin kehittämistä; ja

ii) antamaan suosituksia siitä kuinka UM:n kehitysyhteistyötä voitaisiin parantaa pohjautuen evaluaatioraporteista nouseviin ja yhteenvedettyihin käsityksiin Suomen kehitysyhteistyöstä.

Tästä johtuen tehtävä koostui kahdesta osasta: 1) meta-arvioinnista, jossa arvioitiin tarkasteltavana olevien evaluointiraporttien laatua, ja 2) summatiivisesta meta-analyysistä, jossa koottiin yhteen näiden raporttien sisältö. Meta-evaluaoinnin tavoitteena oli tuottaa:

i) kokonaiskuva arvioitavista evaluaatioista,

ii) hajautettujen arviointiraporttien ja niiden tehtävänkuvausten arviointi,

iii) luotettavien evaluaatiolöydösten synteesi ja

iv) evaluaatioraporteista muita mahdollisesti esille nousevia asioita.

Metaevaluaatio toteutettiin käyttäen kaksivaiheista monimenetelmäistä analyysiä, ja se noudatti osallistavaa lähestymistapaa. Jotta resursseja voitiin hyödyntää tehokkaasti hyvän arviointikäytännön mukaisesti, se perustui:

i) UM:n aiemmin tilaamista meta-arvioinneista saatuihin kokemuksiin ja mitä niistä on opittu,

ii) muiden organisaatioiden tekemiin meta-arviointeihin,

iii) evaluaatiotiimin aikaisemmin suorittamiin samankaltaisiin tehtäviin,

iv) alustavasta asiakirjojen tarkastelusta tehtyihin havaintoihin ja

v) UM:n arviointiprosessin aikana tekemiin havaintoihin.

Arviointimenetelmien osalta analyysin ensimmäinen vaihe sisältää **metodologisen laatuarvioinnin**, jossa käytetään standardoitua arviointivälinettä (eli yksityiskohtaista tarkistuslistaa 51 evaluaatioraportille ja niiden tehtävänkuvauksille). Ensimmäinen vaihe alkoi raporttien ja tehtävänkuvausten laadun lukuisten yksittäisten osa-alueiden arvioinnilla. Tarvittaessa nämä osa-alueet tarkistettiin kyllä/ei-vastausvaihtoehdoilla, ja muuten sovellettiin nelivaiheista asteikkoa, joka sisälsi selkeästi määritellyt kategoriat. Siksi otettiin käyttöön kategoriat ”hyvä tai erittäin hyvä”, ”tyydyttävä”, ”parantamisen tarvetta” ja ”riittämätön”. Tämän asteikon käyttöönotossa oli otettu huomioon aikaisemmat kokemukset ja muista samankaltaisista tehtävistä opittu asia: kun laadultaan paremmat raportit ja erinomaisesti tehty työ on tiivistetty yhteen kategoriaan, voidaan laadultaan alemmilla kategorian tasoilla eriyttää voimakkaammin, mikä mahdollistaa lopuksi yksityiskohtaisten suositusten tekemisen evaluaatiokapasiteetin kehittämiselle.

Seuraavassa vaiheessa tietty määrä yksittäisiä osa-alueita painotettiin niiden suhteellisen tärkeyden mukaan ja koottiin yhdeksi alueeksi (luokiteltuna nelivaiheisella asteikolla). Lopuksi alueet koottiin yhteen kokonaisarviointia varten (jälleen nelivaiheiseen asteikkoon) sisältäen:

i) tehtävän esittelyn ja kontekstianalyysin laatu,

ii) evaluointimetodologian laatu,

iii) evaluoinnin löydösten laatu,

iv) johtopäätösten ja suositusten laatu ja

v) tiivistelmien laatu.

Vaiheittainen menettely estää liiallisen yksinkertaistamisen samalla kun se kattaa yksityiskohtaisesti laajan valikoiman eri osa-alueita. Korkean standardointiasteensa ansiosta menettely on vakaa koskien arvioijan mahdollisten ennakkoasenteiden vaikutusta arviointiin.

Analyysin toinen vaihe koostuu Suomen kehitysyhteistyön (siinä määrin kuin se on katettu tässä meta-evaluaatiossa käsitellyissä arviointiraporteissa) yksityiskohtaisesta sisältöarvioinnista käyttäen semi-standardoitua arviointityökalua. Tämä vaihe edellyttää vähintään minimaalista raportin metodologista laatua, jossa on otettu huomioon saatavilla olevan materiaalin ja samantyyppisten tehtävien konteksti, eikä se siten ole yhtä tiukka kuin jos se olisi tarkoitettu puhtaasti tieteellisiin tarkoituksiin. Näin ollen 50 arviointiraporttia kävi läpi samanlaisen vaiheittaisen menettelyn kuin ensimmäisen vaiheen kohdalla kuvattiin.

Tässä kohtaa evaluaatiotiimi ei enää tarkistanut raportin laatuun liittyviä näkökohtia. Sen sijaan se siirsi evaluaatioraportteihin sisältyvät arvioinnit standardoituihin luokituksiin. Tällöin sovellettiin nelivaiheista asteikkoa, jossa oli vastausvaihtoehdot ”ei”, ”enemmän ei”, ”enemmän kyllä” ja ”kyllä” ja yhteenvetotietoja laskettiin systemaattisten tulosten saamiseksi seuraavista asioista:

i) interventioiden relevanssi,

ii) interventioiden vaikuttavuus,

iii) interventioiden tehokkuus,

iv) interventioiden vaikutus,

v) interventioiden kestävyys ja

vi) interventioiden avun vaikuttavuus ja kolme K:ta (eli koherenssi, koordinointi ja komplementaarisuus).

Lisäksi tietyt arvioinnin taustalla olevat syyt, saadut kokemukset ja mitä niistä on opittu, ja arvioijien esittämät suositukset kerättiin avainsanojen avulla ja niille tehtiin temaattinen koodaus MaxQDA®-ohjelmistopaketilla. Lopuksi laadullinen sisältöanalyysi helpotti yleisten trendien ja esille tulevien asioiden tunnistamista.

Korkealaatuisuuden takaamiseksi kaikki arviointivälineet testattiin laajasti etukäteen. 10 % satunnaisesti valittuja raportteja (viisi laatuarviointia ja viisi sisältöarviointia varten) analysoitiin ristikkäisesti, ja evaluaatiotiimin vetäjät tarjosivat intensiivistä teknistä taustatukea. Lisäksi järjestettiin sisäisiä ja ulkoisia validointityöpajoja, joissa tuloksia tarkistettiin ristiin sekä meta-arviointiryhmän sisällä että UM:ssä. Metaevaluaatiota koskevien rajoitusten osalta on tärkeää ymmärtää, että

i) tätä meta-arviointia ei voida pitää yksittäisten projektien tai ohjelmien uudelleenarviointina, joten tuloksia voidaan tulkita vain koosteena läpikäydyistä evaluaatioraporteista,

ii) tulokset ja johtopäätökset koskevat vain murto-osa Suomen kehitysyhteis-työstä, ja ne perustuvat 51 evaluaointiraporttiin, jotka on tehty kahden-, monen- ja monen-kahdenvälisistä interventioista, eivätkä näin ollen koske muita Suomen kehitysyhteistyöinstrumentteja,

iii) analyysi perustuu vain arviointiraporttien ja tehtävänkuvausten sisältämiin tietoihin, eikä triangulaatio muiden tietolähteiden kanssa ei ollut mahdollista, ja

iv) arviointivälineitä sovellettiin erittäin heterogeenisten interventioiden evaluaatioraportteihin (esim. lukuisat maat, alueet, aihealueet, interventiobudjetit), jolloin arvioiden laatua ja sisältöä painotettiin yhtälailla sekä pienten että suurten interventioiden kohdalla.

Meta-arvioinnin **tärkeimmät löydökset, johtopäätökset ja suositukset** esitellään yhteenvetotaulukossa. Ne esitetään erikseen analyysin jokaiselle vaiheelle (eli raportin laatuarviointi ja sisältöarviointi) ja ryhmitellään temaattisten näkökohtien mukaisesti. Vastaavat arviointikysymykset on määritelty suluissa kunkin temaattisen näkökohdan osalta.

Tärkeimmät löydökset lyhyesti ovat:

i) Tehtävänkuvausten kokonaislaatu on tyydyttävä 60 %:ssa tehtävänkuvauksista. Erityisesti metodologiasta, evaluointimenettelystä, laadunvarmistuksesta ja läpileikkaavista tavoitteista esitetyt tiedot olivat varsin heikkoja.

ii) Tehtävänkuvausten korkeampi laatu johtaa keskimäärin myös arviointiraporttien korkeampaan laatuun. Tehtävänkuvausten osuudet arvioinnin tarkoituksesta, tavoitteista ja laajuudesta, metodologiasta ja arviointiprosessista ovat erityisen tärkeitä raportin kokonaislaadulle.

iii) Raportin kokonaislaatu arvioidaan arvosanalla ”tyydyttävä” kahdessa kolmasosassa raporteista ja ”parantamisen varaa” kolmanneksessa raporteista. Havainnot tehdään usein heikon metodologian pohjalta, mutta vaikuttavat silti suhteellisen luotettavilta. Interventiologiikkaan, oletuksia ja tehtävän rajoituksia koskeva keskustelu puuttuu usein.

iv) Arviointiraporttien mukaan Suomen kehitysyhteistyön kokonaislaatu arvioidaan laadultaan kohtalaiseksi tai paremmaksi 70 %:ssa interventioista. Relevanssia pidetään vahvuutena ja kestävyyttä suurimpana haasteena. Tärkeimmät suositukset kohdistuvat intervention suunnitteluun, laajuuteen, hallintoon, kapasiteettiin ja kestävyyteen.

Vaikka UM:n hajautettujen arviointien adekvaattiudesta ei voida tehdä johtopäätöksiä, johtopäätöksemme on että niin tehtävänkuvausten laadussa kuin evaluaatioraporttien laadun varmistuksessa on parantamisen varaa. Tämä puolestaan tuo ilmi useita kapasiteettiin liittyviä puutteita UM:ssä.

Tähän pohjasimme seuraavat pääsuositukset:

i) Kaikki interventiot tulisi kartoittaa keskeisten piirteidensä perusteella, jotta stratifioitu otanta voidaan toteuttaa tulevaisuudessa.

ii) Evaluointimanuaalia parannetaan merkittävästi, jotta voidaan parantaa arviointimetodologioiden ja -käytäntöjen tuntemusta ja harkitaan olemassa olevien rakenteiden parantamista, esim. keskittämmällä tiedonhallinta ja koordinaatio kehitysevaluaatioyksikköön.

iii) Parannetaan laadunvarmistusta ja varmistetaan metodologinen asiantuntijuus arvioijia rekrytoitaessa.

iv) Varmistetaan, että tämän metaevaluaation tulokset jaetaan laajasti palautteena arviointien toimeksiantajille.

SAMMANFATTNING

Finlands utrikesministerie beställer regelbundet meta-utvärderingar. Detta görs av enheten för utvärdering av utvecklingssamarbetet (EVA-11), en operativt oberoende enhet som är direkt underställd statssekreteraren.

Meta-utvärderingar tjänar vanligtvis två **syften**: de ses som värdefulla verktyg

i) ”för ansvarsutkrävande och ökad öppenhet gentemot samarbetsländer, allmänheten, riksdagen, den akademiska världen, media och de som arbetar med internationellt utvecklingssamarbete utanför utrikesministeriet” (jfr. uppdragsbeskrivning), och

ii) för att bedöma kvaliteten på den övergripande utvärderingsfunktionen genom att sammanställa resultat och lärdomar från ett brett spektrum av insatser inom utvecklingssamarbetet som finansieras av Finland.

Den meta-utvärdering som här redogörs för omfattande 51 utvärderingsrapporter (decentraliserade utvärderingar) som färdigställdes mellan september 2015 och augusti 2017, samt motsvarande 45 uppdragsbeskrivningar. Av dessa var 23 halvtidsutvärderingar och 28 slutgiltiga utvärderingar av enskilda bilaterala, multilaterala eller multi-bi-projekt och program som beställts av olika ämnesenheter och regionavdelningar inom utrikesministeriet, ambassader och multilaterala samarbetspartners.

Det **huvudsakliga syftet** med meta-utvärderingen är som följer: Den syftade till att tillhandahålla

i) kortfattade slutsatser och rekommendationer som gör det möjligt för utrikesministeriet att höja kvaliteten på decentraliserade utvärderingar, förbättra handläggningen av utvärderingar samt att främja uppbyggnad av utvärderingskapacitet, och

ii) övergripande observationer om finskt utvecklingssamarbete som framkommer av utvärderingsrapporterna för att ta fram rekommendationer om hur utrikesministeriets utvecklingssamarbete kan förbättras

Uppdraget bestod av två delar: (1) en meta-utvärdering som granskar kvaliteten på de utvärderingsrapporter som valts ut, och (2) en meta-analys som sammanfattar innehållet i rapporterna på en övergripande nivå. **Målen** med meta-utvärderingen var att presentera:

i) en helhetsbild av utvärderingsportföljen,

ii) en bedömning av olika utvärderingsrapporter (decentraliserade utvärderingar) och motsvarande uppdragsbeskrivningar,

iii) en sammanställning av tillförlitliga utvärderingsresultat, och

iv) andra relevant frågeställningar som tas upp av utvärderingsrapporterna.

Meta-utvärderingen var utformad som en två-stegsanalys baserad på en kombination av olika metoder, och genomfördes på ett sätt som främjade deltagande. För att säkerställa effektivt resursutnyttjande och i linje med god utvärderingspraxis, utgick utvärdering från

i) lärdomar från tidigare meta-utvärderingar beställda av utrikesministeriet,
ii) meta-utvärderingar genomförda/beställda av andra organisationer,
iii) liknande uppdrag som utförs av meta-utvärderingsteamet,
iv) slutsatser från en första dokumentgranskning, och
v) utrikesministeriets observationer under utvärderingsprocessen.

Vad gäller den **metod** som låg till grund för meta-utvärderingen, bestod den första fasen av en **systematisk kvalitetsbedömning** genomförd med hjälp av ett standardiserat analysverktyg (en detaljerad checklista för 51 utvärderingsrapporter och motsvarande uppdragsbeskrivningar). Rapporterna och uppdragsbeskrivningarna bedömdes utifrån ett stort antal del-aspekter och svaren fördes in i checklistan, som innehöll både ja/nej-frågor och frågor som kunde besvaras utefter en skala med fyra olika alternativ. De alternativ som tillämpades var "bra eller mycket bra", "godkänd", "behov av förbättring" och "otillräcklig". Denna kategorisering byggde på lärdomar från liknande uppdrag: genom att inordna bra rapporter i en kategori och tillhandhålla flera alternativ för att rangordna de som inte uppnår samma standard kan koncisa slutsatser dras och detaljerade rekommendationer ges för kapacitetsutveckling av utvärderingsfunktionen.

I nästa steg gjordes en bedömning och rangordning av ett antal enskilda del-aspekter på basis av deras relativa betydelse och dessa jämkades därefter samman i en aspekt med en bredare definition (med hjälp av en fyrgradig skala). Dessa aspekter sammanfördes i en övergripande bedömning (också med hjälp av en fyrgradig skala), innefattande

i) kvalitet på inledningar och kontextanalyser,
ii) kvalitet på utvärderingsmetod,
iii) kvalitet på utvärderingsresultat,
iv) kvalitet på slutsatser och rekommendationer, och
v) kvalitet på sammanfattningarna.

Detta stegvisa tillvägagångssätt gjorde det möjligt att undvika överdriven förenkling och på samma gång täcka in ett brett spektrum av aspekter i detalj. Den högra graden av standardisering har även motverkat otillbörlig påverkan och partiskhet.

Den andra fasen omfattade en djupgående analys av finskt utvecklingssamarbete (i den uträckning som medgavs av innehållet i det urval av utvärderingsrapporter som omfattades av meta-utvärderingen). Analysen gjordes med hjälp av ett delvist standardiserat bedömningsverktyg. Denna fas förutsatte att rapporternas kvalitet uppnådde vissa minimumkrav, baserade på tillgängligt material och liknande uppdrag, och inte av strikt vetenskaplig karaktär. I denna fas bedömdes 50 utvärderingsrapporter på ett liknande, stegvist sätt som i den först fasen.

Bedömningen i denna fas omfattande inte några kvalitetsaspekter. Istället utgick bedömningen från den befintliga analysen och de slutsatser som presenterades i utvärderingsrapporterna, och graderade denna information utefter en satt standard. En fyrstegsskala användes med svarsalternativen „nej“, „snarare nej“, „snarare ja“ och „ja“, och statistik togs fram för att fastställa resultat vad gällde insatsernas

i) relevans,

ii) måluppfyllelse,

iii) kostnadseffektivitet,

iv) effekt,

v) bärkraft, och

vi) biståndseffektivitet (koherens, samordning och komplementaritet).

Motiveringar för särskilda bedömningar, generella lärdomar och rekommendationer kategoriserades med hjälp av nyckelord och kodades därefter med hjälp av mjukvarupaketet MaxQDA®. I ett sista steg gjordes en kvalitativ innehållsanalys för att urskilja trender och nya frågeställningar.

För att säkerställa hög kvalitet testade alla analysverktyg utförligt före användning. Tio procent av slumpmässigt utvalda rapporter utsattes för korsanalys (fem för kvalitet och fem för innehållsbedömning) och både team-ledaren och dennes ställföreträdare bidrog med fackmässigt understöd. Dessutom hölls interna och externa seminarier inom utvärderingsteamet och med utrikesministeriet för att verifiera resultat.

Vad gäller uppdragets viktigaste begränsningar är det viktigt att påpeka

i) att meta-utvärderingen inte skall ses som en ytterligare utvärdering av enskilda projekt eller program och således kan resultaten endast bedömas på aggregerad nivå;

ii) att resultat och slutsatser endast kan anses gälla för en bråkdel av Finlands utvecklingssamarbete baserat på 51 utvärderingsrapporter om bi-, multi- och multi-bilaterala insatser, och därför inte är tillämpliga för andra typer av finska utvecklingssamarbete;

iii) att analysen endast förlitar sig på information från utvärderingsrapporter och uppdragsbeskrivningar, och därmed kunde triangulering gentemot andra informationskällor inte tillämpas; och

iv) att bedömning omfattade utvärderingsrapporter av vitt skilda insatser (omfattande ett stort antal länder, regioner, sektorer, insatser med skiftande budgetar m m), vilket gjorde det nödvändigt att vikta utvärderarnas bedömning av kvalitet och innehåll efter små och stora insatser.

I följande tabell uppsummeras metautvärderingens **viktigaste resultat, slutsatser och rekommendationer**. Dessa presenteras separat för varje steg i analysen (dvs. bedömningen av rapporternas kvalitet och innehåll) och grupperas i olika aspekter. För varje aspekt anges motsvarande utvärderingsfråga inom parantes.

De viktigast slutsatserna kan summeras som följer:

i) 60 % av uppdragsbeskrivningar i urvalet är av godkänd kvalitet. Information on metod, utvärderingsprocess, kvalitetssäkring och hur tvärfrågor skall analyseras är ofta bristfällig.

ii) Det finns i allmänhet ett samband mellan kvaliteten på uppdragsbeskrivningar och kvaliteten på utvärderingsrapporter. Avsnitten om syfte, mål och omfattning av utvärderingen; om metod och om utvärderingsprocessen är särskilt viktiga för den övergripande rapportkvaliteteten.

iii) Två tredjedelar av utvärderingsrapporterna är av godkänd kvalitet. Vi noterar att slutsatser ofta dras på basis av undermåliga metoder, men samtidigt verkar vara någorlunda tillförlitliga. Förändringsteori, grundläggande antaganden, och begränsningar berörs ofta inte tillräckligt i rapporterna.

iv) Enligt utvärderingsrapporterna är den övergripande kvaliteten på Finlands utvecklingssamarbete godkänd eller bättre för 70 % av insatserna. Styrkan ligger i insatsernas relevans. Den största utmaningen är att förbättra bärkraften. De viktigaste rekommendationerna i rapporterna berör områdena ”planering”, ”omfattning”, ”hantering”, ”kapacitet” och ”bärkraft”.

Även om vi inte kan dra några övergripande slutsatser om utvärderingsportföljen står det klart att det finns utrymme för förbättringar vad gäller kvaliteten på uppdragsbeskrivningar och kvalitetssäkringen av utvärderingsrapporter, vilket pekar på kapacitetsbrister inom utrikesministeriet.

Mot denna bakgrund vill vi ge följande huvudrekommendationer:

i) att samtliga insatser inventeras och kategoriseras för att möjliggöra ett stratifierat urval i framtiden.

ii) att väsentligt förbättra utvärderingsmanualen, för att öka kunskapen om utväderingsmetoder och praxis och att överväga en förbättring av befintliga strukturer, t ex genom centralisering av systemet för kunskapshantering och bättre samordning med EVA-11.

iii) att förbättra kvalitetssäkringen och ställa högra krav på metodkompetens då utvärderare upphandlas.

iv) att resultaten av denna meta-utvärdering sprids tillräckligt för att säkerställa återkoppling till de som planerar och genomför insatser.

SUMMARY

The Ministry for Foreign Affairs of Finland (MFA) commissions meta-evaluations on a regular basis. This is done through the Development Evaluation Unit (EVA-11), an operationally independent unit that reports directly to the Under-Secretary of State.

In general, the **rationale** behind meta-evaluations is twofold: They are understood as a valuable tool

i) "for accountability and improved transparency towards partner countries, general public, parliamentarians, academia, media and development professionals outside the MFA" (cf. ToR), and

ii) for the analysis of the quality of its overall evaluation function by synthesising results and lessons learnt from a wide range of development cooperation interventions funded by Finland.

Subject to this meta-evaluation were 51 decentralised evaluation reports and 45 corresponding ToRs developed between September 2015 and August 2017. The documents comprised 23 mid-term and 28 final evaluations of single bilateral, multilateral or multi-bi projects and programmes commissioned by various thematic units, regional departments of the MFA, embassies and multilateral partners.

The **main purpose** of the assignment is as follows: It aimed at providing

i) concise conclusions and recommendations enabling the MFA to enhance the quality of decentralised evaluations, to improve evaluation management practices and to foster evaluation capacity development, and

ii) aggregated insights on Finnish development cooperation emerging from the evaluation reports to derive recommendations on how to improve MFA's development cooperation

Thus, the assignment consisted of two parts: (i) a meta-evaluation assessing the quality of the evaluation reports under consideration and (ii) a summative meta-analysis aggregating the content of these reports. Accordingly, the o**bjectives** of the meta-evaluation comprised the provision of:

i) an overall picture of the evaluation portfolio,

ii) an assessment of different decentralised evaluation reports and their ToR,

iii) a synthesis of reliable evaluation findings, and

iv) other identified issues emanating from the evaluation reports.

The **meta-evaluation was designed** as a two-stage multi-method analysis and followed a participatory approach. To utilise resources efficiently in line with good evaluation practice, it was built on

i) lessons learnt of previous meta-evaluations commissioned by MFA,

ii) meta-evaluations carried out by other organisations,

iii) similar assignments conducted by the meta-evaluation team,

iv) findings from an initial document review, and

v) insights by MFA gained throughout the evaluation process.

With respect to the **evaluation methods**, the first stage of the analysis comprises a **methodological quality assessment** using a standardised assessment tool (i.e. a detailed checklist for 51 evaluation reports and their ToRs). It started with the assessment of a large number of single sub-aspects of report and ToR quality. Whenever appropriate, these sub-aspects were checked against yes/no answer options, otherwise a four-step scale with clearly defined categories was applied. Therefore, the categories "good or very good", "satisfactory", "need for improvement" and "inadequate" were introduced. The introduction of this scale acknowledged a lesson learnt from similar assignments: Summarising better reports and extraordinary work in one category allows stronger differentiation at the lower end to finally derive concise conclusions and detailed recommendations for evaluation capacity development.

In a next step, a number of single sub-aspects was weighted according to their relative importance and summarised to one aspect (graded on a four-step scale). Finally, aspects were summarised to an overall assessment (again, on a four-step scale) comprising

i) quality of introductions and context analyses,

ii) quality of evaluation methodology,

iii) quality of evaluation findings,

iv) quality of conclusions and recommendations, and

v) quality of executive summaries.

This stepwise procedure avoids oversimplification while covering a wide range of different aspects in detail. At the same time, it is highly robust to evaluator biases given its high degree of standardisation.

The second stage of the analysis comprises a detailed content assessment of Finnish Development Cooperation (as far as covered by the evaluation reports under consideration in this meta-evaluation) using a semi-standardised assessment tool. This stage is conditional on minimal methodological report quality, understood in the context of the available material and comparable assignments, and hence, not as strict as for purely scientific purposes. Thus, 50 evaluation reports underwent a similar stepwise procedure as described for the first stage.

Here, the meta-evaluation team no longer checked on aspects related to report quality. Instead, it transferred assessments provided in the evaluation reports into standardised ratings. By doing so, a four-step scale with the answer options "no", "rather no", "rather yes" and "yes" was applied and summary statistics were calculated to derive systematic results for

i) interventions' relevance,

ii) interventions' effectiveness,

iii) interventions' efficiency,

iv) interventions' impact,

v) interventions' sustainability, and

vi) interventions' aid effectiveness and triple C (i.e. coherence, complementarity, coordination).

Further, underlying reasons for a particular assessment, lessons learnt and recommendations presented by the evaluators were collected in key words and underwent thematic coding with the software package MaxQDA®. In a final step, a qualitative content analysis facilitated the identification of general trends and emerging issues.

To ensure high quality, all assessment tools were extensively pre-tested. 10% randomly selected reports (i.e. five for quality and five for content assessment) were cross-analysed and both team leader and deputy provided intensive technical backstopping. Additionally, internal and external validation workshops were conducted to cross-validate the results within the meta-evaluation team as well as with the MFA.

Regarding the main limitations it is important to understand

i) that this meta-evaluation cannot be understood as a re-evaluation of single projects or programmes and thus, results can only be interpreted at an aggregated level;

ii) that results and conclusions only hold for a fraction of Finland's development cooperation portfolio based on 51 evaluation reports of bi-, multi-, and multi-bilateral interventions and thus, are not valid for other instruments of Finnish development cooperation;

iii) that the analysis is only relying on information from evaluation reports and ToRs, and thus no triangulation with other data sources was possible; and

iv) that the assessment tools were applied to evaluation reports of very heterogeneous interventions (e.g. wide range of countries, regions, thematic sectors, intervention budgets) which required weighting quality and content of evaluators' assessments equally for small and large interventions.

Main findings, conclusions and recommendations of the meta-evaluation are presented in the following summary table. They are presented separately for each stage of the analysis (i.e. the report quality assessment and the content assessment) and grouped according to thematic aspects. Corresponding evaluation questions are specified in brackets for each thematic aspect.

At a glance, **main findings** can be summarised as follows:

i) The overall quality of the ToR is satisfactory for 60% of the assessed ToR. Information provided on the methodology, the evaluation process, quality assurance and the cross-cutting objectives is often rather weak.

ii) On average, higher ToR quality is associated with higher report quality. Sections on purpose, objectives and scope of the evaluation; on the methodology, and on the evaluation process are particularly important for overall report quality.

iii) Overall report quality is satisfactory for two thirds of the reports. We observe that findings are often based on weak methodologies but appear to be somewhat reliable. Appropriate discussion of the intervention logic, underlying assumptions and its limitations are often neglected.

iv) According to the evaluation reports, the overall quality of Finnish development cooperation is assessed as of moderate quality or better for 70% of the interventions. Relevance is considered as strength and sustainability as greatest challenge. Major recommendations provided by the evaluators are related to the intervention fields of "Planning", "Scope", "Management", "Capacity" and "Sustainability".

While we cannot conclude on the adequacy of MFA's decentralised evaluation portfolio, we can **conclude** that the quality of the ToRs and the quality assurance of evaluation reports: both leave room for improvement which in turn reveals capacity gaps within MFA.

In consequence, we derive the following **key recommendations**:

i) to run an inventory of all interventions classified by key characteristic to enable stratified sampling in the future.

ii) to improve the evaluation manual substantially, to enhance knowledge of evaluation methodologies and practices and to consider improving existing structures, e.g. via stronger centralisation of the knowledge management system and better coordination with EVA-11.

iii) to enhance quality assurance and to ensure methodological expertise when recruiting evaluators.

iv) to ensure that the results of this meta-evaluation are sufficiently disseminated to feed back this information to implementers.

SUMMARY TABLE

Findings	Conclusions	Recommendations
related to the quality assessment of the evaluation reports (meta-evaluation)		
MFA's decentralised evaluation portfolio (EQ1) • We find a high number of evaluation reports on interventions in the fields of environment/climate, conflict/security and in the partner country Nepal. • Given the lack of information on the whole population of bi-, multi- and multi-bilateral interventions, we cannot assess to which extent this sample of evaluation reports is representative for this part of Finnish development cooperation. • The quality assessment of bi- and multilateral Finnish development cooperation is only based on the 50 decentralised evaluation reports. Self-assessments by the implementers or cross-checks on the interventions were beyond this assignment.	We cannot conclude on the adequacy of MFA's decentralised evaluation portfolio. Triangulation and contextualisation beyond using different evaluation reports as data source was impossible.	**for commissioning future meta-evaluations** *R5.2: Enhance the representativeness of future samples* (i.e. set up and maintain an inventory of all interventions classified by key characteristics to enable stratified sampling) *R5.1: Use the same assessment tools for future meta-evaluations* to allow comparisons over time and sub-group comparisons. *R5.3 Enhance the sources of evidence for future meta-evaluation* (e.g. allow online surveys with implementers or evaluators to obtain information on the evaluation process and to triangulate findings)
Quality of ToR and their linkage to overall report quality (EQ3, EQ6) • The overall quality of ToRs is satisfactory for 60% of the ToRs. • All ToRs could be improved in some ways and more than one third are assessed as in need of significant improvement. In particular information provided on the methodology, the evaluation process, quality assurance and the cross-cutting objectives was rather weak. • On average, a higher quality of ToRs is associated with a higher quality of the subsequent evaluation reports. • The ToR's sections on purpose, objectives and scope of the evaluation; on the methodology, and on the evaluation process are particularly important for overall report quality.	C4: While the overall quality of ToRs can be considered as satisfactory, there is room for improvement with regard to providing methodological and practical advice. C5: A higher quality of ToRs is related to a higher quality of evaluation reports.	**for drafting ToRs (also based on C1, C2, C6)** *R2.1: Be more precise on methodological requirements and on expectations regarding the different OECD DAC criteria* (i.e. addressing evaluation design, underlying sampling strategies, known limitations, e.g. outcome analysis in effectiveness chapter) *R2.2: Amend ToRs by several missing aspects* (i.e. (i) revision of the intervention logic (ii) cross-cutting objectives, (iii) triple C, (iv) implementable recommendations and addressees, (v) users of the report and their expectations, (vi) provision of general lessons learnt, (vii) length and content of the executive summary) *R2.3: Pay particular attention to the quality of ToRs for smaller evaluations (in terms of budget and intervention size)*

Findings	Conclusions	Recommendations
Reliability and quality of the evaluation reports (EQ2, EQ4, EQ5, EQ7) • The overall report quality is assessed as "satisfactory" for two thirds of the reports and in "need for improvement" for one third. • Findings are often obtained based on a weak methodology and there is a great need of improvement. The selection and presentation of evaluation design, sampling strategies and resulting limitations is unclear in about half of the reports. The intervention logic, fundamental for a sound understanding of the intervention and an appropriate analysis, is discussed comprehensively in less than one third of the reports. More than half of the reports do not link their findings to the data sources. • MFA's request to include the context analysis after the methodology chapter is unusual and not often followed by the evaluators. About three quarters of the reports, regardless of who was the commissioning entity, are not in line with MFA's requested structure in any way. • Overall report quality does not vary between i) evaluations commissioned by MFA or others, ii) by individual/independent consultants or teams of consulting firms/institutes; or iii) according to different project budgets.	C1: Most evaluation reports feature considerable weaknesses regarding methodological rigour and transparency. Still, except for one report, findings appear to be somewhat reliable. C2: None of the reports' quality is highly satisfactory. About two thirds feature some, one third substantial quality flaws. C3: The overall report quality does not vary between different sub-groups.	**for evaluation management** (also based on C4, C5, C6) *R4.1: Enhance quality assurance throughout the evaluation process* (i.e. (i) make sufficient resources available for methodological and thematic quality assurance of inception reports, (ii) verify compliance with proposed methodology and MFA's requirements in draft reports, (iii) insist on sources of evidence, triangulation, use of the intervention logic to obtain findings and causal attribution of findings to interventions, (iv) not accept reports considerably failing in the above-mentioned, which do not respond to evaluation questions or which lack complete sections) **for recruitment of evaluators** (also based on C4, C5, C6) *R3.1: Be gender-transformative throughout the recruitment process* This comprises the empowerment of women and LGBT and goes beyond the gender-balancing of evaluation teams. *R3.2: Ensure sufficient methodological expertise* This is at least equally important as thematic and regional expertise and key to improve the quality of evaluation reports.

Findings	Conclusions	Recommendations
Gaps in MFA's evaluation capacity (EQ8) • The aspect is already captured by the key findings on ToR quality as presented above.	C6: The fact that the quality of the ToRs leaves room for improvement reveals capacity gaps within MFA.	**General guidance on decentralised evaluations within the MFA** (also based on C1, C2, and C4) *R1.1: Improve the Evaluation Manual* (i.e. to sensitise for (i) transparency regarding data collection instruments, (ii) contextualise findings with previous evaluation results, (iii) linking evidence to findings, (iv) triangulation, (v) discussion of causal attribution, (vi) guidance on impact and sustainability analyses, (vii) streamlining report structures, (viii) rough estimates on costs, personal and time requirements of different evaluation designs, (ix) their explanatory power, (x) responsibilities of commissioners and evaluators within the evaluation process.) *R1.2: Enhance knowledge of evaluation methodologies and on evaluation practice with EVA-11 as focal point* (i.e. regarding (i) drafting specifications on methodology, evaluation process, quality assurance and cross-cutting objectives for ToRs, (ii) expertise to assess suggested methodologies of inception reports and review draft reports, (iii) knowledge of costs of different evaluation designs, feasibility of tasks, human resource requirements and time frames to keep expectations for evaluations realistic) *R1.3: Consider improving existing structures* (e.g. a centralised knowledge management system and stronger coordination with EVA-11)
related to the content assessment as made by the evaluators		
Quality, strengths and weaknesses of bi- and multilateral Finnish development cooperation according to OECD DAC criteria (EQ10-14, EQ23-EQ25) • The overall quality of bi-, multi- and multi-bilateral interventions is assessed for 70% of the 50 interventions as of moderate quality or better. • As more than one third of the interventions is assessed as being weak with regard to their effectiveness, efficiency or impact and about half of the interventions with regard to their sustainability, there is room for improvement in these areas. • Relevance is a typical strength and sustainability is the greatest challenge of bi- and multilateral interventions. • The overall quality of interventions at regional or global level does not significantly differ from the overall quality of interventions at national level. Similarly, no differences can be detected for different regions, thematic sectors or intervention budgets.	C7: The quality of the bi- and multilateral interventions under consideration is assessed quite positively with their relevance being considered as a particular strength and sustainability as the greatest challenge.	**for evaluation management** *R4.2 Make use of meta-evaluation results from the content assessment* EVA-11 should ensure that there is sufficient and appropriate dissemination and uptake of the meta-evaluation results emanating from the summative analysis. Particular importance should be paid to the synthesised recommendations regarding M&E systems.

Findings	Conclusions	Recommendations
Gender as cross-cutting objective in bi- and multilateral Finnish development cooperation (EQ15-18) • Finnish development cooperation is neither gender-blind nor gender-transformative, but somewhere in between. • Assessment of other cross-cutting objectives was not possible given the lack of analyses in the majority of reports.	C8: Interventions are mostly not gender-transformative.	
Aid effectiveness of bi- and multilateral Finnish development cooperation (EQ19-EQ22) • The assessment of aid effectiveness and triple C (i.e. coherence, coordination and complementarity) is not deeply anchored into Finnish development cooperation evaluation practice. • It remains unclear if and to what extent the interventions under consideration follow one of these concepts.	C9: It remains often unclear if and to what extent the interventions follow the concepts of aid effectiveness and triple C.	
Major recommendations emerging from decentralised evaluation reports (EQ26) • More than half of the evaluation reports contain recommendations related to the intervention fields of "Planning", "Scope", "Management", "Capacity" and "Sustainability". More than three quarters of the reports contain recommendations related to "M&E". • Only 30 out of 50 evaluation reports contain lessons learnt. Just under half of the lessons learnt presented are in fact intervention-specific recommendations. "True lessons learnt" in accordance to the OECD DAC definition are spread over a wide range of different topics. Hence, no "typical" lessons could be identified.	C11: Apparently evaluators regard intervention planning, scope, management, capacity and/or sustainability as improvable.	

1 INTRODUCTION

1.1 Scope, purpose, and objectives of the meta-evaluation

In order to assess Finnish development cooperation and the reliability of evaluation reports, the Ministry for Foreign Affairs of Finland (MFA) commissions meta-evaluations on a regular basis. This is done through the Development Evaluation Unit (EVA-11) which is an operationally independent unit that reports directly to the Under-Secretary of State. The MFA appreciates meta-evaluation as *"a tool for accountability and improved transparency towards partner countries, general public, parliamentarians, academia, media and development professionals outside the MFA."* (cf. Terms of Reference (ToR)). The MFA further understands meta-evaluation as a valuable tool facilitating the analysis of its overall evaluation function and its quality by synthesising results and lessons learnt from a wide range of different development cooperation interventions funded by Finland.

Within the **scope** of this assignment all decentralised evaluation reports and corresponding ToRs conducted between September 2015 and August 2017 were subject to a meta-evaluation. In contrast to larger centralised evaluations at policy level which are directly commissioned byEVA-11, decentralised evaluations cover mid-term reviews, mid-term evaluations and final evaluations of single bilateral or multilateral projects or programmes commissioned by various thematic units or regional departments of the MFA, by embassies or by multilateral partners. Decentralised evaluations which were finalised until August 2015 are covered by earlier meta-evaluations conducted in 2007, 2009, 2012, 2014 or 2016. In this regard, this meta-evaluation is seamlessly connected with earlier efforts.

All decentralised evaluation reports and corresponding ToRs conducted between September 2015 and August 2017 were subject to this meta-evaluation.

The **purpose** of the assignment is as follows:

vi) It aims at drawing concise conclusions and recommendations enabling the MFA to enhance the quality of decentralised evaluations, to improve evaluation management practices and to foster evaluation capacity development.

vii) It aims at providing an overall picture of the current evaluation portfolio disclosing possible gaps in MFA's operations.

vii) It aims at aggregated insights on joint lessons learnt emerging from the evaluation reports and at disclosing strengths and challenges of the analysed portfolio to derive recommendations on how to improve Finnish development cooperation.

Thus, the assignment consists of two parts: (i) a meta-evaluation to assess the quality of the evaluation reports under consideration and (ii) a summative meta-analysis to aggregate the content of these reports. Accordingly, the **objectives** of the meta-evaluation comprise the provision of:

v) An overall picture of the evaluation portfolio,

vi) An assessment of different decentralised evaluation reports and their ToR,

vii) A synthesis of reliable evaluation findings, and

ix) Other identified issues emanating from the evaluation reports.

Originally, it was intended to put the results of this meta-evaluation into perspective to the Meta-evaluation of Project and Programme Evaluations in 2014-2015. However, as the earlier meta-evaluation followed a completely different assessment methodology, the MFA acknowledged in the validation workshop that a systematic comparison is not possible. To enhance the long-term utility of meta-evaluations in future, the MFA plans to standardise the assessment tools with the aim to carry out comparable meta-evaluations every two years.

The remainder of this report is organised as follows. The introductory chapter 0 is complemented by the presentation of the evaluation questions (1.2) and general information on the assignment (1.3). In the method chapter 2 the general approach to this meta-evaluation (2.1), data sources (2.2), assessment tools (2.3), the procedure of aggregation and further analysis (2.4) as well as limitations (2.5) are presented. A context analysis is provided in chapter 3. It gives an overview of Finland's development policies (3.1), the delivery of Finnish aid (3.2) and the evaluation reports under consideration for this meta-evaluation in light of the Finnish development context (3.3).

In chapter 4, the findings of the quality assessment of the evaluation reports are presented. First of all, it analyses the quality of underlying ToRs (4.1) and subsequently provides insights on the introductions and context analyses provided (4.2), evaluation methodologies applied (4.3), the way of deriving evaluation findings (4.4), conclusions and recommendations drawn (4.5), further aspects like cross-cutting themes or formal aspects (4.6) and the executive summaries provided (4.7). Furthermore, the overall quality of the evaluation reports in relation to the ToR quality as well as disaggregated quality for subsample groups according to various characteristics like different commissioners, mid-term vs. final evaluation etc. (4.8) are assessed.

After comprehensive quality assessment, chapter 5 provides a summative content analysis to synthesise the contribution of the fraction of Finnish development cooperation which is captured by this assignment. This includes an assessment along the OECD-DAC criteria relevance (5.1), effectiveness (5.2), efficiency (5.3), impact (5.4) and sustainability (5.5), as well as insights on gender and other cross-cutting themes (5.6), aid effectiveness and on the European Union's triple C (5.7). Furthermore, an analysis of the lessons learnt presented in the evaluation reports (5.8) and the recommendations drawn by the evaluators (5.9) are presented. The chapter is completed by an assessment of the overall quality of Finnish development cooperation in the light of the analysed evaluation reports which also appreciates different sub-groups within the sample e.g. according to geographical scope or different sectors (5.10).

Key objectives were a quality assessment of evaluation reports and ToR and a synthesis of reliable evaluation findings.

Finally, chapter 6 contains the conclusions of the meta-evaluation team and chapter 6 provides recommendations to improve the quality of the evaluation reports and to enhance the contribution of Finland's development cooperation.

1.2 Evaluation questions

In the ToR, the MFA specified the evaluation questions as follows:

"Meta-evaluation:

Assessment and description of MFA's decentralized evaluation portfolio (evaluation reports and their corresponding ToRs) based on the OECD/DAC evaluation principles and standards, classified by countries, sectors, budgets, evaluation types, managing units of MFA, commissioner, etc.

- Assessment of the reliability of evaluation reports
- Are there gaps in evaluation capacity of MFA that need to be strengthened?
- Is there a difference between the quality of MFA commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?

Meta-analysis:

1. What can be said about the Finnish development cooperation based on the reliable decentralized evaluation reports, and related planning documents by each OECD/DAC criteria and other relevant criteria identified in Finnish development policies
2. What are the major issues emerging from the decentralized evaluation reports?
    - Success stories, good practices and challenges"

As these questions are very comprehensive and comprise multiple dimensions, the MFA agreed during the inception phase to the following specifications to simplify structuring of the analysis:

For the **meta-evaluation**:

1. How can MFA's decentralised evaluation portfolio be described?
2. How is the quality of MFA's decentralised evaluation reports?
3. How is the quality of the corresponding ToRs?
4. How is the quality of MFA's decentralised evaluations classified by countries, sectors, evaluation types, commissioner, etc. if applicable?
5. Is there a difference between the quality of MFA-commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?
6. Are there systematic patterns regarding the quality of the evaluation reports and corresponding ToRs?
7. How reliable are the decentralised evaluation reports?
8. Are there gaps regarding MFA's evaluation capacity?

9. What are recommendations to improve the quality of MFA's decentralised evaluations?

For the **summative meta-analysis**:

10. What can be said about the relevance of Finnish development cooperation based on the reliable decentralised evaluation reports?
11. What can be said about the effectiveness of Finnish development cooperation based on the reliable decentralised evaluation reports?
12. What can be said about the efficiency of Finnish development cooperation based on the reliable decentralised evaluation reports?
13. What can be said about the impact of Finnish development cooperation based on the reliable decentralised evaluation reports?
14. What can be said about the sustainability of Finnish development cooperation based on the reliable decentralised evaluation reports?

Gender and other cross-cutting objectives:

15. What can be said about the consideration of gender equality in Finnish development cooperation based on the reliable decentralised evaluation reports?
16. What can be said about the consideration of reduction of inequality/equal opportunities to participate/rights of the most vulnerable in Finnish development cooperation based on the reliable decentralised evaluation reports?
17. What can be said about the consideration of climate sustainability/climate change preparedness and mitigation in Finnish development cooperation based on the reliable decentralised evaluation reports?
18. What can be said about the consideration of the human rights-based approach in Finnish development cooperation based on the reliable decentralised evaluation reports?

Aid effectiveness and triple C:

19. What can be said about the aid effectiveness of Finnish development cooperation based on the reliable decentralised evaluation reports?
20. What can be said about the complementarity of Finnish development cooperation based on the reliable decentralised evaluation reports?
21. What can be said about the coordination of Finnish development cooperation based on the reliable decentralised evaluation reports?
22. What can be said about the coherence of Finnish development cooperation based on the reliable decentralised evaluation reports?

Overall quality, strength and weaknesses:

23. What can be said about the overall quality of Finnish development cooperation based on the reliable decentralised evaluation reports?
24. What are the major strengths emerging from the reliable decentralised evaluation reports?

25. What are the major challenges emerging from the reliable decentralised evaluation reports?

Major recommendations from the evaluation reports:

26. What are the major recommendations to improve Finnish development cooperation emerging from the reliable decentralised evaluation reports?

1.3 The assignment

This meta-evaluation was part of a Framework Contract for providing Evaluation Management Services (EMS) to the Development Evaluation Unit (EVA-11) of the MFA, delivered by a consortium composed of Particip GmbH, as the main contractor, and Indufor Oy. The EMS is a new approach launched by EVA-11to manage centralised evaluations. The purpose of the new concept is to strengthen the quality of outsourced evaluations and to increase flexibility, efficiency and effectiveness of the MFA in planning and commissioning evaluation assignments.

The main difference to the previous procedures is that each evaluation assignment is divided into two service orders. The first service order is kick-started when EVA-11provides draft ToR of the assignment with which the consortium can start searching for potential Team Leaders. The Evaluation Management Services Coordinator recruited by the consortium then shortlists potential candidates for submission toEVA-11. Once the Team Leader has been approved by EVA-11 and recruited, the Team Leader with the assistance of the consortium prepares an evaluation proposal including comments to the ToR, followed by identification of team members and a draft budget. For this assignment Team Leader, deputy and methodological expert from CEval GmbH were selected, the Finnish development policy evaluation expert came from Indufor Oy and the development evaluation generalist from Particip GmbH.

One of the key differences in this process, compared to the previous procedures, is the step where the evaluation Team Leader, an expert of the subject matter, provides his/her inputs to the ToR. After the ToR have been finalised and the team members defined, the actual evaluation begins under the second service order following a normal evaluation procedure. The process is facilitated by the EMS Coordinator, contracted by the Consortium, who acts as an interlocutor and quality assurance expert between the parties.

Therefore, the ToR of this assignment were a result of the close cooperation between EVA-11, the Team Leader and Deputy Team Leader (CEval GmbH), the EMS Coordinator and the Consortium partners. Similarly, during implementation, quality and rigor of analysis as well as deliverables were assured by the EMS Coordinator, the Consortium's internal processes, the team leaders, as well as by a Reference Group established by EVA-11.

2 METHODOLOGY

2.1 Approach

Although EVA-11 has commissioned meta-evaluations at regular intervals, the approaches and methodologies of these assessments have not been standardised. This has created challenges in outlining clear trends in development cooperation over time. This meta-evaluation therefore built on (i) lessons learnt of these past experiences, (ii) meta-evaluations carried out by other organisations, (iii) similar assignments conducted by the meta-evaluation team, (iv) insights from an initial document review and (v) clarifications as well as ideas by MFA gained through meetings.

Overall, we applied a **two-stage approach** to respond to the evaluation questions. The first stage of the analysis provides insights for all evaluation reports and focused on **methodological quality assessment**. In the second stage we delved deeper into detail and focused on **content assessment** against the OECD-DAC criteria and the aid effectiveness agenda.

The checklists with criteria and sub criteria used in this study are based on an approach developed specifically for meta-evaluations and systematic reviews at CEval. They aim at: (i) establishing a robust toolkit for the MFA to evaluate the quality of its decentralised evaluations, (ii) providing reliable insights for accountability purposes, and (iii) drawing emerging issues of projects and programmes from the evaluation reports' point of view. Thereby, it is important to understand that we aimed at developing **practicable tools which build on best practice**.

The two-stage multi-method approach comprised methodologica quality assessment and content assessment.

For the quality assessment, given that the evaluation reports under consideration are heterogeneous with respect to various aspects, a high degree of content-related and methodological heterogeneity had to be taken into consideration. On the one hand the contexts of the interventions differ tremendously: (i) varying context conditions e.g. poverty levels, degree of political stability, etc. in the countries under consideration, (ii) differences among implementing partner organisations e.g. level of operations, financial resources etc., (iii) different thematic focuses, and (iv) varying working approaches e.g. technical, human rights-based etc. On the other hand, the evaluations are characterised by conceptual differences like (i) different scope and scale of the evaluations (e.g. mid-term vs. final, programme vs. project evaluation etc.) and (ii) different evaluation designs with accordingly varying data sources and analysis methods used (i.e. contribution analyses, ex-post facto designs etc.).

The first stage of the analysis started with the assessment of a large number of single aspects related to methodological quality, thus acknowledging heterogeneity of the reports. This step helped to avoid oversimplification, and allowed covering a wide range of different topics in detail. They were whenever appropriate and sufficient, checked against yes/no answer options, otherwise a

four-step scale was applied. A set of single aspects was then weighted and summarised to one sub-criterion (graded on a four-step scale). Finally, criteria were weighted and summarised to an overall assessment (again, graded on a four-step scale).

A grading system with a four-step scale has the advantage to avoid the, in social science well proved, human tendency to centrality. We know from similar experiences that it is helpful to summarise better reports and extraordinary work in one category allowing stronger differentiation at the lower end. Thus, the categories "good or very good", "satisfactory", "need for improvement" and "inadequate" were introduced. This was particularly beneficial for deriving concise conclusions and detailed recommendations for evaluation capacity development. By following this stepwise procedure, we finally identified general trends, displayed heterogeneity, prepared the ground for enhancing the quality of evaluations, and offered concise summarising results tables.

To provide valid, objective and reliable results in the second-stage, evaluation reports, which did not pass a threshold of minimal methodological quality, were excluded from the summative content analysis. Hence, the assessment of the joint contribution of MFA's development cooperation was conditional on methodological standards. However, we understand minimal methodological quality in the context of the available material and comparable assignments and did not apply as strict criteria as would be required for purely scientific purposes.

Similarly, as for the first stage, the **content assessment** followed a stepwise procedure for single aspects, sub-criteria and criteria. Thereby, the meta-evaluation team no longer checked on aspects related to quality. Rather, it transferred the assessments provided in the evaluation reports into standardised ratings. Thus, it was no longer under question if a report addresses for example the OECD-DAC criterion relevance in a methodologically and technically sound manner. Instead, it was asked whether and to which extent an evaluator assessed the intervention analysed as relevant. By doing so a four-step scale was applied with the answer options "no", "rather no", "rather yes" and "yes".

Given the summative character of this second stage analysis, we went beyond standardised assessment, and also captured influencing factors which determine the assessments provided in the evaluation report. We coded such factors for a number of different sub-criteria of the OECD-DAC criteria, aid effectiveness, complementarity, coordination and coherence. The lessons learnt and recommendation of the evaluation reports were subject to a similar analysis. After finalising these steps for all evaluation reports under consideration, we applied qualitative content analysis and summary statistics to derive systematic results. This allowed concise summarising and identification of emerging issues of the fraction of Finnish Development Cooperation under consideration in this meta-evaluation. To further enhance organisational learning and evaluation capacity development important aspects were exemplarily highlighted.

To sum up: we provided a two-stage analysis with **quantitative and qualitative data analysis methods**. The analysis grid (Annex 5) displays in details which data sources and data analysis methods were used to reply to each evaluation question. A **multi-method approach** utilised resources efficiently and is in line

with good evaluation practice. The meta-evaluation provides separate insights for different strata of the heterogeneous sample (e.g. for mid-term vs. final evaluations or for different regions).

Our **participatory approach** fostered exchange with the MFA during all stages of the analysis. In order to ensure high quality, all assessment tools have been extensively pre-tested and 10% randomly selected reports (i.e. five for quality and five for content assessment) were cross-analysed (The following reports were randomly selected for cross-check: For the quality assessment reports No. 16, 23, 32, 39 and 52, for the content assessment reports No. 6, 21, 24, 53, 54). The Team Leader and the deputy, who led the overall assignment, provided intensive technical backstopping during each phase of this meta-evaluation.

2.2 Data sources

For the meta-evaluation and the summative meta-analysis, 56 evaluation reports were the main **source** of information. Although the collection of these evaluation reports is based on a request of EVA-11 to the different regional and thematic divisions and a search from the MFA's electronic archive application AHA, it is possible that not all evaluations have been reported by the divisions, hence that single evaluations are missing.

The **sample** includes all evaluation reports published between September 2015 and August 2017 known to EVA-11, including multi-bi projects and programmes which are completely or partially funded by the MFA. The administration of these interventions as well as their evaluations were done either directly by the MFA or by a partner organisation. In the latter case, the MFA has participated in commenting the ToR and evaluation reports but has not been the commissioner of the evaluation.

The sample includes muti-, bi- and multi-bilateral interventions funded by the MFA or by its partners.

Appraisal reports were not subject to this meta-evaluation as they are considered to be planning documents. Moreover, the evaluation team excluded one report which was in the sample twice, one very brief summary report, two self-evaluations and one report which only looked at the Norwegian contribution to an intervention as displayed in Annex 11. Thus, overall the sample for the meta-evaluation was reduced to 51 reports for the quality assessment. As one report did not comply with minimal methodological standards, this report was excluded for the content assessment, reducing the sample to 50 reports.

To answer to some of the evaluation questions it was necessary to consult the ToR. They were available for 45 of the reports. In Annex 11, all evaluation reports including information on availability of the corresponding ToR, year of writing, responsible MFA unit and budgets for the intervention and the evaluation are specified as received by the meta-evaluation team. Furthermore, the meta-evaluation's ToR requested to include the invitation to tenders (ITT). As they are only available for six evaluation reports, they could not be systematically used for this analysis.

2.3 Assessment tools

As primary and secondary data collection is the basis for an ordinary evaluation, data processing lays the foundation for meta-evaluations. Analytically, this is not always clearly distinguishable from data analysis and could be also seen as first step of the analysis process. Regardless of this scientific discourse, we first present the development and the general structure of the three assessment tools (i.e. for quality, ToR and content), and then explain how grades at section and sub-section level were calculated.

For the quality assessment of the reports and the ToR we developed an analysis tool, which is mainly based on the recent MFA evaluation manual (2013). Especially the checklist for the evaluation report, the outline of the evaluation report in the annex and the list of criteria were important sources of information. As a second source, the existing tool for meta-evaluation by Norad was consulted as proposed by the MFA. Further, the EU-ROM analysis grids for quality assessments were consulted but did only confirm the information already obtained. Importantly, this zero-draft tool was then compared with quality standards for evaluation by OECD-DAC to confirm the coverage of all important aspects and the alignment with international evaluation standards. We observed that the MFA evaluation manual is strongly based on these international standards and varies only occasionally. However, as the meta-evaluation team recognised that some important aspects were missing in the MFA manual, some amendments were made (e.g. request for provision of data collection instruments in the annex, results of previous evaluations, linking evidence, triangulation of findings and causal attribution of the intervention to the findings).

In general, the structure of the quality assessment tool follows the chapters of the evaluation report as suggested by the MFA manual. However, the meta-evaluation team anticipated that relevant information is sometimes not in the respective chapter. Thus, in principle regardless of where information was placed, it was considered by the meta-evaluation team. The structure of the tool runs chronologically, from the introduction, methodology, context and intervention logic, findings, conclusions, and recommendations to the annex to facilitate easy application. Cross-cutting objectives and general issues follow as answered best after the report has been read until the end. Also, the summary is easier to assess when the report is already known by the meta-evaluator. Therefore, these topics have been shifted to the end of the assessment process.

The main sections consist of sub-sections with very specific statements, so-called aspects, which were checked in terms of true or false. For example, the first section *1. Introduction and background* contains the sub-section *1.1 Rationale and purpose*. Within this section there are two statements which have been assessed by the meta-evaluator. For example, one of these statements is: *1.1a Report describes purpose of evaluation*. The meta-evaluation team checked if the original evaluator has described the purpose of the evaluation in the evaluation report and selects one of the answer options "yes" or "no". Most statements in the quality assessment tool could be answered with "yes" or "no", because many aspects refer to checking for existence of certain information in the report. Still, in several cases there are more answer options (on a four-step scale) which either refer to different grades of completeness or to more specific

assessments introduced by the evaluator. Table 2 provides an example to assess whether the sources of information are described.

Example: four scale question

Aspect	Answer options	Guidance for choosing answer
2.2a The sources of information are described.	(1) no, (2) short and incomplete , (3) short and complete, (4) detailed and complete	(1) no information (2) cryptic, incomplete, not naming types of documents or different groups to be interviewed etc., (3) short but naming all sources of information, (4) minimum one paragraph with three or more sentences with all sources of information

For the composition of the different sub-sections please refer to Annex 6. The exact specifications within the sub-sections can be withdrawn from Annex 7 where the instrument is presented in its entire complexity.

For a comprehensive meta-evaluation, it is important to also include an assessment of the underlying ToR. Reports may lack information as some aspects are not requested by the ToR. In order to detect these gaps, to generally determine the quality of the ToR, and to review the compliance of the ToR with MFA guidelines, quality assessment of the ToR is another part of the meta-evaluation. Consequently, a **ToR assessment tool** has been developed based on the MFA manual and the instrument used by NORAD. It is relatively compressed and divided into eight sections: intervention, purpose, objectives and scope of the evaluation, evaluation questions, evaluation criteria, methodology, feasibility, evaluation process and quality assurance as well as overarching and cross-cutting criteria. Again, all sections consist of different sub-sections including various aspects (see Annex 8). As the aspects refer to coverage of the topic in the ToR, the answering options are exclusively "yes" and "no".

For the summative analysis, we developed a separate tool with content-related criteria. The tools by UN Women and EU ROM as well as the ToR of this meta-evaluation laid its foundation. Furthermore, the MFA Evaluation Manual and the Manual for Bilateral Cooperation have been consulted to ensure compliance with MFA's standards. The **content assessment tool** consists of two main sections.

In the first part, the content of the evaluations with respect to the evaluation criteria is assessed. This comprises the five DAC criteria accompanied by additional criteria of aid effectiveness and the EU's triple C, i.e. coherence, complementarity, and coordination. Again, all criteria are further narrowed down into sub-criteria with single aspects. For each of the DAC criteria we first took over the general assessment of the original evaluators. In a next step, we focused on single aspects and captured their assessment. For key aspects we further searched for underlying reasons presented. We differentiated between positive and negative reasons and collected them in key words to be processed in the data analysis. Finally, we asked for each criterion if it is an example of good practice.

In the second part underlying reasons for evaluators' assessment on the OECD DAC criteria, lessons learnt and recommendations were captured in detail. Therefore, we applied thematic coding with the software package MaxQDA and

The standardised assessment tools for ToR and report quality are based on MFA's evaluation manual and lessons learnt from other meta-evaluations.

The content assessment tool focusses on the OECD DAC criteria and goes beyond standardisation when capturing underlying reasons for evaluator's assessment.

allocated both, lessons learnt and recommendations, to different statements. Whenever a lesson learnt or a recommendation did not fit to any category, it was captured under the section "others" to allow possible identification of new categories throughout later analysis steps. The composition of the different sub-sections is presented in Annex 9. For the exact specifications within the sub-sections please refer to Annex 10 where the complete instrument is presented.

2.4 Aggregation and further analysis

After the completion of the three semi-standardised assessment tools, **aggregations** were undertaken. In a first step, a grade was calculated from the results for each aspect under a sub-section. Weights were given to each aspect to balance its influence according to its importance. The default weight was set at "1." Only if some aspects are more important in comparison to other aspects in a particular sub-section or section, the weight was increased accordingly. Considering the weight, the arithmetic mean was then calculated at sub-section and section level. The allocation of weights at aspect, sub-section and section levels are presented in Annex 6.

In general, we focused at sub-section and section levels. Whenever appropriate, single aspects were considered to elaborate on results. In addition, an overall score draws a general picture on the quality of the evaluation reports under consideration. To generate such a score, we opted for an aggregation of all key chapters of the quality assessment as presented in this report. By allocating equal weights we did not overemphasise on a single element. On the other hand, by following key chapters, we underline the importance of grouping different sections to meaningful key topics as follows: (i) quality of introductions and context analyses, (ii) quality of evaluation methodologies, (iii) quality of evaluation findings, (iv) quality of conclusions and recommendations and (v) quality of executive summaries. Due to limited data availability the chapter on further aspects (i.e. integration of cross-cutting objectives, formal reporting aspects, validation and quality assurance and composition of the evaluation team) was not taken into account in the overall quality score.

Single aspects were aggregated to sub-aspects, several sub-aspects to aspects, and aspects to an overall quality assessment.

Furthermore, an overall quality score for the ToRs was generated along the simple weighted main sections of the assessment tool: (i) intervention, (ii) purpose, objective and scope of the evaluation, (iii) evaluation questions, (iv) evaluation criteria, (v) methodology, (vi) evaluation process and quality assurance, and (vii) cross-cutting objectives. Again, due to large data gaps the feasibility assessment of the evaluation did not feed into the overall ToR score.

Similarly, an overall score was developed to display the overall quality of Finnish development cooperation. Therefore, we aggregated the scores of the single OECD-DAC criteria assessed by the evaluator and divided them through the sum of OECD-DAC criteria assessed. Again, due to lack of information, evaluators' assessment of cross-cutting objectives, aid effectiveness and triple C of the intervention were not considered to create this overall score. For the detailed composition of the three overall scores please refer to Annex 6.

Overall scores enabled us to perform sub-group comparisons. Thus, we analysed whether the overall quality of the evaluation reports is different when MFA was the commissioner (vs. other partners), when the evaluation was implemented by individual/independent consultant(s) (vs. a team by consulting firms/institutes), and when it was a mid-term evaluation (vs. final). Moreover, the report quality and ToR assessment tools were linked to each other to detect general patterns (e.g. to check whether low quality ToR led to low quality reports).

Overall assessments allow performing sub-group comparisions and testing for linkages between quality of ToRs and reports.

In addition, we searched for differences regarding the quality of Finnish development cooperation, when comparing the overall score for national vs. regional/global interventions, different regions, different sectors and different project budgets.

Mann-Whitney test statistics were applied in the statistical software package STATA to detect significant differences between two groups and Kruskal-Wallis test statistics to differentiate for several groups. Spearmen's correlation coefficients were employed to analyse potential linkages between the quality of the ToR and the quality of the evaluation reports and ordinary least squares (OLS) regression and ordered logistic regression analysis with robust standard errors was conducted to identify determinants of the overall quality of Finnish development cooperation. For further explanations on the statistical tests used, please refer to https://stats.idre.ucla.edu/other/mult-pkg/whatstat/.

Beyond the standardised analysis, collected reasons, lessons learnt and recommendation captured by the content assessment tool were investigated separately by employing **qualitative content analysis**. Given tremendous variation in terms of quality of the lessons learnt, and the high complexity due to the vast number of recommendations this required some preparatory work.

Qualitative content analysis allows understanding of typical reasons underlying particular assessments and identification of typical recommendations emanating from evaluation reports.

With regard to lessons learnt, each lesson was scored 1, 2 or 3 depending on the quality of the formulation. Lessons that were formulated in line with the OECD-DAC definition (OECD 2010): *"Generalisations based on evaluation experiences with projects, programs, or policies that abstract from the specific circumstances to broader situations. Frequently, lessons highlight strengths or weaknesses in preparation, design, and implementation that affect performance, outcome, and impact."*, were given the score 3 (i.e. high quality). If the text would allow extracting the lesson with a reasonable level of expert judgement, it was scored 2 (i.e. medium quality). If it was not possible to conclude what the lesson would be or if arbitrary interpretation would have been necessary to identify a lesson as such, score 1 was given (i.e. low quality). Thus, lessons scored 1 typically described intervention-level findings or recommendations and were thus not taken into consideration for further analysis.

To provide a meaningful synthesis of rather heterogeneous lessons learnt and recommendations made by various evaluators in different reports, a three-step approach has been utilised. In a first step, the lessons or recommendations found in the evaluation reports were broadly assigned to categories corresponding to the main thematic interests of the meta-evaluation (e.g. the DAC criteria, aid effectiveness or M&E). This allowed identifying first tendencies with regard to the frequency of certain topics. Subsequently in a second step, the lessons or recommendations within each broader category were generalised and clustered to the extent possible. Finally, in a third step their overall frequency and

their importance based on the expert judgement of the meta-evaluation team were assessed. Lessons and recommendations appearing in more than 50% of the reports were synthesised and generalised further; they form the main part of the synthesis. Less frequent lessons or recommendations were treated anecdotally and added as illustrating examples when perceived as relevant.

Overall, aggregation and further analysis enabled us to identify influencing factors and general trends to derive systematic lessons from and recommendations for Finnish development cooperation.

2.5 Limitations and coping strategies

It is important to highlight that the analysis is only relying on information from 51 evaluation reports and above-mentioned documents. Project documents were not fed into the analysis and no original evaluators were consulted to receive further information. Self-assessments by the implementers or cross-checks on the interventions were beyond this assignment. Hence, triangulation and contextualisation in this regard was impossible. Thus, the analysis is limited to the information written down by the original evaluators and their assessments. Information not explicitly reported could not be considered. As the report is the medium that the user (MFA) receives, it should contain all information necessary to understand the evaluation process as well as the results from the evaluation. However, checking on the independence of the original evaluators goes beyond the scope of this assignment. It can be only guaranteed by measures of the MFA to ensure an appropriate selection process of evaluators.

Data sources cannot be triangulated as the evaluation reports are the only source of information for this assignment.

Regarding the sample of evaluation reports under consideration it is important to consider two main limitations: (i) Geographical scope, sectorial affiliation as well as intervention and evaluation budgets vary widely within the sample. Similarly, the nature of the intervention, the nature of the evaluations, their commissioner and the nature of the implementer are mixed. However, given the lack of information on the whole population of bi-, multi- and multi-bilateral interventions we cannot assess to which extent this sample is representative for this fraction of Finnish development cooperation. (ii) That the assessment tools were applied to evaluations of very heterogeneous interventions spread over a wide range of countries, regions, thematic sectors and intervention budgets required simplification. The quality and content of evaluators' assessments were weighted equally for small and large interventions. (iii) Limited information from the reports further obliged us to ground the overall content assessment exclusively on evaluators' assessment of the OECD DAC criteria. These limitations have to be kept in mind to put this meta-evaluation report correctly into perspective.

Given the lack of information on the whole population of bi-, multi- and multi-bilateral interventions the representativeness of this sample cannot be assessed.

The assessment of reports was conducted by different meta-evaluators and complex tools had to be filled out in an objective and unbiased way. To avoid the risk of subjective assessment or different understandings of specifications, huge efforts were undertaken during the development of the tools. Specifications determined the answering options as exact as possible to avoid biased results. Henceforth, many ratings have been limited to yes/no-answers and questions have rather been split-up until a yes/ no-answer was possible. This

helped making the selection for the meta-evaluation team as easy and reliable as possible.

Nevertheless, for some aspects more detailed assessments were considered of relevance for the user. Thus, also a four-scale rating was introduced whenever appropriate and feasible. Here, exact guidance for each category was written down, so that the meta-evaluation team was able to decide according to these determinations. The four-scale grades on sub-section and section level were all calculated from the results and consequently left no room for a biased assessment.

In the content analysis it was even more difficult to ensure that different meta-evaluators come to similar assessments and rate in a congruent way as contents needed to be interpreted correctly. In order to facilitate this assessment, we structured the content analysis along detailed questions to avoid arbitrary answers. In addition, the meta-evaluation team was not entitled to list main factors or reasons based on their own judgement, but they collected all items mentioned in the report which were in the end analysed at a general level. This reduced room for subjective assessments tremendously.

Whenever it comes to weightings throughout the aggregation process results are based on heavy expert judgements which are prone to subjectivity. To minimise subjectivity weights were discussed within the meta-evaluation team. However, appreciating this limitation, we refrain from overemphasising on overall scores and also present insights on section and sub-section levels. Thus, overall aggregates were only developed for the sake of linking different assessment tools and performing an economically efficient analysis for different subsample groups.

Another weakness of the overall evaluation report quality score consists in the failure of integrating further aspects like cross-cutting objectives, formal reporting aspects, quality assurance or composition of the evaluation team. Information presented in the chapter "further aspects" only grounds on selective reports because (i) aspects were not requested by the ToR (e.g. quality assurance), (ii) evaluators did not explicitly report on a matter (e.g. stakeholder validation), or (iii) interventions did not capture certain aspects (e.g. climate sustainability). Thus, missing values would have disturbed the analysis or required arbitrary decisions. Therefore, we perceived the exclusion of these aspects as the methodologically most robust alternative.

To cope with missing information regarding the treatment of some OECD-DAC criteria within the single reports, we decided to punish such reports which were obliged by the ToR to capture a OECD-DAC criterion but did not do so. Whenever the ToRs were not available (n=6) we refrained from such punishment. To abstain from punishing interventions for failures of the evaluators, punishment was limited to the evaluation report quality score.

For the overall score to assess the quality of Finnish development cooperation, the limitation centres around the exclusion of the assessments on cross-cutting objectives, aid effectiveness and triple C. Due to the severe lack of assessments, again exclusion was the only way to develop a consistent overall score.

Yes/no-answers and clearly defined four-step scales reduced room for biased assessements considerably.

Regarding the synthesis of lessons learnt and recommendations, the analysis faced the limitation that most of the lessons and recommendations drawn in the individual evaluation reports are tailored to the evaluated intervention and can only be red within the specific context to which they apply. Moreover, they highly depend on varying priority areas demarcated by the nature of the intervention, the ToR and the preferences of the evaluators. As mentioned above, in such cases lessons were no more true lessons learnt according to the OECD-DAC definition and were hence excluded from the analysis to avoid biased results. When it comes to recommendations, intervention-specificity is often a key characteristic. Hence, we applied the three-step approach for appropriate generalisation as presented above. Consequently, the analysis was limited to point out observable commonalities.

Cross-checking of randomly selected evaluation reports confirmed high consistency among team members. Internal and external workshops supported cross-validation of results.

Besides testing the instrument and its specifications, the pre-test facilitated alignment of answering behaviour and eliminated final arbitrary aspects inherent to the tools. Furthermore, a cross checking procedure of a randomly selected 10% of the evaluation reports confirmed high consistency among the individual team members. Additionally, internal and external validations workshops were conducted to cross-validate the results within the meta-evaluation team as well as with the MFA.

Given the nature of a desk study this assignment cannot be understood as a re-evaluation of single interventions.

However, it is important to understand that this assignment cannot be understood as a re-evaluation of single projects or programmes. This is the nature of a meta-evaluation desk study. Hence, we have to emphasise that results can be only interpreted at aggregated level. Please note that an interpretation of assessments at individual project or programme level is not possible due to methodological reasons.

On a different note, it has to be kept in mind that results and conclusions only hold for the fraction of Finland's development cooperation portfolio as they are based on 51 evaluation reports of bi-, multi-, and multi-bilateral interventions. Hence, they are not valid for other instruments of Finnish development cooperation. Moreover, it remains unclear to which extent the sample of evaluation reports at hand is representative for the whole bi-, multi- and multi-bilateral portfolio of Finnish development cooperation as further discussed in chapter 3.3. At least, discussions during inception and validation workshops suggest that the sample is perceived as an acceptable proxy.

3 CONTEXT ANALYSIS

3.1 Finland's development policies

Finland's development cooperation dates back to the 1960s when the government decided to start disbursing Official Development Assistance (ODA) to developing countries. In 1975, Finland became member of the OECD-DAC but it was only in 1996 when the MFA published its first policy guidance documents for the implementation of development assistance. The first Development Policy Programme was published in 2004. For a more detailed discussion on the history of Finland's development cooperation, see the report on the Evaluation of Finland's Development Cooperation Country Strategies and Country Strategy Modality by MFA of Finland (2016c).

After the first policy in 2004, the MFA has launched three different development policies; Development Policy Programme 2007-2011, Development Policy Programme 2012-2015, and the Government Report on Development Policy 2016-2019. The following provides an overview of the key characteristics of each policy and a brief discussion on how the policies have developed along the years.

The main objective of the **2007–2011 Finland's Development Policy Programme** – Towards a Sustainable and Just World Community is the *"eradication of poverty and ecologically sustainable development according to the Millennium Development Goals agreed jointly in the United Nations"* placing emphasis on climate and the environment (MFA of Finland, 2008). It also stresses *"crisis prevention and support for peace processes as an important element in promoting socially sustainable development"*. The policy outlines key cross-cutting themes to be mainstreamed in all development cooperation, which are:

- Promotion of the rights and the status of women and girls, and promotion of gender and social equality,
- Promotion of the rights of groups that are easily excluded; and the promotion of equal opportunities for participation, and
- Combating HIV/AIDS.

Table 1 summarises key goals, themes, cross-cutting objectives, geographic priorities and partner countries of Finland's Development Policy from 2007-2011.

Table 1: Summary of Finland's Development Policy 2007–2011

Development Policy 2007-2012
Key goals – Poverty eradication – Sustainable development.
Themes – Promoting ecologically, economically and socially sustainable development in accordance with Millennium Development Goals – Climate and environment – Respect for and promotion of human rights – Links between development, security and human rights.
Cross-cutting objectives – Gender equality, women and girls – Social equality and equal opportunities for participation – Combating of HIV/AIDS as a health and social problem.
Geographic priorities – Least developed countries.
Partner countries – Ethiopia – Kenya – Mozambique – Nepal – Nicaragua – Tanzania – Vietnam – Zambia.

Source: MFA of Finland 2017a.

In February **2012, Finland's Development Cooperation Policy** was revised adopting a new Human Rights-Based Approach (HRBA) to development, while the overarching goal remained *"eradication of extreme poverty and securing a life of human dignity for all people in accordance with the UN Millennium Development Goals"* (MFA of Finland, 2012a). The policy focused on five *"working methods"* of democratic ownership, accountability, openness, effectiveness, coherence and concentration (on least developed countries). The cross-cutting *"themes"* were upgraded to *"objectives"* including gender equality, reduction of inequality, and climate sustainability.

The priority areas of the policy were defined as:

- Democratic and accountable society that promotes human rights,
- An inclusive green economy that promotes employment,
- Sustainable management of natural resources and environmental protection, and
- Human development.

Table 2 summarises key goals, themes, cross-cutting objectives, geographic priorities and partner countries of Finland's Development Policy from 2012-2015.

Table 2: Summary of Finland's Development Policy 2012–2015

Development Policy 2012-2015
Key goals – Poverty reduction – Human rights and societal equity.
Themes – Democratic and accountable society – Inclusive green economy that promotes employment – Sustainable management of natural resources and environmental protection – Human development.
Cross-cutting objectives – Gender equality – Reduction of inequality – Climate sustainability.
Geographic priorities – Least developed countries – Fragile states.
Partner countries – Ethiopia – Kenya – Mozambique – Nepal –Tanzania – Vietnam – Zambia.

Source: MFA of Finland 2017a.

The Government published **Finland's current Development Policy in February 2016**, which is, in fact, a Government Report on Development Policy (MFA of Finland, 2016a). It is aligned with the 2030 Agenda for Sustainable Develop-

ment, the core goal remaining as the eradication of extreme poverty and reduction of poverty and inequality. The priority areas of the policy are:

- Enhancing the rights and status of women and girls
- Improving the economies of developing countries to ensure more jobs
- Livelihood opportunities and well-being
- Democratic and better-functioning societies
- Increased food security and better access to water and energy, and
- Sustainability of natural resources

Table 3 summarises key goals, themes, cross-cutting objectives, geographic priorities and partner countries of Finland's Development Policy from 2007-2011.

Table 3: Summary of Finland's Development Policy 2016-2019

Development Policy 2016-2019
Key goals – Poverty reduction – Reduction of inequality – Realisation of human rights – Support for the Sustainable Development Goals.
Themes – Rights of women and girls – Reinforcing economies to generate more jobs, livelihoods and well-being – Democratic and well-functioning societies – Food security, access to water and energy, and the sustainable use of natural resources.
Cross-cutting objectives – Gender equality – The rights of the most vulnerable – Climate change preparedness and mitigation.
Geographic priorities – Least developed countries, the most fragile states and those suffering from conflicts or climate and natural disasters.
Partner countries – Afghanistan – Ethiopia – Kenya – Mozambique – Myanmar – Nepal – Somalia – Tanzania – Zambia.

Source: MFA of Finland 2017a.

The key underlining philosophy of this current development policy continues being the HRBA to development cooperation. Another key characteristic is the emphasis on climate change, which is stated as being *"one of mankind's greatest challenges"*. The policy stipulates that all activities undertaken will be geared towards mitigating climate change and supporting climate change adaptation and preparedness.

3.2 Delivery of Finnish aid

Finland delivers aid through a number of different channels and modalities. According to the development policies (MFA of Finland 2008; 2012a; 2016a, see also MFA of Finland 2016c), these can be classified into seven main categories:

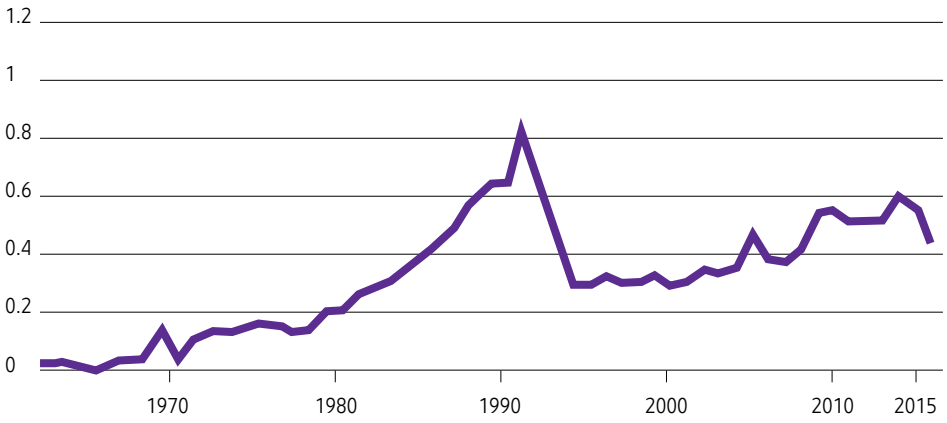
1. **Multilateral cooperation.** This instrument includes financing for a number of United Nations (UN) organisations, such as United Nations Population Fund (UNFPA), United Nations Children's Fund (UNICEF), United Nations Economic and Social Council (ECOSOC), United Nations Conference on Trade and Development (UNCTAD), and Food and Agriculture Organization of the United Nations (FAO), among others. Finland channels multilateral aid also through some of the main development banks and financing institutions, such as The World Bank Group (WBG), the African Development

Bank (AfDB), the Asian Development Bank (ADB), the Inter-American Development Bank (IDB), and the Nordic Development Fund (NDF).

2. **Bilateral and regional cooperation.** An important part of Finland's development cooperation is channelled through bilateral projects and programmes. The country-level interventions are guided by specific country strategies since 2012. These initiatives are supported by regional (multi-country) initiatives that are often channelled through international organisations or financing institutions. Similarly, the European Union (EU) is an important partner for Finland in development cooperation, e.g. through providing funding to EuropeAid.
3. **Humanitarian aid.** Finland's humanitarian aid is delivered mainly in collaboration with UN organisations, the International Red Cross and the Red Crescent, Finland's Red Cross and Finnish Church Aid.
4. **Cooperation with Civil Society Organizations (CSOs).** Support for CSOs was increased during the policy 2012–2015, but declined again in 2016. Overall, CSO cooperation has constituted an important part of Finland's development cooperation portfolio in the past years. The types of modalities include programme-based support (PBS), project-based support, direct support to CSOs in developing countries (Fund for Local Cooperation), including some other types of support such as travel and project formulation grants. Furthermore, Finland has provided long-term programme-based support for CSO umbrella organisations such as Kepa and Kehys ry.
5. **Private sector cooperation.** Finland has supported private sector development directly in developing countries and by encouraging collaboration between Finnish companies and their partners in the target countries. The Finnish Fund for Industrial Cooperation (Finnfund) and the Finnish Business Partnership Programme (Finnpartnership) as well as the Business with Impact (BEAM) programme, are among the key mechanisms to stimulate private sector activities in developing countries.
6. **Cooperation with higher education institutes and research on development policy.** International mobility of students and teachers is in the centre of the collaboration with higher education institutes in the field of development cooperation. The Higher Education Institutions Institutional Cooperation Instrument (HEI ICI) instrument supports capacity strengthening of higher education institutes in developing countries.
7. **Climate finance.** Finland channels climate finance through international mechanisms such as the Global Environment Facility (GEF) and the Green Climate Fund (CGF). Part of the support for Finnfund is also classified as climate finance.
8. **International non-governmental organizations (INGOs).** The MFA can also fund international non-governmental organisations (INGOs) for activities that are in line with Finland's development policy priorities and goals, and when the interventions are complementary to the other types of support.
9. **Finnish civil society organisations' communications and global education projects.** The funds are meant to be used in Finland for development communications and global education in the context of development cooperation or development policy.

The volumes of Finnish development aid increased steadily from the initial years in the 1970s until the economic depression in 1990s (OECD, 2017). Figure 1 shows, in the new millennium, the share of Finnish ODA as percentage of Gross National Income (GNI) started growing again slowly until 2015 when the government decided to implement important cuts into the development cooperation budget.

Figure 1: Finnish ODA as percent of Gross National Income (GNI)



Source: OECD 2017a.

3.3 Evaluation reports in light of the Finnish development context

This meta-evaluation covers only evaluation assignments carried out between September 2015 and August 2017. Thus, all evaluations are implemented under the framework of the Development Evaluation Norm established in 2015, which provides the definition and the legal basis for evaluation of development policy and cooperation.

Development evaluation serves a dual purpose in the MFA, accountability and organisation-wide learning. In terms of accountability, evaluation of relevance, efficiency, effectiveness and impact is a responsibility set for the MFA by the State Budget Act and State Budget Decree. The learning aspect aims at constant improvement of the quality of development cooperation through the provision of independent and impartial knowledge on the activities (MFA of Finland, 2015).

Evaluations carried out by the MFA are also guided by the Evaluation Manual (MFA of Finland, 2013), which sets out the key contents and quality standards of both decentralised and centralised evaluations. The Manual for Bilateral Cooperation provides detailed guidance on how development partners should take into account Result Based Management (RBM) and the HRBA to development during various phases of the project cycle. The first version of the Manual was published in 2012 (MFA of Finland, 2012b). It was updated in 2016 (MFA of Finland, 2016b). Additionally, a report template with detailed information on the content of the different sections is handed out to the evaluators to write their report accordingly.

Developemt evaluation serves a dual purpose in MFA: accountability and organisation-wide learning.

The implementation period of the interventions which underlie this meta-evaluation falls between 2005 and 2017, apart from three exceptions (one project started in 2001 and two in 2004). In other words, nearly all interventions started in 2005 or later when the first Finnish Development Policy had already been endorsed. In fact, more than half of the interventions (57%, 29 out of 51) fall under the validity of the 2007-2012 policy (classifying the interventions based on their start date). For 13 interventions the start date was not accessible to the meta-evaluation team. On the other hand, 80% of the interventions (41) were under implementation in 2014, which links these interventions to the Development Policy Programme 2012-2015.

However, these linkages have to be taken with care as some interventions built on previous phases or their design has taken place long time before the implementation has started. As mentioned earlier, the types of interventions that are covered by this meta-evaluation only include bilateral and multilateral and in some cases so-called multi-bi interventions.

The MFA has recently finalised a large evaluation on the development cooperation with CSOs (MFA of Finland, 2017b). Similarly, the other instruments of Finnish development cooperation are evaluated mainly through centralised evaluations commissioned directly by EVA-11, and not by the regional units as it is the case for project and programme evaluations. Therefore, the **conclusions and emerging issues that will be identified as a result of the meta-evaluation hold only for a fraction of Finland's development cooperation** portfolio.

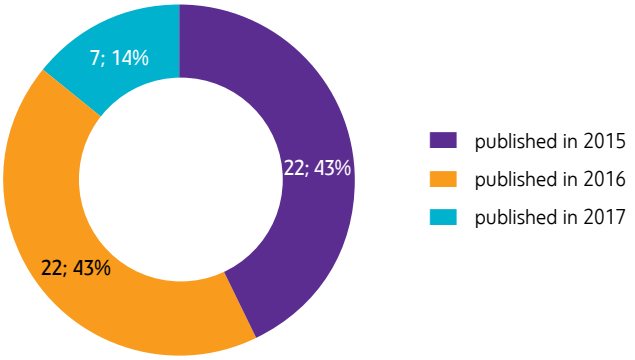
In addition, the interventions that have been evaluated by the regional units do not necessarily represent the whole portfolio of bi-, multi-, and multi-bilateral interventions. According to the MFA, there has been no clear sampling strategy developed to select a representative set of interventions to undergo mid-term or final evaluation at a specific point in time. Moreover, there is no systematic list of all bi-, multi-, and multi-bilateral interventions with their key characteristics (e.g. geographical scope, budget range, nature of the intervention, implementation dates etc.) to test ex-post the representativeness of the sample of evaluation reports at hand. Therefore, the findings of the analysis have to be understood in this limited context. However, according to MFA staff the available sample of evaluation reports is perceived as nearly complete and fairly illustrative of the whole portfolio of bi-, multi-, and multi-bilateral interventions of Finnish development cooperation.

MFA has not yet developed a sampling strategy to select a representative set of interventions to be evaluated at a specific point in time.

To respond to the first evaluation question EQ1 *"How can MFA's decentralised evaluation portfolio be described…?"*, the sample of evaluation reports is presented according to different characteristics of the evaluations and the underlying interventions.

Figure 2 displays, out of overall 51 evaluation reports, 22 (43%) were published in 2015, another 22 (43%) in 2016 and the remaining 7 (14%) in 2017. If not stated otherwise the sample size referred to is the total sample (51 in the meta-evaluation, 45 for the ToRs and 50 for summative analysis). When we refer to a different sample size we include it in brackets (eg. 20 out of 44, 45%) or in the beginning of the paragraph. Furthermore, for sample sizes <40 we do not provide percentages to avoid generalisations as the statistical explanatory power is limited.

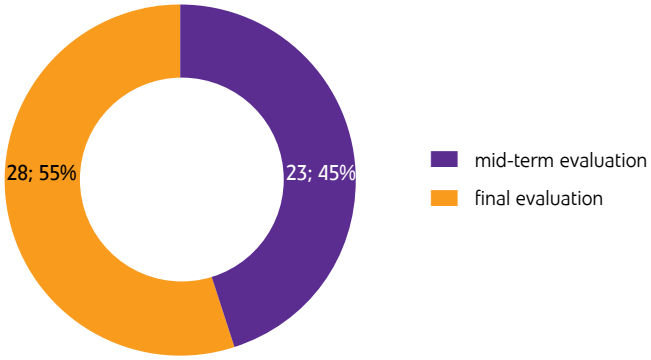
Figure 2: Year of publication of the evaluation report (n=51)



Source: own statistics based on analysis of reports

Figure 3 shows that the sample consists of 23 (45%) mid-term evaluations or mid-term reviews and 28 (55%) final evaluations.

Figure 3: Nature of the evaluation (n=51)

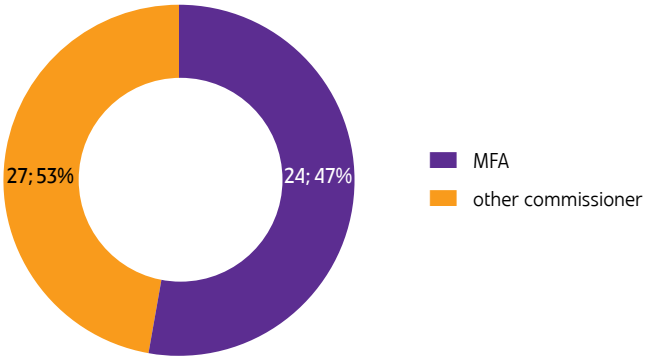


Source: own statistics based on analysis of reports

Among reports from 2016, the share of final evaluations (15 out of 22) is highest, followed by an almost equal share in 2015 (12 mid-term vs. 10 final), and a clearly lower share for 2017 (5 mid-term vs. 2 final). Given these unequal shares sub-group comparisons according to the year of publication are highly biased by the nature of the evaluation (mid-term vs. final) and are hence not conducted throughout further analysis.

With regards to the commissioner of the evaluation, Figure 4 displays roughly half of the evaluations (24, 47%) that were commissioned by the MFA.

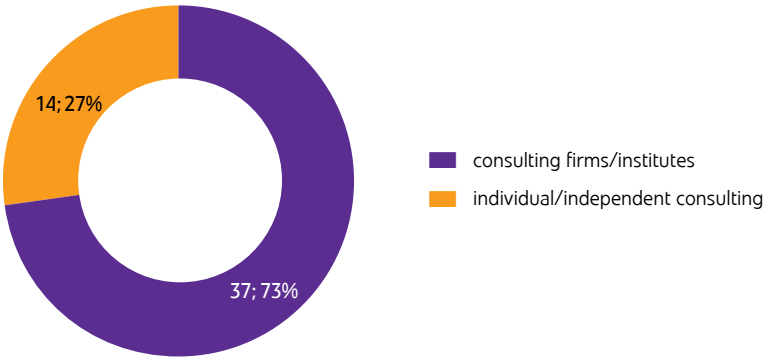
Figure 4: Commissioner of the evaluations (n=51)



Source: own statistics based on analysis of reports

Figure 5 shows that a bit less than three quarters of the evaluations (37, 73%) were implemented by evaluation teams from consulting firms or institutes, whereas 14 (27%) were conducted by individuals or independent consultants. By this we mean that only a single person was hired to implement the evaluation or that a team of two independent consultants was recruited by other commissioners. The MFA does not contract individuals without institutional affiliation.

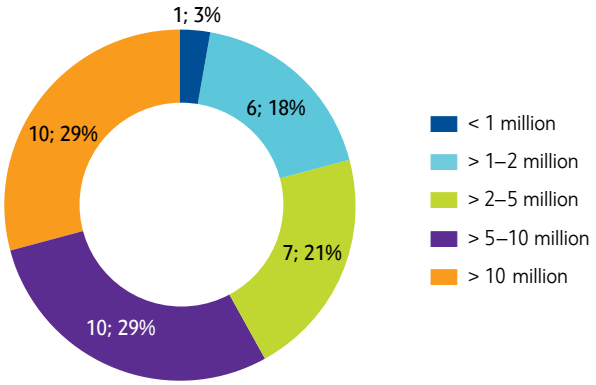
Figure 5: Implementer of the evaluation (n=51)



Source: own statistics based on analysis of reports

Information on the interventions' budget from Finland is only available for 34 out of 51 reports. It ranges from 0.4 million up to 22 million Euro (with a mean of roughly 7 million and a median of roughly 6 million). Figure 6 displays that interventions with a budget of less than one million Euro are rather the exception for bi- or multilateral interventions in Finnish development cooperation, while a fair amount of interventions is filed in all other budget ranges (i.e. >1–2, >2–5, >5–10 and > 10 millions).

Figure 6: Finland's budget of the intervention (n=38)

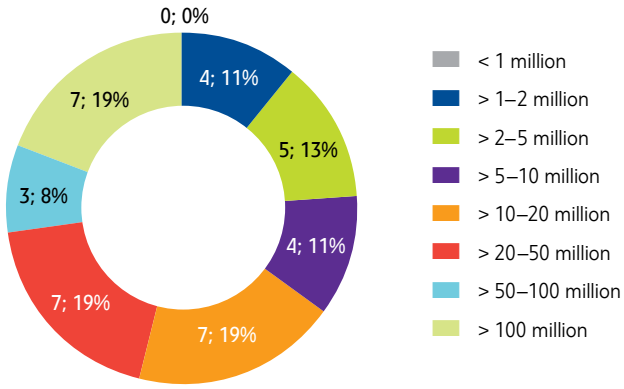


Source: own statistics based on analysis of reports

Our statistics have shown that the lower Finland's budget for an intervention, the higher the possibility that the evaluation has been conducted by an independent evaluator. The mean of Finland's project budget (These figures are available only for 34 interventions in the sample of evaluation reports.) for evaluations conducted by individual/independent consultants is almost half (5,564,218 €) of those conducted by other evaluation entities (9,249,545 €). Similarly, it turned out that Finland's budget for the intervention is significantly higher, when MFA is the commissioner with a similar difference as described above. For further details please refer to Annex 13).

Figure 7 shows the overall budget for the interventions for 37 cases. It ranges from roughly one million up to roughly 750 million Euro (with a mean about 77 million and a median of roughly 13 million) pointing to the fact that Finland is contributing to a number of multilateral large-scale efforts. All budget ranges (i.e. <1-2, >2-5, >5-10, >10-20, >20-50, >50-100, and > 100 million, with the exception of a budget below one million) characterise some of the interventions under consideration.

Figure 7: Overall budget for interventions (n=37)

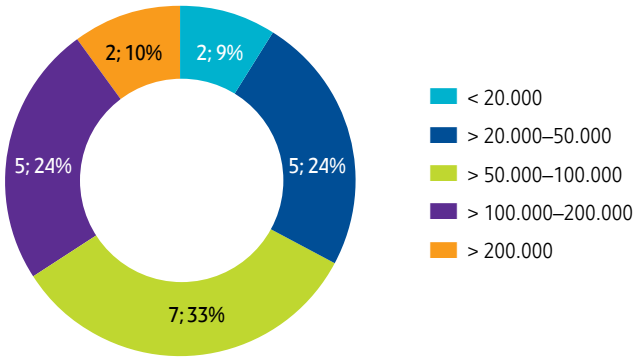


Source: own statistics based on analysis of reports

This wide range of intervention budgets suggests that evaluation budgets vary accordingly as in line with good evaluation practice between 1 and 2% of the overall intervention budget should be dedicated to monitoring and evaluation.

Figure 8 discloses that the net evaluation budget is only available for 21 interventions. Thus, figures have to be taken with care. The net evaluation budget ranges from 10.000 up to 340.000 Euro (with a mean of roughly 97.000 and a median of 80.000). The majority of the interventions is spread over a range from 50.000-100.000 Euro (7 out of 21), while a considerable number is found in the next lower (20.000-50.000 Euro) and the next higher (100.000-200.000 Euro) budget ranges (5 out of 21 each). Net evaluation budgets of less than 20.000 Euro or more than 200.000 Euro are rather rare (2 out of 21 each).

Figure 8: Net evaluation budget of the interventions (n=21)



Source: own statistics based on analysis of reports

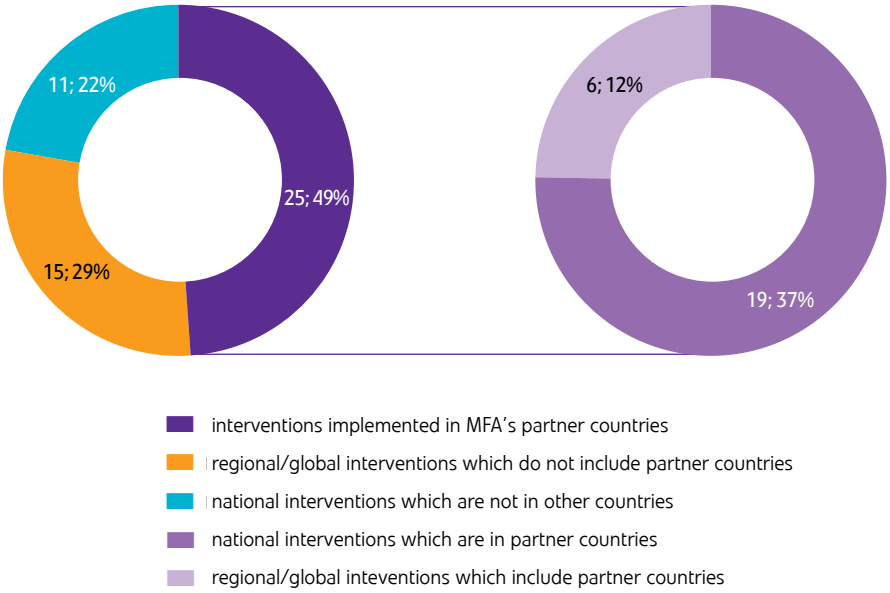
A significant positive relation has been found between the net evaluation budget and the overall budget of the intervention. As expected, the higher the overall budget of the intervention, the higher the net evaluation budget. Not surprisingly, it turned further out that evaluations with a smaller net evaluation budget are significantly more likely conducted by individual/independent consultants (see Table 16 in Annex 13). The evaluation budget of individual/independent consultants is roughly a quarter compared to those conducted by companies and institutes. This needs to be borne in mind when conclusions are drawn in later chapters.

Regarding the geographical scope, 30 interventions (58%) are operating at the national or sub-national level, while 21 (42%) are either regional or global level interventions. Logical reasoning suggests that this is highly correlated to the nature of the intervention. Thus, interventions at the (sub-)national level are assumed to be rather bilateral, while interventions at the regional or global level seem to be rather multilateral. However, as the nature of the intervention does often not become clear from the evaluation reports at hand, this aspect is not further assessed. Thus, whenever sub-group comparisons according to the geographical scope are performed throughout the analysis, results can be taken as proxy for the nature of the intervention.

Beyond the differentiation of national vs. regional/global interventions, it is interesting to note that half of the interventions (25, 49%) are implemented in MFA's partner countries (see figure 9). With seven interventions there is a strong focus on Nepal within our sample, followed by Ethiopia (4), Vietnam (3), Zambia (3), Mozambique (2) and Tanzania (1). The remaining five interventions address multiple partner countries at a time.

Figure 9 further shows that out of those interventions six are at the regional/global level and 19 are at the national level. In turn, this means, that eleven interventions at national level (22% of the whole sample) are not directed according to Finland's geographical priority area as defined in the three Finnish Development Policies presented above. For the remaining 15 regional/global interventions, it is, however, not clear whether they did not include MFA's partner countries or whether the evaluation reports do not disclose that the intervention also addresses one or more of MFA's partner countries.

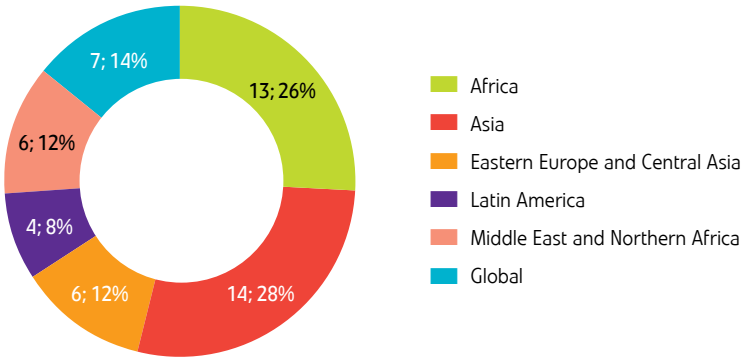
Figure 9: Geographical scope of interventions in partner countries (n=51)



Source: own statistics based on analysis of reports

With regard to the regional distribution, Figure 10 illustrates that the majority of the interventions focusses on Asia (14, 28%) and Africa (13, 26%). Several interventions (7, 14%) are implemented in multiple regions. Clearly less interventions are implemented in Eastern Europe and Central Asia and the Middle East and Northern Africa (each 6, 12%), with Latin America only targeted by four interventions (8%).

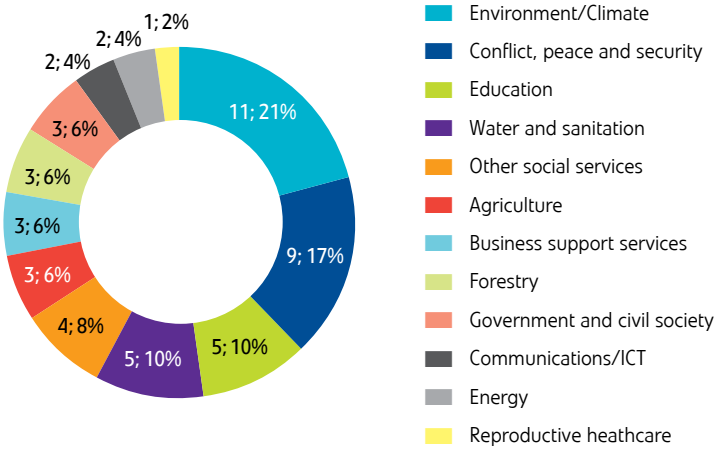
Figure 10: Regional distribution of interventions (n=43)



Source: own statistics based on analysis of reports

Finally, Figure 11 discloses that the four sectors (i) environment and climate, (ii) conflict prevention, resolution, peace and security, (iii) education and (iv) water and sanitation contain 58% of the interventions, while the remaining 42% are distributed over as many as eight sectors.

Figure 11: Sectorial distribution of interventions (n=51)



Source: own statistics based on analysis of reports

Thus, the sectorial distribution within our sample is rather fragmented. This hampers sub-group comparisons seriously, hence, in the remainder we only use the four prominent sectors to control for sectoral specificities when performing disaggregated analyses and further take agriculture and forestry together as one sector and government and civil society as well as other social services as another joint sector.

4 FINDINGS OF THE META-EVALUATION

This chapter presents the findings of the quality assessment of 51 evaluation reports. First, in chapter 4.1 we respond to EQ3 on the quality of the ToR.Then, we analyse different sections of the evaluation reports as follows: Chapter 4.2 on the quality of introductions and context analyses, chapter 4.3 on the appropriateness of evaluation methodologies, chapter 4.4on how evaluation findings were obtained and the coverage of the OECD DAC criteria, chapter 4.5 on the quality of conclusions and recommendations, and chapter 4.6 on further aspects like cross-cutting objectives, validation, and quality assurance. In chapter 4.7 the quality of the executive summaries is assessed. Accordingly, these sections provide answers to evaluation questions EQ2 (on the quality of MFA's decentralised evaluation reports), EQ4 (on quality classified by different aspects like evaluation type or implementer) and EQ5 (on differences of quality between MFA-commissioned evaluations vs. evaluations commissioned by other institutions).

Finally, chapter 4.8 provides insights on linkages between the quality of the ToR and the quality of the reports, and hence answers to EQ6 (on systematic patterns). EQ7 (on the reliability of the decentralised evaluation reports and EQ8 (on gaps regarding MFA's evaluation capacity) and EQ9 (on recommendation by the meta-evaluation team) are answered in the concluding chapter 6 and in the recommendations' chapter 7which go beyond insights of the quality assessment and also draws on findings from the content assessment provided in chapter 5.

4.1 Quality of underlying ToRs

Highlights of the chapter addressing EQ 3:

- The overall quality of ToRs is satisfactory for 60% of the ToRs.
- All ToRs could be improved in some ways and more than one third are assessed as in need of significant improvement.
- The most valuable information provided concerns the evaluation description, the evaluation questions, and the evaluation criteria.
- The least valuable information provided concerns the methodology, the evaluation process, quality assurance and on the cross-cutting objectives.
- ToRs by the MFA are in general of higher quality than those of other commissioners (on a scale from 1-4: 2.64 vs. 2.37).

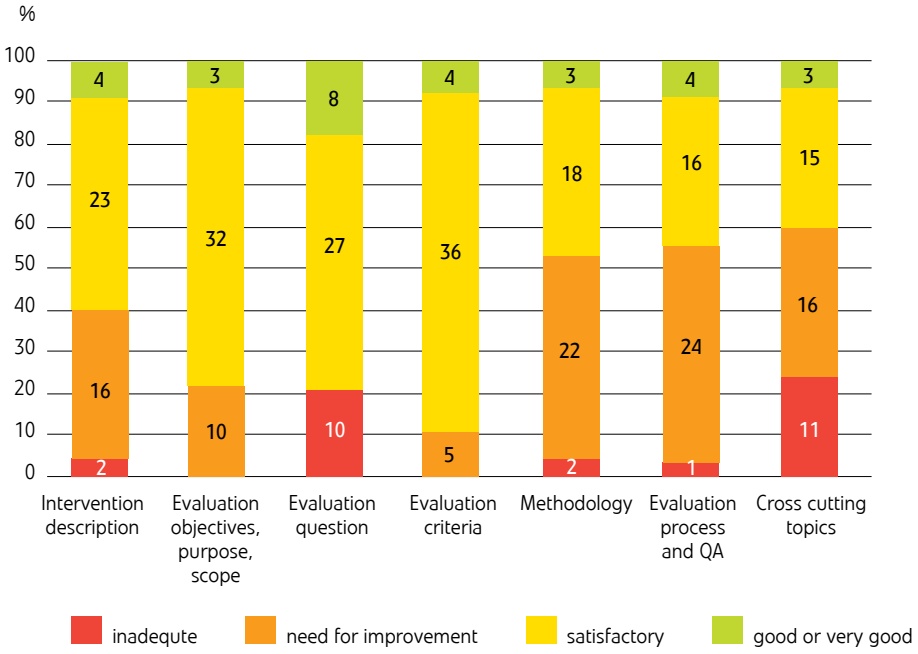
The ToRs determine how the evaluation should be implemented. They serve also as a guideline for formal and structural aspects of the report. In this regard, the quality of the ToRs is assessed regarding (i) the description of the intervention,

Overall, the sections on methodology, evaluation process and quality assurance, and cross-cutting objectives are of lower quality than other sections of the ToR.

Nearly a quarter of the ToRs is assessed inadequately regarding specifying evaluation questions and cross-cutting objectives.

(ii) the evaluation objectives, purpose and scope, (iii) evaluation questions, (iv) evaluation criteria, (v) methodology, (vi) evaluation process and quality assurance, and (vii) cross-cutting objectives as illustrated in Figure 12. In total, 45 ToRs from 51 reports were available for the analysis.

Figure 12: ToR assessments (n=45)



Source: own statistics based on analysis of reports

Most ToRs provided a "good or very good" (4, 9%) or "satisfactory" (23, 51%) **description of the intervention** which comprises the **context of the intervention and objectives, strategies and implementation of the intervention**. Several ToRs are assessed of lower quality in this regard: 16 (36%) are in "need for improvement" and 2 (4%) are inadequate.

The **evaluation** itself is much better described regarding its **objectives, purpose and scope**; 32 ToRs (71%) are rated as "satisfactory" and three ToRs (7%) as "good or very good" while 10 ToRs (22%) are in "need for improvement". While **rationale and purpose** are presented "satisfactory" by 32 ToRs (71%) and "good or very good" by 10 ToRs (22%), the **scope** is rated for 19 ToRs (42%) as in "need for improvement" or "inadequate".

The sections on **evaluation questions** comprise the **adjustment the questions to the needs of the commissioner** and **limiting the number of questions**. More than half of the ToRs are rated as satisfactory (27, 60%) and another 8 (18%) are assessed as "good or very good". Ten ToRs (22%) are "in need for improvement". The vast majority of ToRs that formulated evaluation questions adapted them to the needs of the interventions (34 out of 38, 89%). The larger problem is the number of evaluation questions; about one quarter of the ToRs (9 out of 38, 24%) limit themselves to 12 questions as requested. Three quarters (29, 76%) have a much higher total, in single cases up to 70 questions.

The section on the **evaluation criteria** is rated quite well, 36 ToRs (80%) are providing evaluation criteria to a "satisfactory" extent and four ToRs (9%) are

assessed as "good or very good". Only five ToRs are "in need for improvement". Exceptionally, the OECD DAC criteria "relevance", and "effectiveness" are left out of the ToRs (each once). Neglecting "impact" or "sustainability" is more common (8 and 4 ToRs, respectively). The **"Triple C" criteria** of EU coherence (13, 29%), complementarity (8, 18%) and coordination (8, 18%) or **aid effectiveness** (7, 16%) are requested to a much smaller extent.

Overall, the **methodology** section is of lower quality than other sections of the ToR. The assessment is mixed: About half of the ToRs (22, 49%) are rated with "need for improvement" and two (4%) as inadequate whereas 18 ToRs (40%) have been assessed as "satisfactory" and only three (7%) as "good or very good". The sub-sections include whether the commissioners requested the usage of **qualitative and quantitative methods** (25, 56%), **triangulation of sources** (17, 38%) or a **disaggregated analysis** (8, 18%). Furthermore, it was analysed if the ToRs specified **available materials** (23, 53%), **envisaged data collection techniques** (34, 76%) or **envisaged data analysis techniques** (8, 18%).

The description of the **evaluation process** has an equally large room for improvement given the mixed results: 24 ToRs (53%) assessed as in "need for improvement" and one ToR (2%) as "inadequate", 16 ToRs (36%) were identified as "satisfactory" and 4 ToRs (9%) as "good or very good". A look at underlying sub-sections reveals, the **deliverables** are described by all but one ToR (98%) as well as the **phases of the evaluation process** are almost always illustrated (41, 91%). **Information on the approximate duration of activities, place of work, roles and responsibilities** within the evaluation is often missing (14, 32%; 22, 49% and 15, 33% respectively). Only 17 ToRs (38%) refer to the kind of quality assurance desired. In contrast, 28 ToRs (62%) do not refer to **quality assurance** at all.

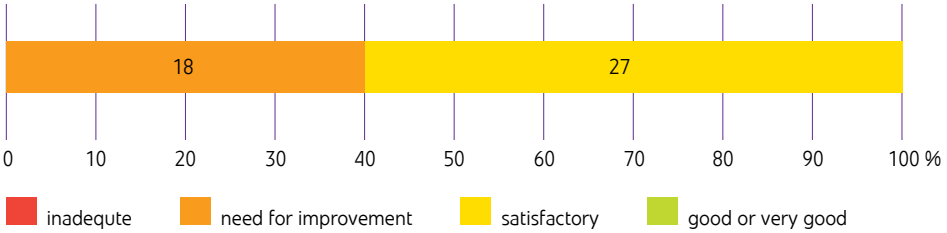
Regarding **cross-cutting objectives (gender equality, reduction of inequality, combat against HIV/Aids, climate sustainability and HRBA)** the results of the analysis are quite diverse. While three ToRs (7%) integrate the cross-cuttings in a "good or very good" manner and 15 ToRs (33%) integrate them to a "satisfactory" extent, eleven ToRs (24%) do not integrate them at all and 16 ToRs (36%) only incomplete. Gender equality is the objective integrated in two thirds of the ToRs (30, 67%). Other objectives are requested by less than half of the ToRs (reduction of inequality: 18, 40%; climate sustainability: 17, 38%; HRBA: 21, 47%). Combating HIV/AIDS has been requested only in four cases (9%).On a different note, we attempt to check on the **feasibility of the ToRs** taking the scope requested and the number of working days foreseen as well as the evaluation budget into consideration. The budget was often not provided, hence **budgetary feasibility** could be only assessed for 20 reports. In six cases (30%) it was regarded as too low for the envisaged tasks in the evaluation. The working days or time period was more often given, for 42 ToRs the **feasibility in terms of time resources** has been assessed. For 12 reports out of 42 (29%) the working days or time period provided to implement the evaluation accordingly was judged as inadequate.

To allow an assessment of the **overall quality of ToRs** the quality of the different sections (intervention, evaluation, evaluation questions, evaluation criteria, methodology, evaluation process and cross-cutting objectives) were aggregated. Feasibility was not integrated due to serious limitations on data availability.

The aggregate on the overall quality of the ToRs disclose that no ToRs are assessed "good and very good" and neither as "inadequate". Most ToRs are assessed as in "satisfactory" (27, 60%) followed by those assessed as "in need for improvement" (18, 40%)

Figure 13: Overall quality of ToRs

Source: own statistics based on analysis of reports



The overall quality of the ToRs from MFA-commissioned evaluations is statistically significantly on average higher. The mean for the overall ToR rating of MFA commissioned evaluations is at 2.64 and for evaluation not commissioned by MFA at 2.37 (see Table 18 in Annex13). This refers especially to the assessment of the description of the intervention (means: 2.81 vs. 2.26), evaluation criteria (means: 3.00 vs. 2.67) and cross-cutting objectives (means: 2.50 vs. 1.67). For methodology in turn, evaluations MFA commissioned are on average lower (means: 2.03 vs. 2.50).

The overall ToR assessment is better for MFA-commissioned evaluations than for evaluations by other commissioners.

The overall quality of ToR for independent consultants are of lower quality than for consulting firms.

Further, overall quality of the ToRs for individual/independent consultants are statistically significantly of lower quality than ToRs for consulting firms or institutes (means: 2.37 vs. 2.56). Thus, inferior ToRs are also a plausible explanation for the weaker quality delivered by individual/independent consults as assessed further above.

4.2 Quality of introductions and context analyses

Highlights of the chapter addressing EQs 2 and 7:

- The quality of introductions is adequate for 43 out of 51 reports (84%).
- However, information on the scope of the evaluations and the reporting of the evaluation questions is missing in more than one third of the reports. Missing evaluation questions are particularly alarming as they frame the assignment and reflect the commissioner's demands.
- The quality of the context analysis lags behind as more than one third of the reports are in need of improvement.
- Sometimes information related to the context analysis is integrated within the introduction section or the relevance chapter.

Regarding the introductions of the evaluation reports, the provision of six different aspects were analysed: (i) the rationale and the purpose of the evaluation, (ii) the objectives of the evaluation, (iii) the evaluation object, (iv) the scope of the evaluation, (v) the evaluation questions and (vi) the results of previous evaluations if any. As described above, each aspect consists of sub-aspects, which are provided in detail in Annex 7.

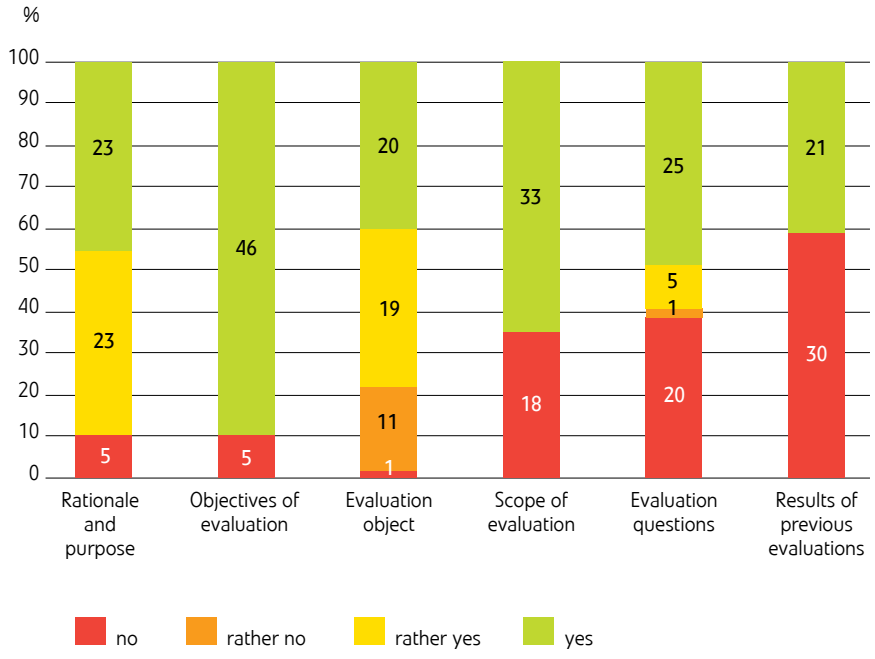
Figure 14 displays results that can be summarised as follows: A bit less than half of the evaluation reports (23 out of 51, 45%) describe the **rationale and the intended user** of the evaluation, another 23 reports only contain the purpose but not the user and five reports (18%) do not provide any of this information.

In contrast, in almost all reports (46, 90%) the **objectives of the evaluation** are reported. Regarding the **evaluation object** the picture is equally positive. On a four-step scale, an aggregate for capturing sub-aspects like evaluation budgets, time resources and detailed objectives reveals a "good or very good" assessment for a bit more than one thirds of the reports 20 (39%) and a "satisfactory" assessment for about another third (19, 37%). Over 70% of the evaluators provide relatively detailed information on the intervention, especially on its objectives and the time period of the intervention.

Figure 14 shows in turn that the description of the **scope of the evaluation** and the evaluation questions are assessed of lower quality as this information is often lacking. Alarmingly, more than one third of the reports (20, 39%) do not report or reference to **evaluation questions**, and hence leave their work without proper contextualisation to their assignment.

However, 21 reports (42%) refer to previous evaluations. Whether this figure is large or small cannot be assessed as the existence of **previous evaluations** is unknown to the meta-evaluation team. Not surprisingly, final evaluations (16 out of 28) refer much more often to previous evaluations than mid-term evaluations (5 out of 23): the probability of an existing midterm evaluation is higher. Still, only occasionally reports are directly building upon the formerly obtained results and are using them for their analysis.

Figure 14: Contents of introduction (n=51)



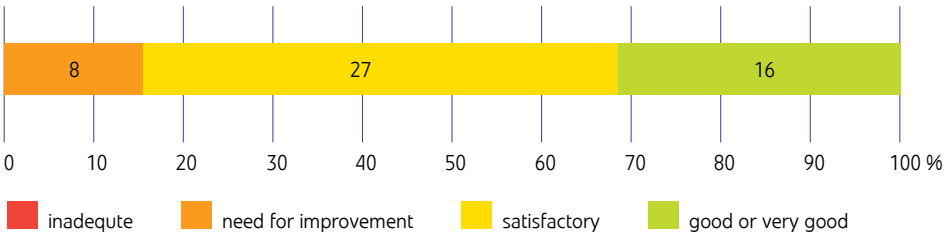
Source: own statistics based on analysis of reports

More than one third of the reports lack evaluation questions. This is particularly alarming as evaluation questions frame the assignment.

Overall, the quality of the introductions is quite good.

Overall, regarding the provision of general information on the intervention and the evaluation itself, reports perform quite well. The aggregated assessment at section level reveals that no report is rated as inadequate and over 80% of reports are assessed as "good or very good" (16) and "satisfactory" (27) whereas eight reports are rated as in "need for improvement".

Figure 15: Overall ratings of introductions (n=51)

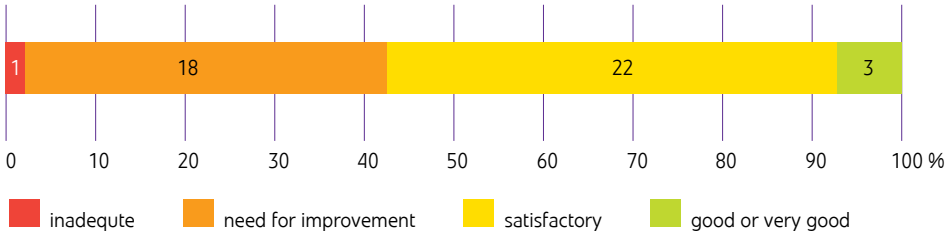


Source: own statistics based on analysis of reports

The assessment of the context analyses paints a more negative picture. Seven reports do not provide a context analysis at all. Figure 16 shows that out of the 44 reports providing a context analysis, half (22 out of 44, 50%) describe the context of the intervention to a "satisfactory" degree and more than one third (18 out of 44, 40%) with "need for improvement".

The quality of the context analysis lacks behind.

Figure 16: Overall rating of context analysis (n=44)

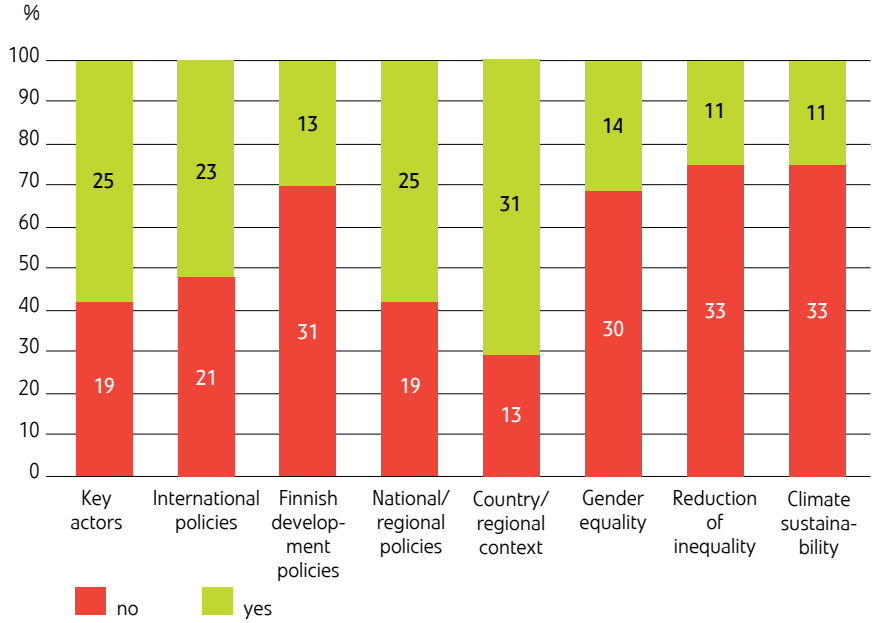


Source: own statistics based on analysis of reports

This overall rating is based on the eight different sub-aspects shown in the figure below as well as on the linkage between context analysis and the intervention. The individual aspects are (i) key actors, (ii) international policies/strategies, (iii) Finnish development policies/strategies, (iv) national/regional policies, (v) country/regional context, (vi) gender equality, (vii) reduction of inequality and (viii) climate sustainability.

Especially the **cross-cutting objectives** are mostly not covered by the context analysis. Gender equality is referred to in 14 reports and reduction of inequality and climate sustainability each in 11 reports. But reference to **Finnish development policy** is only provided in the context analyses in 13 reports. Evaluation reports that were not commissioned by the MFA do not often pay attention to Finnish policies (21 out of 23). Surprisingly also half of the MFA-commissioned evaluations do not refer to Finnish policies in their context analyses (10 out of 21).

Figure 17: Contents of context analysis (n=44)



Source: own statistics based on analysis of reports

In contrast, the **socio-economic, political and/or cultural country context** is discussed in more than two thirds of the reports featuring a context analysis (31 out of 44, 70%). Less often, for about half of those reports, **international policies** (23 out of 44, 52%) and the **national policies** (25 out of 44, 57%) are discussed. Furthermore, the content analysis is mostly put into perspective given the intervention (yes: 23 out of 44, 52%, rather yes: 15 out of 44, 34%). However, in six reports the linkage is not or not always obvious.

Evaluators locate the context analysis at different places in their reports, sometimes after the introduction following commissioners' standards and sometimes after the methodology section as requested by the MFA. At times it is integrated in the introduction section and not provided separately.

References to Finnish development policies and cross-cutting objective are made in less than one third of the reports.

4.3 Quality of evaluation methodologies

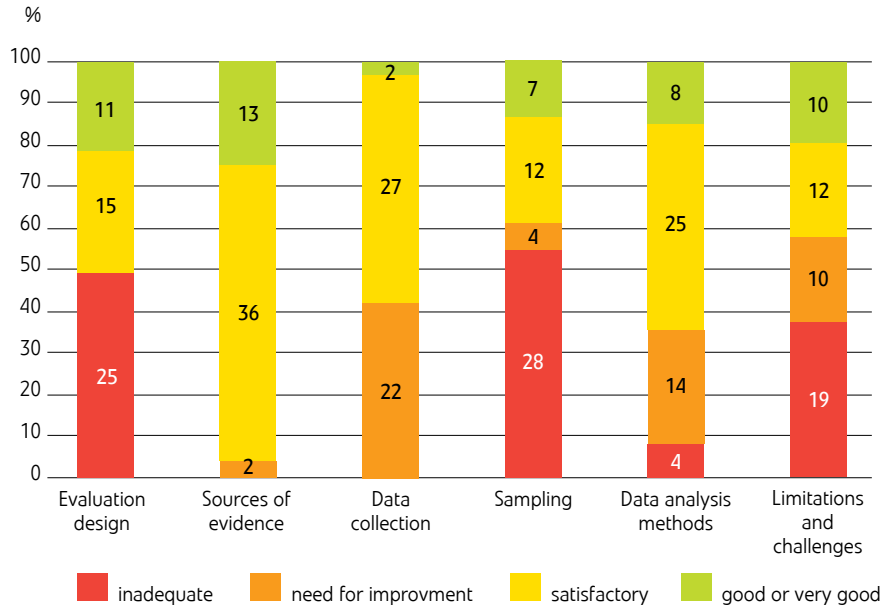
Highlights of the chapter addressing EQs 2 and 7:

- The assessment of the evaluation methodologies reveals that about half of the 51 reports are in need of improvement.
- Evaluators appropriately present and treat sources of evidence in nearly all reports and data collection methods in two thirds of the reports.
- The selection and presentation of evaluation design, sampling strategies and resulting limitations is unclear in about half of the reports.
- In general, the methodologies applied by individual/independent consultants are of lower quality than those by consulting firms or institutes (on a scale from 1-4: 2.19 vs. 2.57). Possible causes are threefold: lack of capacity, lower evaluation budgets and lower quality of the ToR.
- No other significant differences between the quality of MFA-commissioned reports and those by other commissioners or over time could be found in the disaggregated analysis.
- As evaluation methodologies lay the foundation for findings, conclusions and recommendations, this assessment is of particular importance.

Sources of evidence and data collection methods are presented appropriately, whereas elaborations on the evaluation design, the sampling strategy and underlying limitations are often missing.

The analysis of the evaluation methodologies comprises six aspects as shown by Figure 18: (i) evaluation design, (ii) sources of evidences, (iii) data collection methods, (iv) sample, (v) data analysis methods and (vi) limitations. Similarly, as above each aspect consists of sub-aspects, which are presented in detail in Annex 7.

Figure 18: Description and appropriateness of methods (n=51)



Source: own statistics based on analysis of reports

The analysis shows that only half of the reports (26 out of 51, 51%) **describe the evaluation design** and/or at least the general approach to the assignment, whereas 25 reports (49%) do not describe design at all. The evaluation approach (e.g. participatory) is more often provided than the design (e.g. comparison groups) (23, 45% vs. 14, 28%).

The sources of evidence are most frequently presented within the methodology chapters or sections. Thus, almost one third of the reports are assessed as "good or very good" (13, 26%) and 71% of the reports (36) are rated as "satisfactory". A closer look at sub-aspects shows that "need for improvement", if any, centres around failure to use (i) the intervention's M&E data, (ii) additional literature going beyond the intervention's documentation, and (iii) including representatives of the institutional environment as interview partners. However, the latter has to be treated with caution as from the reports it was often not clear whether interview partners were directly involved in the implementation, benefited from the intervention or if they belonged to the institutional environment.

The mixture of information sources is in general assessed as quite adequate. For a bit more than two thirds of the reports (36, 71%) three or more data sources were used with a mixture of primary and secondary data. Although for the remaining third (14, 27%) evaluators accessed also three or more data sources, they did not draw on both secondary and primary data. Only one report stands out for the questionable practice of disclosing the use of only two different data sources. Concerning the transparency of underlying information sources,

reports are performing quite well. A list of persons interviewed is provided in the great majority of reports (44, 86%) as is a list of the documents consulted (42, 82%).

Data collection methods were specified in the majority of the reports. In about half of the reports (27, 53%) their provision is assessed as "satisfactory" and in another two reports (4%) as "good or very good". No report provides inadequate information in this regard and thus the remaining 22 reports (43%) are assessed as in "need for improvement". Table 4 shows that according to the reports, interviews have been conducted in all evaluations. Furthermore, for more than half of the reports evaluators carried out focus group discussions (28, 56%), for 41% (21) a survey was implemented and for one third (17, 33%) it was explicitly mentioned that participatory observations had been used.

Table 4: Data collection methods used (n=51)

	No.	Share
Interviews	51	100%
Focus group discussions	28	56%
Survey	21	41%
Participatory observation	17	33%
Other	10	20%

In general, the evaluators implemented a diverse range of data collection methods and thereby fulfilled the quality criterion of an appropriate mixture of data collection methods. In more than three quarters of the evaluations (42, 82%) at least two different data collection methods have been applied. However, nine reports (18%) are limited as findings are grounded only on a single data collection method.

An observed weakness is the fact that the validity and reliability of the data are discussed only in a very few reports (8, 16% and 7, 14% respectively). In addition, in five reports (10%) we found evidence for severe failures with respect to data collection. For example, a survey was deemed an inadequate instrument as the sample size of the population to be surveyed was too small for meaningful standardised analysis at a later stage.

Elaborating on the data collection methods used and providing the data collection instruments in annexes are important factors contributing to transparency and allowing for methodological quality checks and a content-related revision of the evaluators' work. Only one third of the reports (18, 35%) provide at least partially the data collection instruments employed. This was, however, not explicitly requested by the MFA.

By far the weakest aspect in the methodology chapters or sections is **the information provided on the sample**. It comprises information on (i) the sample, (ii) the sampling strategy, and (iii) the justification of the sampling strategy. In more than half of the reports (28, 55%) this information is "inadequate" and in four reports (8%) it is in "need for improvement" due to incompleteness. Only about a third of the reports (19, 37%) are assessed as "satisfactory" (12, 24%) or "good or very good" (7, 14%) in this category.

The great majority of the reports (43, 84%) do not justify the chosen sampling strategy. Thus, they do not provide information why the selection of information sources is appropriate. Even worse, in about two thirds of the reports (32, 63%) sampling strategy is not presented at all. This means, for example, that the selection of interview partners was completely arbitrary. Most alarming, in about half of the reports (24, 47%) the sample composition is not presented. It therefore remains, for example, completely unclear how many interview partners in each category (e.g. beneficiaries, project staff, institutional environment) were spoken with and whether this seems plausible given the assignment.

There is significant statistical evidence that the aspect of sampling is dealt with worse in MFA-commissioned evaluation reports (mean: 1.63) than in reports commissioned by other partners (mean: 2.31). For details please see Annex 7.

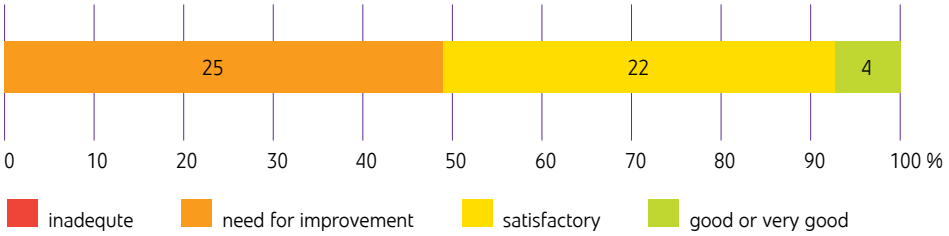
Another aspect of the methodology chapters or sections refers to the description of **data analysis methods**. Only eight reports (16%) provide comprehensive information, 25 reports (49%) are rather incomplete and 14 reports (27%) are very incomplete. In four reports (8%) information on the analysis methods is missing altogether.

For the appropriateness of analysis methods it is important that qualitative as well as quantitative data analysis methods are applied. In two thirds of the reports (34, 67%) the usage of qualitative as well as quantitative analysis methods is described. They were thus assessed as appropriate. However, in 11 reports (22%), severe failures regarding the application of data analysis methods were detected. For example, names of interview partners were disclosed in the analysis, hence violating the standard of anonymity (in five reports), figures were not contextualised giving a wrong impression of the real situation because of inappropriate scaling, or individual opinions were generalised.

Finally, the presentation of **limitations and challenges** was quite mixed. Only ten reports (20%) stand out with "good or very good" discussions of this aspect, followed by 12 (24%) assessed as "satisfactory". On the other hand, ten reports (20%) are in "need for improvement" and one third (19, 37%) of the reports are assessed as "inadequate" in this regard. Most of the limitations described relate to data collection (32, 63%) or refer to the evaluation process (21, 41%). Only in few cases (6, 12%) are challenges regarding the data analysis methods provided.

Overall, as shown by Figure 19, even though the methodological assessment reveals that there is only one report rated as "inadequate", half of the reports under consideration (25, 49%) are rated with "need for improvement". There are only four reports (8%) assessed as "good or very good" and 22 (43%) achieve "satisfactory" results.

Figure 19: Overall rating on methodology (n=51)



Source: own statistics based on analysis of reports

The overall assessment of the evaluation methodology shows statistically significantly lower scores for individual/independent consultants. While individual/independent consultants achieve a mean assessment of 2.19, other firms or institutes score 2.57 on average. For further details please refer to Annex 7. This points to a higher share of methodological limitations among individual/independent consultants but may be also caused by on average significantly lower net evaluation budgets for individual/independent consultants. The three aspects where the individual/independent consultants in general score lower are sampling, data analysis methods and limitations. Further analyses do not reveal any other significant differences between sub-groups (such as MFA commissioned evaluations) or over time.

4.4 Quality regarding evaluation findings

Highlights of the chapter addressing EQs 2 and 7:

- More than half of the 51 reports do not link their findings to the data sources.
- The intervention logic, fundamental for a sound understanding of the intervention and an appropriate analysis, is discussed comprehensively in less than one third of the reports.
- Furthermore, in one third of the reports, evaluators mix findings with conclusions and recommendations.
- The logical flow from the data to the findings, conclusions and recommendations is thereby weakened.
- Taken the above-mentioned points together: Findings are often obtained based on a weak methodology and there is a great need of improvement.
- The coverage and quality of the sections on relevance, effectiveness and efficiency are satisfactory or better for about two thirds of the reports.
- The coverage and quality of the sections on sustainability is a bit weaker; about 40% of the reports are in need of improvement or inadequate. They often lack a three-dimensional approach of economic, social and environmental sustainability.
- If assessed (for 40 reports only), the sections on impact are unstructured in more than half of the reports.

Whether the quality regarding the evaluation findings is appropriate has been assessed from three angles. We analysed how findings have been obtained, whether evaluators presented, discussed and reviewed the intervention logic and which content evaluators captured under the different OECD DAC criteria.

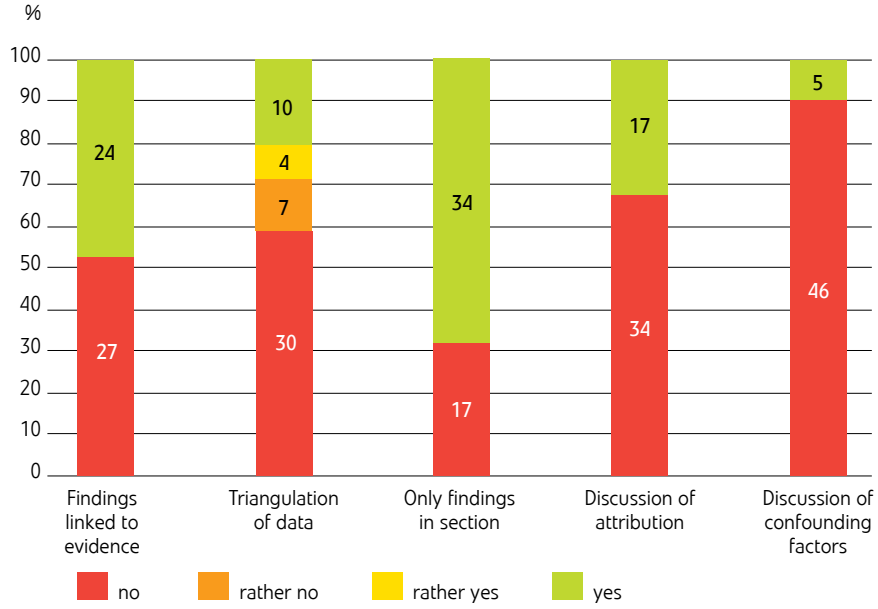
For about half of the evaluation reports methodological quality is acceptable.

Methodologies of independent consultants are of lower quality than those of consultancies.

Findings are often obtained on weak methodologies.

With regards to obtaining the findings, (i) the linkage of the findings to the data by providing data sources, (ii) the triangulation of findings, (iii) the presentation of findings clearly separated from conclusions and recommendations, and (iv) the causal attribution of the intervention to the results are assessed as presented in Figure 20.

Figure 20: Quality of findings (n=51)



Source: own statistics based on analysis of reports

Regarding the **linkage of findings to the analysed data**, the assessment reveals that more than half of the reports (27, 53%) do not clearly refer to the sources of information when reporting findings.

We did not expect that reference to data sources be made for every sentence, but rather for every few paragraphs as exemplary shown in the following box.

Good practice example:

"Throughout the interviews the research component contribution was a recurring theme. According to the interviewees' responses, this component brought out the rights of the Amazonian communities to their view of society, empowering them by legitimizing their ancestral knowledge, culture and language." (Report No.03)

Similarly, for **triangulation** (often promised in the methodological sections), we searched for evidence that evaluators discussed conflicting or confirming results from multiple sources. Evidence for the actual triangulation of data sources was rarely found. In only ten reports (20%) were the great majority of results put into perspective with reference to different data sources, while in more than half of the reports (30, 59%) evaluators failed to do so at all. Seven reports (14%) sometimes refer to different data sources and four reports (8%) often put the results into perspective. The following box presents exemplary good practice on this matter.

Good practice example:

"In a survey conducted with 98 respondents for this evaluation 81 (82.62%) of total respondents told that their social status was increased as people have started listening and respecting them. (…) Focus group discussions conducted at various levels also mentioned that their social status was largely increased as their voices are better heard and that they are respected at home and community." (Report No. 27)

Despite this negative finding, results seem mostly plausible and might be derived from the data even though the link is not explicitly highlighted. Occasionally, the meta-evaluation team has had severe doubts whether findings were really based on the collected data. By neglecting the indication of sources and the discussion of results from different sources, evaluators violate the principle of transparency and undermine the credibility of their own evaluation.

Another major issue is the fact that **findings** are often **intermingled with conclusions and recommendations** without a clear separation. In one third of the reports (17, 33%) evaluators do not only present findings in the findings sections but also intersperse recommendations. This is a major quality constraint. It leaves serious doubt about the credibility of the results, especially when no linkage between the data and the reported findings can be observed.

Furthermore, the **causal inference** of outcomes and impacts is mostly not discussed. This is for example the case when outcome objectives are presented and achieved outputs are listed below on the assumption that the outcome is achieved by default when outputs were generated. In only one third of the reports (17, 33%) the evaluators discuss whether the results can be attributed to the intervention. Further discussion on possible confounding factors is in general not provided and only presented in five reports (10%). The following box shows exemplarily how good practice looks.

Good practice example:

"Whether it may be attributed to the Programme, to the Ministry of Natural Resources and Environment or to the general development in the population as the information wave hits the country is hard to measure but it is a fact that environmental awareness at all levels is on the rise and that environmental issues now are taken seriously in virtually all transparent planning decisions." (Report No. 4)

Following the **intervention logic** (i.e. the programme theory, logical framework, results model etc.) is crucial to understand the intervention being evaluated and to structure the effectiveness and impact analysis. We checked (i) whether the logical framework was described, (ii) whether the evaluators in their analysis of the results models, if any, clearly referred to inputs, outputs, outcomes and impact and (iii) whether a discussion on its validity and underlying limitations was provided.

Only 15 reports (29%) include a comprehensive **description** of the intervention logic and (13, 29%) a partial description, while seven (14%) describe the intervention logic incompletely and 16 (31%) do not describe it at all.

A **results model** defining input, output, outcome and impact is provided in eight reports. Out of these, four reports where already published in 2015, arguing against the hypothesis of increased usage of results models by evaluators in recent years due to stronger orientation to results-based management over time.

In about half of the reports (30, 59%) evaluators **assess the intervention logic** and point out shortcomings if applicable. In seven reports (14%) evaluators do this assessment without describing the intervention logic in a first step. Eleven reports (22%) provide a further review of underlying assumptions of the intervention logic. Thus, even though the intervention logic is often not described in full, shortcomings are discussed more frequently.

The OECD DAC criteria set an important standard in the evaluation of development cooperation. To evaluate Finnish development cooperation, the MFA Manual (2013) specifies what exactly should be covered under each of the criteria.

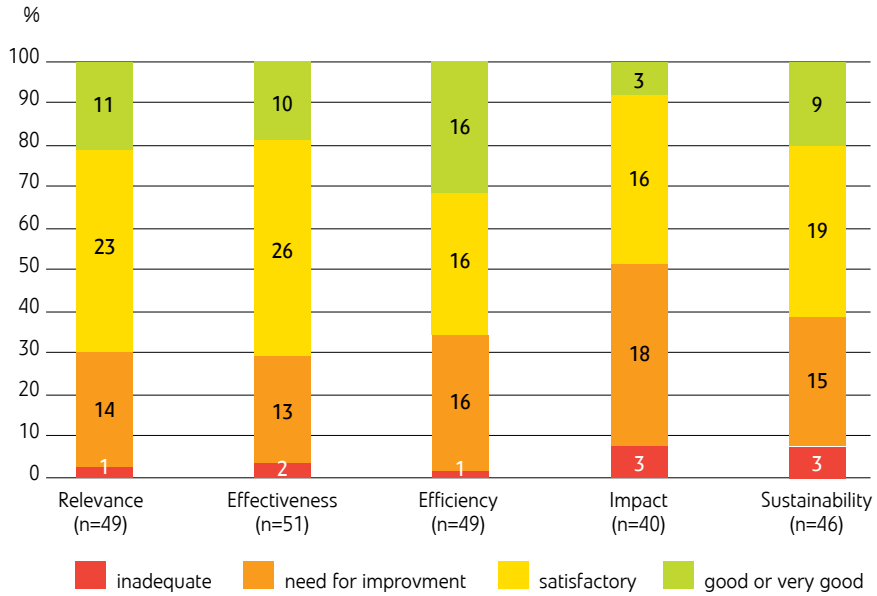
We assessed whether the criterion in general was discussed and if so, which aspects the evaluators treated. Afterwards an aggregate for each DAC-criterion was built taking the coverage of the requested aspects and the detail and quality of what was provided into account.

In 49 out of 51 reports evaluators discuss the **relevance** of the underlying intervention. The relevance regarding the needs of the target group is discussed in 39 out of 49 reports (80%) and the relevance regarding the needs of the final beneficiaries is discussed by 33 out of 49 reports (67%). For the differentiation between target groups and final beneficiaries it is important to understand that the notion "target group" summarizes very heterogeneous groups such as senior government staff or school children, whereas final beneficiaries exclusively refer to the poor population.

Furthermore, the relevance of the intervention with regard to its consistency with and support to national policies is assessed in the great majority of the evaluation reports (42 out of 49, 86%). In contrast, the consistency with MFA policies is less often discussed (20 out of 49, 41%). International conventions, policies, strategies or goals are addressed in 26 out of 49 reports (53%).

The overall overview provided in Figure 21 displays that, out of those reports which treat relevance, about two thirds (34 out of 49, 69%) are rated as either "good or very good" (11) or "satisfactory" (23). Another 14 reports (29%) are in "need for improvement" and one report is assessed as "inadequate". Interesting, in mid-term evaluations the quality of the relevance chapter or section is rated statistically significantly better. While final evaluations score on average 2.65 in this chapter, mid-term evaluations achieve 3.17 on average.

Figure 21: Are the DAC Criteria appropriately captured in the report?



Source: own statistics based on analysis of reports

In all reports, evaluators discuss the **effectiveness** of the intervention. Even though, according to the OECD guidelines (OECD 2017b), the analysis should focus on outcome achievement; almost all of the reports (46, 90%) provide a discussion on outputs. This does not come as a complete surprise, as many ToRs request evaluators to do so and evaluators then tend to wrongly include this aspect in the effectiveness rather than in the efficiency chapter. Notwithstanding this, 44 reports (86%) discuss the outcomes of the intervention, while only seven (14%) fail to do so.

Unfortunately, the terminology of outputs, outcomes and impacts is often not correctly used throughout evaluation reports. Therefore, in addition to outputs being analysed, and outcomes being labelled as outputs, outcomes are also sometimes confused with impacts and analysed in the impact chapter. This might be caused by limited methodological knowledge of the correct definitions and their application by the evaluator and/or by incorrect intervention logics.

Content-wise, more than half of the evaluation reports (29, 57%) contain a discussion of results for the target groups and an equally large number (28, 55%) include discussion of the results for the final beneficiaries. Similarly, in more than half the reports (28, 55%) evaluators refer to gender aspects or provide disaggregated results for women and men in the effectiveness section.

As Figure 21 shows, more than two thirds of the reports (71%) are rated either as "good or very good" (10) or "satisfactory" (26) with regard to the quality of the effectiveness chapter, whereas a quarter of the reports (13, 25%) are in "need for improvement" and two are judged as "inadequate".

Efficiency of the intervention is discussed in 49 of 51 reports (96%). Topics covered in more than three quarters of the reports are time efficiency of the intervention (37 out of 49, 76%), cost-efficiency (40 out of 49, 82%) and the efficiency of the implementation management (40 out of 49, 82%).

The coverage and quality of the sections on relevance, effectiveness and efficiency are appropriate for about two thirds of the reports. The sections on impact and sustainability lack a bit behind.

The efficiency of the personnel and the conversion of inputs into high quality outputs are only discussed in about half of the reports (28 out of 49, 57% and 27 out of 49, 55% respectively). The latter is often addressed only implicitly and/or in a separate paragraph. Thus, aspects important for efficiency assessment are not only covered by this chapter but partly also in chapters called "performance analysis". This is often caused by similar specifications of the ToRs which are thereby clearly inconsistent with the OECD DAC guidelines (OECD 2017b).

Figure 21 illustrates that, out of the 49 reports which treat efficiency, the quality of this chapter is assessed as "good or very good" (16) or "satisfactory" (16) for two thirds of the reports (32, 65%). The remaining third is in "need for improvement" (16) with the exception of one report assessed as "inadequate".

Impact is the least covered criterion in the reports. Eleven reports (22%) do not report on the impact of the intervention. As expected, a higher share of final evaluations (24 out of 28) than mid-term evaluations (16 out of 23) contain a discussion on this criterion. On the positive side, all but one report commissioned by the MFA report on impact (23 out of 24). In three quarters of the reports which capture impact (30 out of 40, 75%), evaluators discuss if the intervention contributed to its overall objective. In only seven reports (18%) did evaluators analyse whether there have been any unintended impacts by the intervention.

Roughly half of the reports discuss whether the intervention has contributed to enhance the quality of life (21 out of 40, 53%), whether there has been any contribution to enhance institutional quality (25 out of 40, 63%) and whether the intervention has contributed to changes in the partner countries policies or to sector reforms (18 out of 40, 45%).

Overall, the quality of the impact chapter is rather negatively assessed with more than half of the reports capturing this criterion (21 out of 40, 53%) judged as either in "need for improvement (18) or inadequate (3). Only three reports (8%) are assessed as "good or very good" and 16 out of 40 (40%) as "satisfactory".

Sustainability of the intervention is assessed in 46 reports (90%). Evaluators focus most often on economic sustainability (31 out of 46, 67%), followed by social sustainability (24 out of 46, 52%) and less frequently environmental sustainability (10 out of 46, 22%). Only five reports (11%) apply the three-dimensional concept of economic, social and environmental sustainability. However, two thirds of the evaluators (31 out of 46, 67%) perceive sustainability as a multi-faceted concept and hence analysed sustainability regarding multiple dimensions such as institutional and economic sustainability.

More than two thirds of the reports (33 out of 46, 72%) discuss whether the benefits of the intervention are likely to continue. Furthermore, a similar number of reports discuss the capacity of target groups (including the final beneficiaries) and of the implementing agencies to make the intervention sustainable (34 out of 46, 74% and 29 out of 46, 63% respectively). The financial capacity of the target group (including final beneficiaries) and of the implementing agencies is less often discussed (19 out of 46, 41% and 22 out of 46, 48% respectively).

Overall, for more than half of the reports capturing sustainability (28 out of 46, 61%), the quality of the sustainability chapter is assessed as "good or very good" (9) or "satisfactory" (19). About one third (15 out of 46, 33%) are in "need for improvement" and three are "inadequate".

In comparison to one another, the quality of the chapters on relevance, effectiveness and efficiency is rated better than the quality of the chapters on impact and sustainability. Both these chapters often suffer from relatively short and unstructured analyses, with the quality of the impact chapter often being particularly low.

4.5 Quality of conclusions and recommendations

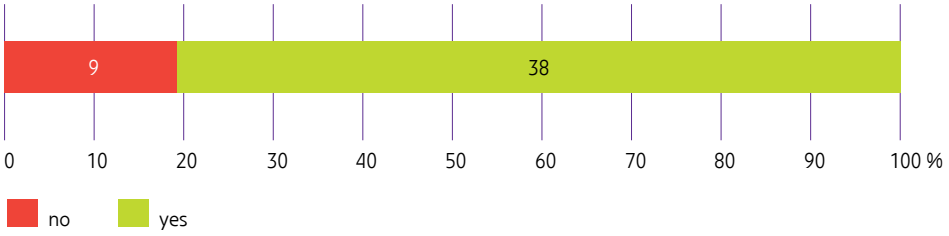
Highlights of the chapter addressing EQs 2 and 7:

- For more than 80% of the 51 reports, conclusions and recommendations appear to be logically derived from findings (as far as this could be assessed by the meta-evaluation team).
- An assessment of the relevance and the usability of the conclusions and recommendations is not possible from an external perspective.
- In one quarter of the reports, the quality of the conclusions is unacceptable.
- Some of the evaluation reports have been accepted by MFA or other commissioners without conclusions or recommendations.
- Recommendations are directed to actors in about two thirds of the reports, but in more than 80% of the reports no prioritisation, direction to specific actors and timeline for implementation is provided.
- Only about half of the reports provide lessons learnt although it is generally requested by MFA's Evaluation Manual and by more than half of the ToRs.

Regarding the quality of the conclusions, four reports do not have a section on conclusions. From a methodological point of view this is a severe failure of report quality as it seriously hampers the usability of the evaluation results. For the remaining 47 reports Figure 22 shows whether (i) conclusions are derived from findings and (ii) whether conclusions refer to the OECD DAC criteria.

Figure 22 shows that in more than three quarters of the reports with a conclusion chapter (38 out of 47, 75%), the conclusions are **derived from findings**. However, in nine reports (19%) we found new information in the conclusions which was not presented in the findings. Either a new, not yet presented data source was revealed or, even worse, new findings were presented without reference to any data. The obvious inconsistency between findings and conclusions is a severe failure. Putting these nine reports together with the four reports missing conclusions, we judge that one quarter of the reports (13 out of 51, 25%) are seriously deficient in quality.

Figure 22: Conclusions are derived from findings (n=47)



Source: own statistics based on analysis of reports

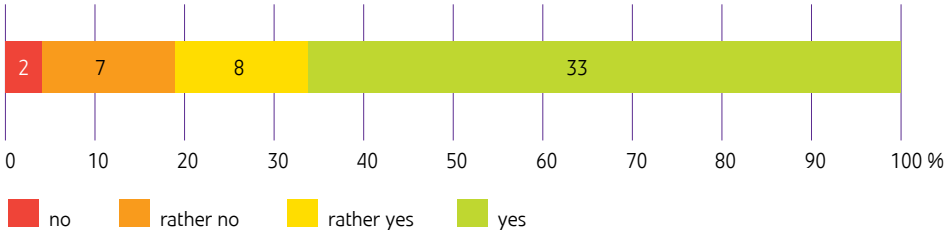
Conclusions are mainly derived from findings and recommendations are mainly derived from findings and conclusions.

With respect to **reference to the OECD DAC criteria**, conclusions in one third of the reports (15 out of 47, 32%) cover all five criteria. About half of the reports (24 out of 47, 51%) cover only some of the criteria and six reports (12%) refer to none of the criteria in the conclusions. Similar to the findings chapter, impact is the least covered topic. Less than half of the reports (22 out of 47, 47%) refer to impact in the conclusions.

With regard to the **recommendations** we only find one report that does not provide any recommendations. Even though the reports often focus more on an exit strategy, some practicable recommendations should have been given that go beyond general statements. Thus, for 50 reports, we assessed whether (i) recommendations are derived from findings and conclusions. In addition, the analysis asks whether recommendations are (ii) directed to actors in general, (iii) prioritised, (iv) addressed to specific actors, and (v) time bound as well as whether (vi) lessons learnt were derived.

Two thirds of the reports (33 out of 50, 66%) derive their **recommendations from findings and conclusions** as shown in Figure 23. In contrast, eight reports (16%) are flawed by inconsistencies between recommendations and findings and/or conclusions. This raises serious concerns about the credibility of the recommendations.

Figure 23: Recommendations are derived from findings and conclusions (n=50)

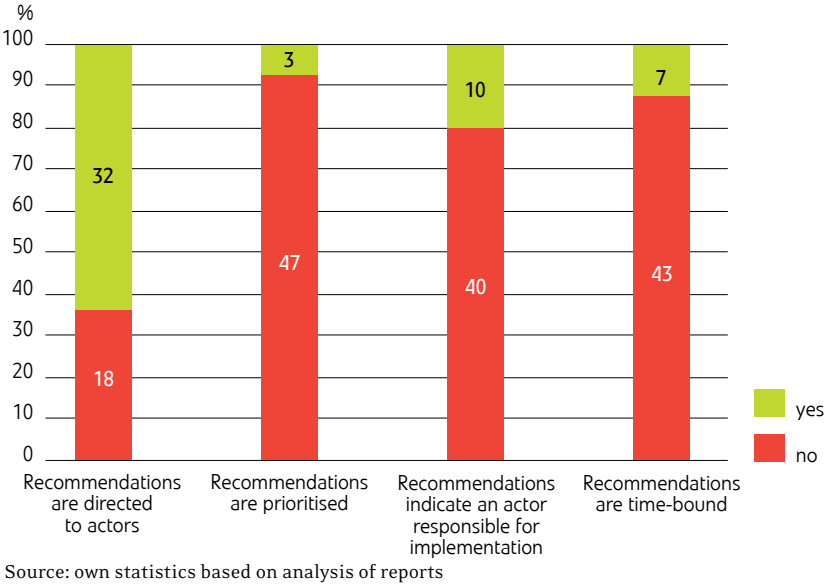


Source: own statistics based on analysis of reports

Recommendations often lack preciseness.

Again, recommendations in about two thirds of the reports (32 out of 50, 64%) are **directed to actors** (e.g. MFA, implementing agency, UN) as illustrated in Figure 24.

Figure 24: Recommendations (n=50)



Source: own statistics based on analysis of reports

In contrast, only very few reports **prioritise the recommendations** (3 out of 50, 6%), indicate a **specific actor responsible** for the implementation (10 out of 50, 20%) or set **schedules for the implementation of recommendations** (7 out of 50, 14%). Thus, there are several areas where recommendations turn out to be hardly pragmatic.

Lessons learnt are presented in only a bit more than half of the reports (29 out of 51, 57%). The provision of lessons learnt is requested in the MFA Evaluation Manual and furthermore also by more than half of the corresponding ToRs (30 out of 45, 67%). Nevertheless, the request for lessons learnt in the ToRs did not lead to a higher share of evaluations integrating lessons learnt in the report. Thus, there is frequently a lack of ability to generalise, as evaluation reports fail to go beyond intervention-specific recommendations.

4.6 Further aspects

Highlights of the chapter addressing EQs 2 and 7:

- About two thirds of the reports include gender equality and reduction of inequality as cross-cutting objectives. Less than half of the 51 reports integrate climate sustainability and HRBA as cross-cutting objectives. Thus, in turn, cross-cutting objectives have not been assessed in a number of reports.
- If assessed, quality of cross-cutting analyses is better for gender equality and climate sustainability (51% and 60% of the reports assessed as good or very good vs. about 40% for other cross-cutting objectives).
- MFA's request to include the context analysis after the methodology chapter is unusual and not often followed by the evaluators. About three quarters of the reports, regardless of who was the commissioning entity, are not in line with MFA's requested structure in any way.
- For about 80% of the evaluation reports annexes are complete.
- Some reports (6 out of 51) show weaknesses with regards to writing and editing.
- Insights on validation of findings and quality assurance are not provided in more than three quarters of the reports.
- The composition of the evaluation team regarding gender quality, thematic knowledge, evaluation capacity and local expertise remains unclear for 80% of the reports.
- At least one quarter of the evaluation reports were produced by gender unbalanced teams.

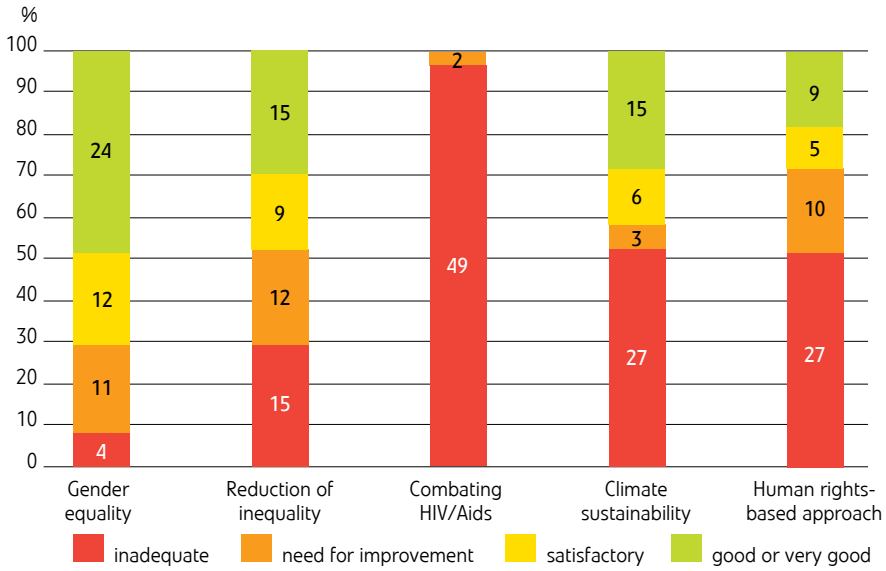
In this chapter the quality of reports is assessed with respect to further aspects: (i) the integration of cross-cutting objectives, (ii) the structure, style and annexes of the report, (iii) validation and quality assurance, and (iv) the composition of the evaluation team.

As described in section 3.1, the **cross-cutting objectives** are deeply anchored within Finnish development policies. However, given policy changes over time, they differ for the interventions under consideration for this meta-evaluation. As interventions and hence evaluation reports cannot clearly be linked to a particular development policy, the analysis process was cumbersome. We analysed the reports for the five cross-cutting objectives (i) gender equality, (ii) reduction of inequality/focus on vulnerable groups, (iii) combating HIV/AIDS, (iv) climate sustainability and the (v) HRBA. In a first step, we assessed whether a topic was covered by the report and, if yes, the level of detail with which it was discussed.

Apart from gender equality, cross-cutting objectives are quite often not treated in evaluation reports.

Most reports deal at least with one of the cross-cutting objectives of Finnish development cooperation as shown in Figure 25.

Figure 25: Integration of cross-cutting topics (n=51)



Source: own statistics based on analysis of reports

As expected, **gender equality** is most often discussed in the reports and often integrated to a very high degree. About two thirds of the reports (36 out of 51, 71%) integrate gender equality in findings, conclusions and recommendations, nearly half of the reports (24, 47%) do this in a "good or very good" manner, another quarter (12, 24%) to a "satisfactory" extent. Among these were four evaluation reports on interventions with a principal focus on gender. However, 11 evaluation reports (22%) are in "need for improvement" and four (8%) were assessed as "inadequately" which means that they do not deal with gender equality at all. Considering that all relevant Finnish development policies place great emphasis on gender equality and that, as a consequence, all evaluations should have treated this aspect, it is striking that roughly one out of three evaluators in our sample failed to do so.

Reduction of inequality or the focus on vulnerable or marginalized groups is similarly integrated into more than two thirds of the reports (36, 71%) but the level of detail is on average much lower than for gender equality. Fifteen reports (29%) capture the objective to a "good or very good" extent and nine (18%) in a "satisfactory" way, while 12 reports (24%) are in "need for improvement" and 15 (29%) have not captured the aspect. However, these results are less conclusive, as it remains unclear whether underlying interventions were obliged to integrate reduction of inequality as a cross-cutting objective and, thus whether the evaluators should have integrated it into the analyses.

The least considered cross-cutting objective is **combating HIV/AIDS**. It has only been a key aspect on the Finnish development agenda for a few years and has also received less attention on the global level. Thus, not surprising, only two reports mention it, and also do not provide a deeper analysis of the intervention in this regard. This is further reinforced by a low number of interventions with a direct connection to the health sector; only one intervention in reproductive healthcare and five water and sanitation interventions.

Even though in general only less than half of the reports (24, 47%) include **climate sustainability**, it is often covered in a "good to very good" way (15, 30%) when integrated. Six reports (12%) are dealing with it in a "satisfactory" manner and three with "need for improvement". The remaining half of the reports (27, 53%) did not capture climate sustainability. For the same reasons as mentioned in conjunction with reduction of inequality, the analysis is not conclusive.

One explanation for the vast number of reports assessed positively with regard to the mainstreaming of environment and climate change could be the rather large number of interventions directly focusing on these topics (9). In addition, three interventions in the forestry sector, two in the energy sector and three in the agricultural sector cannot afford to ignore this cross-cutting objective neither. On the other hand, there are several interventions in sectors like education, governance or conflict prevention where a connection to climate sustainability is not as obvious and consequently the issue is often left out by the evaluators. Once again, it is important to note that here we do not discuss the integration of cross-cutting objectives into the interventions, but rather explanatory factors for their low coverage by the evaluation reports under consideration.

Climate sustainability is addressed significantly more often in MFA-commissioned reports (15 out of 24 vs. 9 out of 27) than in those commissioned by other institutions. This can only be partially explained by a higher number of interventions with stronger linkages to the objective. One possible explanation might be that awareness raising through Finnish policies and evaluation guidelines and the integration of cross-cutting climate objectives in the ToRs might have had a positive influence on evaluation practice.

Regarding the integration of the **Human-Rights Based Approach (HRBA)** the picture is very similar. In nearly half of the reports (27 out of 51, 47%) evaluators do not cover this objective. If treated, in general, an in-depth analysis is often lacking. Only nine reports (18%) were assessed as "good or very good" in this regard, 5 (10%) as "satisfactory" and 10 (20%) as in "need for improvement". Similarly, MFA-commissioned evaluation reports integrate the HRBA more frequently than others; 15 of 24 MFA-commissioned reports integrate it while only nine of 27 reports of other commissioners refer to the HRBA.

Structure, style and annexes of the evaluation reports have been assessed. In particular, attention was given to (i) the reports' structure, (ii) the inclusion of ToRs, (iii) the attachment of a list of people interviewed, (iv) documents consulted (v) a two-pager as communication instrument, and (vi) proper editing and writing style.

In most cases the **structure** does not follow the MFA's manual (only 13 out of 51 reports, 26%). While some reports lack important chapters such as context analysis or the methodology, others do not comply with the specification to put the context analysis behind the methodology chapter. This is often handled the other way around by other commissioners and by international standards (e.g. UNICEF 2010, USAID 2012). Thus, it is no surprise that particularly, those reports not commissioned by MFA (18 of 24), are not in line with MFA's structure. It can be assumed that the considerable number of MFA-commissioned

reports (7 of 27) which did not comply with MFA's request thus gave higher priority to international conventions.

Another aspect of the structure is the annexes. In this regard the **ToRs** are requested by the MFA and often by other commissioners to be annexed to the report. In general, they are provided in more than three quarters of the reports (39 out of 51, 77%). As described already in the methodology section the requested **lists of people interviewed** and of **documents consulted** can be found in more than 80% of the reports. The newly requested **two-pager communication tool** has been introduced only recently and is not often annexed. It has been provided only by four out of the 22 reports completed in 2016 and 2017.

Even though many reports contain paragraphs that are difficult to understand or have spelling or grammar errors, they are mostly comprehensible. Only six reports (12%) are assessed as not properly **written and edited**.

Assessments regarding **validation of findings and quality assurance** are somewhat inconclusive. In some reports evaluators mention **validation** meetings with stakeholders (10), MFA (6) or both (10). The only insight from this finding is that the majority of reports do not provide information regarding this topic. Whether and with whom validation workshops have taken place throughout the evaluation process remains unclear. However, the schedules within the ToRs often suggest that especially for MFA-commissioned evaluations validation activities are requested.

Validation of findings and quality assurance cannot be comprehensively assessed from the evaluation reports only.

The assessment of **quality assurance** is similar. Three quarters of the reports (38, 75%) do not mention any mechanism of quality assurance. Thus, it remains unclear whether and how quality was assured. In five reports evaluators elaborate on external as well as internal quality assurance, in four reports they address only external, and in three reports only internal quality assurance.

Finally, we assessed the appropriateness of the **composition of the evaluation team**. Often only the names on the cover page are given and at times only the company conducting the evaluation is specified. This provides at best some hints about **gender-balance**. Out of the 33 evaluation reports which disclose the names of the evaluators, seven have been conducted by an individual. Thus, 26 reports were produced by teams. Out of these, 12 are performed by gender-balanced team which means that there is either an equal distribution of gender (e.g. 1:1, 2:2) or a small gender difference (e.g. 1:2, 2:3 etc.), whereas 14 have been produced mainly by male-dominated teams or rarely by female-dominated teams. Although for 15 reports we do not know anything on the evaluators, the analysis discloses that at least about a quarter of all reports (14 out of 51, 27%) has been produced by teams which are not gender-balanced.

At least one quarter of the evaluation teams is not gender-balanced.

Beyond this, more than three quarters of the reports (41, 80%) do not provide detailed information on the evaluation team at all. Thus, a comprehensive assessment of the team composition with reference to **gender equality, thematic knowledge, evaluation capacity and local expertise** could only be conducted for ten cases. Hence, further analysis remains inconclusive.

Finally, as data on the evaluation team was often missing, assessments regarding **lack of independence** of the evaluators were in general not possible. However, in one report the subjective assessments by the evaluator were so numer-

ous throughout the narratives that a lack of independence has to be suspected. Therefore, we excluded the evaluation from further analysis.

4.7 Quality of executive summaries

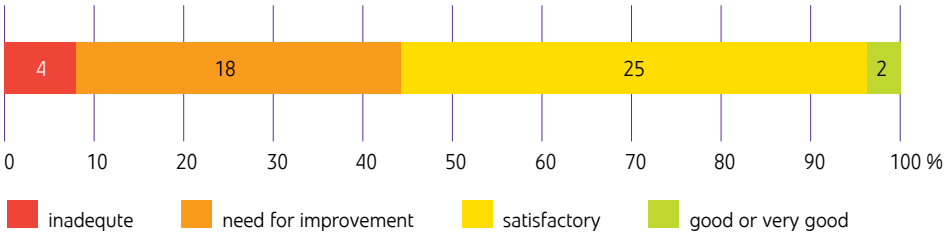
Highlights of the chapter addressing EQs 2 and 7:

- Just over half of the executive summaries are adequate. Thus, there is considerable room for improvement.
- Only two out of 51 reports provide a fully comprehensive executive summary.
- Two reports do not provide an executive summary and, hence should not have been accepted by the commissioner.
- The most valuable information provided concerns the description of the intervention, findings, conclusions and recommendations.
- About one third of the executive summaries do not include information on the evaluation design and methodology.

The executive summary is a key feature of the evaluation report. This is often the only part of the report that is read by a broader audience. Thus, it is important that it is of high quality and that all core information of the evaluation report is included. Nevertheless, two reports of 51 (4%) do not provide a summary. As is lack of conclusions and recommendations, this is grounds for rejection of the evaluation report.

Further analysis of summaries was undertaken for the remaining 49 reports. Assessments of (i) the completeness, (ii) the style and (iii) the coherence with the report have been part of the analysis. With respect to **completeness**, the summary should resemble an evaluation report and provide **rationale, objectives, scope, design and methods of the evaluation, describe the intervention and include findings, conclusions, recommendations and lessons learnt**. Figure 26 illustrates that only two of the summaries (4%) cover all these topics and are rated as "good or very good" and roughly half (25 out of 49, 51%) have "satisfactory" summaries. In contrast, four summaries (8%) are rated with the lowest quality category "inadequate" (providing only one or two of the eleven components) and more than one third (18 out of 49, 37%) are in "need for improvement" (lacking five to seven of eleven components).

Figure 26: Completeness of Summary (n=49)



Source: own statistics based on analysis of reports

Just over half of the executive summaries are adequate.

The **description of the intervention** (38 of 49, 78%), **findings** (42 of 49, 82%), **conclusions** (37 of 49, 76%) and **recommendations** (45 of 49, 92%) are included by most evaluators. **Methods** (34 of 49, 69%) and the **evaluation design** (27 of 49, 55%) are often not included. Furthermore, the **table requested by MFA to summarise findings, conclusions and recommendations** is most frequently lacking. Only eight summaries (16%) provide it completely and seven (14%) incompletely. Reports providing a complete table are exclusively commissioned by MFA. This is no surprise given the specific request which is often not known from other commissioners. **Lessons learnt** are also mostly not included in the summary. Half of the reports providing lessons learnt in the report (29 of 51) include them as well in the summary (15 of 29).

Individual/independent consultants are scoring lower regarding the completeness of the summary than other evaluation units. The mean is 2.12 for individual/ independent consultants in contrast to 2.67 for other evaluation entities, pointing again to higher methodological knowledge of the latter, a better resource endowment for the evaluations reports produced by the latter, or a mixture of both.

In two summaries (4%) we observed **inconsistencies** with the report. For example, in one case, new information was provided that did not appear in the report.

With regard to the **writing style**, summaries are in general well written. Only two summaries (4%) have been assessed negatively. Thus, also two of six reports which have been assessed negative regarding the writing style perform better with respect to summary.

4.8 Linkages between quality of ToRs and quality of reports

Highlights of the chapter addressing EQ 6:

- 16% of reports fail to adequately respond to the evaluation questions and hence miss the point of the exercise.
- Nearly all of the reports cover the OECD DAC criteria requested by the ToR with the exception that in 10% of the reports impact is not discussed although it has been requested. Impact is the OECD DAC criterion with the highest omission in the ToRs. 16% of the ToRs do not request it.
- The overall report quality is assessed as "satisfactory" for two thirds of the reports and in "need for improvement" for one third.
- On average, introductions and context analyses, methodologies, and conclusions and recommendations are rated as "satisfactory" whereas findings and summaries are assessed as in "need for improvement".
- Overall report quality does not vary as between MFA-commissioned evaluations and evaluations by other commissioners, or between evaluations conducted by individual/independent consultants and those of consulting firms/institutes; or according to different project budgets.
- On average, a higher quality of ToRs is associated with a higher quality of the subsequent evaluation reports.
- The ToR's sections on purpose, objectives and scope of the evaluation; on the methodology, and on the evaluation process are particularly important for overall report quality.

It is expected that the quality of ToRs and reports are highly connected. In this chapter we first assess (i) whether the reports are answering the evaluation questions formulated in the ToRs and (ii) whether the reports captured those OECD DAC criteria request by the ToRs. Furthermore (iii) the overall quality of the reports is assessed to enable (iv) linking the overall report quality to the overall quality of the ToR.

On a more general level it is of interest for the quality of an evaluation report **if the evaluators answer the evaluation questions by the commissioner**. Although this question is of utmost interest for the commissioner, it is very cumbersome to assess by the meta-evaluation team. As the ToRs were missing for seven evaluation reports (inclusive of one ToR without evaluation questions), the analysis was limited to 44 reports. Furthermore, the fact that some ToRs provide a huge number of questions (several reports have over 40 questions) and the fact that most reports do not structure their findings and conclusions according to the questions hampers the analysis. As assessment on a four-step scale lacked unambiguous categories, we could only reveal a tendency with a simplified yes/no answer. Accordingly, it turned out that seven out of 44 reports (16%) do rather not comprehensively answer the evaluation questions stipulated in the ToRs. Thus, on average one out of six reports tends to miss the point of the exercise.

Regarding the **coverage of requested OECD DAC criteria**, out of 45 reports with ToRs all but one were supposed to assess the relevance of the intervention and all did so. For effectiveness, one report was not requested to discuss effectiveness but did so nonetheless. Regarding efficiency, all reports were obliged to treat it and all but one did so. A considerable gap is observed for impact. This was requested to be discussed only in 37 reports, but four reports did not present findings for it. On the other hand, three reports not requested to cover impact did, in fact, include it. For sustainability, all 41 reports supposed to capture it did so. In addition, one included it without having been requested to do so. Thus, the largest divergence can be seen for impact, which is at the same time the criterion most likely to be omitted in the ToRs. However, the great majority of the reports cover the OECD DAC criteria as requested.

For the **overall rating** of report quality, the quality of the executive summary, the introduction and context analysis, the methodology, the evaluation findings and the conclusions and recommendations have been taken into account (as explained in chapter 2.4). For further details on the aggregate please refer to Annex 7.

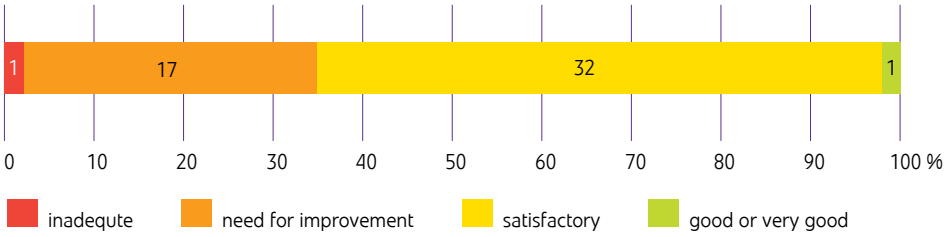
As would be expected from the previous sections, the overall rating of report quality is mediocre and cause for some concerns. While about two thirds of the reports are rated as overall "satisfactory" (32, 63%), one third (17, 33%) are rated with an overall "need for improvement". Even reports rated "satisfactory" sometimes display flaws and can be improved. Only one report is assessed as "good or very good" and one report is assessed "inadequate". The latter has been excluded from further analysis. This report is in large parts anecdotal and violates against several evaluation standards like anonymity and independence.

From all of these reports it is possible to derive some useful information on the intervention for the implementing agencies and/ or for the commissioner.

On average, a higher quality of ToRs is associated with a higher quality of the subsequent evaluation reports.

Overall report quality does not vary between sub-groups.

Figure 27: Overall quality of reports (n=51)



Source: own statistics based on analysis of reports

While introduction and context analysis, methodology, and conclusions and recommendations are on average rated as "satisfactory", findings and summaries are on average assessed as in "need for improvement".

Differences at an overall level for MFA-commissioned evaluations vs. evaluations by other commissioners, for evaluations conducted by individual/independent consultants vs. those of consulting firms/institutes, and by different project budgets (in a reduced sample according to data availability) are insignificant. Thus, at an overall level no systematic differences can be detected between sub-groups of the sample.

Assessing the **linkages between the overall quality of the evaluation reports and the overall quality of the ToRs**, we find that the overall report quality and overall ToR quality are statistically significantly correlated. Thus, higher overall quality of ToRs is associated with higher overall quality of the evaluation reports. Symmetrically, weaker ToR quality is associated with weaker report quality. While quality of the ToRs is not the only factor at play, a causal linkage running to quality of the final report can plausibly be inferred. As the ToRs are always first, reverse causality can be excluded. High quality of the ToR section on purpose, objectives and scope of the evaluation; the section on methodology and the section on the evaluation process are positively correlated and hence, particularly important for high report quality of the subsequent report.

5 SUMMATIVE ANALYSIS

After assessing the quality of the evaluation reports and the associated ToRs, the summative analysis focusses on a content assessment. We aggregate the assessments provided in each evaluation report which passed minimal quality standards (50 reports). Thus, it is important to note that the meta-evaluation team does not assess the interventions themselves, but rather synthesises the findings of the evaluators as presented in their evaluation reports. Hence, a fraction of Finnish development cooperation portfolio comprising selected single bilateral or multilateral interventions is assessed based on the reliable decentralised mid-term and final evaluation reports.

In chapters 5.1to 5.5 we provide answers to the evaluation questions EQ10 to EQ14 on relevance, effectiveness, efficiency, impact and sustainability of Finnish development cooperation. Chapter 5.6 responds to EQ15 (gender) and discusses EQ16 to EQ18 (other cross-cutting objectives), while chapter 5.7 is dedicated to EQ19 (aid effectiveness), EQ20 (complementarity), EQ21 (coordination) and EQ 22 (coherence). Chapters 5.8 and 5.9 synthesise lessons learnt and recommendations drawn by the evaluators to respond to EQ26 (recommendations to improve Finnish development cooperation by the evaluation reports). Finally, chapter 5.10 provides insights on the overall quality of Finnish development cooperation and differentiates according to various characteristics like geographical scope or different thematic sectors. Thereby, it provides answers to EQ23 (on the overall quality of Finnish development cooperation), EQ24 (on strengths) and EQ25 (on weaknesses).

5.1 Relevance

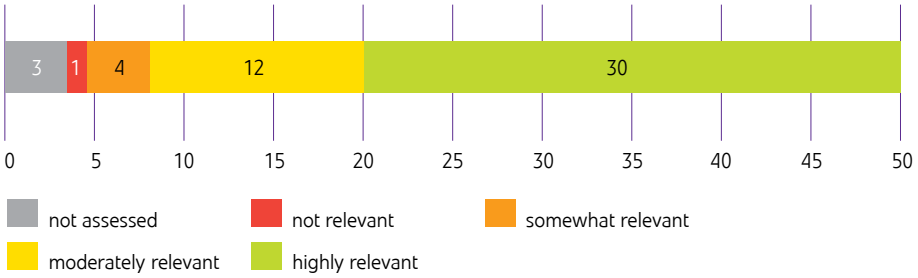
Highlights of the chapter addressing EQ 10:

- About two thirds of 47 interventions are assessed as highly relevant.
- Roughly one quarter of the interventions is assessed as moderately relevant.
- Eight (9%) of the interventions is assessed as only somewhat relevant.

The OECD-DAC criterion Relevance assesses the extent to which the intervention is suited to the priorities and policies of the target group, recipient and donor (OECD, 2017b).

As Figure 28 illustrates, according to the evaluation reports interventions were in most cases (42 out of 50, 84%) assessed as moderately relevant (12) to highly relevant (30). Five reports (10%) considered their assessed intervention as somewhat (4) to not at all relevant (1), while three reports (6%) did not assess relevance.

Figure 28: Relevance according to the evaluation reports (n=50)



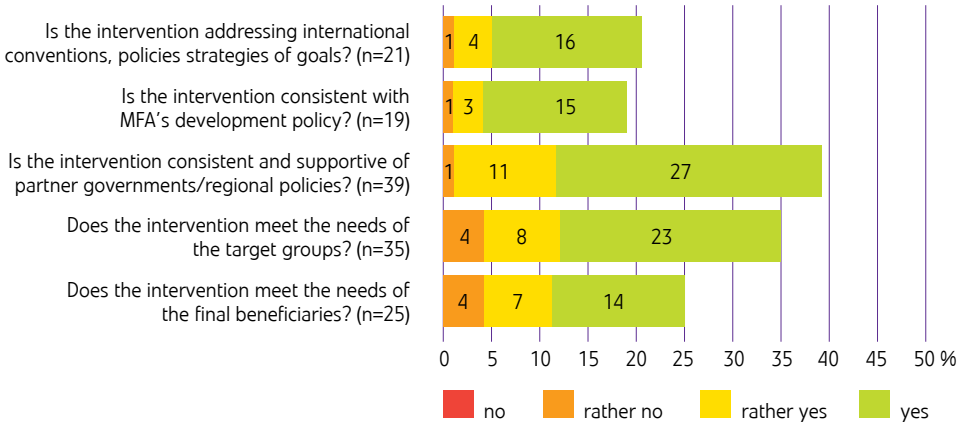
Source: own statistics based on analysis of reports

Not all authors of the reports which treat relevance base their assessment on the same aspects of this criterion. As illustrated in Figure 29, in the majority of reports, evaluators focused strongly on the consistency of an intervention with national/regional policies and paid rather little or no attention to its consistency with MFA development policies or with international conventions.

27 out of 39 interventions were assessed as consistent with national/regional policies, another 11 as rather consistent and only 1 as rather not consistent. We refrain here from presenting percentages to avoid the impression of possible generalisation. This procedure is always followed when assessments are only available for a limited number of reports. Whether the intervention addresses international conventions was only assessed for 21 interventions, but again with generally positive findings: 16 interventions were assessed as consistent and 4 as rather consistent. Similarly, 15 out of 19 reports were assessed as consistent with MFA's policies and another 3 as rather consistent. Thus, in general it can be observed: if any kind of consistency was assessed, evaluation reports were mostly positive about it.

Irrespective of which aspect of relevance is assessed, a vast majoritiy of evaluators rated the intervention as relevant in this regard.

Figure 29: Evaluators' assessment of different aspects of relevance



Source: own statistics based on analysis of reports

This holds also true for the questions whether the intervention met the needs of (i) the target groups in general and (ii) the final beneficiaries in particular. Please refer to chapter 4.4, p. 34 for a definition of target group and final beneficiaries. For better than half of interventions, both questions were answered with "yes" and with "rather yes" (respectively, 14 out of 25 and 23 out of 35). However, meeting the needs of the final beneficiaries has been only appropriately evaluated in half of the reports under consideration (25 out of 50). Hence, results have to be taken with care. Meeting the needs of the target group has at least been appropriately assessed for 70% of our sample (35 out of 50).

Interventions were assessed as meeting the needs of their target group and/or their final beneficiaries for several reasons: They were either

(i) successfully aligned with national, regional of MFA policies,

(ii) directly responsive to the demands of their stakeholders, or

(iii) actively reached out to the stakeholders during the development of the intervention's design.

If an intervention

(iv) acknowledged the specific situation of the recipient region or country, e.g. through needs assessment, or if it

(v) initiated a development-enhancing innovation, e.g. provision of tools to control and protect livelihood resources,

it received positive ratings as well. In contrast, evaluators assessed interventions as failing to meet the needs of the target groups and/or final beneficiaries mostly because of

(i) inappropriate design, e.g. too vague, too broad, insufficiently focused or not adapted to specific country/regional conditions, and

(ii) exclusion of relevant stakeholders from the intervention design.

Furthermore, evaluators highlighted in a few cases

(iii) inadequate selection of the target group,

(iv) ignorance of the diversity of the target group,

(v) mismatch of target group and intervention (e.g. if a targeted government already had sufficiently developed capacities in the field of the intervention), or

(v) the mere failure of interventions

as explanatory factors for their assessments. Examples of reasons for the assessment can be found in Box 1.

Box 1. Examples of reasons for the assessment of relevance

- "A project like … that contributes to the development of the irrigation sector with a focus on rural small- scale farmers is highly relevant for the beneficiaries and fully in line with Finnish and Zambian development strategies and priorities." (Report No 2)
- "The bottom to top scheme for project design marked an important methodological pathway for external cooperation, and was highly praised by most of our interviewees who expressed appreciation to Finland and the MFA for the respect they showed for the interests of indigenous groups and for their human rights." (Report No 3)
- "The FFF [Forest and Farm Facility] approach is highly aligned with the national policies of participating countries. Its model of directly supporting FFPO proposals financially and technically is highly relevant to the needs and priorities of target forest and farm smallholders, who view it as filling the gaps in rural development cooperation that other actors do not usually address." (Report No 35)
- "The Programme's design clearly addresses the global and regional challenges of deforestation and forest degradation. It highlights activities that are aimed to improving governance 2of forest resources, enhancing institutional capacity and developing systems for monitoring forest resources and national forest carbon stocks. In particular, the Programme is a relevant response to UNFCCC negotiations and the emerging REDD+ agenda. It is therefore adding value as far as addressing global/regional challenges and priorities is concerned." (Report No 20)
- "(T)he overall implementation approach can be characterized as "one size fits all". All activities under the three components have been the same for all 312 Farmers' Clubs irrespective of their specific conditions, needs, requirements or priorities. Conditions differ substantially from area to area including farming system (rice-based or maize-based), market access, soil type, rainfall, road infrastructure, availability of money, average land tenure, water availability and access to urban." (Report No 51)
- "MHM [Menstrual Hygiene Management] activities under the … programme met actual needs of adolescent girls only in a very small way: MHM facilities have not been built in all schools; where they have been built, they have not been built well; and even where they have been built well, they are not always used – with a lack of trained teachers being the main constraint to reaching adolescent schoolgirls with information and guidance on MHM, although such counselling was found to be very useful." (Report No 56)
- "While there is a growing consensus that STI [Science, Technology and Innovation] has an important role to play in contributing to poverty reduction, the optimal ways of achieving this are still emerging. At the same time, Mozambique has extremely limited resources related to S&T expertise, infrastructure and finances, which places severe constraints on what is possible in the short term.' The mechanisms for using STI to reduce poverty are left 'in the air." (Report No 5)
- "The project objectives relate well with the rights and needs of target groups (right-holders), in terms of social and economic empowerment, advancing women's rights, social reconciliation of the former combatants and CAW&Gs. [Conflict Affected Women and Girls] While participatory bottom up approach was applied during the project planning and implementation phase to identify beneficiaries from target groups (e.g. IPWA [Inter-party Women Alliances], CAW&G, former combatants, VAW [Violence against women] survivors), MTE [mid-term evaluation] noted from its field communications that the project benefits in few cases went to those who weren't directly affected by the conflict. Some KII [key informant i2nterview] respondents expressed that those women who were not directly affected by conflict have benefited from the project." (Report No 24)

5.2 Effectiveness

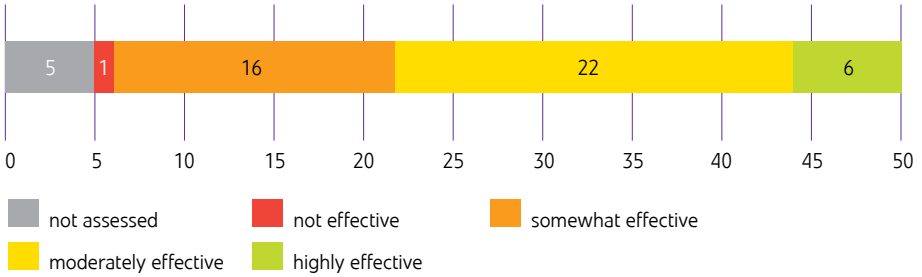
Highlights of the chapter and summary answer to EQ 11:

- Six out of 45 interventions (13%) are assessed as highly effective.
- A bit less than half of the interventions are assessed as moderately effective.
- About one third of the interventions is assessed as only somewhat effective.

The DAC criterion Effectiveness describes the extent to which the development intervention attains its objectives (OECD 2017b).

Figure 30 shows that, according to the evaluation reports, over half of the interventions (28 out of 50, 56%) were considered as moderately effective (22) or highly effective (6). In 16 of the cases (32%), the evaluators ranked the interventions as somewhat effective. One intervention was considered as not effective at all and for 5 interventions (10%) effectiveness was not assessed.

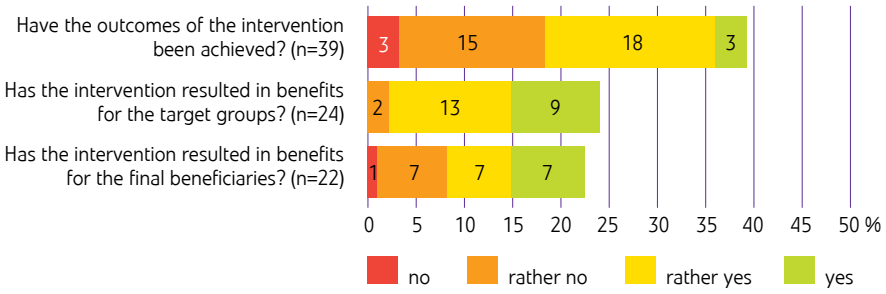
Figure 30: Effectiveness according to evaluation reports (n=50)



Source: own statistics based on analysis of reports

Beyond the overall assessment, we looked at the evaluators' judgement on the achievement of outcomes. In addition, answers on the attainment of benefits for (i) the final beneficiaries and (ii) the target group were synthesised. Figure 31 illustrates that 39 out of 50 reports (78%) provide an appropriate assessment on outcome achievement, while in only 22 out of 50 reports (44%) benefits for the final beneficiaries and in 24 out of 50 reports (48%) benefits for the target groups were adequately assessed. Again, as the latter two aspects have only been found in less than half of the evaluation reports under consideration for this assignment, the findings are limited to some broad tendencies.

Figure 31: Evaluators' assessment of different aspects of effectiveness



Source: own statistics based on analysis of reports

If assessed, just over half of the evaluators are positive regarding outcome achievement.

With respect to creating benefits for the target groups or final beneficiaries evaluators are more positive.

In 21 out of 39 cases, the question whether the intervention has achieved its outcomes has been answered with "rather yes" (18) or "yes" (3). For another 15 reports the evaluators responded with "rather no" and for three reports with "no". Thus, on a general note outcome achievement is rather mixed (21 positive vs. 18 negative cases).

The analysis of underlying reasons for positive assessment reveals three justifications emerging most frequently:

(i) improved capacities of target groups,

(ii) positive influence at policy level, and

(iii) improved conditions in the specific case of water and sanitation.

Other reasons include

(iv) improved cooperation among stakeholders,

(v) enhanced communication and partnerships,

(vi) better education, and

(vii) improved environmental management.

Common reasons for a negative assessment include

(i) shortcomings in the monitoring and evaluation (M&E) system (e.g. monitoring not implemented, existing baseline assessment not followed up, and lack of data collection),

(ii) poor project design,

(iii) political instability in the target countries (mentioned a few times), and

(iv) natural disaster (mentioned once).

Several interventions were perceived by the evaluators as rather not achieving their outcomes because means of verification were missing to allow assessment. Box 2 lists some examples how evaluators justified their assessment of outcome achievement.

Box 2. Examples of reasons for the assessment related to outcome achievement

- "Capacity development is a focus for the Community Led Accelerated WASH [Water supply, Sanitation and Hygiene] in … (COWASH) implementation approach at all levels. The effectiveness of capacity building by COWASH is considered as very good by all stakeholders met and also by the Training Impact Research commissioned by the Project to evaluate the impact of training and capacity development in … regions." (Report No 7)
- "The Mid-Term Evaluation (MTE) found that the Forest and Farm Facility (FFF) implementation is on track in achieving its outcomes. The supported Forest and Farm Producer Organisations (FFPOs) are engaging through their apex organizations, and are able to include their issues on political agendas. FFPOs also made notable progress in strengthening their capacity to engage in business and to participate in forest and farm based value chains through inclusive business model." (Report No 35)

- "Project records show that 71% of Village Development Committees/Municipalities in the project districts are declared open defecation free." (Report No 43)
- "By the time of the evaluation in February 2016, most … achievements can be characterized as outputs rather than outcomes. Consequently, there is a gap between what has been produced and the expected impact. Project actors seem to be aware of this and underline the importance of activities in 2016 in closing the gap towards strategic results and impact. The assessment of the result gap is complicated because the Project has not systematically monitored the performance at the outcome level." (Report No 26)

As a number of evaluators did not follow the input-output-outcome-impact logic and hence did not report on outcome achievement, the meta-evaluation team further searched for evaluators' assessment of benefits produced. Results with regard to benefits for the target group are overall positive, as illustrated in Figure 31. Out of 24 reports, 9 interventions are assessed as beneficial and 13 as rather beneficial, while only 2 were judged as rather not beneficial.

Underlying reasons for the positive assessments are the following:

(i) Improved technical, institutional and/or managerial capacity (by far the most frequent explanation),

(ii) empowerment, for interventions where the target groups were at the same time final beneficiaries (e.g. increased sense of pride, self-esteem and visibility),

(iii) economic and/or financial improvements (e.g. positive welfare outcomes, additional resources leveraged, growth of micro/small/medium entrepreneurs),

(iv) improved service delivery,

(v) deepened partnerships,

(vi) enhanced knowledge management, and

(vii) improved governance.

Reasons for negative assessment were

(i) uneven achievement of results within project components or among geographical regions and

(ii) insufficient and/or inappropriate planning of activities (e.g. lack of capacity development plans or gender analyses).

Box 3 displays some examples.

Box 3. Examples of reasons for the assessment related to benefits for the target groups

- "Climate change negotiators have improved understanding of United Nations Framework Contract on Climate Change (UNFCCC) high profile topics." (Report No 33)
- "There are real differences between Member States, as to how they perceive the issues of HIV prevention and drug demand reduction. In this area, the … Programme needs to be more effective at advocating change and then supporting participating Member States." (Report No 8)

Those reports that provide a discussion of benefits for the final beneficiaries (22) suggest that roughly one third of the interventions rather did not generate benefits, one third rather produced benefits and another third is assessed to have been definitely beneficial (7 each). Only one intervention was assessed as not at all beneficial. Figure 31 thus underlines a clearly positive assessment for two thirds of the interventions evaluated.

Among the underlying reasons for positive assessment, evaluators mentioned

(i) changes in attitudes and awareness either by the beneficiaries themselves or by other stakeholders that have an influence on them, (e.g. improved commercial attitude, growing cultural acceptance, or positive masculinity),

(ii) empowerment (e.g. increased participation of indigenous peoples or increased self-confidence to stand for elected position),

(iii) improved service provision (e.g. better education, improved services, and access to electricity),

(iv) improved financial and/or technical support to vulnerable groups (less prominent),

(v) enhanced health (few interventions),

(vi) improved water and sanitation situation (few interventions), and

(vii) improved literacy (few interventions).

Reasons for negative assessment include that results were

(i) geographically or demographically unevenly distributed (e.g., the most remote areas neglected, women or vulnerable groups addressed less than others) (most common),

(ii) insufficient duration of the intervention, and

(iii) use of inadequate technologies or approaches (e.g. poor operation and maintenance of facilities, inability of final beneficiaries to afford offered services, or use of loans rather than grants in that specific context).

Some examples are provided in Box 4.

Box 4. Examples of reasons for the assessment related to benefits for the final beneficiaries

- "Client in-depth interviews in Balkh indicated that … has good relationships with the community and with religious leaders and that through these leaders 'our men can get important information about family planning which is not against Islam'. This was reiterated by a Ministry of Religious Affairs official who emphasised close links between … and religious leaders: 'The main thing about … is that they have got the religious leaders' support. When the mobile clinics of … are going somewhere, the religious leaders are there to help them especially in the case of resistance from community members. The community members accept everything said by a religious leader.'" (Report No 8)
- "Most informants suggested there are no issues of discrimination for clients accessing … services – that all people are able to benefit equally from their services. The main barriers to access identified were cost (of the service and/or transport) and geographical location of services. Almost all informants requested that … services be extended to more rural and remote areas and other provinces, as currently they are only available in more urban and 'wealthier' areas." (Report No 6)

5.3 Efficiency

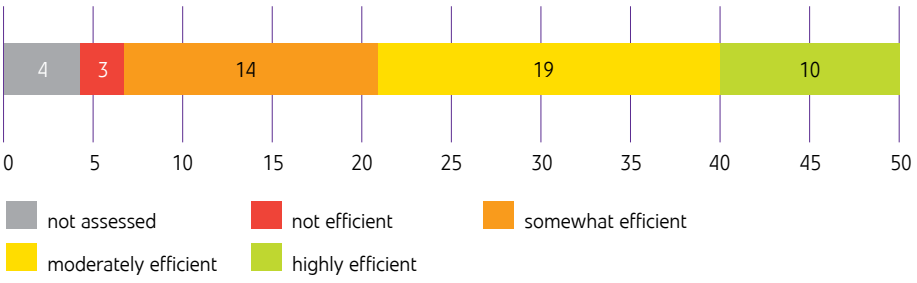
Highlights of the chapter and summary answer to EQ 12:

- Ten out of 47 interventions (21%) are assessed as highly efficient.
- 40% of the interventions (19) are assessed as moderately efficient.
- Just under one third of the interventions is assessed as only somewhat efficient.

The DAC criterion Efficiency measures quantitative and qualitative outputs in relation to the inputs used. Put differently, it measures whether the aid uses the least costly resources possible in order to achieve the desired resu2lts (OECD, 2017b).

As Figure 32 highlights, according to the evaluation reports, interventions were considered "moderately" or "highly" efficient in the majority of cases (29 out of 50, 58%). In contrast, a considerable number of interventions (14, 28%) were assessed as only "somewhat" efficient and in three cases the judgement of the evaluators was "not efficient at all". For 4 interventions efficiency has not been assessed (8%).

Figure 32: Efficiency according to evaluation reports (n=50)



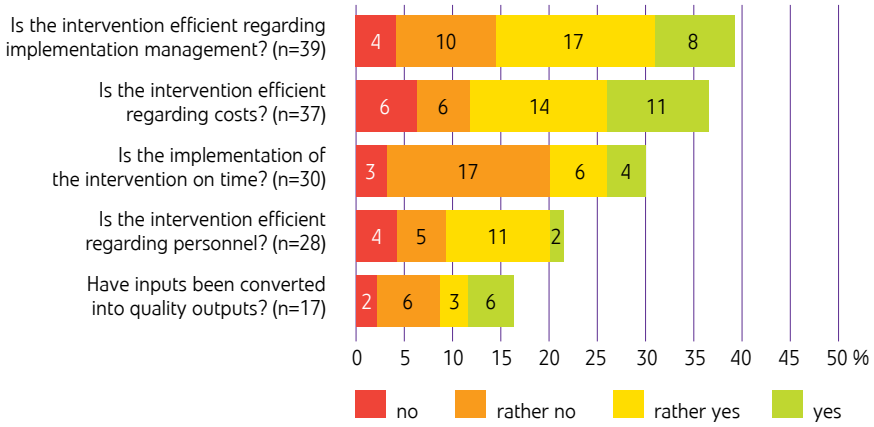
Source: own statistics based on analysis of reports

Regardless of which aspect of efficiency has been assessed a considerable number of interventions turn out to be (rather) inefficient.

In contrast if assessed, roughly two thirds of the interventions are (rather) efficient in terms of implementation management, costs and personnel.

Not all of the evaluators who appropriately evaluated efficiency assessed the same aspects of this criterion. Figure 33 shows whether and how the evaluators answered a set of questions. In general, reports mostly focused on management, cost and time efficiency aspects; less frequently on issues related to personnel and the quality of outputs. Overall, the different aspects display a mixed picture regarding efficiency. Figure 33 shows that time efficiency is rated rather low, whereas cost and management efficiency score significantly better.

Figure 33: Evaluators' assessment of different aspects of efficiency



Source: own statistics based on analysis of reports

Figure 33 further illustrates that out of 39 reports which provide an answer to the question whether the intervention is efficient regarding implementation management, 17 interventions were assessed as rather management efficient and eight as management efficient. Thus, roughly two out of three assessments were at least rather positive. 10 interventions were judged as rather not management efficient and four as not management efficient at all.

The analysis of underlying reasons reveals that the management of interventions was often rated as efficient when evaluators found

(i) solid management structures in place (e.g., a clear set of responsibilities and tasks, ideally coupled with a results-based management approach),

(ii) when there was good oversight through functioning steering committees and reporting (M&E) systems,

(iii) when qualified staff was recruited, and worked in small and efficient teams, and

(iv) when there was a good communication and coordination with the commissioner and other partners.

On the other side, management was found to be inefficient in cases where

(i) plans or structures were contradictory and non-transparent plans (at times those deficiencies where already found in the early project documents of the interventions),

(ii) in interventions which lacked oversight, control and strategic guidance (e.g., when steering committees were not yet in place or not functioning),

(iii) when there was no or only a limited M&E system in place, and,

(iv) in cases where coordination with partners was difficult.

Box 5 provides an overview of the broad range of factors and reasons highlighted in the evaluation reports.

Box 5. Examples of reasons for the assessment of efficient management

- "The extent to which the Steering Committee actually provides meaningful strategic direction is unclear. There are many other meetings and briefings occurring at the CND [Commission on Narcotic Drugs] at the same time. The meetings are relatively short and the agenda is driven more by UNDOC Staff than Member States. However, the chance to provide political oversight is extremely helpful and greatly appreciated by the Member States, and should be considered as good practice." (Report No 8)
- "The inception phase of three months has been very efficient and as a core result the SNE II implementation and management structure is in place and functioning. Especially the core teams at the REBs, woredas, and RCs levels are working effectively." (Report No 25)
- "Flexibility was also called for at several stages of the project due to substantial changes in funding. The project has shown excellent adaptive management capacity in the way it has dealt with these unexpected setbacks. This has also been evident in the way the project reorganised the original five components of the project (which are overlapping in nature) in work streams that relate to more discrete activities such as the work on LCAs in livestock, policy support activities, and the work related to gender in CSA. The link with the original sub-components, however, was maintained in both the Project Implementation Plans and the semi-annual progress reports." (Report No 41)
- "The project has successfully involved its target population in designing, planning, implementing and monitoring of project activities. Project's efficiency is also increased because complaints and feedback are taken positively and resolved promptly. The project is efficient because it also adopted tried-and-tested approaches from Phase I, thereby saving time and resources and reducing the risks of failure." (Report No 43)
- "The Steering Committee (SC) and Supervisory Board (SVB) performed less than expected. Perhaps the biggest problem was the lack of motivation to supervise and monitor project implementation and the inefficiencies shown in the decision making process." (Report No 5)
- "The option of indefinite service contracts or alternately the use of a company contract to supply the positions would have offered significant operational efficiency gains relative to the approach using individual contracts." (Report No 10)
- "The development of the SSU-CCO offices has not led to gains in efficiency but has, despite some good staff members, to confusion around the relationship between the IA and the SSU. There are instances of SSU staff being involved in the direct interview of IA staff which really blurs boundaries." (Report No 11)
- "The weight of UN procedures is a constraint on an otherwise very well managed programme." (Report No 46)

The assessment of cost efficiency is, in general, similarly positive. Out of 36 interventions for which cost efficiency was appropriately evaluated, 14 have been assessed as rather cost efficient and 11 as cost efficient. In turn, one third of the interventions was assessed as rather not cost efficient or not cost efficient (six each).

Evaluators grounded their positive assessments on factors which led ultimately to reduced costs for the interventions such as

(i) optimised procedures,

(ii) solid management systems (in particular with regard to financial management and controlling), and

(iii) successful recruitment of qualified and competent staff or

(iv) attracting additional resources, be it from other donors, implementing partners or even from target groups or beneficiaries (less prominent).

On the other hand, negative assessments were often caused by the fact that interventions had financial problems, sometimes from the start. These include, but are not limited to,

(i) unexpected or higher costs,

(ii) budget cuts,

(iii) a low budget indicating suboptimal resource planning from the beginning,

(iv) the lack of a sound financial management or auditing system to detect and correct financial bottlenecks,

(v) overspending, and

(vii) in a few cases spending high amounts of money to achieve relatively limited outputs (e.g. only benefitting a few or a restricted target group).

The examples presented in Box 6 showcase some of the aspects mentioned.

Box 6. Examples of reasons for the assessment of cost efficiency

- "COWASH employed high community contribution to optimize the use of available resources and reduce the cost per beneficiary while keeping the same level of outcomes. Community contribution is in terms of unskilled labour, local materials provision, road construction, venue provision for drilling crews. Apart from building local capacity and enhancing ownership, the use of Woreda offices and their technical personnel to undertake capacity building training to the WASHCOs was a commendable approach to reduce cost." (Report No 7)
- "Despite delayed beginning and some confusions among the stakeholders at the beginning on how the program should be implemented, it has demonstrated itself as a successful program managed primarily by Nepali institutions, with significantly lower management costs (6%) compared to expert-dominated models of the past, and also allocating 80% of the money to local level." (Report No 11)
- "According to the evidence obtained through the analysis of project documents and interviews, careful assessment of forthcoming interventions has allowed the Project to avoid unnecessary expenditures. In fact, the Project has achieved significant

savings that have allowed it to extend Project duration and fund additional activities. One example of such efficiency is the experimental approach to cultivating saplings for reforestation purposes, whereby two different methods (planting vs. sowing) were tested for highest survival rates. By identifying the most effective method with minimal costs, the Project avoided the risk of possible failure and respective loss of Project funds. Likewise, the Project has opted to target the most damaged plot with the lowest probability of self-restoration by natural processes, increasing the value per dollar invested." (Report No 23)

- "The programme has consequently been efficient by triggering funds from other donors, triggering counterpart funds for initiatives of interest (for example with UN Women), and generating capacities which others can then use without incurring the original investment. This has made it, from a donor perspective, cost-efficient, and highly relevant to a number of other donor programmes." (Report No 46)
- "Private sector service delivery and technical assistance has focused on establishment of youth and women's groups. Numbers of beneficiaries are low (98 youth and 19 women). As an indication of efficiency it cost approx. Euro 50,000 to support 98 youth as service providers (not counting the technical support). Assuming all youth are successful, this represents an average cost of approximately Euro 500 per individual. This figure is considered high." (Report No 17)
- "Financial record keeping has not yet been computerised and is done in hand-writing. This procedure delays the preparation of the financial reports at all levels and is prone to mistakes." (Report No 44)
- "With both the initial project design and the initial composition of the PIU very technically oriented, it is perhaps no surprise that the project management failed to see the need for more attention for marketing and management during phase I. It is nevertheless a point of great concern that an irrigation project can continue for several years on the basis of a purely technical approach without neither the PIU [Project Implementation Unit] nor the PSC [Project Steering Committee] fully realising the need to address the marketing and management issues. Apart from the bureaucratic delays, it is one of the main reasons why the cost effectiveness of the project is so low." (Report No 2)

In contrast, the question "Is the implementation of the intervention on time?" has been answered negatively for two thirds out of 30 reports. For 17 interventions the answer is "rather no" and for another three it is "no", whereas for only six interventions it is "rather yes" and for only four it is "yes".

Overall, evaluators found interventions to be on time when

(i) funds were disbursed easily and

(ii) when the available financial resources were adequate to the nature and challenges of the intervention.

On the other hand, reasons for a negative assessment turned out to be

(i) deficiencies in the areas above,

(ii) high administrative burdens, overly bureaucratic or inefficient procedures and structures,

(iii) delays or unsatisfactory results in staff recruitment (less prominent), and

(iv) factors beyond the control of the intervention such as the political context or natural disasters.

Box 7 provides some illustrative examples.

Box 7. Examples of reasons for the assessment of time efficiency

- "Flexible disbursement procedures allowed to respond in time to emerging issues and helped to reduce micromanagement by DPs [Development Partners]." (Report No 44)
- "[The programme] in Nepal in general, is widely acknowledged for its timely allocation and disbursement of budget resources and consequent implementation and completion of projects as planned. Due to Finnish budgets, this has been possible in spite of systematic delays in the availability of GoN [Government of Nepal's] budget resources." (Report No 43)
- "Contributing factors to delays are related to capacity in the different implementing organizations, the institutional arrangements of the programme (i.e. many implementing partners and the combination of having separate implementation and technical assistance budgets) and the selection of commodity VCs [value chains] (onion and potato) where processing and market potential is questioned." (Report No 17)
- "Planned expenditure was off target almost every year, revealing low capacity to organize activities and budget. It was not clear why this happened in a recurrent way every year; lack of skills or expertise at the MCT [Ministry of Science and Technology] and lack of initiative from the TA [technical assistance] component to adjust planning to real expenditure, could have been the reasons of these discrepancies." (Report No 5)
- "The low cost- and time effectiveness can be attributed to 3 main factors that are largely outside the control of the … project management: (i) the initial project design which was very technically oriented without much consideration for marketing and management aspects; (ii) the initial budgets for construction were up to 80% under-estimated; and (iii) the bureaucratic procedures for procurement and other important decisions, the direct result of a hybrid management system that had to comply with both GRZ [Government of Zambia] and AfDB rules and regulations. It is clear that for future projects in support of the irrigation sector, these 3 issues require specific attention if the projects are to be cost and time effective." (Report No 2)
- "However, delay in execution due to multiple factors including compliance with reporting obligations and subsequent disbursements as well as political and contextual factors caused an increase in project expenses for some implementing partners. For example; the mapping and training of aspirants were delayed as political parties struggled to provide the list of aspirants to the implementing partners." (Report No 55)

The majority out of 22 interventions were assessed as rather efficient as to staffing (11) or efficient (two), while five interventions were assessed as rather not efficient and another 4 as not efficient. As this aspect is discussed in less than half of the evaluation reports (22 out of 50), results are limited to provide some hints.

Finally, assessments regarding the conversion of inputs into high quality outputs were synthesised. Unfortunately, only one third of the evaluation reports (17 out of 50) provide insights on this aspect. Similarly, results cannot go beyond providing exemplary insights. For those 17 reports, assessments are mixed, with roughly half of the interventions being assessed (rather) positive and the other half as (rather) negative (nine vs. eight).

Due to the limited number of cases, the analysis of underlying reasons for assessments regarding efficiency of personnel and quality of outputs is inconclusive. Nevertheless, a few examples are presented as anecdotal evidence in Box 8.

Box 8. Examples of reasons for the assessment of efficiency of personnel and quality of outputs

Efficiency of personnel

- "Although the core team members each have their specific areas of responsibility, they are all well informed about the other components of the project and can, where necessary, contribute to activities that are not part of their core area of expertise. […] The team also doesn't hesitate to call in external expertise from other FAO divisions or from outside of the organization when they feel they don't have the right qualifications to provide the support themselves." (Report No 41)
- "Many expressed concern however regarding the frequent turnover of staff at all levels, which slows down implementation as new staff need time to get up to speed and existing staff are required to provide repeated briefings." (Report No 53)

Quality of outputs

- "The FE [final evaluation] believes that in Ecuador and Tanzania where countries' projects have already completed their planned activities, the Programme has achieved reasonably good value for money. In Viet Nam and Zambia, respective projects have a reasonable likelihood of high efficiency, but more for physical results that for their timeliness. The Programme was efficient in making available resources to the five partner countries projects in conformity with their work plans. The resources disbursed allowed projects to achieve high activity exec8ution rates." (Report No 20)
- "The continuous disruptions from its original plan and conception to the weak presence of qualified human resources have influenced its capacity to transform the available resources into the required output/results, both from quality and quantity point of views." (Report No 5)
- "Action plans, business plans and applications are generally of a low quality suggesting poor conversion of available resources." (Report No 17)

5.4 Impact

Highlights of the chapter and summary answer to EQ 13:

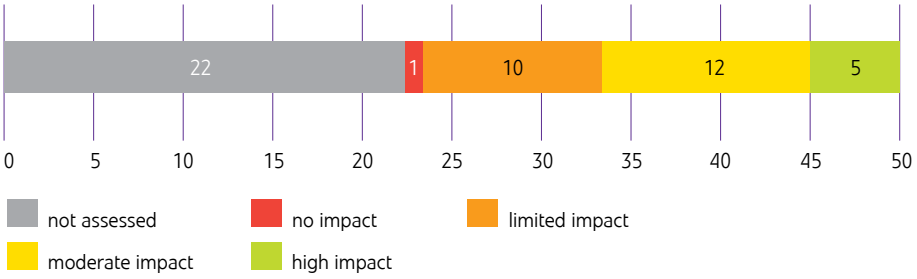
- Impact is assessed for just over half of the interventions (28 out of 50).
- Five out of 28 interventions (17%) are assessed as having a high impact.
- 42% of the interventions (12) are assessed as having only a moderate impact.
- About a third of the interventions is assessed as only having limited impact.

The DAC criterion Impact measures the positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators (OECD, 2017b).

Figure 34 shows that only 28 out of 50 reports (56%) provide an appropriate assessment of impacts. Thus, results of this section have to be taken with care as they are not representative of the sample for this assignment. For nearly half of the interventions under consideration, we do not know anything regarding impact.

Among those ones which do include an impact analysis, only one intervention was considered as having no impact at all. However, more than one third of the cases (ten out of 28) were assessed as having only some impact. For 12 cases, the evaluators reported moderate impact, and for five cases high impact. Thus, of 28 reports assessing impact a slight tendency to the positive side can be observed.

Figure 34: Impact according to the evaluation reports (n=50)



Source: own statistics based on analysis of reports

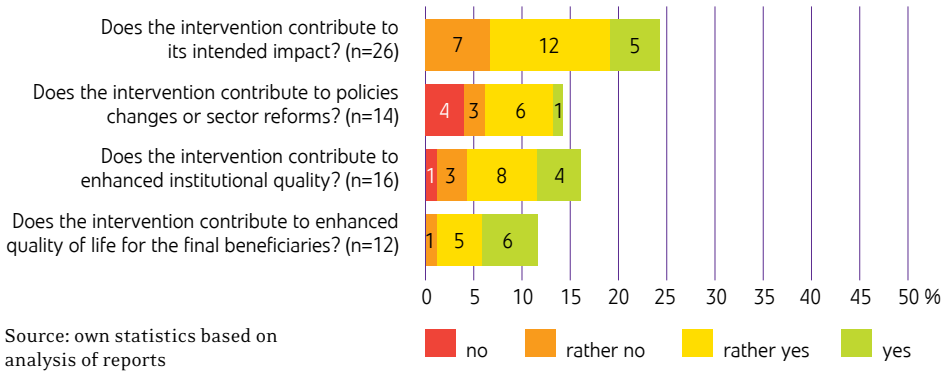
As displayed in Figure 35, we differentiated four aspects of impact: contribution to intended impacts, contribution to policy changes or reforms, contribution to enhanced institutional quality, and contribution to enhanced final beneficiaries' quality of life. Once again, results are indicative at best given the small number of evaluation reports assessing these aspects.

In general, from the 24 evaluation reports which assess whether or not interventions have contributed to their intended impact, five provide a positive answer. Half (12) state that the interventions evaluated have rather contributed to their impact, whereas seven interventions did rather not contribute according to the evaluators.

Enhanced institutional quality is the most successful kind of impact, with 12 out of 16 interventions assessed as having rather contributed (eight) or contributed (four) on it. Contribution to policy changes or reforms and contribution to enhanced quality of final beneficiaries' lives present a mixed picture, with equal shares of (rather) positive and (rather) negative assessments (7 vs. 7, 6 vs. 6).

Only some evaluators assessed different aspects of impact. If assessed, ratings are more often on the positive side.

Figure 35: Evaluators' assessment of different aspects of impact



Source: own statistics based on analysis of reports

As reasons for all three aspects were provided very rarely and as the number of interventions which have been assessed is very low, further analysis is not very productive. However, some reasons are presented as examples of the variety of explanatory factors.

When the reports provided information on how intervention contributed to changes in the partner country's/region's policies or to sector reforms, reported changes related to agricultural, land, and education policies. In addition, gender equality and food security agenda were mentioned among others. The reasons explaining these achievements included factors such as another donor's additional support, government ownership and political commitment, as well as citizen engagement.

Positive factors related to contribution to enhanced institutional quality included integration of indigenous languages (in the case of interventions related to national education systems) and inclusion of environmental aspects (in the case of interventions related to municipal planning systems). Among reasons for negative assessment were cited lack of inter-institutional coordination and failure to translate institutional changes into concrete action.

Positive factors related to impact on final beneficiaries' quality of life included improved economic situation and better health. Negative reasons included lack of financial stability at household level and uneven distribution of results within the final beneficiary population. Box 9 presents some examples.

Box 9. Examples of reasons related to the assessment of different aspects of impact

Contribution to policy changes or reforms

- "The strategy of … managed to introduce the topic of food and nutrition security in the agendas of the institutional actors at three levels: regional, national and local. It benefited from favourable circumstances in 2012, supported by the political legitimacy of the issue of food and nutrition security and a consolidation of regional integration." (Free translation from Spanish, Report No 30)

Contribution to enhanced institutional quality

- Another important educational decision related to … was the creation of the "Institutos de Lengua y Cultura" for each indigenous nation. There are sixteen Culture and Language Institutes (ILC) currently working to rescue the knowledge and culture of the indigenous groups. Many indigenous researchers trained by … at the Universidad de San Simón, Cochabamba (Bolivia), are now part of the Institutes of Language and Culture that work on promoting Intercultural and Bilingual Education." (Report No 3)

Contribution to enhanced quality of life of final beneficiaries

- "The likelihood of FFF's rural poverty impact can be assessed by considering the extent to which FFF small grants, trainings, and other interventions are likely to contribute to improved livelihoods of target groups from forest and farm management. To this end, the main livelihood "building blocks" that are analyzed for likelihood of impact relate to human, social, and political capital, and to natural, financial and physical assets. By improving these building blocks, the FFF improves the long-term resilience of target smallholder farmers and communities. Field level observations by the MTE team revealed impressive progress made in these domains." (Report No 35)

5.5 Sustainability

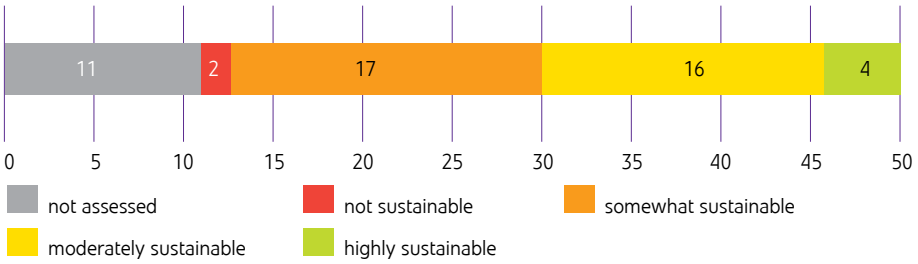
Highlights of the chapter and summary answer to EQ 14:

- On average, one out of three interventions is assessed as moderately sustainable.
- On average, one out of three interventions is assessed as only somewhat sustainable.

The DAC criterion Sustainability is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn (OECD, 2017b).

Figure 36 illustrates that in 20 out of 50 reports (40%) interventions were considered highly sustainable (4) or moderately sustainable (16). Seventeen of the interventions (34%) were evaluated as moderately sustainable. Two projects (4%) were considered to be not at all sustainable and 11 evaluations (22%) did not assess sustainability. Thus, the synthesis of those reports which assess the criterion suggests a mixed picture with a nearly equal number of moderately sustainable and only somewhat sustainable interventions (16 vs. 17).

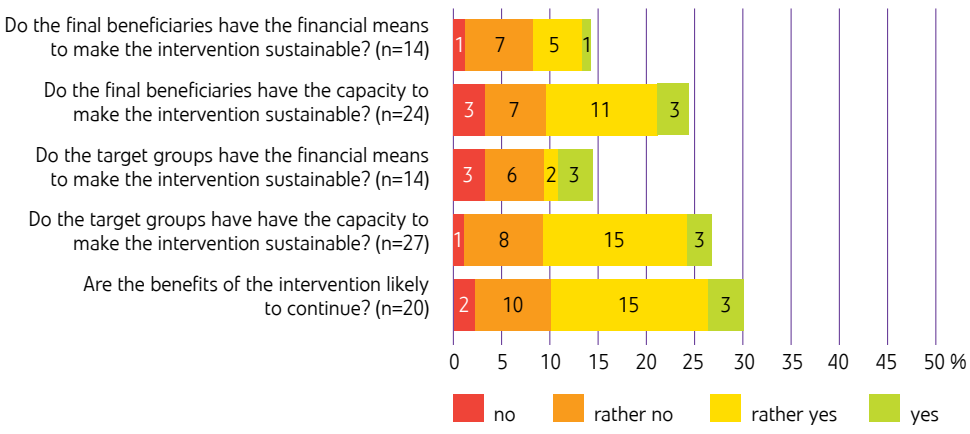
Figure 36: Sustainability according to the evaluation reports (n=50)



Source: own statistics based on analysis of reports

Not all of the reports which treat sustainability evaluate the same aspects of this criterion. Figure 37 illustrates, aspects related to "beneficiaries' capacity, "target groups' capacity," and overall "likelihood of continuation of benefits" were assessed more frequently (24, 27 and 30 out of 50, respectively) than "availability of financial resources to make the intervention sustainable" (14 out of 50). Once again, due to these low numbers, results are limited to provide some hints.

Figure 37: Evaluators' assessment of different aspects of sustainability



Source: own statistics based on analysis of reports

If assessed, financial means of the target groups or final beneficiaries are more often assessed as threat to sustainability than capacity.

Still, if assessed evaluators are clearly more often positive than negative regarding the continuation of benefits.

Out of 30 reports providing an answer to the question whether the benefits of the intervention are likely to continue, the great majority is (rather) positive (18). About one third of the evaluators (10) respond to the question with "rather no" and in two cases the benefits are assessed as ceasing with the termination of donor support.

A look at the capacity of target groups to ensure intervention sustainability reveals a similarly positive picture, with two thirds (18 out of 27) of the interventions assessed as (rather) positive and one third (9 out of 27) as (rather) negative. Although a bit less distinct, an overall view at the final beneficiaries' capacity turns also out positive with 14 out of 24 interventions judged as (rather) sustainable and 10 out of 24 as (rather) not sustainable.

In contrast, the synthesis of the financial means shows that for the majority of interventions which have been assessed in this regard, target groups (9 out of 14) and final beneficiaries (8 out of 14) are assessed as (rather) not having the financial means to make the intervention sustainable. In turn, for 5 out of 14 interventions' target groups were assessed as (rather) having the financial means and for 6 out of 14 interventions 'final beneficiaries were assessed as (rather) having it. Thus, data suggests a lack of financial means among target groups and final beneficiaries seems to be more frequently threatening the sustainability of an intervention than a lack of technical capacity.

Reasons provided in evaluation reports for positive assessment of overall sustainability include:

(i) engagement of government counterparts (e.g., implementation through local structures or alignment with government priorities),

(ii) stakeholder participation (e.g., beneficiary involvement in decision-making and clear demand from beneficiaries),

(iii) adequate arrangements with implementing partner (e.g., committed organisations, intervention integrated in the organisation, stipulation that trained staff will remain after intervention ends),

(iv) good market demand for products (few cases),

(v) embeddedness of interventions in ongoing activities (few cases),

(vi) improved legal frameworks (few cases), and

(vii) enabling environment (few cases).

Negative reasons provided can be systematised as follows:

(i) dependence on continued external financial and technical assistance (by far most frequent),

(ii) high staff turnover (in government or implementing partner institutions),

(iii) too short-term interventions, and

(iv) a top-down approach to implementation.

Box 10 provides some examples related to sustainability assessment.

Box 10. Examples of reasons related to the assessment of the sustainability of interventions

- "With regards to specific Programme activities, the sustainability of the Programme is partially encouraging is certain respects, although questions still remain. For example, a number Information Desks have been established with the support of the RoLHR Programme and one staff cost for each of the 15 pilot district Information Desks is funded through the Programme. The government funds all additional operational costs associated with the Information Desks. The government will gradually absorb the staffing costs for the Information Desks, since this is one of the priority activities of the Third Five Year Strategic Plan of the Judiciary (2014/15–2018/19)." (Report No 42)
- "However, the results in many partner countries remain fragile and a continued technical and financial support will be still needed. The main obstacle to sustainability of results achieved up to now is the absence of modalities to ensure long-term financing for addressing continuous inventory, particularly in countries with decentralized political systems, where forest resource management responsibilities may be strongly decentralized." (Report No 20)

5.6 Gender and other cross-cutting objectives

Highlights of the chapter and summary answer to EQs 15–18:

- Finnish development cooperation is neither gender-blind nor gender-transformative, but somewhere in between.
- Eight interventions were assessed as gender-mainstreamed and four focus on gender equality and women's rights, whereas 15 interventions were only assessed as gender-aware
- Assessment of other cross-cutting objectives was not possible given the lack of analyses in the majority of reports.

We limited our assessment of integration of cross-cutting objectives to the gender equality and women's empowerment (GEWE) aspect as only for this cross-cutting objective more than half of the reports (36 of 50) integrated the objective throughout the report to a degree that allows a viable and systematic review (see discussion on other cross-cutting objectives in the end of the chapter 5.6). The concept underlying GEWE is sufficiently inclusive to adequately capture MFA's philosophy of gender equality mainstreaming and women's and girls' rights as well as to cover the variety of terms used in the evaluation reports at hand.

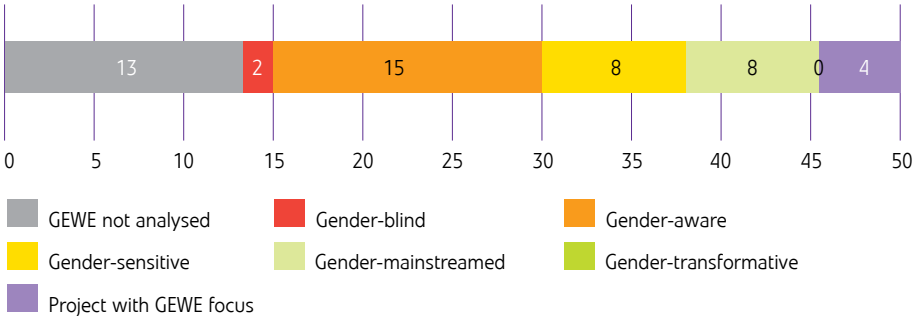
As mentioned earlier in chapter 3.3, GEWE has been one of the key priorities in Finland's development cooperation in all published development policies. Hence, it is valid to conclude that all interventions considered here should have placed emphasis on the topic, and that all evaluators should have acknowledged GEWE in their assessments. It has already been shown in the quality assessment in chapter 4.6 that the latter is not the case. It was impossible for the meta-evaluation team to determine whether in such cases the evaluators or the interventions failed to integrate GEWE. As a consequence, 13 evaluation

reports have been excluded from further analysis. Furthermore, we excluded four evaluation reports of interventions with a main focus on GEWE as in such cases this is no more a cross-cutting objective.

The 33 remaining reports were classified into five different categories according to the level of inclusion of GEWE in the interventions as assessed by the evaluators: gender-transformative, gender-mainstreamed, gender-sensitive, gender-aware and gender-blind following the definitions applied commonly by a wide range of development actors, including a recent report by the Independent Evaluation Office of the Global Environment Facility (GEFIEO, 2017).

As displayed by Figure 38, roughly half of the interventions were assessed as only gender-aware (15 out of 33), whereas the other half (16 out of 33) were assessed in equal shares as either gender-sensitive or gender-mainstreamed (8 apiece). Two of the interventions were classified as gender-blind and none as gender-transformative.

Figure 38: Classification of GEWE (n=50)



Source: own statistics based on analysis of reports

Thus, the analysis reveals a mixed picture: on average Finnish development cooperation is neither gender-blind nor gender-transformative, but somewhere in between. Given the prominent GEWE focus in Finnish development policy, the share of solely gender-aware interventions is rather high. In contrast, the fact that roughly one quarter of the interventions are assessed as gender-mainstreamed and that Finland designs a number of interventions with a main GEWE focus reflects the strong attention given to GEWE aspects. As the analysis was limited to roughly two thirds (33) of the 50 evaluation reports, results interpretation has to be taken with care.

With regard to the other cross-cutting objectives, they were not integrated at all or only integrated sporadically by more than half of the 50 reports. More specifically, reduction of inequality/equal opportunities to participate/rights of the most vulnerable has been only integrated by 23 reports, climate sustainability/climate change preparedness and mitigation by only 21 reports and the human rights-based approach (HRBA) by only 14 reports. This result from the quality analysis already complicated a possible analysis as it does not allow generalisation within the sampling. Even worse, for interventions whose evaluation reports do not cover a particular cross-cutting objective it is unclear whether these interventions ignored the topic or whether the evaluators did not pay attention to it.

In conjunction with the above and in contrast to GEWE, the inclusion of the other cross-cutting themes or objectives has not been as continuous and systematic in MFA's policy guidance. Given the fact that there was no practical way of defining which particular policy framed every single intervention under concern for this analysis, we had to refrain for methodological reasons from assessing them. Such assessments would have been biased in many directions caused for example through arbitrary assignment of interventions to development policies or small sub-sample sizes with low explanatory power.

5.7 Aid effectiveness and triple C

Highlights of the section and summary answer to EQs 19–22:

- The assessment of aid effectiveness and triple C (i.e. coherence, coordination and complementarity) is not deeply anchored into Finnish development cooperation evaluation practice.
- It remains unclear if and to what extent the interventions under consideration follow one of these concepts.
- Thirty-one of the 38 interventions assessed promote ownership and 26 of the 29 interventions assessed align priorities with national or regional policies.
- For 23 of the 32 interventions assessed coordination is evaluated as rather positive.
- The promotion of management for results is about as often (rather) neglected as (rather) supported (19 vs. 17) in the 36 interventions where it is assessed.

The aim of this section is to provide some insights on aid effectiveness and on the implementation of the European's Union triple C. Although they share similarities with one another, the MFA and the meta-evaluation team agreed to look at both concepts. Thereby, we appreciate that the evaluation reports at hand are on interventions which were designed and implemented over a time span where first the one and later the other concept figured more prominently on the international development agenda.

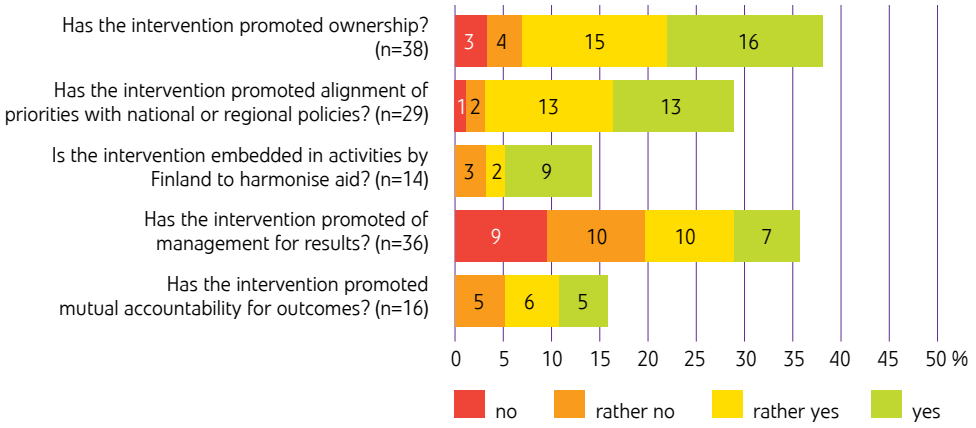
In line with the aid effectiveness agenda, we synthesised the insights from the evaluation reports according to the key dimensions as presented in Figure 39. It is important to keep in mind that roughly one quarter of the 50 evaluation reports (12, 24%) do not deal with aid effectiveness at all.

While roughly three quarters provide insights whether the intervention under consideration has promoted ownership and management for results, less than one third assess whether the intervention has been embedded in activities in order to harmonise Finnish aid and whether it has promoted mutual accountability for outcomes. Whether the intervention has promoted alignment of priorities with national or regional policies is answered in roughly 60% of the reports. Given the small number of reports capturing these aspects, findings are limited to some tendencies.

Promotion of management for results appears to be mediocre. In a quarter of the reports (nine out of 36) the intervention has been assessed as not promoting management for results and in another quarter (ten) as rather not. In contrast, according to the evaluation reports the great majority of interventions (31 out of 38) rather promote (15) or promote (16) ownership. A similar picture

can be drawn regarding the promotion of alignment of activities. For 26 out of 29 interventions the evaluators answer the question with "rather yes" or "yes" (13 each). If the embeddedness in activities by Finland to harmonise aid is assessed, results seem to be as well rather positive. This is also the case for the promotion of mutual accountability for outcomes. As pointed out earlier, particularly these last two insights are not representative due to the small number of reports providing evidence.

Figure 39: Evaluators' assessment of aid effectiveness

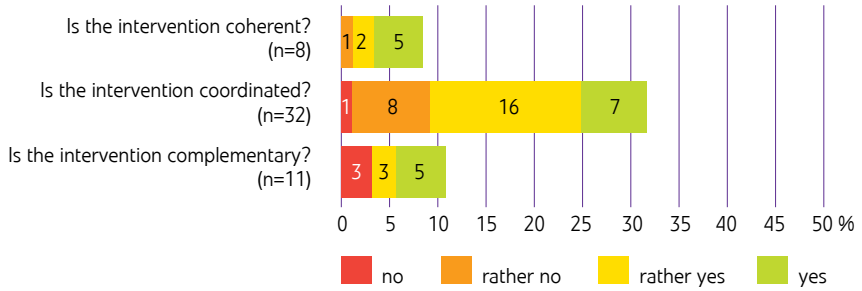


Source: own statistics based on analysis of reports

A closer look at the assessment of triple C reveals comparable challenges with regards to the coverage of this concept in the evaluation reports, as shown in Figure 40. Only eight out of 50 reports give an answer to the question whether the intervention is coherent. Only one intervention is assessed as rather not coherent, two are judged as rather coherent and the remaining five as coherent. In 11 out of 50 reports the complementarity of the intervention is assessed as follows: three interventions are judged as not complementary, another three as rather complementary and the remaining five as complementary. Once again, from the low number of reports treating these aspects, it does not become clear whether this rather positive picture also holds true for other interventions of Finnish development cooperation.

The empirical base to assess coordination is considerably better. This aspect is assessed for 32 out of 50 reports and again, the tendency is clearly positive. While only one intervention out of 32 is assessed by the evaluators as not coordinated and another eight as rather not coordinated, more than two thirds are judged as rather coordinated (16) or coordinated (seven).

Figure 40: Evaluators' assessment of triple C



Source: own statistics based on analysis of reports

5.8 Lessons learnt presented in the evaluation reports

Highlights of the chapter addressing EQ 26:

- Only 30 out of 50 evaluation reports contain lessons learnt.
- Just under half of the lessons learnt presented are in fact intervention-specific recommendations.
- "True lessons learnt" in accordance to the OECD DAC definition are spread over a wide range of different topics. Hence, no "typical" lessons could be identified.

In the course of the summative analysis a total of 211 lessons learnt in 30 out of 50 reports was identified. The remaining 20 reports did not include any lessons learnt. Nearly half of the lessons learnt presented in the evaluation reports (i.e. lessons that were given the score 1; 98; 46% as described in chapter 2.4) were formulated in a manner that would not allow using them in contexts extending beyond the specific intervention concerned. Thus, they are per definition no lessons, and were hence excluded from further analysis. The methodology with regards to the aggregation and synthesis of lessons learnt is described in more details in chapter 2.4. The remaining 113 lessons (54% of all lessons identified) are lessons that provide an added value for learning purposes in the wider development cooperation context (score 2: 30, 14%, score 3: 83, 40%).

Table 5 presents how many reports include lessons learnt by thematic category. The table extends from "Planning," for which a lesson is found in 11 out of the 50 reports assessed (22%), to "Efficiency", "Relevance", and "Time," for each of which a lesson could only be found in one report each. The table further shows that there is no single category prominently presented. The most prominent categories "Planning", "Sustainability" and "Participation" are only captured in 22%, 16% and 12% of the reports. Put differently, there can be no typical lessons identified in the course of this summative analysis. Nevertheless, the most relevant categories in terms of numbers and lessons perceived as interesting for Finnish Development Cooperation are presented, in the spirit of examples, in the following table.

Table 5: Number of reports including lessons learnt categorised under different themes (n=50)

Category	Number of reports	In % of all reports
Planning	11	22
Others	10	20
Sustainability	8	16
Participation	6	12
Capacity	5	10
Communication	5	10
Coordination	5	10
M&E	5	10
Management	5	10
Scope	5	10
Aid effectiveness	4	8
Effectiveness	4	8
Impact	3	6
Exit strategy	2	4
Financial	2	4
Gender	2	4
Efficiency	1	2
Relevance	1	2
Time	1	2

Note: Ten reports include lessons learnt that did not fit into a generalised category. Hence, they are summarised in the category "Others".

The analysis did not reveal typical lessons learnt.

The lessons under the category "Planning" typically encourage better engaging experts on substantive aspects of project planning, keeping expectations realistic and analysing risks as well as underlining assumptions in proportion to their importance. Other aspects include the importance of adapting projects to local situations and having a so-called "Plan B" or other flexible arrangements for adaptive management; involving stakeholders at planning phase, and allowing sufficient time for project preparation.

Most of the lessons captured under the category "Others" referred to technical lessons; e.g., regarding types of latrines that function well or the role of aboveground biomass in forest inventories. Evaluations also mention that in multi-country operations managed by UN organisations, in-country presence is important for successful implementation.

The lessons in the category "Sustainability" commonly referred to the importance of using existing structures, bottom-up planning and implementation, and avoiding dispersion of activities. Similarly, the lessons in the category "Participation" related to being realistic about time, scope, and ambition, as well as the importance of adaptive management. Two lessons stand out from the group as interesting examples. One report (No 24) mentions that local "verifiers" and village committees were used to ensure appropriate village-level project beneficiary selection. Another report (No 37) emphasised the importance

of recognising and building on the capacity and enthusiasm of local leading champions to ensure that change will happen.

5.9 Recommendations drawn in the evaluation reports

Highlights of the chapter addressing EQ 26:

- More than three quarters of the reports contain recommendations related to "M&E".
- More than half of the evaluation reports contain recommendations related to the intervention fields of "Planning", "Scope", "Management", "Capacity" and "Sustainability".

Recommendations of the evaluators often focus on M&E, management, scope, sustainability, capacity and planning of an intervention.

Throughout the sample of evaluation reports under consideration, sound recommendations are much more common than proper lessons learnt. The following table presents how many reports include recommendations specific to a given thematic category. The table spans from the category "M&E" which is found in 38 reports out of the 50 assessed (76%) to "Coherence" and "Complementarity" which could only be found in three reports (6%).

Table 6: Frequency of recommendations by broader category (n=50)

Category	Number of reports	In % of all reports
M&E	38	76
Management	27	54
Scope	27	54
Sustainability	26	52
Capacity	25	50
Planning	25	50
Coordination	24	48
Gender	23	46
Communication	20	40
Aid effectiveness	17	34
Personnel	17	34
Financial	15	30
Exit strategy	15	30
Effectiveness	15	30
Efficiency	13	26
Participation	11	22
Relevance	10	20
Time	9	18
Others (not captured above)	9	18
Equipment	5	10
Impact	5	10
Coherence	3	6
Complementarity	3	6

Categories of recommendations appearing in more than half of the evaluation reports are considered as typical and were further synthesised. The methodology with regards to the aggregation and synthesis of recommendations is

described in more details in chapter 2.4. They comprise "M&E", "Management", "Scope", "Sustainability", "Capacity" and "Planning".

Recommendations on planning are presented in 25 out of 50 reports. Furthermore, 27 reports are relating to the **scope** of Finnish interventions. Due to the interdependence of the two categories (scope being determined during the planning phase of interventions in most of the cases), both categories were jointly analysed.

In 15 reports evaluators recommended to review the planning of ongoing interventions or to improve planning activities of subsequent interventions, mostly in terms of project design and the Theory of Change (ToC). This suggests that the design of these interventions either had flaws and gaps from the beginning or that contextual changes required a review of the design over time.

This is very much in line with another cluster of recommendations which calls for raising awareness of the importance of planning in general, for institutionalising and better structuring of the planning process and for better supporting implementing partners and related institutions during the planning phase. The following recommendation exemplarily summarises the importance of institutionalised planning processes to avoid or mitigate problems from the start of the interventions: *"The current Manual for Bilateral Programs contains procedures, which if followed appropriately, would avoid many of the failings noted in the programming of …. This Manual is therefore in general recommended for its current task."* (Report No 5)

Furthermore, several evaluators recommended that planning and project design should be based on thorough situational analyses and risk assessments beyond mere formalities, as pointed out in one report: *"Particular attention must be paid to the risk analysis in a project document. They must be realistic and systematic assessments, instead of checklists routinely filled out. This may imply methodological development work from MFA's part."* (Report No 26)

In a few reports evaluators highlighted that planning should be realistic, especially with regard to budgets for the individual activities and phases. This is also in line with several recommendations made on the interventions' scope: ten reports include recommendations to narrow (or at least not increase) the geographical scope or the interventions' scope of activities. Furthermore, in several reports evaluators recommended to carefully assess whether an extension of the scope is actually in the best interests of the intervention, for example: *"While it is appreciated that the Programme has been extended to all provinces, an exit strategy shall carefully consider whether it is feasible to achieve sustainable results in all targeted provinces or whether it is better to achieve complete and solid results in a few provinces so that the Government may replicate these visible successes. […] For the integrated spatial planning, Technical Assistance focus should be on what can be completed fully and thus used as demonstration for those that may lack behind."* (Report No 4)

However, recommendations to decrease or at least maintain the scope of interventions appear to be the minority within the sample. As many as 16 reports contain recommendations to extend the scope of activities in terms of content;

seven reports include recommendations to extend the geographical scope and three reports call for extending the activities to other target groups or beneficiaries, as highlighted in the following example: *"Unmarried young people also need attention; current emphasis on 'young married women' should be expanded."* (Report No 6). Last but not least, in one report it was recommended to include and conceive measures for scaling up already in project design to ensure realistic planning while anticipating broadened scope over time.

Recommendations on the management of interventions are made in 27 out of 50 reports. Aspects analysed in this category are related to some sub-sections of the efficiency assessment (see chapter 5.x), particularly to implementation management. The recommendations provided are rather intervention-specific and hence, can be rarely generalised.

Broadly, two kinds of recommendations were identified. In 15 reports evaluators recommended changes to the organisational structure of the intervention, e.g. by creating new positions, merging or splitting units or shifting responsibilities and tasks. Another eight reports comprise recommendations on functional improvements of specific bodies within the interventions. A selection of these recommendations is presented as anecdotal evidence in Box 11 below.

Box 11. Examples for recommendations on the management of interventions

- "Develop a simple business and staffing plan for Pakse laboratory based on specific ESIA [economic and social impact assessment] monitoring needs in the province to initiate a minimum level of commercial sampling required for basic laboratory services and operationality." (Report No 4)
- "Risk analysis must include indicators and contingencies, which can trigger a warning and a response. The process of 'ex ante' risk assessment produces a minimized risk matrix, which gives rise to complacency that risks have been taken into account. This is the exact opposite of what risk management should do and indicators or trigger events should be built into programme results to be monitored." (Report No 5)
- "When UNDOC [United Nations Office on Drugs and Crime] introduces change, a change management plan should be included, which is supported by both internal and external communication plans and by Senior Management to alleviate potential barriers to implementation." (Report No 8)
- "A revised structure is proposed that seeks clear lines of authority for administrative, financial and technical decision making with accountability." (Report No 10)
- "In order to better inform Program Council members about the project selection process it is recommended that after each Project Selection Committee meeting a brief report is produced summarising (inter alia) the main reasons why some proposals were unsuccessful." (Report No 12)
- "The Secretariat should likewise consider appointing a full-time experienced knowledge management specialist to lead this work and help coordinate it with the broader PMR [Partnership for Market Readiness] Technical Work Program. In addition, the Secretariat should explore more effective ways of managing and disseminating relevant knowledge that exists outside the PMR and continue using external specialists for preparation of demand-driven Technical Notes and other knowledge products. Finally, the PA [Partnership Assembly] may want to consider establishing a specific Working Group to help guide and oversee the PMR's knowledge management and sharing activities." (Report No 16)

- "Clarify Roles and Responsibilities within the PSU [Programme Support Unit]: AgroBIG needs to review the roles and responsibilities of staff within the PSU. The roles of the Programme Director and the Chief Technical Advisor require review based on current implementation experience. The MFA and BoFED [Bureau of Finance and Economic Development] should facilitate this process. One role should lead the PSU and the other should bring and be responsible for appropriate technical advice and support. The TA [Technical Assistance] team should also work to support planning and reporting of the overall programme, as this will have positive feedback – in this way it is likely that delays will be minimised, and the TA team will be seen by the GoE [Government of Ethiopia] to have a more active role." (Report No 17)
- "When the contract is signed with the new Lead Consultant, his/her continuous presence at the project site should be ensured. This entails: a) that the contract is made on a 10/12 months basis, and b) that regarding actual working days and time, the contract is aligned with normal international practices." (Report No 25)
- "Project management and monitoring arrangements must include an appropriate role for the MFA, which enables its participation in timely decision making as well as receiving information." (Report No 26)
- "It is recommended that the donor agency should accept lump sum contracts for this type of grant programs. These contracts would be easier to manage by each party. Payments, e.g. in 3 – 4 instalments, could be made against milestones defined beforehand in the contract." (Report No 39)

Recommendations with regard to the capacity of implementing partners and beneficiaries are made in 25 out of 50 reports. Stakeholders' capacity is an important determinant for effectiveness, impact and sustainability as is also the degree of cooperation with and ownership by national counterparts.

In a considerable number of reports (12), the capacity of implementing partners is assessed as weak and therefore it is recommended to develop their capacity, in particular in terms of technical or thematic knowledge. Improving the capacity of final beneficiaries to make better use of the services delivered is also recommended, though only in a few reports. The same holds true for empowering beneficiaries and raising awareness for specific issues related to beneficiaries and/or vulnerable groups.

Recommendations to improve the quality of capacity development and training activities were raised in eight reports. Most of these recommendations concentrate on technical measures to improve capacity development activities via improved methodologies, better time management and better trainers or equipment. The following recommendation provides a good summary of general difficulties and limits of capacity development and complements this section on partners' capacity and ways to improve it: *"Recognise that capacity building in general takes time, and that capacity building for highly complex themes like climate change, where firmly entrenched development patterns need to shift, is very process based and immersed in a plethora of socio-political factors that a project cannot influence directly, and as such, requires capacity building approaches better synched to the timelines of these processes and their key actors, and with realistic expectations of what impact can be expected."* (Report No 33)

On a more general level, in earlier reports (published in 2015), evaluators tend to focus on the needs to strengthen partners' capacity and to compensate lack

of skills or knowledge related to it, while later reports more often contain recommendations on how capacity development could be made more effective.

Recommendations on sustainability are presented in 32 out of 50 reports. An overarching issue which is taken up by many reports (19) is the recommendation to develop exit or sustainability strategies. Such strategies or plans to accompany the phasing out are being recommended in the MFA Manual for Bilateral Cooperation, yet a fair number of interventions have not developed them.

Another major concern is the capacity of implementing partners or target groups to continue their work or to continue enjoying benefits without the continuing support of Finnish interventions. On the one hand, this is related to knowledge and technical capacity and in 14 reports evaluators recommended improvements in that respect to bolster sustainability. The recommendations made range from classical capacity development (e.g. provide more or better training to teachers) to innovative ideas such as to make use of trained beneficiaries as staff, trainers or facilitators for future projects in the field: *"As a measure of future efficiency and sustainability, the Evaluation recommends using the trained beneficiaries as the human resource base for multidisciplinary programmes and projects of regional scope. Thus, the Rural Farmers' Association Green Valley can be used as a base for targeting local communities in the areas of disaster risk reduction, local area development, business incubators, and the like. Trained farmers and guesthouse owners, as well as eco-club members and school representatives, can be used as trainers and educators, for replicating the project model in other regions."* (Report No 23)

In five reports it is also recommend that technical or administrative issues (e.g. clarification on responsibilities of different government bodies or agencies, ensuring technical feasibility of chosen approaches or assisting with the installation of new equipment) be best resolved by the intervention before the support ends.

Capacity to continue work or to enjoy benefits also has a financial dimension. Eight reports include recommendations to identify new sources of funding, to assist in the creation of revenue or to support the development of financial resources appropriate to the partners, target groups or beneficiaries. In that sense, one evaluation report provides an interesting recommendation. Although it cannot be easily operationalised, it flags an important aspect to take into account when trying to achieve sustainable change: *"Especially in the case of communities or [community-based organisations] CBOs: Do not persuade the communities to abandon their previous livelihoods activities before the new one is economically sustainable."* (Report No 18)

Several other recommendations made in the reports did not fit into any of the synthesised categories above, such as the recommendation to improve the dissemination of success stories. The most important in terms of numbers (for six reports) is the recommendation to extend support beyond the initial period (at least in a minor form). Reasons for this recommendation vary and are largely dependent on the context of the intervention. However, several evaluators pointed out the existing need and demand for continued support by target groups and beneficiaries. In most cases, this recommendation is made to com-

pensate shortcomings of the initial intervention and to increase the likelihood of a positive impact.

Recommendations on monitoring and evaluation (M&E) are presented in 38 out of 50 reports. M&E is a fundamental issue which is closely linked to several other topics, such as efficiency, effectiveness or aid effectiveness (i.e. management for results) via the need to closely monitor interventions' performance. Overall, throughout the evaluation reports, the assessment of these categories is mixed at best and the large number of reports recommending improvements in M&E is in line with this assessment.

Recommendations on M&E can be allocated to two categories: those pointing at the establishment of an M&E system (17 reports) and those focussing on the improvement of existing systems (28 reports). Given these figures, a considerable number of interventions did not have any functioning M&E system. In several cases, the evaluators explicitly recommend to create a results-oriented M&E system.

When an M&E system exists, in many cases the evaluators recommended adapting indicators or increasing coverage by including specific topics such as compliance, social accountability or private sector development. In several cases, the involvement of other actors (such as government agencies, ministries or research institutes) is also recommended to increase the relevance and reach of the system. In this regard, several evaluators also recommended a closer cooperation with other donors' interventions on M&E-related activities as highlighted in the following example: *"Proactively support inclusion of indicators for CCOs of WASH in GTP II and OneWASH (DFID supported M&E consultancy), and contribute to performance measurement accordingly".* (Report No 7)

Furthermore, several recommendations aim at the improvement of data quality and enhancing the efficiency of data collection as well as the M&E system as a whole. A recurrent aspect in that regard is the use of modern technologies for M&E, as underlined by the following example: *"Enhance record-keeping systems by speeding up access to the CLIC system which will allow a shift to digital systems, enable more comprehensive client-focused support and information, will improve the interface with the MoPH and will facilitate potential use for a wider range of monitoring, evaluation and research applications."* (Report No 6) Finally, one mid-term evaluation recommended to assess the implementation of its recommendations at the end of the project.

5.10 Overall quality, strenghts and weaknesses of the interventions

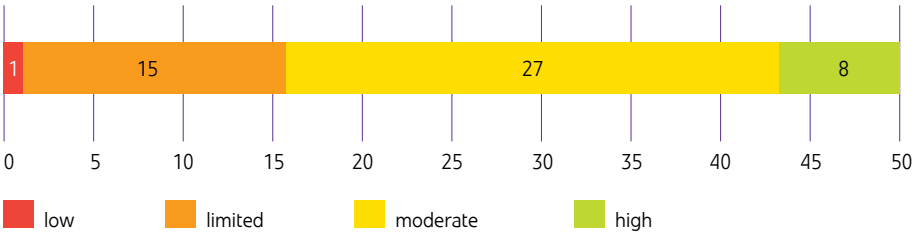
Highlights of the chapter addressing EQs 23–25:

- 70% of the 50 interventions (35) are assessed as of moderate quality or better. Hence, the overall quality of the bi-, multi- and multi-bilateral interventions under consideration is quite good.
- The overall quality of interventions at regional or global level does not significantly differ from the overall quality of interventions at national level. Similarly, no differences can be detected for different regions, thematic sectors or intervention budgets.
- Relevance is a systematic strength of bi- and multilateral interventions.
- Sustainability is the greatest challenge of bi- and multilateral interventions.
- As more than one third of the interventions is assessed as being weak with regards to their effectiveness, efficiency and impact and about half of the interventions with regards to their sustainability, there is considerable room for improvement in these areas.

To analyse the **overall quality** of an intervention, the sum of the assessments of all OECD DAC criteria captured in the evaluation report was divided by the total number of OECD DAC criteria covered. Due to limited data availability, assessments on cross-cutting objectives, aid effectiveness and triple C were not used in the overall aggregate.

As illustrated in Figure 41, the bi-, multi- and multi-bilateral interventions under consideration are quite positively assessed by the evaluators. About two thirds of the interventions (35, 70%) are assessed as of "high quality" (8, 16%) or as of "moderate quality" (27, 54%). In contrast, about one third of the interventions (16, 32%) was seen as being only of "limited quality" or of "no quality at all (1, 2%).

Figure 41: Overall quality of bi- and multilateral interventions (n=50)



Source: own statistics based on analysis of reports

In the course of our **disaggregated analysis**, comparisons of sub-groups remain statistically insignificant. For the 50 interventions under consideration, no differences have been detected between national level vs. regional/global level interventions and according to different regions or sectors. However, these results have to be taken with caution as regional and sectorial sub-groups within our sample are very small (e.g. only five interventions in the educational sector, only six interventions in Northern Africa and Middle East).

In addition, correlation coefficients turn out insignificant when testing for linkages between the overall intervention budget and the quality of the intervention, and the overall Finnish budget of an intervention and its quality.

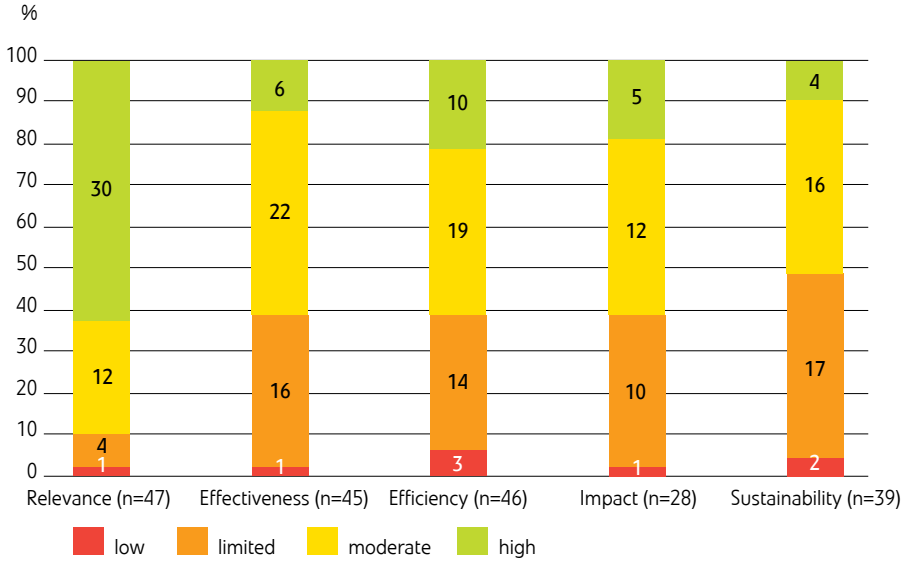
Disaggregated analyses at the level of single OECD DAC criteria did also not reveal any significant differences with the exception of one finding: Regional/global level interventions turn out to be of higher relevance than national level interventions. They are on average assessed as "highly relevant" (mean: 3.79) while interventions on the national level are on average assessed as "moderately relevant" (mean: 3.32). It is obvious that interventions on global level have much higher resource endowment. When controlling for budget differences, the result does not hold and whether an intervention was designed for the national or regional/global level remains statistically insignificant.

Drawing instead on "typical" recommendations of the evaluators as identified in chapter 5.9., it is of greater importance for the quality of an intervention whether planning, scope, and management are appropriate, whether the intervention suits the technical and financial capacities of the target groups and final beneficiaries, and whether challenges for the sustainability of potential changes are anticipated right from the beginning of an intervention.

To identify **strengths and weaknesses** of bi- and multilateral interventions of Finnish development cooperation as assessed by the evaluators, (i) a comparison of the quality assessments on single OECD DAC criteria, (ii) a review of different aspects assessed under each OECD DAC criteria, (iii) a review of different aspects of aid effectiveness and (iv) a review of the gender analysis have been performed.

Figure 42 allows a comparison of the **quality of each OECD DAC criteria**: Interventions' quality is particularly strong with respect to relevance. For 90% of the interventions it is assessed as "high" or "moderately". In contrast, sustainability of the interventions is weakest as compared to other OECD DAC criteria. Fewer than half of the interventions are assessed as "moderately" or "highly" sustainable. Interventions' quality regarding effectiveness, efficiency and impact is better with about 60% of the interventions assessed as "moderately" or "highly" successful in this regard.

Figure 42: Quality on single OECD DAC criteria



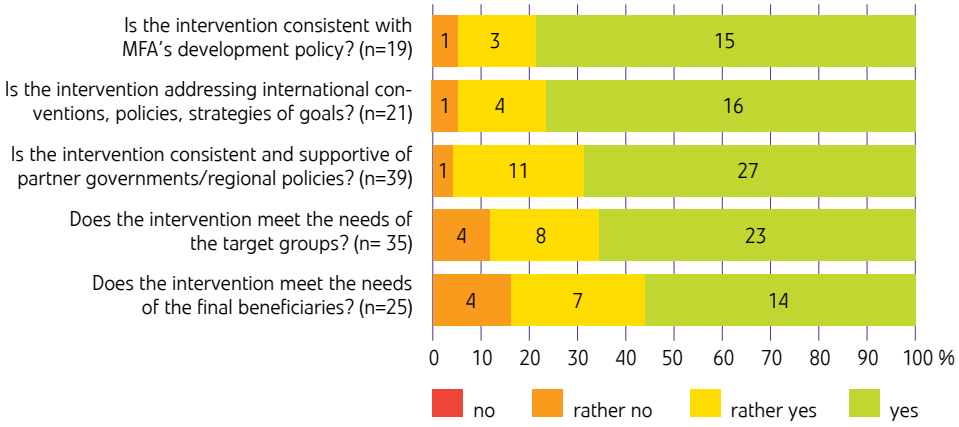
Source: own statistics based on analysis of reports

Relevance turns out as a strengths of Finnish development cooperation, while sustainability is rather challenging.

Deriving systematic strengths and weaknesses regarding **sub-aspects of the OECD DAC criteria** is challenging as some of these have only been assessed in a fraction of the evaluation reports under consideration. Hence, all following results within this chapter are limited to some tendencies.

Figure 43 illustrates whether assessed interventions are strong with regard to each of the different aspects on relevance i.e. consistency with MFA's development policy: addressing international goals, supporting partner/regional policies, meeting the needs of the target groups and final beneficiaries.

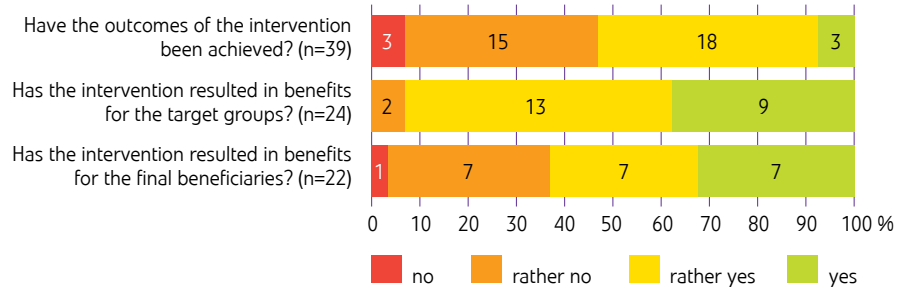
Figure 43: Quality of different aspects on relevance



Source: own statistics based on analysis of reports

From Figure 44, on effectiveness, a weakness is that about half of the interventions did rather not or not achieve their outcomes. However, the figure also shows that nearly all interventions (that is, those for which this aspect was assessed in the reports) resulted in benefits for the target group, which is a strength. Unfortunately, this does not necessarily mean that this is also the case for final beneficiaries.

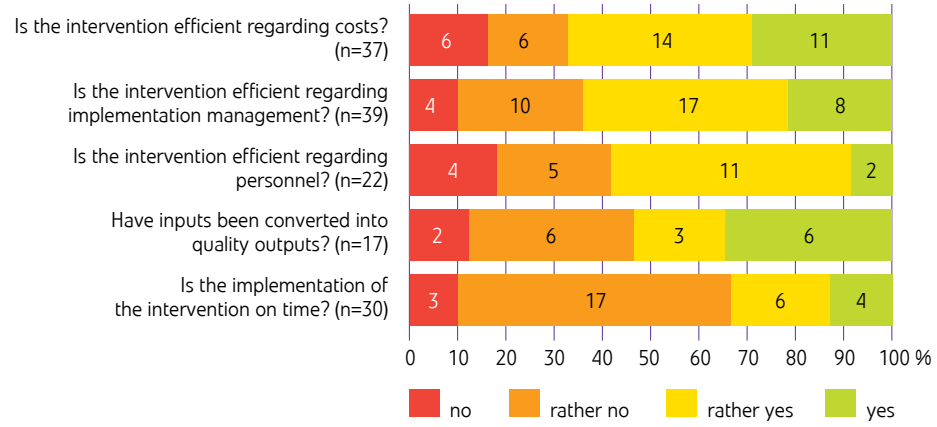
Figure 44: Quality of different aspects on effectiveness



Source: own statistics based on analysis of reports

Figure 45 illustrates that time inefficiency is a frequent weakness of the interventions (at least according to the reports that assess this issue). Findings indicate that two out of three interventions are delayed. The picture with regard to cost efficiency, implementation management and efficiency of staffing is considerably better. Only about one out of three interventions is inefficient in this regard. With respect to conversion of inputs into quality outputs shares are about fifty-fifty. Although exact differences between different aspects are difficult to detect given the different number of assessments, implementation on time stands out as the most serious challenge.

Figure 45: Quality of different aspects on efficiency

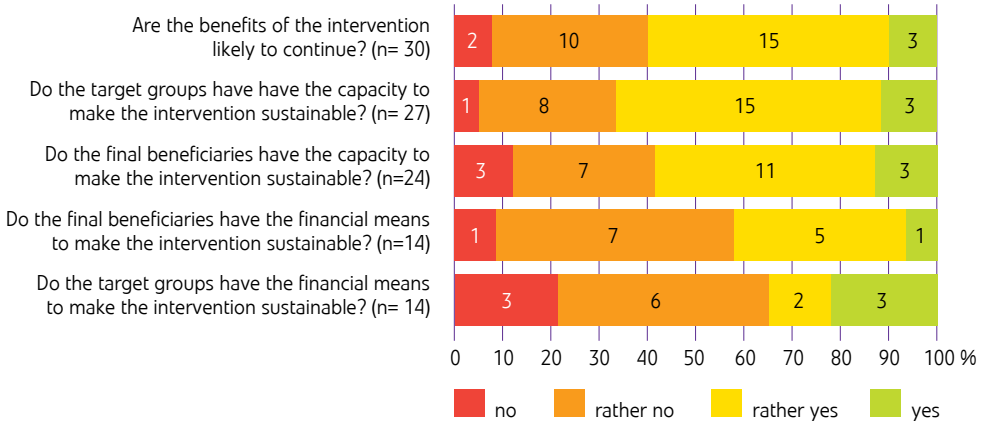


Source: own statistics based on analysis of reports

Regarding impact, the small number of reports does not allow for a credible analysis of strengths or weaknesses.

Figure 46 on sustainability illustrates that two out of five interventions benefits are unlikely to continue after the intervention's end. A deeper look suggests that this can be caused by a lack of capacity among target groups and final beneficiaries, as well as by a lack of their financial means. The lack of financial means is the more important factor.

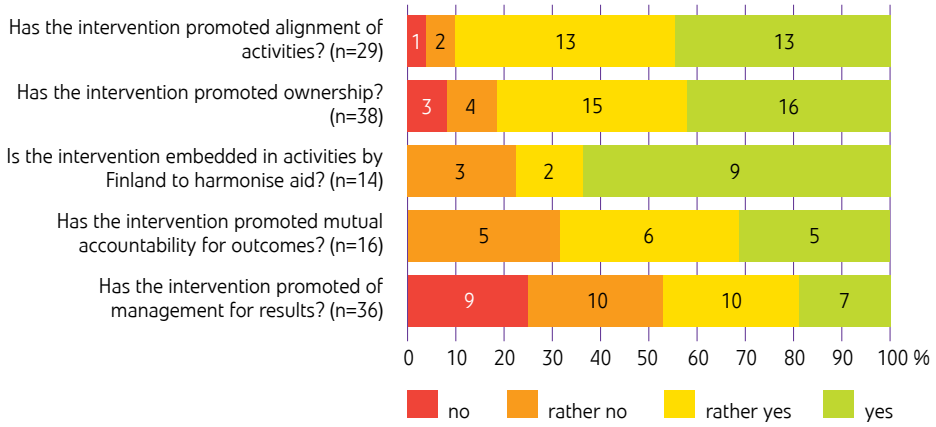
Figure 46: Quality of different aspects on sustainability



Source: own statistics based on analysis of reports

Beyond the OECD DAC criteria, Figure 47 on **aid effectiveness** suggests that promotion of alignment of activities, promotion of ownership and embeddedness of the intervention in activities by Finland to harmonise aid are rather strengths of the interventions. In contrast, it is a weakness that about half of the interventions for which this aspect was evaluated fail to promote management for results.

Figure 47: Quality of different aspects on aid effectiveness



Source: own statistics based on analysis of reports

Similarly, as for impact, insights on triple C cannot be further analysed, as the number of reports addressing this aspect is too small within our sample.

Finally, the gender analysis showed that a considerable number of analysed interventions are assessed as gender-mainstreamed or gender-sensitive. This is identified as a strength. In addition to this, some interventions focus exclusively on gender equality and women's rights. However, at the same time, more interventions have been assessed as only gender-aware or in rare cases even as gender-blind, which is a weakness. Thus, considerable room for improvement also remains in this regard.

6 CONCLUSIONS

Opening remark on MFA's decentralised evaluation portfolio (EQ1) and limitations of the findings of this meta-evaluation

- Geographical scope, sectorial affiliation as well as intervention and evaluation budgets vary widely among the evaluation reports considered. Similarly, the nature of the intervention, the nature of the evaluations, their commissioner and the nature of the implementer are mixed. In comparison to other sectors, we find a high number of evaluation reports on interventions in the fields of environment/climate and conflict/security. Evaluation reports on interventions in the partner country Nepal are also more common in the sample than reports on interventions in other MFA partner countries.
- However, given the lack of information on the whole population of bi-, multi- and multi-bilateral interventions we cannot assess to which extent this sample of evaluation reports is representative for this fraction of Finnish development cooperation. Thus, we cannot conclude on the adequacy of MFA's decentralised evaluation portfolio.
- Furthermore, the quality assessment of bi- and multilateral Finnish development cooperation is only based on the 50 decentralised evaluation reports subject to this meta-analysis. Self-assessments by the implementers or cross-checks on the interventions were beyond this assignment. Hence, triangulation and contextualisation in this regard was impossible.
- The fact that the assessment tools were applied to evaluations of very heterogeneous interventions spread over a wide range of countries, regions, thematic sectors and intervention budgets required simplification. The quality and content of evaluators' assessments were weighted equally for small and large interventions. Limited information from the reports further obliged us to ground the overall content assessment exclusively on evaluators' assessment of the OECD DAC criteria.
- These limitations have to be kept in mind in order to put our following conclusions correctly into perspective.

6.1 Reliability and quality of the evaluation reports (EQ2, EQ4, EQ5, EQ7)

Conclusion 1: *Most evaluation reports feature considerable weaknesses regarding methodological rigour and transparency. Still, except for one report, findings appear to be somewhat reliable.*

Methodological weaknesses particularly include (i) lacking reference to the intervention logic when presenting the findings, (ii) an insufficient explanation of how observed effects were attributed to the intervention, (iii) missing (logical) links between findings, conclusions and recommendations or (iv) findings that are mixed with conclusions and recommendations, and (v) not even providing conclusions and recommendations at all. **Transparency is mostly compromised** by missing (i) sources of evidence and (ii) discussions of how data sources and methods were triangulated when presenting findings.

Overall, fifteen evaluation reports include at least one of these serious weaknesses. While the comprehensive data collection efforts in many evaluations at least suggest that results are founded on quite a substantial database and thus can be regarded as **somewhat reliable**, in most cases these weaknesses

threaten the credibility of the evaluation reports and question the appropriate use of data by the evaluators. It also indicates a **substantial lack of transparency** in terms of informing the reader about how the evaluators came to their conclusions.

Conclusion 2: *None of the reports' quality is highly satisfactory. About two thirds feature some, one third substantial quality flaws.*

The **reports' structure** feature considerable weaknesses including (i) lacking evaluation questions in the introduction, and (ii) a rather uncommon chapter sequence which puts the context analysis behind the methodology chapter and which often is not followed by the evaluators.

With regard to the results presentation common deficits comprise (i) lacking appropriate outcome and impact analyses, which are furthermore often incomplete and not adequately structured (ii) focusing on arbitrary selected aspects of efficiency, preventing a comprehensive assessment at intervention level (which is attributable in part to weak ToRs), and (iii) a lack of a three-dimensional (economic, social, and environmental) approach to sustainability.

While in two out of three reports the **discussion of cross-cutting objectives** includes gender equality, only less than half discuss reduction of inequality climate sustainability and the HRBA. So, in accordance with its prominence in Finnish development policy, gender equality is the most anchored cross-cutting objective in evaluation practice. At least, if climate sustainability is discussed, it is mostly done in a systematic way.

In one out of four reports, the **presentation of conclusions and recommendations** is not acceptable. As to conclusions, they are either (i) not available at all, or (ii) reflect data not presented and validated previously in the report, or (iii) are comingled with new findings. Recommendations are often weak regarding prioritisation, direction to specific actors and timeline for implementation. Eventually, only about half of the reports provide lessons learnt.

Some of the reports have been accepted by MFA and other commissioners without conclusions or recommendations. This raises the question whether MFA's feedback mechanisms in the review process are always functional.

The **composition of the evaluation team** regarding gender quality, thematic knowledge, evaluation capacity and local expertise remains unclear for the great majority of reports. However, according to the names presented in the reports, at least one quarter of the evaluation reports are produced by gender unbalanced teams.

The **quality of summaries** is not alarming, but there is room for improvement, as many executive summaries lack information on the evaluation design and/or the methodology applied.

Given the small sample size, a disaggregated analysis of countries, sectors and regions was not possible. At the same time, it is also not clear to the meta-evaluation team why report quality should differ among these characteristics. Beyond this limitation, disaggregated analysis revealed:

Conclusion 3: *The overall report quality does not vary between different sub-groups.*

Concerning the overall report quality, **no considerable differences between mid-term and final evaluations** could be identified. Final evaluations refer more often to previous evaluations, which is logical as the possibility that earlier evaluations exist is higher. In mid-term evaluations the quality of the relevance chapter is rated better. This is also plausible, as at this point in time, commissioners and evaluators often have a stronger focus on the relevance of an intervention.

Teams from consulting firms or institutes provide slightly better quality reports than individual/independent consultants. The former usually provide better methodology sections. Individual/independent consultants also score lower regarding sampling, data analysis methods, and discussion of limitations. This is a drawback, as evaluation methodologies lay the foundation for reliable findings, conclusions and recommendations. Furthermore, the summaries by individual/independent consultants are weaker in comparison to those of consultancy firms/institutes.

Possible causes are threefold: lack of capacity, lower evaluation budgets and lower quality of the ToRs. Correlation analysis suggests that lower performance of individual/independent consultants is more likely to be caused by lower methodological knowledge and by lower quality of the ToRs.

Still, the **overall report quality and evaluation budget are not significantly correlated**. This is also the case for project budget, as a proxy for evaluation budget to increase the sample size.

While overall **no substantial quality differences between reports commissioned by the MFA and those commissioned by others** could be identified, the former are regarded as being more comprehensive. This is however, partly related to MFA's ToRs, which more often request coverage of the OECD DAC criterion "impact", and the cross-cutting objectives climate sustainability and human-rights based approach. This suggests that superior ToRs, awareness rising through Finnish policies and evaluation guidelines might have a positive influence on evaluation practice. However, although **MFA-commissioned evaluations score better on the structure requested by MFA**, they mostly do not fulfil their requirements. This shows in turn that full compliance with MFA guidelines is not yet reached.

In the methodology section of the reports the sampling is assessed significantly lower for MFA-commissioned evaluations than for evaluations by other commissioners. Thus, **MFA-commissioned reports show inferior quality in one out of several methodological aspects**. This might point either to a capacity gap of hired evaluators, a weaker methodological quality assurance by the MFA, or a mixture of both.

6.2 Quality of ToR and their linkage to overall report quality (EQ3, EQ6)

Conclusion 4: *While the overall quality of ToRs can be considered as satisfactory, there is room for improvement with regard to providing methodological and practical advice.*

All ToRs contain comprehensive information about the evaluation background, subject, scope and objective, the stakeholders of the evaluation, the evaluation questions and criteria. In contrast, specifications on the methodology, the evaluation process, quality assurance and on cross-cutting objectives are of lower quality. Thereby, it has to be highlighted that **ToRs by the MFA are in general of higher quality than those of other commissioners**.

Conclusion 5: *A higher quality of ToRs is related to a higher quality of evaluation reports.*

The sections on (i) purpose, objectives and scope of the evaluation, (ii) methodology and (iii) evaluation process are particularly important for overall report quality. The relative weakness of methodology and process sections just pointed out suggests that **ToRs are failing to adequately support evaluation quality**.

6.3 Gaps in MFA's evaluation capacity (EQ8)

With the Evaluation Manual (2013), the Development Evaluation Norm (2015) and the Manual for Bilateral Cooperation (2012) MFA provides guidance to its staff and to external evaluators. These documents are fully in line with OECD's guidance and international standards. However, we identified some important gaps, which may have a negative impact on evaluation practice:

The Evaluation Manual (i) does not request for provision of data collection instruments in the annexes, (ii) it does not urge to contextualisation of evaluators' findings by reference to previous evaluation results, (iii) does not emphasise linking evidence to findings, (iv) it does not explicitly require triangulation of data and methods to obtain reliable findings and (v) it does not request discussing causal attribution of the intervention to the findings.

Conclusion 6: *The fact that the quality of the ToRs leaves room for improvement reveals capacity gaps within MFA.*

As mentioned above specifications on the methodology, the evaluation process, quality assurance and on cross-cutting objectives feature considerable gaps. Furthermore, **aspects regarding feasibility are an issue**. Budgets, time resources and numbers of working days (or their non-specification) are not in line with the number and content of the evaluation questions, suggesting that often authors of ToRs do not have a sound knowledge of (i) costs of different evaluation designs, (ii) feasibility of tasks, (iii) human resource requirements and (iv) realistic time frames within the evaluation process. This can lead to unrealistic expectations and evaluation failure.

Finally, partly weak evaluation report quality, particularly with regard to missing key sections, raises **concerns regarding MFA's evaluation capacity** with respect to the steering of the inception and the report reviewing phases.

6.4 Quality, strengths and weaknesses of bi- and multilateral Finnish development cooperation according to OECD DAC criteria (EQ10-14, EQ23-EQ25)

Conclusion 7: *The quality of the bi- and multilateral interventions under consideration is assessed quite positively with their relevance being considered as a particular strength and sustainability as the greatest challenge.*

According to the evaluators, roughly **two out of three interventions are of moderate quality or better**, whereby their quality does not vary significantly among geographical regions, thematic sectors or budgets. It also apparently does not matter whether an intervention is at national, regional or global level. However, these results have to be taken with caution as regional and sectorial sub-groups within the sample are very small.

The quality of an intervention depends mainly on (i) the appropriateness of its planning, scope, and management; (ii) its conformity with the technical and financial capacities of the target groups and final beneficiaries, and (iii) the anticipation of challenges for the sustainability of intended benefits at its beginning.

90% of the interventions were assessed as highly or moderately relevant. Interventions' quality regarding effectiveness, efficiency and impact is clearly lower with about 60% rated in the upper categories. Eventually, their sustainability scores lowest with just under 50% rated in the upper categories.

Limited evidence on single aspects suggests that interventions are strong with regard to (i) their consistency with MFA's development policy, (ii) addressing international goals, (iv) supporting partner/regional policies, (v) meeting the needs of the target groups and final beneficiaries, and (vi) resulting in benefits for the target groups. On the other hand, a look at the weaknesses discloses (i) that the latter benefits are not necessarily transformed to benefits for the final beneficiaries, (ii) that about half of the interventions did rather not or not achieve their outcomes, (iii) that two out of three interventions are delayed, and (iv) that two out of five interventions' benefits are unlikely to continue after the interventions end.

6.5 Gender as cross-cutting objective in bi- and multilateral Finnish development cooperation (EQ15-18)

Conclusion 8: *Interventions are mostly not gender-transformative.*

For more than a decade, Finnish development policies have been paying close attention to gender equality and women's rights. However, the cross-cutting objective has not yet been fully integrated in evaluation practice as more than one third of the interventions were not assessed in this regard. Limited evidence from the summative analysis reveals a mixed picture.

While some interventions were assessed as gender-mainstreamed and some focus on gender equality and women's rights exclusively, we found also a considerable number of solely gender-aware interventions.

Evaluation reports providing insights on the other cross-cutting objective (i.e. (i) reduction of inequality, (ii) climate sustainability and (iii) the human rights-based approach) are rare and did not allow a systematic analysis.

6.6 Aid effectiveness of bi- and multilateral Finnish development cooperation (EQ19–EQ22)

Conclusion 9: *It remains often unclear if and to what extent the interventions follow the concepts of aid effectiveness and triple C.*

Aid effectiveness and triple C are only rarely covered in the evaluation reports. The relatively large number of reports that have not taken these concepts into consideration suggests that they are not deeply anchored in the evaluation practice of Finnish development cooperation.

We, on the other hand, did find some hints, suggesting that interventions often (i) promote ownership and (ii) consist of aligned activities, (iii) tend to be coordinated with other interventions and (iv) are moderately successful in promoting management for results.

6.7 Major recommendations emerging from decentralised evaluation reports (EQ26)

Typical recommendations relate to the issues of (i) monitoring and evaluation, (ii) planning and scope of the intervention, (iii) implementation management, (iv) capacity of beneficiaries and other stakeholders, and (v) sustainability of the intervention.

Conclusion 10: *Most interventions lack functioning M&E systems.*

It is of concern that, according to the evaluators, **one third of the interventions does not have any functioning M&E system**. Accordingly, in such cases the introduction of such a system is recommended in the reports. Just under half of the recommendations on M&E centres around improving the M&E system by (i) adapting indicators, (ii) extending the coverage of the M&E system, (iii) improving data quality, (iv) increasing the efficiency of data collection, and/or (v) using modern technologies for data collection, management and analysis. Since M&E is a precondition to allow management for results, these recommendations are considered as highly beneficial for improving Finnish development cooperation.

Conclusion 11: *Apparently evaluators regard intervention planning, scope, management, capacity and/or sustainability as improvable.*

Every second report includes recommendations on the issues of planning, scope, management, capacity and/or sustainability. With regard to planning, such recommendations include (i) reviewing the project design and developing/adapting the Theory of Change (ToC), (ii) raising awareness of the importance

of planning in general, for institutionalising and better support of implementing partners, (iii) performing and exploiting situational analyses and risk assessments for planning purposes, and (iv) realistically taking into account budget and time constraints.

While recommendations regarding the scope are not consistently geared in one direction or another (i.e. narrowing vs. broadening the intervention's scope), those on intervention management centre around (i) improving the intervention's organisation structure and (ii) enhancing its functionality.

Typical recommendations in the field of capacity development are (i) enhancing technical or thematic knowledge, (ii) increasing the capacity of final beneficiaries to better use the services delivered, (iii) empowering beneficiaries and raising awareness of the challenges they face and of the interests of vulnerable groups, and (iv) improving the quality of capacity development activities through appropriate methodologies, adequate time management, and use of well-skilled trainers.

Recommendations on sustainability generally have a strong focus on exit strategies. In several reports, evaluators call for (i) acknowledgement of/ improvements in implementing partners' or target groups' capacities to make the benefits of the intervention sustainable, (ii) working towards clearly shared responsibilities among different actors before the support ends, and (iii) identification of new sources of funding. Although strategies to accompany phasing out are being recommended in the MFA Manual for Bilateral Cooperation, a fair number of interventions does not seem to have developed them.

Taken together, these recommendations provide valuable insights for each stage of the project management cycle. While typical lessons learnt could not be identified, they eventually lead to the conclusion, that **project planning and implementation is regarded as an essential field for improvement**.

7 RECOMMENDATIONS

EQ9 specifies: *"What are recommendations to improve the quality of MFA's decentralised evaluations?"* In this regard we structure our recommendations according to the following aspects: (i) general guidance on decentralised evaluations within the MFA, (i) drafting ToRs, (iii) recruiting evaluators, (ii) evaluation management, and (v) commissioning future meta-evaluations.

7.1 General guidance on decentralised evaluations within the MFA

R1.1: Improve the Evaluation Manual

The Evaluation Manual should be updated to close existing gaps regarding the commissioners' capacity (i) to draft ToRs and (ii) to assure the quality of evaluation deliverables and regarding evaluators' capacity (iii) to produce high quality evaluation reports. These gaps concern, but are not limited to, (i) increased transparency regarding data collection instruments, (ii) placing findings in the context of previous evaluation results, (iii) the linking of evidence to findings, (iv) the triangulation of findings and (v) the discussion of causal attribution of the intervention to the findings. Additionally, (vi) impact as well as sustainability analyses need to be further guided or standardised in order to receive more meaningful results. (vii) The structure of reports should be made consistent with that of other donors (i.e., the context analysis should figure within or directly after the introduction). Furthermore, it would be beneficial to provide guidance on (viii) very rough estimates on costs, personal and time requirements of different evaluation designs, (ix) their explanatory power and (x) associating tasks and responsibilities of commissioners and evaluators within the evaluation process.

- This recommendation is mainly linked to conclusions 1, 2, 4 and 6.
- Main implementation responsibility: MFA, in particular EVA-11
- Urgency: high, immediate action required, should be completed in 2018.
- Priority: high

R1.2: Enhance knowledge of evaluation methodologies and on evaluation practice

The quality of evaluations benefits from enhanced knowledge of evaluation methodologies on the commissioner's side. This includes, but is not limited to, subjects such as drafting specifications on the methodology, the evaluation process, quality assurance and on cross-cutting objectives for ToRs as well as expertise to assess suggested methodologies of inception reports and review draft reports.

In addition, the EVA-11 should ensure sound knowledge of all commissioners of the (i) costs of different evaluation designs, (ii) feasibility of tasks, (iii) human resource requirements and (iv) realistic time frames within the evaluation process. This would greatly help keeping the expectations for evaluations realistic.

Finally, the EVA-11 should ensure that commissioners are aware of their responsibilities and tasks and possess the necessary knowledge and skills to provide structured and constructive feedback on all evaluation deliverables (in particular, the inception and draft reports). The EVA-11 should be the focal point in coordinating or delivering training to other commissioners within Finnish development cooperation.

- This recommendation is mainly linked to conclusions 1, 2, 4 and 6.
- Main implementation responsibility: MFA, in particular EVA-11
- Urgency: medium, consecutive after implementation R1.1, should start in 2019 and be understood as continuous task
- Priority: high

R1.3: Consider improving existing structures

Given the before-outlined numerous shortcomings of the evaluations as regards to their contents and methodologies, and the apparently insufficient capacities of their commissioners to provide for sufficient quality evaluation reports, a centralised knowledge management system and stronger coordination with the EVA-11 should be considered. While promulgating 'evaluative thinking' in the entire organisation is surely beneficial for every practitioner in order to make his/her intervention evaluable and thus can be highly recommended, awareness (e.g. about the requirements for reliable data, methods and results) alone does not guarantee the exigency of professionally designed and implemented evaluations according to international scientific standards.

A greater stake of EVA-11 within decentralised evaluation practice would also allow the MFA establishing and adhering to a coherent evaluation strategy with regard to sampling of interventions, putting focus on particular sectors that are most relevant for the ministry, or even developing an overarching M&E system, which is a crucial prerequisite for an institution-wide management for results anyway.

- This recommendation is mainly linked to conclusions 1, 2, 4, 5 and 6.
- Main implementation responsibility: MFA leadership
- Urgency: high, should be kicked-off by meta-evaluation results
- Priority: medium

7.2 Recommendation for drafting ToRs

R2.1: Be more precise on methodological requirements and on expectations regarding the different OECD DAC criteria

Methodological requirements such as evaluation design, underlying sampling strategies and known limitations should be clearly identified and addressed

in the ToRs. Further, the ToRs should formulate clear expectations regarding the assessment of the different DAC criteria in order to prevent incomplete or unstructured analyses (e.g., absence of a detailed outcome analysis, arbitrary assessment of efficiency or lack of a three-dimensional approach to sustainability).

- This recommendation is mainly linked to conclusions 1, 2, 4, 5, and 6.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: medium, consecutive after provision of R1.1, should by systematically implemented from 2019 onwards
- Priority: high

R2.2: Amend the ToR by several missing aspects

Commissioners should amend the ToR by the following important aspects: (i) discussion and revision of the intervention logic (ii) systematic integration of cross-cutting objectives into the evaluations, (iii) integration of triple C (i.e. coherence, coordination and complementarity), (iv) formulation of implementable recommendations and identification of who should be tasked with implementation, (v) identification of the users of the evaluation report and formulation of implementable expectations with regards to practicable recommendations, and (vi) provision of general lessons learnt to foster learning beyond intervention-specifics. Further, commissioners should provide information with regards to the expected length, level of detail and content of the executive summary already in the ToRs.

- This recommendation is mainly linked to conclusions 1, 2, 4, 5, 6, 8 and 9.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: medium, consecutive after provision of R1.1, should by systematically implemented from 2019 onwards
- Priority: high

R2.3: Pay particular attention to the quality of ToRs for smaller evaluations (in terms of budget and intervention size)

As the ToRs for individual/independent consultants tend to be of lower quality, we recommend to pay particular attention to the quality of ToRs (and therefore to all the recommendations mentioned in this sub-chapter) for evaluations with small budget or scope.

- This recommendation is mainly linked to conclusions 1, 2, 4, 5 and 6.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: medium, consecutive after provision of R1.1, should by systematically implemented from 2019 onwards
- Priority: medium

7.3 Recommendations for recruitment of evaluators

R3.1: Be gender-transformative throughout the recruitment process

Commissioners should set a good example for gender-transformative recruitment of evaluation teams in both international and local contexts. This comprises the empowerment of women and LGBT and goes beyond the gender-balancing of evaluation teams.

- This recommendation is mainly linked to conclusion 2.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: high, immediately and continuously
- Priority: high

R3.2: Ensure sufficient methodological expertise

Methodological knowledge and skills should be regarded at least as equally important as thematic and regional expertise when recruiting evaluation experts. In light of the methodological shortfalls observed in many of the evaluation reports, this recommendation is considered key to improve the overall quality.

- This recommendation is mainly linked to conclusions 1,2,4,5, and 6.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: high, immediately and continuously
- Priority: high

7.4 Recommendation on evaluation management

R4.1: Enhance quality assurance throughout the evaluation process

Commissioners should make available sufficient time and human resources for thorough methodological and thematic quality assurance of the inception report and should verify the compliance with the proposed methodology in the draft report. When reviewing draft reports, pay attention to MFA's requirements regarding structure, editing and writing standards and make sure that evaluators comply with them. Further, make sure that evaluators (i) display their sources of evidence, (ii) elaborate on triangulation of sources and methods when presenting results, (iii) make use of the intervention logic to obtain findings and (iv) discuss the causal attribution of findings to the intervention. Do not accept reports that (i) considerably fail in any of the above-mentioned categories,(ii) which are not referring and not responding to the evaluation questions, (iii) where no clear link from findings to conclusions to recommendations is established, or (iii) with seemingly arbitrary or missing conclusions and recommendations.

- This recommendation is mainly linked to conclusions 1 and 2.
- Main implementation responsibility: MFA regional units & other commissioners of evaluations
- Urgency: high, immediately, should be systematically integrated in 2019
- Priority: high

R4.2 Make use of meta-evaluation results from the content assessment

EVA-11 should ensure that there is sufficient and appropriate dissemination and uptake of the meta-evaluation results emanating from the summative analysis. Particular importance should be paid to the synthesised recommendations regarding M&E systems.

- This recommendation is mainly linked to conclusions 7, 8, 9, 10 and 11.
- Main implementation responsibility: MFA regional units, EVA-11 & other commissioners of evaluations
- Urgency: high, immediately in 2018
- Priority: medium

7.5 Recommendations for commissioning future meta-evaluations

R5.1: Using the same assessment tools for future meta-evaluations

To allow comparisons over time, it is of utmost importance to maintain the same assessment tools in future meta-evaluations. As over the years the number of evaluation reports increases, sub-group comparisons, for example regarding different evaluation budget ranges, different regions or thematic sectors, will be possible.

- This recommendation is mainly linked to the opening remarks of the conclusions section. Main implementation responsibility: MFA EVA-11
- Urgency: low
- Priority: high

R5.2: Enhance the representativeness of future samples

The explanatory power of future meta-evaluation will increase when the underlying sample can be considered representative of the whole population of bi-, multi- and multi-bilateral interventions. Therefore, we recommend setting up and maintaining an inventory of all interventions classified by key characteristics (i.e. budget, duration, sector, region, nature of the intervention, commissioner). This would enable the MFA to make a selection of interventions to be evaluated based on these key characteristics and later allow the meta-evaluation team to assess the representativeness of their sample and to adjust the sample composition if necessary.

- This recommendation is mainly linked to the opening remarks of the conclusions section.
- Main implementation responsibility: MFA statistic department, EVA-11, regional units
- Urgency: high
- Priority: medium

R5.3 Enhance the sources of evidence for future meta-evaluation

To allow for triangulation and contextualisation of findings, we recommend the use of additional data sources. Online surveys with implementers are an efficient way to collect a self-assessment on the interventions and gain further information on the evaluation process and the usage of evaluation results. Furthermore, evaluators could be consulted regarding their perspective on the evaluation process.

- This recommendation is mainly linked to the opening remarks of the conclusions section.
- Main implementation responsibility: MFA leadership, EVA-11
- Urgency: low
- Priority: medium

REFERENCES

MFA of Finland (2008). Development Policy Programme 2007 Towards a Sustainable and Just World Community: Government Decision-in-Principle 2007. Helsinki: MFA of Finland.

MFA of Finland (2012a). Finland's Development Policy Programme: Government Decision-in-Principle 16 February 2012. Helsinki: MFA of Finland.

MFA of Finland (2012b). Manual for Bilateral Cooperation 2012. Helsinki: MFA of Finland.

MFA of Finland (2013) Evaluation Manual. Helsinki: MFA of Finland.

MFA of Finland (2015). Development Evaluation Norm. Helsinki: MFA of Finland.

MFA of Finland (2016a). Finland's Development Policy: One World, common future - towards sustainable development. Helsinki: MFA of Finland.

MFA of Finland (2016b). Manual for Bilateral Cooperation 2012. Helsinki: MFA of Finland.

MFA of Finland (2016c). Evaluation of Finland's Development Cooperation Country Strategies and Country Strategy Modality. Helsinki: MFA of Finland.

MFA of Finland (2017a). Final Evaluation of Regional Forest Projects in Mekong, Andean and Central America. Helsinki: MFA of Finland.

MFA of Finland (2017b). Evaluation on Programme-based Support to Civil Society Organizations – Part 3. Helsinki: MFA of Finland

Independent Evaluation Office of the Global Environment Facility GEFIEO (2017) Gender Mainstreaming in the GEF. The GEF. Retrieved from: http://www.gefieo.org/sites/default/files/ieo/evaluations/files/gender-study-2017.pdf, 21.12.2017

OECD (2010). Glossary of Key Terms in Evaluation and Results Based Management. Retrieved from https://www.oecd.org/dac/evaluation/2754804.pdf (21.12.2017)

OECD (2017a). Compare your country. Official Development Assistance 2016. ODA 1960-16 Trends. Retrieved from http://www2.compareyourcountry.org/oda?cr=18&cr1=oecd&lg=en&page=1 (21.12.2017)

OECD (2017b): DAC Criteria for Evaluating Development Assistance. Retrieved from http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm (21.12.2017).

THE META-EVALUATION TEAM

This meta-evaluation is conducted by a team of five persons. Dr. Stefan Silvestrini and Dr. Susanne Johanna Väth act as Team Leader and Deputy Team Leader. They were substantially involved in developing the design of the meta-evaluation and the summative analysis and coordinated the work of three meta-evaluators: Dr. Cornelia Römling, a methodological expert, Petra Mikkolainen, a Finnish development policy evaluation expert and Michael Lieckefett a development evaluation generalist. The multidisciplinary and gender-mixed team benefited from complementary competencies while fulfilling the standards set in the tender. The tight meta-evaluation schedule justified the size of the evaluation team.

Dr. Stefan Silvestrini: As team leader Stefan Silvestrini took the overall responsibility for the assignment and was involved in all stages of the analysis. He will led the initial document review for the context analysis and drafting the inception report. In the implementation phase, he was responsible for backstopping the quality and the content assessment and also conducted analyses of randomly selected reports to cross-check the assessments. To facilitate a joint analysis Stefan Silvestrini was be in close contact to all team members and guided an internal synthesis workshop. Finally, he supervised the reporting phase and ensured proper presentation of meta-evaluation results.

Dr. Susanne Johanna Väth: As Deputy Team Leader Susanne Johanna Väth worked in close cooperation with Stefan Silvestrini. She took the lead during service order one and was responsible for presenting the general meta-evaluation approach and the methodology to the reference group. During the inception phase she guided the finalisation of the methodology, and led the development and operationalisation of the quality and content assessment tools as well as their pre-test and adjustment. In the implementation phase, she was mainly involved in the quality assessments of the reports to be analysed. Moreover, she took responsibility and supports Stefan Silvestrini in the course of the joint analysis and led drafting of the meta-evaluation report.

Dr. Cornelia Römling: As meta-evaluator with a strong methodological background, Dr. Cornelia Römling supported Susanne Johanna Väth in the development of the meta-evaluation design during the inception phase. Furthermore, she was substantially involved in the quality assessment of the reports to be analysed during the implementation phase.

Petra Mikkolainen: As meta-evaluator with in-depth knowledge of Finnish development cooperation, Petra Mikkolainen supported Stefan Silvestrini in the inception phase. Her tasks comprised reviewing Finnish documents and contributing to the context analysis. In addition, she took a major stake in the content assessment of the reports to be analysed.

Michael Lieckefett: As meta-evaluator with strong analytical skills and sound knowledge of developing contexts, Michael Lieckefett was primarily involved in the content assessment of the reports to be analysed. Furthermore, he supported the Team Leader with regard to evaluation management throughout all phases of the assignment. In this regard, he took minutes of meetings, safeguards data base maintenance and supported overall time management.

ANNEX 1: TERMS OF REFERENCE

Ministry for Foreign Affairs of Finland

Terms of References

1. BACKGROUND TO THE EVALUATION

The Ministry for Foreign Affairs of Finland (MFA) assesses Finnish development cooperation by carrying out two types of evaluations. One type is the comprehensive, policy level evaluations (centralized evaluations) commissioned by the Development Evaluation Unit (EVA-11). Second type is the project and program evaluations (decentralized evaluations) commissioned by the unit or department responsible for the project or program in question.

EVA-11 commissions regularly meta-evaluations in order to synthesize the findings, explore the issues and assess the reliability of the decentralized evaluations. This is the Terms of Reference (ToR) for the meta-evaluation of project and program evaluations (decentralized evaluations) carried out between September 2015 and August 2017. The evaluation will be based on the assessment of the decentralized evaluation reports and corresponding Terms of References (ToR) documents.

Meta-evaluation can provide a clear account of the evaluation function of MFA during a certain period of time by classifying decentralized evaluation reports by commissioner, country, sector etc. and by assessing the reports. Meta-analysis of decentralized evaluations can also bring together otherwise scattered evaluation findings on the results of development cooperation projects and programmes funded by MFA.

Meta-evaluation is also seen as a tool for accountability and improved transparency towards partner countries, general public, parliamentarians, academia, media and development professionals outside the MFA.

2. RATIONALE, PURPOSE AND OBJECTIVES OF THE EVALUATION

The purpose of the meta-evaluation is twofold: first, the meta-evaluation helps the MFA to improve the evaluation reports, the evaluation management practices and the overall evaluation capacity development. It also provides an overall picture of the current evaluation portfolio which helps the MFA to identify possible gaps. Second, the meta-analysis is expected to aggregate data and bring forward issues and lessons learned emerging from the evaluation reports as well as give recommendations which will help the MFA to improve the development cooperation. The meta-analysis will sum up what kind of strengths and challenges regarding Finnish development cooperation are identified in different evaluation reports.

The objective is also twofold: first, the meta-evaluation assesses different decentralized evaluation reports and related planning documents. It will also draw an overall picture of the evaluation portfolio in 2015-2017. Second, the meta-analysis synthesizes reliable evaluation findings and issues rising from the evaluation reports on Finland's development cooperation.

The results of this meta-evaluation will be compared to the Meta-evaluation of Project and Programme evaluations 2014-2015 in order to compare possible differences between these two meta-evaluations.

In order to enhance the long-term utility of Meta-evaluations the assessment tools will be standardized and meta-evaluations will be carried out regularly in every two years.

3. SCOPE OF THE EVALUATION

The meta-evaluation consists of two parts:

1) Meta-evaluation of the decentralized evaluation reports and their corresponding terms of references. The meta-evaluation will also produce an overview of MFA's decentralized evaluation activities classified by countries, sectors, budgets, evaluation types, managing units of MFA, etc.

The assessment of the evaluation reports (mid-term evaluations, final evaluations, ex-post evaluations and impact evaluations) will include all decentralized evaluation reports conducted between January 2015 and June 2017, their corresponding ToRs as well as ITTs and Inception Reports if they are available for the majority of reports under consideration allowing systematic exploitation of the material.

The sample includes evaluation reports of so called multi-bi projects/programmes funded partly by MFA. The administration of these projects and their evaluations may have been done by a partner organization in which case MFA has participated in commenting ToRs and evaluation reports but has not been the commissioner of the evaluation. During the assessment also a comparison of the quality between MFA commissioned evaluations and evaluations commissioned by MFA's partners will be made.

Meta-evaluation will assess the reliability of the reports and their ToRs applying the OECD/DAC evaluation principles and standards. The second part of this assignment is a summative meta-analysis based on all evaluation reports that have been assessed as reliable during the meta-evaluation.

Appraisal reports will be excluded from this meta-evaluation altogether as they are considered to be planning document instead of evaluations.

2) Meta-analysis of reliable evaluation findings on Finland's development cooperation verified against the OECD-DAC evaluation criteria demonstrating how Finnish development policy goals have been achieved based on findings in different reports. Meta-analysis will also sum up the major issues evident in current development cooperation emerging from the decentralized evaluation reports. The synthesis will conclude what are the main reasons for success or challenges in development cooperation projects and programs and what are the lessons learned based on the findings from evaluation reports.

4. EVALUATION QUESTIONS

Meta-evaluation:

1. Assessment and description of MFA's decentralized evaluation portfolio (evaluation reports and their corresponding ToRs) based on the OECD/DAC evaluation principles and standards, classified by countries, sectors, budgets, evaluation types, managing units of MFA, commissioner, etc.
    - Assessment of the reliability of evaluation reports
    - Are there gaps in evaluation capacity of MFA that need to be strengthened?
    - Is there a difference between the quality of MFA commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?

Meta-analysis:

2. What can be said about the Finnish development cooperation based on the reliable decentralized evaluation reports, and related planning documents by each OECD/DAC criteria and other relevant criteria identified in Finnish development policies
3. What are the major issues emerging from the decentralized evaluation reports?
    - Success stories, good practices and challenges

5. GENERAL APPROACH AND METHODOLOGY

The main method used in the meta-evaluation will be document review.

Assessment tools for both phases will be developed utilizing already existing tools. The methodology for both meta-evaluation and meta-analysis will be clearly described as well as the criteria based on which the reliability of evaluation reports is assessed.

The main sources of information will be the evaluation reports (mid-term evaluations, final evaluations, ex-post evaluations, impact evaluations) and their corresponding ToRs as well as Development Policy Programme documents, guidelines, earlier meta-evaluations, Government Reports to the Parliament and administrative in-house norms.

As evaluation reports under consideration considerably vary with regard to thematic focuses, context conditions, implementing partner organizations, scope and scale of the evaluation as well as evaluation designs and data sources, a high degree of content-related and methodological heterogeneity has to be taken into consideration for the quality assessment.

A checklist with criteria and sub-criteria enabling a fair and adequate grading has to be developed and based on insights from MFA's earlier meta-evaluations, clarifications provided by MFA during the inception phase, similar assignments conducted by the evaluation team and other meta-evaluations like those of UN Women and Norad as well as the EU ROM system. Criteria comprise but are not limited to credibility, completeness, adequacy of documentation and appropriateness of evaluation methods applied.

The consultant is expected to develop a four-step grading system with unambiguous grades to facilitate objective rating. The assessment tool has to be pre-tested and adjusted in line with MFA's feedback. Findings of the quality assessment will be aggregated and presented in summarizing results tables to identify general trends, display heterogeneity and prepare the ground for enhancing the quality of evaluations.

In a second-stage a content assessment provides insights on the joint contribution of MFA's development cooperation and will be conditional on minimal methodological standards in the context of the available material and comparable assignments. The evaluation team will also identify any emerging issues, both positive and negative, from the material.

The evaluation team is expected to cross-analyze approximately 10 % of all reports using random selection in order to avoid subjective bias.

The consultant is encouraged to raise issues that are important to the evaluation but are not mentioned in this ToR. Similarly, in consultation with EVA-11, the consultant might exclude issues that are in the ToR but may not be feasible and those remarks will be presented by latest in the inception report.

The evaluation must be gender and culturally sensitive and respect the confidentiality, protection of source and dignity of those interviewed.

6. MANAGEMENT OF THE EVALUATION

EVA-11 will be responsible for overall management of the evaluation process. EVA-11 will work closely with other units/departments of the MFA and other stakeholders in Finland and abroad.

A reference group for the evaluation will be established and chaired by EVA-11. The use of a reference group is a key step in guaranteeing the transparency, accountability and credibility of an evaluation process and plays a key role in validating the findings.

The mandate of the reference group is to provide advisory support and inputs to the evaluation, e.g. through participating in the planning of the evaluation and commenting deliverables of the consultant.

The members of the reference group will include:

Suvi Virkkunen Advisor on Development Policy/KEO

Jussi Karakoski Advisor on Development Policy/ALI

Sanna Takala Advisor on Development Policy/ASA

The tasks of the reference group are to:

- act as source of knowledge for the evaluation;
- participate in the planning of the evaluation (providing input to the ToR);
- participate in the relevant meetings (e.g. inception meeting and possible debriefing and validation meeting);
- comment on the deliverables of the consultant (i.e. inception report, draft final report, final report) to ensure that the evaluation is based on factual knowledge about the subject of the evaluation and
- play a key role in disseminating the findings of the evaluation and support the implementation, dissemination and follow-up on the agreed evaluation recommendations.

7. EVALUATION PROCESS, TIMELINES AND DELIVERABLES

The evaluation will tentatively start in August 2017 and end in February 2018. The evaluation consists of the following phases and will produce the respective deliverables. During the process particular attention should be paid to strong inter-team coordination and information sharing within the team. It is highlighted that a new phase is initiated only when the deliverables of the previous phase have been approved by EVA-11. All the reports have to be sent with an internal quality assurance note and the revised reports have to be accompanied by a table of received comments and responses to them.

It should be noted that internationally recognised experts may be contracted by EVA-11 as external peer reviewer(s) for the whole evaluation process or for some phases/deliverables of the evaluation process, e.g. final and draft reports (inception report, draft final and final reports). In case of peer review, the views of the peer reviewers will be made available to the Consultant.

The language of all reports and possible other documents is English. Time needed for the commenting of different reports is 2-3 weeks. The timetables are tentative, except for the final report.

A. START-UP PHASE

The administrative meeting regarding the administration, methodology and content of the evaluation will be held with the contracted team leader and EMS coordinator in Helsinki in August 2017. The purpose of the meeting is to go through the evaluation process, related practicalities and to build common understanding on the ToR.

Participants in the administrative meeting in Helsinki: EVA-11, Team Leader and the EMS coordinator of the Consultant. Other team members may participate.

The start-up meeting regarding the second service order will be held in September 2017 via Skype. The purpose is to get to know the whole evaluation team and go through the second service order and related administrative matters.

Participants in the start-up meeting: EVA-11 (responsible for inviting and chairing the session), Evaluation Team and EMS coordinator. Meeting will be arranged as a Skype session.

Deliverable: Agreed minutes of the meeting by the consultant.

B. INCEPTION PHASE

Inception report

The Inception phase includes testing and finalizing the assessment tools and preparation of detailed evaluation plan.

The inception report consists of the detailed meta-evaluation plan and finalized assessment tools including:

- finalization of the methodology and assessment tools
- final work plan and division of work between team members
- tentative table of contents of final report
- data gaps

The inception report will be presented, discussed and the needed changes agreed in the inception meeting in October 2017. The inception report must be submitted to EVA-11 two weeks prior to the inception meeting. Purpose of the inception meeting is to establish a community to enable dialogue and learning together as well as to get to know the evaluation team and the reference group.

Participants to the inception meeting: EVA-11, reference group and the Team Leader (responsible for chairing the session), evaluation team and EMS coordinator in person.

Venue: Kirkkokatu 12, Helsinki.

Deliverables: Inception report including the evaluation plan, finalized assessment tools and the minutes of the inception meeting by the Consultant

C. IMPLEMENTATION PHASE

The Implementation phase will start in October 2017.

Direct quotes from interviewees and stakeholders may be used in the reports, but only anonymously ensuring that the interviewee cannot be identified from the quote.

A debriefing/validation meeting of the initial findings (not yet conclusions or recommendations) may be arranged in Helsinki in December. The purpose of the possible seminar would be to share initial findings and also validate them.

The MFA will not organise interviews or meetings with the stakeholders on behalf of the evaluation team, but will assist in identification of people and organizations to be included in the evaluation.

Deliverables/meetings: Debriefing/validation workshop supported by PowerPoint presentation on the preliminary results. A workshop on initial findings in Helsinki.

Participants in the MFA workshop: EVA-11, reference group, other relevant staff/stakeholders, the Team Leader in person (responsible for chairing the session) and the evaluation team (can be arranged via Skype).

D. REPORTING AND DISSEMINATION PHASE

The reporting and dissemination phase will take place in January 2018 and produce the Final report. Dissemination of the results is organized during this phase.

The report should be kept clear, concise and consistent. The report must follow writing instructions and template provided by EVA-11 and it should contain inter alia the evaluation findings, conclusions and recommendations. The logic between those should be clear and based on evidence.

The final draft report will be sent for a round of comments by the parties concerned. The purpose of the comments is only to correct any misunderstandings or factual errors. The time needed for commenting is 3 weeks.

The final draft report must include abstract and summaries (including the table on main findings, conclusions and recommendations). It **must be of high and publishable quality**. It must be ensured that the translations use commonly used terms in development cooperation. The consultant is responsible for the editing, proof-reading and quality control of the content and language.

The report will be finalised based on the comments received and **must be ready in February 2018**. The final report must include abstract and summaries (including the table on main findings, conclusions and recommendations) in Finnish, Swedish and English. The final report will be delivered in Word-format (Microsoft Word 2010) with all the tables and pictures also separately in their original formats. Online translators cannot be used with MFA document materials.

As part of reporting process, the Consultant will submit a methodological note explaining how the quality control has been addressed during the evaluation. The Consultant will also submit the EU Quality Assessment Grid as part of the final reporting.

In addition, the MFA requires access to the evaluation team's interim evidence documents, e.g. completed matrices, although it is not expected that these should be of publishable quality. The MFA treats these documents as confidential if needed.

Deliverables: Final report (draft final report and final report).

A management meeting on the final results may be organized in Helsinki tentatively in January 2018 and the Team Leader must be present in person.

A public presentation on the results will be organized on the same visit as the possible management meeting. It is expected that at least the Team leader is present.

A public Webinar will be organized and recorded by EVA-11. Team leader will give short presentation of the findings in a public Webinar. Presentation can be delivered from distance. Only a sufficient internet connection is required.

The MFA will prepare a management response to the recommendations.

8. EVALUATION TEAM

The Team Leader will lead the work and will be ultimately responsible for the deliverables. The competencies of the team members shall be complementary. All team members shall have fluency in English and at least one team member must have fluency in Finnish, because part of the documentation is available only in Finnish. The Team Leader and the team have to be available until the reports have been approved by the Development Evaluation Unit (EVA-11), even when the timetables change.

9. BUDGET

The evaluation will not cost more than 200 000 € (VAT excluded).

10. MANDATE

The evaluation team is entitled and expected to discuss matters relevant to this evaluation with pertinent persons and organizations. However, it is not authorized to make any commitments on behalf of the Government of Finland or the Ministry. The evaluation team does not represent the Ministry for Foreign Affairs of Finland in any capacity.

All intellectual property rights to the result of the Service referred to in the Contract will be exclusive property of the Ministry, including the right to make modifications and hand over material to a third party. The Ministry may publish the end result under Creative Commons license in order to promote openness and public use of evaluation results.

11. AUTHORISATION

Helsinki, 1.9.2017

Jyrki Pulkkinen

Director

Development Evaluation Unit

Ministry for Foreign Affairs of Finlan

ANNEX 2: DOCUMENTS CONSULTED

EU Commission. (2015). ROM Handbook. Results Oriented Monitoring. Brussels, Belgium.

Independent Evaluation Group-World Bank. (2007). Sourcebook for Evaluating Global and Regional Partnership Programs. Indicative Principles and Standards. Washington, D.C., USA.

NORAD. (2017). The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation (01). Oslo, Norway.

OECD. (2010). Quality Standards for Development Evaluation. DAC Guidelines and Reference Series. Paris, France.

United Nations Evaluation Group. (2016). Norms and Standards for Evaluation. New York, USA

UN Women. (2017). What can we learn from UN-Women Evaluations? A meta-analysis of evaluations managed by UN-Women in 2016. (UNW/2017/CRP.10). New York, USA.

ANNEX 3: ANALYSIS GRID

Evaluation question	Data sources used	Data analysis method
For the meta-evaluation:		
1. How can MFA's decentralised evaluation portfolio be described?	51 evaluation reports, List of project implementation as of 2014, 3 Finnish Development Policies	Descriptive statistics, light touch qualitative content analysis
2. How is the quality of MFA's decentralised evaluation reports?	51 evaluation reports	Quality assessment tool, descriptive statistics
3. How is the quality of the corresponding ToRs?	45 ToRs	ToR assessment tool, descriptive statistics
4. How is the quality of MFA's decentralised evaluation portfolio classified by countries, sectors, evaluation types, commissioner, etc. if applicable?	51 evaluation reports	Quality assessment tool, descriptive statistics
5. Is there a difference between the quality of MFA-commissioned evaluations and the quality of evaluation that are commissioned by MFA's partners?	51 evaluation reports	Quality assessment tool, descriptive statistics
6. Are there systematic patterns regarding the quality of the evaluation reports and corresponding ToRs?	51 evaluation reports, 45 ToRs	Quality assessment tool, ToR assessment tool, descriptive statistics
7. How reliable are the decentralised evaluation reports?	51 evaluation reports	Quality assessment tool, descriptive statistics
8. Are there gaps regarding MFA's evaluation capacity?	51 evaluation reports, 45 ToRs, MFA Manual, Manual for Bilateral Programmes	Quality assessment tool, ToR assessment tool, descriptive statistics, qualitative content analysis,
9. What are recommendations to improve the quality of MFA's decentralised evaluations?	Findings of the Meta-evaluation	Expert judgement
For the summative meta-analysis:		
10. What can be said about the relevance of Finnish development cooperation based on the reliable decentralised evaluation reports?	47 evaluation reports	Content assessment tool, descriptive statistics
11. What can be said about the effectiveness of Finnish development cooperation based on the reliable decentralised evaluation reports?	45 evaluation reports	Content assessment tool, descriptive statistics
12. What can be said about the efficiency of Finnish development cooperation based on the reliable decentralised evaluation reports?	46 evaluation reports	Content assessment tool, descriptive statistics
13. What can be said about the impact of Finnish development cooperation based on the reliable decentralised evaluation reports?	28 evaluation reports	Content assessment tool, descriptive statistics
14. What can be said about the sustainability of Finnish development cooperation based on the reliable decentralised evaluation reports?	39 evaluation reports	Content assessment tool, descriptive statistics
15. What can be said about the consideration of gender equality in Finnish development cooperation based on the reliable decentralised evaluation reports?	50 evaluation reports	Content assessment tool, descriptive statistics

Evaluation question	Data sources used	Data analysis method
16. What can be said about the consideration of reduction of inequality in Finnish development cooperation based on the reliable decentralised evaluation reports?	50 evaluation reports	Content assessment tool, descriptive statistics
17. What can be said about the consideration of climate sustainability in Finnish development cooperation based on the reliable decentralised evaluation reports?	50 evaluation reports	Content assessment tool, descriptive statistics
18. What can be said about the consideration of the human rights-based approach in Finnish development cooperation based on the reliable decentralised evaluation reports?	50 evaluation reports	Content assessment tool, descriptive statistics
19. What can be said about the aid effectiveness of Finnish development cooperation based on the reliable decentralised evaluation reports?	23-36 evaluation reports	Content assessment tool, descriptive statistics
20. What can be said about the complementarity of Finnish development cooperation based on the reliable decentralised evaluation reports?	11 evaluation reports	Content assessment tool, descriptive statistics
21. What can be said about the coordination of Finnish development cooperation based on the reliable decentralised evaluation reports?	32 evaluation reports	Content assessment tool, descriptive statistics
22. What can be said about the coherence of Finnish development cooperation based on the reliable decentralised evaluation reports?	8 evaluation reports	Content assessment tool, descriptive statistics
23. What can be said about the overall quality of Finnish development cooperation based on the reliable decentralised evaluation reports?	50 evaluation reports	Content assessment tool
24. What are the major strengths emerging from the reliable decentralised evaluation reports?	Findings of the summative analysis	Expert judgement
25. What are the major challenges emerging from the reliable decentralised evaluation reports?	Findings of the summative analysis	Expert judgement
26. What are the major recommendations to improve Finnish development cooperation emerging from the reliable decentralised evaluation reports?	50 evaluation reports	Qualitative content analysis

ANNEX 4: METHODOLOGICAL DETAILS OF THE QUALITY ASSESSMENT TOOL

In the following, the different sub-sections of the quality assessment tool are introduced. For the exact specifications within the sub-sections please refer to Annex 7 where the instrument is presented in its entire complexity.

The first section on the introduction and the background contains five sub-sections. Documents of all agencies and organisations named above confirm them as important elements of an evaluation report (see table 3).

Table 7: Quality analysis tool, section 1

1. Introduction and background	
1.1 Rationale and purpose	Purpose, intended user
1.2 Objectives of the evaluation	Objectives of evaluation
1.3 Evaluation object	Time period, budget, intervention area, components of the intervention, target group, objectives of intervention, stakeholders, implementation arrangements, changes in implementation
1.4 Scope of evaluation	Scope, coherence of scope with ToR
1.5 Evaluation questions	Evaluation questions
1.6 Results of previous evaluations	Results of previous evaluations reported

The second and largest section refers to methodological aspects. It comprises key elements to decide upon the credibility of the evaluation (see table 4). With this part, amongst others, the meta-evaluation team undertook assessments of the methods applied (e.g. triangulation) and their correct application.

Table 8: Quality analysis tool, section 2

2. Methodology	
2.1 Evaluation design	Evaluation approach, evaluation design
2.2 Sources of evidence	Sources of information, triangulation of information sources
2.3 Data collection	Data collection techniques, mix of data collection techniques, assessment of correct application, validity & reliability of data
2.4 Sampling	Sample, sampling strategy & justification, assessment of sampling strategy
2.5 Data analysis methods	Data analysis methods, triangulation of methods, correct application of methods
2.6 Limitations and challenges	Limitations regarding: data collection, evaluation process, data analysis methods; influence of limitations, scoping strategies

The third section comprises the context and the intervention logic. The MFA manual does not provide much specification on this chapter, but emphasises the need to establish a connection between the context and the intervention. In addition, we derived further aspects from other sources and structured them as shown in table 5.

Table 9: Quality analysis tool, section 3

3. Context and intervention logic	
3.1 Context	Context analysis: key actors in the sector, international policies or strategies, Finnish development policies or strategies, national policies, country context, cross-cutting topics
3.2 Intervention Logic	Intervention logic, results model, underlying assumptions

The section on findings is another centre piece of the evaluation report. As a first part within this section, the soundness of the analysis and the usage of the sources mentioned in the methodological chapter were analysed. The second part refers to causal inference and its critical discussion. Afterwards, the content of the paragraphs on the DAC criteria was analysed in detail to check whether the right content is treated under the different sub-criteria of the DAC criteria (see table 6).

Table 10: Quality analysis tool, section 4

4. Findings	
4.1 Findings	Evidence-based findings, application of triangulation
4.2 Causal Inference	Discussion of attribution, confounding factors
4.3 Relevance	Existence in report, correct thematic coverage
4.4 Effectiveness	Existence in report, correct thematic coverage
4.5 Efficiency	Existence in report, correct thematic coverage
4.6 Impact	Existence in report, correct thematic coverage
4.7 Sustainability	Existence in report, correct thematic coverage

For the next two sections on conclusions and recommendations, the logical reasoning from subsequent chapters forms an important aspect (see table 7). This means that conclusions should be derived from findings and recommendations should be informed by conclusions. According to the MFA manual, conclusions should be structured along the DAC criteria and recommendations need to be as concrete as possible to ease their implementation. Hence, we developed some criteria which facilitated assessment in this regard. Please note the relevance of the recommendations could not be assessed by the meta-evaluation team as further programme or project specific details would have been necessary for such an assessment.

Table 11: Quality analysis tool, sections 5 & 6

5. Conclusions	Derived from findings, DAC Criteria
6. Recommendations	Derived from conclusions, directed to actors, prioritised, responsible actor, time bound, lessons learned

The seventh section refers to the annexes at hand (see table 8). Even though this section was sometimes not available to the meta-evaluators (as it was sometimes neglected by original evaluators or stored with ambiguous titles in MFA's archives), it adds important information to the analysis. It strengthens the credibility of the report and proves the methodological sound implementation of the evaluation. The first parts of this section refer to the original evaluation team and its composition. Afterwards, the ToR and other annexes are covered. Additionally, we included a check for data collection instruments provided in the annex. From our perspective these are vital for full and transparent reporting. However, none of the consulted sources requested to include this aspect. Hence, we did not punish evaluation reports by giving a poor overall assessment based on lacking the respective annexes. We rather aimed at identifying good practices and at sensitising the MFA for the importance of providing the data collection instruments.

Table 12: Quality analysis tool, section 7

7. Annex	
7.1 Evaluation Team	Presentation, justification, gender balance, thematic expertise, evaluation expertise, local expertise, lack of independence
7.2 ToR	ToR
7.3 Other Annexes	List of people interviewed, documents consulted, internal quality assurance, external quality assurance, two pager, data collection instruments

With assessing the annexes, we completed the chronological review of the reports and furthermore looked at aspects covering the report as whole. First, the integration of the four cross cutting topics "gender equality", "reduction of inequality" "combating HIV/Aids" "climate sustainability" were assessed and additionally we checked for the presence of the human rights-based approach which is closely connected to Finnish development cooperation policy (see table 9). Thereby, we acknowledged that different policies refer to different cross-cutting objectives or thematically close concepts with a different wording in conjunction with an earlier policy. At this stage, we only checked whether the ToR requested to treat a cross-cutting objective and whether the evaluation report covers the topic.

Table 13: Quality analysis tool, section 8

8. Cross-cutting topics	
8.1 Gender equality/rights of women and girls	Topic required by ToR, Integration of Topic
8.2 Reduction of inequality/equal opportunities to participate/rights of the most vulnerable	Topic required by ToR, Integration of Topic
8.3 Combating HIV/Aids	Topic required by ToR, Integration of Topic
8.4 Climate sustainability/climate change preparedness and mitigation	Topic required by ToR, Integration of Topic
8.5 Human rights-based approach	Topic required by ToR, Integration of Topic

The next section covers further general issues. Aspects combined in this section are highly diverse regarding topics and are often not connected to each other. They comprise the documentation of the evaluation process, the structure and style of the report and the coverage of evaluation questions (see table 10).

Table 14: Quality analysis tool, section 9

9. General issues	
9.1 Documentation on evaluation process	Deviations from planned evaluation, validation by stakeholders
9.2 Structure and style	Structure, editing, readability
9.3 Evaluation questions	Rather comprehensive coverage of evaluation questions

The last aspect was very difficult to detect as the original evaluator did not always state the question first and then answer it. Hence, we only looked at a tendency whether the evaluation questions are rather comprehensively captured. For the readability of the document, the application of additional readability apps was considered but due to data protection concerns not pursued. Similarly, as above we looked at a tendency and provide a yes/no answer.

The quality assessment ends with the analysis of the executive summary (see table 11). We put it at the end of the assessment as only then it can be decided if the summary is consistent with the report. We checked for its existence, completeness, style and language.

Table 15: Quality analysis tool, section 10

10. Summary	
10.1 Executive summary	Deviations from planned evaluation, validation by stakeholders
10.2 Completeness of summary	Rationale, objectives, intervention, scope of evaluation, evaluation design, methods, findings, conclusions, recommendations, summarising table, lessons learned
10.3 Style	Clear language of summary
10.4 Consistency	Consistency of summary with report

Finally, the meta-evaluation team provided an indication on whether the report is a potential example of best practice or whether it discloses severe quality problems which would lead to elimination from the following content analysis.

ANNEX 5: QUALITY ASSESSMENT TOOL

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
13	**Introduction and Context**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(1.1+1.2+1.3+1.4+1.5 + 3.2*2)/7**			**10**	**34**	**7**	**51**
1	**Introduction and background**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(1.1+1.2+1.3+1.4+1.5)/5**	**0**	**0**	**8**	**27**	**16**	**51**
1.1	**Rationale and purpose**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(1.1a*2+1.1b)/3		5		23	23	51
1.1a	Report describes **purpose** for evaluation.	no (1), yes (4)	general statement on rational/purpose		5			46	51
1.1b	Report describes **intended user(s) of evaluation**.	no (1), yes (4)	Organizations/divisions/persons are described that will use the results of the evaluation.		28			23	51
1.2	**Objectives of the evaluation: Report describes objectives of evaluation.**	no (1), yes (4)	statement on objectives		5			46	51
1.3	Evaluation object	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(1.3a+1.3b+1.3c+1.3d+1.3e+1.3f+1.3g+1.3h+1.3i)/9		1	11	19	20	51
1.3a	The description of the **intervention** includes **time period**.	no (1), yes (4)	Start AND end of intervention		7			44	51
1.3b	The description of the **intervention** includes **budget**.	no (1), yes (4)			13			38	51
1.3c	The description of the intervention includes **intervention area**.	no (1), yes (4)	Description where exactly the intervention takes places in the country/region.		15			36	51
1.3d	The description of the intervention includes **components of the intervention**.	no (1), yes (4)	Different components of the intervention are described		10			41	51
1.3e	The description of the intervention includes **target groups**.	no (1), yes (4)	Who is going to benefit from the intervention?		15			36	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
1.3f	The description of the intervention includes **objectives of the intervention**.	no (1), yes (4)			6			45	51
1.3g	The description of the intervention includes **stakeholders**.	no (1), yes (4)	(4) different stakeholder groups are mentioned e.g. (N)Go's, implementers, external experts, (secondary) beneficiaries		17			34	51
1.3h	The description of the intervention includes **implementation arrangements (incl. organizational set-up)**.	no (1), yes (4)	(4) Which partners are involved in the project/program? What is their labour division? With whom was the project negotiated?		16			35	51
1.3i	The description of the intervention includes **changes regarding implementation**.	no (1), yes (4)			36			15	51
1.4	**Scope of evaluation**	no (1), yes (4)	1.4a		18			33	51
1.4a	The **scope** of the evaluation is described.	no (1), yes (4)	What is evaluated? Time, area, components		18			33	51
1.4b	The **scope is coherent with ToR**, otherwise justification is given.	no w/o justification (1), no w/ justification or yes (4), no ToR available, n.T., n.a.	In case of large differences ask MFA for IR.	23 (no ToRs or no scope given)	2			26	28
1.5	**Evaluation questions are reported.**	no eq reported (1), few eq are reported (2), more than half of eq or the main eq are reported (3), all eq are reported (4)	(2) only few eq are reported, the selection seems arbitrary, (3) given a different priorities, the main eq e.g. heading eqs are reported or at least half of the eq are reported, also in annex ok with reference.		20	1	5	25	51
1.6	**Results of previous evaluations are mentioned.**	no (1), yes (4)			30			21	51
2.	**Methodology**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(2.1+2.2+2.3+2.4+2.5+2.6)/6**			**25**	**22**	**4**	**51**
2.1	**Evaluation design**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(2.1a+2.1b)/2		25		15	11	51
2.1a	The **general evaluation approach** is described.	no (1), yes (4)	participatory, theory-based, formative, exploratory, empowerment etc. mixed methods		28			23	51
2.1b	The **evaluation design** is described.	no (1), yes (4)	A design is development. I.e. is there a strategy on how to answer the evaluation questions e.g. pre-post design, comparison groups, contribution analysis,		37			14	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
2.2	**Sources of evidence**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(2.2a*2+2.2b+2.2c+2.2d+2.2e+2.2f+2.2g+ 2.2h+2.2i*2+7.3a + 7.3b/13			2	36	13	51
2.2a	The **sources of information** are described.	no (1), short and incomplete (2), short and complete (3), detailed and complete (4)	(2) one sentence, cryptic, incomplete, not naming types of documents or different groups to be interviewed etc., (3) at least two sentences and naming all sources of information, (4) minimum one paragraph with three or more sentences with all sources of information		2	9	10	30	51
2.2b	**Project documents** have been used in the evaluation.	no (1), yes (4)			1			50	51
2.2c	**M&E** data has been used in the evaluation.	no (1), yes (4)			28			23	51
2.2d	**Additional literature** has been used in the evaluation.	no (1), yes (4)			21			30	51
2.2e	The **implementing organisation(s)** has/ have been used as source of information for the evaluation.	no (1), yes (4)			2			49	51
2.2f	The **beneficiaries** have been used as source of information of the evaluation.	no (1), yes (4)			10			41	51
2.2g	The **institutional environment** e.g. external experts, (N)GOs have been used as source of information in the evaluation.	no (1), yes (4)			26			25	51
2.2h	**Other source(s)** of information has/have been used	no (1), yes -specify- (4)			49			2	51
	specify:	free input							
2.2i	The **mix of sources** of information is **appropriate** (data triangulation).	completely inappropriate (1), rather inappropriate (2), rather appropriate (3), completely appropriate (4)	(1) only secondary data or only one source, (2) two sources, (3) three sources, (4) three or more source with mixture of primary and secondary data.			1	14	36	51
2.3	**Data collection**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(2.3a*2+2.3b+2.3c+2.3d+2.3e+2.3f+2.3g* 2+2.3h*2+2.3i*2+2.3j*2+7.3f)/16			22	27	2	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
2.3a	**Data collection techniques** are described in the report.	no (1), short and incomplete (2), short and complete (3), detailed and complete (4)	(2) one sentence, cryptic, incomplete, not naming techniques etc., (3) at least two sentences and naming all techniques, (4) minimum one paragraph with three or more sentences with all techniques		3	5	17	26	51
2.3b	**Interviews** have been conducted in the evaluation.	no (1), yes (4)	If not method section, indications from findings can be considered					51	51
2.3c	**Focus group discussions** have been conducted.	no (1), yes (4)	If not method section, indications from findings can be considered		23			28	51
2.3d	**Participatory observation** has been conducted.	no (1), yes (4)	If not method section, indications from findings can be considered		34			17	51
2.3e	A **survey(s)** has been conducted.	no (1), yes (4)	If not method section, indications from findings can be considered		30			21	51
2.3f	**Other data collection method(s)** has/have been used	no (1), yes -specify- (4)			41			10	51
	specify:	free input							
2.3g	A **mix of data collection techniques** is applied.	no (1), yes (4)	(1) only one, (4) two or more		9			42	51
2.3h	**Data collection techniques are applied without severe failures.**	no (1), yes (4)	(1) e.g. extreme size of focus group discussions, survey population size smaller than 50		5			46	51
2.3i	**Validity** of data is **assessed** by the evaluators.	no (1), yes (4)	There is a paragraph discussing the validity. Measure the instruments what they want to measure? Discussion of internal vs. external validity.		43			8	51
2.3j	**Reliability** of data is **assessed** by the evaluators.	no (1), yes (4)	There is a paragraph discussing the reliability of data e. g. would a repetition of the study yield the same results?		44			7	51
2.4	**Sampling**	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)	(2.4a*2+2.4b*2+2.4c)/5		28	4	12	7	51
2.4a	The **sample** is described.	no (1), brief and incomplete (2), moderate but incomplete (3), complete (4)	(1) no information at all, (2) very incomplete information (e.g. total number of persons involved), (3) incomplete information (e.g. number of persons involved and affiliations but information not connected to the data collection instruments), (4) detailed information (number of persons and affiliation for each data collection technique are provided)		24	7	10	10	51
2.4b	The **sampling strategy** is described.	no (1), yes (4)	Methods or criteria to select the persons from whom to collect data are described.		32			19	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
2.4c	The evaluators **justify the sampling strategy**.	no (1), yes (4)	Reasons for the sampling strategy are described.		43			8	51
2.4d	**Data collection acknowledges all groups of key stakeholders.**	no (1), yes (4), n.a.	Compare purpose and sampling strategy. Are groups involved who are key stakeholders given the purpose of the evaluation?	51					51
2.5	**Data analysis methods**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(2.5a+2.5b+2.5c)/3		4	14	25	8	51
2.5a	**Data analysis methods** are described.	no (1), brief and incomplete (2), moderate but incomplete (3), complete (4)	(1) no information at all, (2) very incomplete information (for few data the data analysis method is described), (3) incomplete information (for most data the data analysis method is described), (4) detailed information (for each data the data analysis method is described)		21	16	6	8	51
2.5b	The **mix of data analysis methods** is **appropriate** (triangulation of methods).	no (1), yes (4)	Qualitative and quantitative analysis methods are used e.g. content analysis, grounded theory, summary statistics, correlations, cross tabulations. Focus on mixture of qualitative analysis and quantitative analysis (tables with figures). This does not mean primary quantitative data has to be collected, but at least secondary data like project documents have to be analysed quantitatively.		17			34	51
2.5c	**Data analysis methods are applied without severe failures.**	no (1), yes (4)	e.g. ignoring basic statistics like mixing up pure numbers and causal effects, generalizing based on single interviews etc.		11			40	51
2.6	**Limitations and challenges**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(2.6a*2+2.6b+2.6c+2.6d*2+2.6e)/7		19	10	12	10	51
2.6a	Limitations regarding **data collection** are described.	no (1), yes (4)			19			32	51
2.6b	Limitations regarding the **evaluation process** are described.	no (1), yes (4)			30			21	51
2.6c	Limitations regarding **data analysis methods** are described.	no (1), yes (4)			45			6	51
2.6d	**Possible influence of limitations on the evaluation is discussed.**	no (1), yes (4)			32			19	51
2.6e	**Coping strategies** for limitations are described.	no (1), yes (4)			38			13	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
3.	**Context and intervention logic**		**No aggregation: Context combined with introduction section, intervention logic integrated in findings**						
3.1	**Context**	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)	(3.1b+3.1c+3.1d+3.1e+3.1f+3.1g+3.1h+3.1i+3.1j*2)/10	7	1	18	22	3	51
3.1a	A **context analysis** is provided in the report.	no (1), yes (4)			7			44	51
3.1b	In the context analysis it is referred to **(inter) national key actors** in the sector.	no (1), yes (4), n.a.		7	19			25	51
3.1c	In the context analysis it is referred to **international policies or strategies**.	no (1), yes (4), n.a.		7	21			23	51
3.1d	In the context analysis it is referred to **Finnish development policies or strategies**.	no (1), yes (4), n.a.		7	31			13	51
3.1e	In the context analysis it is referred to **national/regional policies** (e.g. sector strategies, poverty reduction policies).	no (1), yes (4), n.a.		7	19			25	51
3.1f	In the context analysis it is referred to the **country/regional context (socio-economic, political, cultural factors if applicable)**.	no (1), yes (4), n.a.		7	13			31	51
3.1g	In the context analysis it is referred to **gender (in)equality**.	no (1), yes (4)		7	30			14	51
3.1h	In the context analysis it is referred to **(reduction of) inequality**.	no (1), yes (4)		7	33			11	51
3.1i	In the context analysis it is referred **to climate (sustainability)**.	no (1), yes (4)		7	33			11	51
3.1j	Overall, the **context description is in relation with intervention**.	no (1), rather no (2), rather yes (3), yes (4), n.a.	(1) not at all in relation, (2) few parts in relation, (3) most parts in relation, (4) all parts in relation (direct reference important)	7	3	3	15	23	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
3.2	**Intervention logic**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(3.2a*2+3.2b+3.2c*2+3.2d)/6		13	15	13	10	51
3.2a	The **intervention logic** (IL), logical framework (LF), program theory (PT) or the theory of change (ToC) is described.	no (1), brief and incomplete (2), moderate (3), complete and comprehensive (4), n.a.	(1) not at all, (2) one-two sentences, rather cryptic, incomplete (3) paragraph or table, giving an idea but program does not become fully clear or table is not described in the text, (4) minimum one paragraph with three sentences and very comprehensive table with explanation or very detailed description without table, logic of the programme becomes clear, overall comprehensive and easy to understand, (n.a.) if evaluators mentions the lack of an (appropriate) framework		16	7	13	15	51
3.2b	A **results model (IOOI)** is provided.	no (1), yes (4)	Input, expected output, outcome and impact are in the report.		43			8	51
3.2c	The IL, **LF, PT, ToC or the (IOOI) is assessed** by the evaluator as appropriate, otherwise shortcomings are disclosed.	no (1), yes (4), n.a.			21			30	51
3.2d	**Underlying assumptions** of the intervention logic are **reviewed** by evaluator.	no (1), yes (4), n.a.			39			11	50
4.	**Findings**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(4.1*4+4.2+3.2+4.34567)/5**		**4**	**27**	**20**		**51**
4.1	**Findings**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(4.1a+4.1b+4.1c)/3		11	20	8	12	51
4.1a	Findings are **evidence-based**.	no (1), yes (4)	The findings refer clearly to the data collected.		27			24	51
4.1b	**Results are put into perspective with referral to different data sources.**	no (1), rather no (2), rather yes (3), yes (4)	(1) not put at all into perspective, (2) very rarely put into perspective e.g. only two, three times within the report, (3) often parts put into perspective e.g. around half of the results, (4) vast majority put into perspective (e.g. interviews showed xx but the focus groups came to different results. Or in the survey respondents showed xx which was confirmed by the interviews.)		30	7	4	10	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
4.1c	**Only findings are presented.** (No conclusions, no recommendations)	no (1), yes (4)			17			34	51
4.2	**Causal Inference**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(4.2a+4.2b)/2		34		12	5	51
4.2a	**Attribution** of intervention to results is discussed.	no (1), yes (4)	Evaluators critically discuss the ability of the intervention to attribute to the results.		34			17	51
4.2b	**Confounding factors** are discussed.	no (1), yes (4)			46			5	51
4.34567	DAC Criteria		(4.3+4.4+4.5+4.6+4.7)/5			16	33	2	51
4.3	**Relevance**		4.3b	2	1	14	23	11	51
4.3a	**Relevance** is discussed.	no (1), yes (4)	RELEVANCE IS ALWAYS LINKED TO THE INTERVENTION		2			49	51
4.3b	Relevance is **appropriately captured**.	no (1), rather no (2), rather yes (3), yes (4)	READ THE SECTION AND RATE 4.3c-g, AFTERWARDS ASSESS SECTION IN GENERAL CONSIDERING THESE ASSESSEMENTS.	2	1	14	23	11	51
4.3c	Does the report discuss, if the intervention **meets the needs of the target group**?	no (1), yes (4), n.a.	n.a. if there is no target group (i.e. only final beneficiaries), CODE GOOD PRACTICE	4	8			39	51
4.3d	Does the report discuss, if the intervention **meets the needs of the final beneficiaries** (population)?	no (1), yes (4)	CODE GOOD PRACTICE	2	16			33	51
4.3e	Does the report discuss, if the intervention is consistent and supportive of the partner government/ regional policies?	no (1), yes (4)	CODE GOOD PRACTICE	2	7			42	51
4.3f	Does the report discuss, if the intervention is **consistent with the MFA development cooperation policy**?	no (1), yes (4)	CODE GOOD PRACTICE	2	29			20	51
4.3g	Does the report discuss, if the intervention is **addressing international conventions, policies, strategies or goals**?	no (1), yes (4)	CODE GOOD PRACTICE	2	23			26	51
4.4	**Effectiveness**		4.4b		2	13	26	10	51
4.4a	**Effectiveness** is discussed.							51	51
4.4b	Effectiveness is **appropriately captured**.	no (1), rather no (2), rather yes (3), yes (4)	READ THE SECTION AND RATE 4.4c-g AFTERWARDS ASSESS SECTION IN GENERAL CONSIDERING THESE ASSESSEMENTS.		2	13	26	10	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
4.4c	Does the report discuss, if the **outputs** of the intervention **have been achieved**?	no (1), yes (4)	CODE GOOD PRACTICE		5			46	51
4.4d	Does the report discuss, if the **outcomes** of the intervention **have been achieved**?	no (1), yes (4)	CODE GOOD PRACTICE		7			44	51
4.4e	Does the report discuss, if the intervention has resulted in **benefits for the target group**?	no (1), yes (4), n.a.	n.a. if there is no target group (i.e. only final beneficiaries), CODE GOOD PRACTICE	3	19			29	51
4.4f	Does the report discuss, if the intervention has resulted in **benefits for the final beneficiaries**?	no (1), yes (4)	CODE GOOD PRACTICE		23			28	51
4.4g	Does the report discuss, if the **results are different for men and women**? (differentiate between men and women?)	no (1), yes (4)	CODE GOOD PRACTICE		23			28	51
4.5	**Efficiency**		4.5b	2	1	16	16	16	51
4.5a	**Efficiency** is discussed.	no (1), yes (4)			2			49	51
4.5b	Efficiency is **appropriately captured**.	no (1), rather no (2), rather yes (3), yes (4)	READ THE SECTION AND RATE 4.5c-f, AFTERWARDS ASSESS SECTION IN GENERAL CONSIDERING THESE ASSESSEMENTS.	2	1	16	16	16	51
4.5c	Does the report discuss, if the implementation of the **intervention is/was on time**?	no (1), yes (4)	CODE GOOD PRACTICE	2	12			37	51
4.5d	Does the report discuss, if the **inputs have been converted into high quality outputs**?	no (1), yes (4)	CODE GOOD PRACTICE	2	22			27	51
4.5e	Does the report discuss, if the intervention is **efficient regarding costs**?	no (1), yes (4)	CODE GOOD PRACTICE	2	9			40	51
4.5f	Does the report discuss, if the intervention is **efficient regarding personnel**?	no (1), yes (4)	CODE GOOD PRACTICE	2	21			28	51
4.5g	Does the report discuss, if the **implementation management is efficient**?	no (1), yes (4)	CODE GOOD PRACTICE	2	9			40	51
4.6	**Impact**		(4.6b*2+4.6c)/3	11	3	18	16	3	51
4.6a	**Impact** is discussed.	no (1), yes (4)			11			40	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
4.6b	Impact is **appropriately captured**.	no (1), rather no (2), rather yes (3), yes (4)	READ THE SECTION AND RATE 4.6c-g, AFTERWARDS ASSESS SECTION IN GENERAL CONSIDERING THESE ASSESSEMENTS.	11	3	18	16	3	51
4.6c	Does the report discuss, if the intervention **contributed to its overall objective, reach its intended impact**?	no (1), yes (4)	CODE GOOD PRACTICE	11	10			30	51
4.6d	Does the report discuss, if the intervention has any **unintended impacts**?	no (1), yes (4)	only unintended impacts not distinguished between positive and negative CODE	11	33			7	51
4.6e	Does the report discuss, if the intervention **contributes to enhance the quality of life of the final beneficiaries**?	no (1), yes (4)	CODE GOOD PRACTICE	11	19			21	51
4.6f	Does the report discuss, if the intervention **contributes to enhance institutional quality** (i.e. institutions/services in the partner country/ region have been improved)?	no (1), yes (4), n.a.	n.a. if the intervention does not address the institutional level, CODE GOOD PRACTICE	12	14			25	51
4.6g	Does the report discuss, if the intervention **contributed to changes in the partner country's/ region's policies/ to sector reforms**?	no (1), yes (4)	CODE GOOD PRACTICE	11	22			18	51
4.7	**Sustainability**		4.7b	5	3	15	19	9	51
4.7a	**Sustainability** is discussed.	no (1), yes (4)			5			46	51
4.7b	Sustainability is **appropriately captured**.	no (1), rather no (2), rather yes (3), yes (4)	READ THE SECTION AND RATE 4.7c-f, AFTERWARDS ASSESS SECTION IN GENERAL REFLECTING THESE ASSESSEMENTS.	5	3	15	19	9	51
4.7c	Does the report discuss the **economic sustainability** of the intervention?	no (1), yes (4)		5	15			31	51
4.7d	Does the report discuss the **social sustainability** of the intervention?	no (1), yes (4)	e.g. intervention is accepted by population, approach useful for population etc.	5	22			24	51
4.7e	Does the report discuss the **environmental sustainability** of the intervention?	no (1), yes (4), n.a.		5	36			10	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
4.7f	Does the report discuss the sustainability as a **multifaceted concept**?	no (1), yes (4)		5	15			31	51
4.7g	Does the report discuss if the **benefits** of the intervention are **likely to continue after the completion of the intervention**? (i.e. Do the final beneficiaries further benefit after the intervention ends?)	no (1), yes (4)	CODE GOOD PRACTICE	5	13			33	51
4.7h	Does the report discuss, if the **target group/beneficiaries has the capacity to make the intervention sustainable**?	no (1), yes (4), n.a.	n.a. if there is no target group (i.e. only final beneficiaries), CODE GOOD PRACTICE	5	12			34	51
4.7i	Does the report discuss, if the **target group/beneficiaries has the financial means to make the intervention sustainable**?	no (1), yes (4), n.a.	n.a. if there is no target group (i.e. only final beneficiaries), CODE GOOD PRACTICE	5	27			19	51
4.7j	Does the report discuss, if the **implementing partner organisations / intermediaries have the institutional capacity to make the intervention sustainable**?	no (1), yes (4)	often the same as target group, but can be different e.g. International NGO, Consulting etc. CODE GOOD PRACTICE	6	16			29	51
4.7k	Does the report discuss, if the **implementing partner organisations / intermediaries have the financial means to make the intervention sustainable**?	no (1), yes (4)	often the same as target group, but can be different e.g. International NGO, Consulting etc. CODE GOOD PRACTICE	6	23			22	51
56.	**Conclusions and Recommendations**		**(5.a+6.)/2**		**3**	**10**	**32**	**6**	**51**
5.	**Conclusions**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(5.a*4+5.b+5.c+5.d+5.e+5.f)/10**		**7**	**7**	**17**	**20**	**51**
5.a	**Conclusions are derived from findings.**	no (1), yes (4)	not necessarily direct reference but perceived as consistent with findings.	4	9			38	51
5.b	**Relevance** is discussed.	no (1), yes (4)		4	15			32	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
5.c	**Effectiveness** is discussed.	no (1), yes (4)		4	12			35	51
5.d	**Efficiency** is discussed.	no (1), yes (4)		4	18			29	51
5.e	**Impact** is discussed.	no (1), yes (4)		4	25			22	51
5.f	**Sustainability** is discussed.	no (1), yes (4)		4	20			27	51
6.	**Recommendations**	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**	**(6.a*2+6.b+6.c+6.d+6.e+6.f)/7**		**6**	**23**	**18**	**4**	**51**
6.a	**Recommendations are derived from findings and conclusions.**	no (1), rather no (2), rather yes (3), yes (4)	(1) no logical conjunction to conclusions, (2) very rarely logical conjunction to conclusions e.g. only two, three times, (3) often logical conjunction to conclusions e.g. around half, (4) for vast majority logical conjunction to conclusions or findings	1	2	7	8	33	51
6.b	Recommendations are **directed to actors**.	no (1), yes (4)	majority of recommendations is clearly directed to actors	1	18			32	51
6.c	Recommendations are **prioritised**.	no (1), yes (4)		1	47			3	51
6.d	Recommendations indicate an **actor responsible for implementation**.	no (1), yes (4)	More concrete indication than 'directed to actor'	1	40			10	51
6.e	Recommendations are **time-bound**.	no (1), yes (4)		1	43			7	51
6.f	**Lessons learned are derived.**	no (1), yes (4)			22			29	51
7.	**Annex**								
7.1	**7.1 Evaluation Team**		7.1h	41		1	1	8	51
7.1a	**Team members** are presented.	no (1), yes (4)			26			25	51
7.1b	Team **composition is justified**.	no (1), yes (4)			48			3	51
7.1c	Team is **gender-balanced**, according to report.	no (1), yes (4); n.a.		18	21			12	51
7.1d	Team has **thematic expertise**, according to report.	no (1), yes (4); n.a.		41	1			9	51
7.1e	Team has **evaluation expertise**, according to report.	no (1), yes (4); n.a.		42	1			8	51
7.1f	Team has **local expertise**, according to report.	no (1), yes (4); n.a.		35	1			15	51
7.1g	There is incidence in the report for **lack of independence**.	no (1), yes (4)			50			1	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
7.1h	Team composition is **appropriate**. (agg)	completely inappropriate (1), rather inappropriate (2), rather appropriate (3), completely appropriate (4), n.a.	*summary indicator from above, (1) three or more of the following; incidence for lack of independence, no local expertise, no evaluation expertise, no thematic expertise, and no gender-balance, (2) if max. three of the former, (3) only gender-balance and one other item can be missing but not lack of independence, (4)only gender-balance can be missing*	41		1	1	8	51
7.2	**Report contains ToRs.**	no (1), yes (4)			12			39	51
7.3	**Other annexes**								
7.3a	Report contains **list of people interviewed**.	no (1), yes (4)			7			44	51
7.3b	Report contains **documents consulted**.	no (1), yes (4)			9			42	51
7.3c	Report addresses **internal quality assurance**.	no (1), yes (4)			43			8	51
7.3d	Report addresses **external quality assurance**.	no (1), yes (4)			41			10	51
7.3e	Report contains a **two-pager as communication tool**.	no (1), yes (4)			47			4	51
7.3f	**Data collection instruments are provided.**	no (1), rather no (2), rather yes (3), yes (4)	(1) no data collection instruments are provided, (2) one data collection instrument, (3) most data collection instruments, (4) all data collection instruments		33	5	5	8	51
8.	**Cross-cutting topics**				**6**	**31**	**14**	**0**	**51**
8.1	**Gender equality/ rights of women and girls** is integrated in the report.	no (1), rather no (2), rather yes (3), yes (4)	(1) no integrated at all, (2) integrated only sporadically in few (e.g. only two chapters) (3) reference to topics in findings, conclusions and recommendations but not comprehensively, (4) integrated in findings, conclusions and recommendation with separate sections or paragraphs		4	11	12	24	51
8.2	**Reduction of inequality/equal opportunities to participate/ rights of the most vulnerable** is integrated in report.	no (1), rather no (2), rather yes (3), yes (4)	(1) no integrated at all, (2) integrated only sporadically in few (e.g. only two chapters) (3) reference to topics in findings, conclusions and recommendations but not comprehensively, (4) integrated in findings, conclusions and recommendation with separate sections or paragraphs		15	12	9	15	51
8.3	**Combating HIV/Aids** is integrated in report.	no (1), rather no (2), rather yes (3), yes (4)	(1) no integrated at all, (2) integrated only sporadically in few (e.g. only two chapters) (3) reference to topics in findings, conclusions and recommendations but not comprehensively, (4) integrated in findings, conclusions and recommendation with separate sections or paragraphs		49			2	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
8.4	**Climate sustainability/climate change preparedness and mitigation** is integrated in report	no (1), rather no (2), rather yes (3), yes (4)	(1) no integrated at all, (2) integrated only sporadically in few (e.g. only two chapters) (3) reference to topics in findings, conclusions and recommendations but not comprehensively, (4) integrated in findings, conclusions and recommendation with separate sections or paragraphs		27	3	6	15	51
8.5	**Human rights-based approach** is integrated in report	no (1), rather no (2), rather yes (3), yes (4)	(1) no integrated at all, (2) integrated only sporadically in few (e.g. only two chapters) (3) reference to topics in findings, conclusions and recommendations but not comprehensively, (4) integrated in findings, conclusions and recommendation with separate sections or paragraphs		27	10	5	9	51
9.	**General issues**								
9.1	**Documentation on evaluation process**								
9.1a	**Deviations from planned** implementation of evaluation are described.	no (1), yes (4)			41			10	51
9.1b	Report mentions **validation by stakeholders**, i.e. validation workshop.	no (1), yes (4)	Project staff, representatives of beneficiaries, implementing organisation		31			20	51
9.1c	Report mentions **validation by MFA** or other commissioners.	no (1), yes (4)			35			16	51
9.2	**Structure and style**								0
9.2a	Report is **structured according to MFA template**. (check annex)	no (1), yes (4)	Summary, Introduction, Methodology, Context Analysis, Findings, Conclusions, Recommendations, References, Evaluation Team, ToR, People Interviewed, Documents Consulted, xxx, if chapters missing, specify in comments		38			13	51
9.2b	Report is **properly edited**.	no (1), yes (4)	Clear labelling of graphs and tables. Clear headlines and visual structure.		6			45	51
9.2c	Report is written in **clear language**.	no (1), yes (4)			6			45	51
9.3	**Evaluation questions**								
	The evaluation report answers evaluation questions defined in the ToR.	no (1), yes (4)	comment and be rather generous. n.a. of ToR or evaluation questions missing.	7	7			37	51

No.	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
10.	**Summary**	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(10.2*4+10.3+10.4*2)/9**		**3**	**14**	**31**	**3**	**51**
10.1	**Report contains executive summary.**	no (1), yes (4)			2			49	51
10.2	**Completeness of summary**	inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)	(10.2a+10.2b+10.2c+10.2d+10.2e+10.2f+ 10.2g+10.2h+10.2i+10.2j*4+10.2h)/14	2	4	18	25	2	51
10.2a	Summary describes **rationale/purpose** of evaluation.	no (1), yes (4)		2	17			32	51
10.2b	Summary describes **objectives of evaluation**.	no (1), yes (4)	Look for elaborations.	2	24			25	51
10.2c	Summary describes the **intervention**.	no (1), yes (4)		2	11			38	51
10.2d	Summary describes the **scope of the evaluation**.	no (1), yes (4)	time, area, components	2	23			26	51
10.2e	Summary describes the **evaluation design**.	no (1), yes (4)		2	34			15	51
10.2f	Summary describes the **methods**.	no (1), yes (4)		2	27			22	51
10.2g	Summary describes the **findings**.	no (1), yes (4)		2	7			42	51
10.2h	Summary describes the **conclusions**.	no (1), yes (4)		2	12			37	51
10.2i	Summary describes **recommendations**.	no (1), yes (4)	Also within summarising table ok	2	4			45	51
10.2j	Summary contains a **summarising table (incl. findings, conclusions and recommendations)**.	no (1), very incomplete (2), partly incomplete (3), complete (4)	(1) no table at all, (2) incomplete table with only findings, conclusions OR recommendations, (3) incomplete table with only two of this three elements, (4) complete table.	2	34	3	4	8	51
10.2h	Summary describes **lessons learned**.	no (1), yes (4)		2	32			17	51
10.3	**Style**								
	Summary is written in **clear language**.	no (1), yes (4)		2	2			47	51
10.4	**Consistency**								
	Summary is **consistent** with report.	no (1), yes (4)		2	2			47	51
	OVERALL RATING of the Evaluation Report	**inadequate (1), need for improvement (2), satisfactory (3), good or very good (4)**	**(13+2+4+56+10)/5**		**1**	**17**	**32**	**1**	**5**

ANNEX 6: TOR ASSESSMENT TOOL

	Specification	Rating 1–4	Guidance	Missing/ not appli-cable/no ToRs	1	2	3	4	Total
1. Intervention		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**2**	**16**	**23**	**4**	**51**
1.1	Context of the development intervention	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)		6	8	23	11	3	51
	Finnish policy context	no (1), yes (4)		6	32			13	51
	international policy context	no (1), yes (4)		6	26			19	51
	target area's policy context	no (1), yes (4)		6	17			28	51
	development context	no (1), yes (4)		6	27			18	51
	context with respect to cross-cutting issues	no (1), yes (4)		6	35			10	51
1.2	reference to relevant issues of previous evaluations	no (1), yes (4)		6	33			12	51
1.3	Objectives, strategies and implementation of the Intervention	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)		6	4	6	14	21	51
	description of intervention objectives	no (1), yes (4)		6	5			40	51
	description of implementation strategies of the intervention	no (1), yes (4)		6	11			34	51
	description of resources for implementation of the intervention	no (1), yes (4)		6	23			22	51
	reference to cross-cutting issues relevant for intervention	no (1), yes (4)		6	38			7	51
	description of stakeholders and their role	no (1), yes (4)		6	16			29	51
	description of period of the intervention	no (1), yes (4)		6	7			38	51
2. Purpose, objectives, and scope of the evaluation		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**		**10**	**32**	**3**	**51**
	Rationale and purpose	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		6	2	1	32	10	51
	rationale for evaluation	no (1), yes (4)		6	3			42	51
	rationale for point of time of evaluation	no (1), yes (4)		6	31			14	51
	intended users of evaluation	no (1), yes (4)		6	17			28	51

	Specification	Rating 1–4	Guidance	Missing/ not appli-cable/no ToRs	1	2	3	4	Total
	intended use of evaluation	no (1), yes (4)		6	6			39	51
	Objectives	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)		6	2		39	4	51
	objectives of the evaluation	no (1), yes (4)		6	2			43	51
	prioritization of objectives	no (1), yes (4)		6	41			4	51
	Scope	inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)		6	4	15	25	1	51
	intervention dimensions to be evaluated	no (1), yes (4)		6	21			24	51
	stakeholder groups involved	no (1), yes (4)		6	17			28	51
	geographical area	no (1), yes (4)		6	22			23	51
	time span	no (1), yes (4)		6	12			33	51
	connection of evaluation to other supporting sectors or themes	no (1), yes (4)		6	41			4	51
3. Evaluation questions		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**10**		**27**	**8**	**51**
	evaluation questions adapted to the specific information needs	no (1), yes (4), n.a. (no questions)		13	4			34	51
	maximum of 12 evaluation questions	no (1), yes (4), n.a.		13	29			9	51
4. Evaluation criteria	**relevant criteria for the evaluation (OECD/DAC, and coherence and aid effectiveness)**	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**		**5**	**36**	**4**	**51**
	relevance	no (1), yes (4)		6	1			44	51
	effectiveness	no (1), yes (4)		6	1			44	51
	efficiency	no (1), yes (4)		6				45	51
	impact	no (1), yes (4)		6	8			37	51
	sustainability	no (1), yes (4)		6	4			41	51
	coherence	no (1), yes (4)		6	32			13	51
	complementarity	no (1), yes (4)		6	37			8	51
	coordination	no (1), yes (4)		6	37			8	51
	aid effectiveness	no (1), yes (4)		6	38			7	51
5. Methodology		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**2**	**22**	**18**	**3**	**51**
	request for mix of qualitative and quantitative methods	no (1), yes (4)		6	20			25	51
	request for triangulation	no (1), yes (4)		6	28			17	51
	request for disaggregated analysis	no (1), yes (4)		6	37			8	51
	specification of available materials	no (1), yes (4)		6	22			23	51
	specification of envisaged data collection techniques	no (1), yes (4)		6	11			34	51

	Specification	Rating 1–4	Guidance	Missing/ not appli-cable/no ToRs	1	2	3	4	Total
	specification of envisaged data analysis techniques	no (1), yes (4)		6	37			8	51
6. Feasibility	**Scope of work and given timeframe and resources are feasible.**	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**							
	evaluation budget in ToR	no (1), yes (4)		6	25			20	51
	feasible scope of evaluation given budget	no (1), yes (4)		31	6			14	51
	feasible scope of evaluation given time resources	no (1), yes (4)		9	12			30	51
7. Evaluation Process and QA	**The evaluation process is clearly explained in the ToR.**	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**1**	**24**	**16**	**4**	**51**
Evaluation process		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		6		3	27	15	51
	outline of phases of evaluation process	no (1), yes (4)		6	4			41	51
	outline of sequencing of activities	no (1), yes (4)		6	6			39	51
	outline of approximate dura-tion of activities	no (1), yes (4)		6	14			31	51
	place of work for activities	no (1), yes (4)		6	22			23	51
	specification of roles and responsibilities of commissioner and evaluator(s)	no (1), yes (4)		6	15			30	51
Deliverables		inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)		6	1		20	24	51
	specification of deliverables	no (1), yes (4)		6	1			44	51
	specification of milestones with timeline	no (1), yes (4)		6	21			24	51
Quality assurance	reference to what kind of qual-ity assurance is desired	no (1), yes (4)		6	28			17	51
8. Overarching and cross-cutting criteria		**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**11**	**16**	**15**	**3**	**51**
Gender equality	pointing to gender equality as cross-cutting issue	no (1), yes (4)		6	15			30	51
	requested to by analysed by evaluator	no (1), yes (4)		6	16			30	52
Reduction of Inequality	pointing to reduction of ine-quality as cross-cutting issue	no (1), yes (4)		6	27			18	51
	requested to by analysed by evaluator	no (1), yes (4)		6	27			18	51
HIV/AIDS	pointing to HIV/AIDS as cross-cutting issue	no (1), yes (4)		6	39			6	51

	Specification	Rating 1–4	Guidance	Missing/ not applicable/no ToRs	1	2	3	4	Total
	requested to by analysed by evaluator	no (1), yes (4)		6	41			4	51
Climate sustainability	pointing to climate sustainability as cross-cutting issue	no (1), yes (4)		6	28			17	51
	requested to by analysed by evaluator	no (1), yes (4)		6	28			17	51
Human rights based approach	pointing to HRBA as cross-cutting issue	no (1), yes (4)		6	24			21	51
	requested to by analysed by evaluator	no (1), yes (4)		6	24			21	51
Ethics	request for ethical considerations	no (1), yes (4)		6	38			7	51
	Overall Rating:	**inadequate (1), need for improvement (2), satisfatcory (3), good or very good (4)**		**6**	**18**			**27**	**51**

ANNEX 7: METHODOLOGICAL DETAILS ON THE CONTENT ASSESSMENT TOOL

The following tables 13-17, display the general structure of the sections related to the DAC criteria exclusive of detailed formal aspects. They can be found in Annex 10 were the instrument is presented in its entire complexity.

TABLE 2. Content analysis tool, section 1

1. Relevance
Assessment of relevance of the intervention by the evaluators
According to the evaluators, does the intervention meet the needs of the target groups?
Reasons provided for the positive / negative assessment
According to the evaluators, does the intervention meet the needs of the final beneficiaries?
Reasons provided for the positive / negative assessment
According to the evaluators, is the intervention consistent and supportive of the partner government / regional policies?
According to the evaluators, is the intervention consistent with the MFA development cooperation policy?
According to the evaluators, is the intervention addressing international conventions, policies, strategies or goals?
Is this section of the report a success story?

Content analysis tool, section 2

2. Effectiveness
Assessment of effectiveness of the intervention by the evaluators
According to the evaluators, have the outputs of the intervention been achieved?
Reasons provided for the positive / negative assessment
According to the evaluators, have the outcomes of the intervention been achieved?
Reasons provided for the positive / negative assessment
According to the evaluators, has the intervention resulted in benefits for the target group?
Reasons provided for the positive / negative assessment
According to the evaluators, has the intervention resulted in benefits for the final beneficiaries?
Reasons provided for the positive / negative assessment
According to the evaluators are results different for men and women?
Is this section of the report a success story?

Content analysis tool, section 3

3. Efficiency
Assessment of efficiency of the intervention by the evaluators
According to the evaluators, is / was the implementation of the intervention on time?
Reasons provided for the positive / negative assessment
According to the evaluators, have the inputs been converted into high quality outputs?
Reasons provided for the positive / negative assessment
According to the evaluators, is the intervention efficient regarding costs?

Reasons provided for the positive / negative assessment
According to the evaluators, is the intervention efficient regarding personnel?
Reasons provided for the positive / negative assessment
According to the evaluators, is the implementation management efficient?
Reasons provided for the positive / negative assessment
Is this section of the report a success story?

Content analysis tool, section 4

4. Impact
Assessment of impact of the intervention by the evaluators
According to the evaluators, did the intervention contribute to its overall objective / reach its intended impact?
Reasons provided for the positive / negative assessment
According to the evaluators, does the intervention have any unintended positive impacts?
Specification of unintended impacts, Reasons provided for the positive / negative assessment
According to the evaluators, does the intervention have any unintended negative impacts?
Specification of unintended impacts, Reasons provided for the positive / negative assessment
According to the evaluators, does the intervention contribute to enhance the quality of life of the final beneficiaries?
Reasons provided for the positive / negative assessment
According to the evaluators, does the intervention contribute to enhance institutional quality (i.e. institutions / services in the partner country / region have been improved?
Reasons provided for the positive / negative assessment
According to the evaluators, has the intervention contributed to changes in the partner country's / region's policies or contributed to sector reforms?
Reasons provided for the positive / negative assessment
Specification of change, Reasons provided for the positive / negative assessment
Is this section of the report a success story?

Content analysis tool, section 5

5. Sustainability
Assessment of sustainability of the intervention by the evaluators
According to the evaluators, are benefits of the intervention likely to continue after the completion of the intervention? (i.e. Do the final beneficiaries further benefit after the intervention ends?)
Reasons provided for the positive / negative assessment
According to the evaluators, does the target group have the capacity to make the intervention sustainable? (i.e. knowledge, know-how)
According to the evaluators, does the target group have the financial means to make the intervention sustainable?
According to the evaluators, do the implementing partner organisations have the institutional capacity to make the intervention sustainable?
According to the evaluators, do the implementing partner organisations have the financial means to make the intervention sustainable?
According to the evaluators, does the intervention have an exit strategy?
Is this section of the report a success story?

In a subsequent step, aid effectiveness and triple C were analysed from a Finnish perspective. All aspects covered by the sub-sections listed in table 18 should ideally underlie the interventions and lead as a consequence to higher aid effectiveness. For more details please refer to Annex 10.

TABLE 3. Content analysis tool, Section 6

6. Aid effectiveness and Triple C (Coherence, Complementarity, Coordination)
Assessment of **aid effectiveness** of the intervention by the evaluators
According to the evaluators, has the intervention promoted • Ownership? • Alignment of priorities? • Harmonisation of aid? • Management for development results? • Mutual accountability for outcomes?
Assessment of the **complementarity** of the intervention with EU member states' or other donors' interventions by the evaluators
Assessment of **coordinating** activities connected to the intervention by the evaluator (i.e. Was the intervention coordinated with other initiatives implemented by the same organisation by other donors?)
How do the evaluators assess the **coherence** of the intervention with other Finnish policies beyond development cooperation in the evaluation report?

The second part of the content assessment tool is connected to learning from conclusions and recommendations of the reports. Thus, in the next two sections lessons learnt and recommendations were captured in detail. Therefore, we applied thematic coding in MaxQDA and allocated both - lessons learnt and recommendations - to different statements. Whenever a lesson learnt or a recommendation did not fit to any category, it was captured under the section "others". Table 19 shows the different main categories which were the same for lessons learnt and recommendations.

TABLE 4. Content analysis tool, section 7 and 8

7. Lessons learnt and 8. Recommendations
Financial
Personnel
Time
Capacity
Equipment
Management
Communication
Scope
Participation
Outreach
M&E
Relevance
Effectiveness
Efficiency
Impact
Sustainability
Aid effectiveness

The content tool ended with a question whether the evaluation report describes an exemplary success story.

ANNEX 8: CONTENT ASSESSMENT TOOL

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
1. Relevance								
1.1a Relevance is discussed.	no (1), yes (4)	Transferred from quality tool		2			48	50
1.1b Relevance is methodologically appropriately captured.	n.a., no (1), rather no (2), rather yes (3), yes (4)	Transferred from quality tool, if 'no' or 'n.a.' no analysis of this subsection possible, all 'n.a.'	2	1	13	22	12	50
1.2 **How do the evaluators assess the relevance** of the intervention in the evaluation report?	n.a., not relevant at all (1), somewhat relevant (2), moderately relevant (3), highly relevant (4)	n.a. report is not analysing this aspect, (1) all aspects analysed in relevance are assessed negatively, (2) most aspects analysed in relevance are assessed negatively, (3) most aspects analysed in relevance are assessed positively, (4) all aspects analysed in relevance are assessed positively PLEASE ONLY REFER TO THE REPORT NOT TO THE EXECUTIVE SUMMARY, PLEASE ONLY CODE THE OVERALL ASSESSMENTS PROVIDED IF ANY. THIS IS NOT THE PLACE TO CODE ALL DETAILS; THEY ARE CAPTURED BELOW.	3	1	4	12	30	50
1.3a According to the evaluators, does the intervention meet the **needs of the target group** (i.e. those for whom the intervention has been designed)?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention does not meet the needs of the target group, (2) intervention does mostly not meet the needs of the target group, (3) intervention does somehow meet the need of the target group, (4) intervention does mostly meet the needs of the target group PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE AS E.G. LINK TO POLICIES IS MADE, THEN RATE THIS SUBQUESTION WITH N.A. IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 1.4a (ONLY IN THE EXCEL)	15		4	8	23	50
1.3b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words. PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 1.4b (ONLY IN THE EXCEL)						
1.3c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words. PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 1.4c (ONLY IN THE EXCEL)						

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
1.4a According to the evaluators, does the intervention meet the **needs of the final beneficiaries** (i.e. the local/ poor people)?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention does not meet the needs of the final beneficiaries, (2) intervention does mostly not meet the needs of the final beneficiaries, (3) intervention does somehow meet the need of the final beneficiaries, (4) intervention does mostly meet the needs of the final beneficiaries PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY E.G. ONLY DISCUSSION ON THE LOCAL GOVERNMENT AS TARGET GROUP, THEN RATE THIS SUBQUESTION WITH N.A. THUS, THIS IS NOT ABOUT GUESSING YOURSELF HOW IMPROVED LOCAL GOVERNMENT IS RELEVANT FOR POOR PEOPLE.	25		4	7	14	50
1.4b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words. PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
1.4c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words. PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
1.5 According to the evaluators, is the intervention consistent and supportive of the partner government/ regional policies?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is inconsistent with partner government policies, (2) the intervention is mostly not consistent and supportive of partner government policies, (3) the intervention is mostly consistent and supportive of partner government polices, (4) the intervention is fully consistent and supportive partner government policies PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED. FOR NATIONAL INTERVENTIONS LOOK AT PARTNER GOVERNMENT FOR REGIONAL INTERVENTIONS, AT REGIONAL POLICIES E.G. AU POLICIES.	11		1	11	27	50
1.6 According to the evaluators, is the intervention **consistent with the MFA development cooperation policy**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is inconsistent with the MFA development cooperation policy, (2) the intervention is mostly not consistent with the MFA development cooperation policy, (3) the intervention is mostly consistent with the MFA development cooperation policy, (4) the intervention is fully consistent with the MFA development cooperation policy PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED	31		1	3	15	50
1.7 According to the evaluators, is the intervention **addressing international conventions, policies, strategies or goals**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is not addressing international conventions, policies, strategies or goals, (2) the intervention is mostly not addressing international conventions, policies, strategies or goals, (3) the intervention is mostly addressing international conventions, policies, strategies or goals, (4) the intervention is strongly addressing international conventions, policies, strategies or goals PLEASE LOOK AT ECPLICIT ANSWER, IF THE ASPECT IS NOT DISCUSSED RATE N.A. THIS IS NOT THE PLACE TO MENTION THAT CONSISTENCY WITH MFA POLICY DOES INDIRECTLY MEAN CONSISTENCY WITH INTERNATIONAL GOALS. PLEASE KEEP IN MIND IF REGIONAL INTERVENTIONS ONLY REFER TO REGIONAL POLICIES THIS WOULD BE N.A.	29		1	4	16	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
1.8 Is this section of the report a success story?	no (1), yes (4)	Do you have the impression that this section is a very good example for a very successful project? Then select yes. PLEASE FOCUS ON EXTRAORDINARY WORK.		34			16	50
2. Effectiveness								
2.1a Effectiveness is discussed.	no (1), yes (4)	Transferred from quality tool					50	50
2.1b Effectiveness is appropriately captured.	n.a., no (1), rather no (2), rather yes (3), yes (4)	Transferred from quality tool, if 'no' or 'n.a.' no analysis of this subsection possible, all 'n.a.'		2	11	27	10	50
2.2 **How do the evaluators assess the effectiveness** of the intervention in the evaluation report?	n.a., not effective at all (1), somewhat effective (2), moderately effective (3), highly effective (4)	n.a. report is not analysing this aspect, (1) all aspects analysed in effectiveness are assessed negatively, (2) most aspects analysed in effectiveness are assessed negatively, (3) most aspects analysed in effectiveness are assessed positively, (4) all aspects analysed in effectiveness are assessed positively PLEASE ONLY REFER TO THE REPORT NOT TO THE EXECUTIVE SUMMARY, PLEASE ONLY CODE THE OVERALL ASSESSMENTS PROVIDED IF ANY. THIS IS NOT THE PLACE TO CODE ALL DETAILS; THEY ARE CAPTURED BELOW.	5	1	16	22	6	50
2.3a According to the evaluators, have the (short-term) **outputs of the intervention been achieved**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. outputs are not analysed in report, (1) no outputs have been achieved, (2) most outputs have not been achieved, (3) most outputs have been achieved, (4) all outputs have been achieved PLEASE RATE THIS IF THE ASSESSMENT IS AT THE LEVEL OF DIRECT OUTPUTS OF THE INTERVENTION. FOR THE MAJORITY OF REPORTS YOU HAVE TO DECIDE WHETHER TO ANSWER 2.3 OR 2.4. ONLY FOR ANALYTICALLY SOUND REPORTS, AN ASSESSMENT OF BOTH MAY BE POSSIBLE.	10		10	23	7	50
2.3b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IT IS PARTICULARLY IMPORTANT THAT WE DO NOT MIX OUTPUTS AND OUTCOMES HERE.						
2.3c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IT IS PARTICULARLY IMPORTANT THAT WE DO NOT MIX OUTPUTS AND OUTCOMES HERE.						
2.4a According to the evaluators, have the **outcomes of the intervention been achieved**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. outcomes are not analysed in report, (1) no outcomes have been achieved, (2) most outcomes have not been achieved, (3) most outcomes have been achieved, (4) all outcomes have been achieved PLEASE RATE THIS IF THE ASSESSMENT IS RATHER AT THE LEVEL OF LONGTERM OUTCOMES OF THE INTERVENTION.	11	2	15	18	3	49

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
2.4b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IT IS PARTICULARLY IMPORTANT THAT WE DO NOT MIX OUTPUTS AND OUTCOMES HERE.						
2.4c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL. IT IS PARTICULARLY IMPORTANT THAT WE DO NOT MIX OUTPUTS AND OUTCOMES HERE.						
2.5a According to the evaluators, has the intervention **resulted in benefits for the target group** (i.e. those for whom the intervention was designed)?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. results for the target group are not analysed in report, (1) no benefits for the target group have been achieved, (2) very few benefits for the target group have been achieved, (3) moderate benefits for the target group have been achieved, (4) many benefits for the target group have been achieved PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A. IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 2.6a (ONLY IN THE EXCEL)	26		2	13	9	50
2.5b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 2.6b (ONLY IN THE EXCEL)						
2.5c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words IF THE TARGET GROUP IS THE POOR/LOCAL POPULATION PLEASE COPY YOUR RATING TO 2.6c (ONLY IN THE EXCEL)						
2.6a According to the evaluators, has the intervention **resulted in benefits for the final beneficaires** (i.e. the local/ poor people)?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. results for the final beneficiaries are not analysed in report, (1) no benefits for the final beneficiaries have been achieved, (2) very few benefits for the final beneficiaries have been achieved, (3) moderate benefits for the final beneficiaries have been achieved, (4) many benefits for the final beneficiaries have been achieved PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	28	1	7	7	7	50
2.6b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
2.6c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
2.7 What are the main **results of the gender-analysis** provided by the evaluator?	no (1), yes (4)	no, if no gender-analysis.PLEASE CODE MAIN GENDER RESULTS.	14	3			33	50
2.8 Is this section of the report a success story?	no (1), yes (4)	Do you have the impression that this section is a very good example for a very successful project? Then select yes. PLEASE FOCUS ON EXTRAORDINARY WORK.		44			6	50
3. Efficiency								
3.1a Efficiency is discussed.	no (1), yes (4)	Transferred from quality tool		2			48	50
3.1b Efficiency is appropriately captured.	n.a., no (1), rather no (2), rather yes (3), yes (4)	Transferred from quality tool, if 'no' or 'n.a.' no analysis of this subsection possible, all 'n.a.'	2	1	15	15	17	50
3.2 **How do the evaluators assess the efficiency** of the intervention in the evaluation report?	n.a., not efficient at all (1), somewhat efficient (2), moderately efficient (3), highly efficient (4)	n.a. report is not analysing this aspect, (1) all aspects analysed in efficiency are assessed negatively, (2) most aspects analysed in efficiency are assessed negatively, (3) most aspects analysed in efficiency are assessed positively, (4) all aspects analysed in efficiency are assessed positively PLEASE ONLY REFER TO THE REPORT NOT TO THE EXECUTIVE SUMMARY, PLEASE ONLY CODE THE OVERALL ASSESSMENTS PROVIDED IF ANY. THIS IS NOT THE PLACE TO CODE ALL DETAILS; THEY ARE CAPTURED BELOW.	4	3	14	19	10	50
3.3a According to the evaluators, is/was the **implementation of the intervention on time**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is/was not at all on time, (2) the intervention is/was mostly not on time, (3) the intervention is/was mostly on time, (4) the intervention is/was on time or ahead schedule.	20	3	17	6	4	50
3.3b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.3c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
3.4a According to the evaluators, have the **inputs been converted into high quality outputs**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) inputs have not been converted into high quality outputs, (2) the inputs have mostly not been converted into high quality outputs, (3) the inputs have mostly been converted into high quality outputs, (4) all inputs have been converted into high quality outputs PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A. CHECK FOR EXPLICIT ASSESSMENTS ON THE QUALITY, THIS IS NOT THE SAME AS WHETHER SOMETHING HAS BEEN REACHED. HOWEVER, THIS IS EXPECTED TO BE OFTEN N.A.	33	2	6	3	6	50
3.4b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.4c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.5a According to the evaluators, is the **intervention efficient regarding costs**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is not at all cost-efficient regarding costs, (2) the intervention is mostly not cost-efficient, (3) the intervention is mostly cost-efficient, (4) the intervention is fully cost-efficient	13	6	6	14	11	50
3.5b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.5c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.6a According to the evaluators, is the **intervention efficient regarding personnel**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is not efficient regarding personnel, (2) the intervention is mostly not efficient regarding personnel, (3) the intervention is mostly efficient regarding personnel, (4) the intervention is fully efficient regarding personnel PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	28	4	5	11	2	50
3.6b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
3.6c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.7a According to the evaluators, **is the implementation management efficient**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention is not efficient regarding implementation management, (2) the intervention is mostly not efficient regarding implementation management, (3) the intervention is mostly efficient regarding implementation management, (4) the intervention is fully efficient regarding implementation management	11	4	10	17	8	50
3.7b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.7c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
3.8 Is this section of the report a success story?	no (1), yes (4)	Do you have the impression that this section is a very good example for a very successful project? Then select yes. PLEASE FOCUS ON EXTRAORDINARY WORK.		43			7	50
4. Impact								
4.1a Impact is discussed.	no (1), yes (4)	Transferred from quality tool		10	1		39	50
4.1b Impact is appropriately captured.	n.a., no (1), rather no (2), rather yes (3), yes (4)	Transferred from quality tool, if 'no' or 'n.a.' no analysis of this subsection possible, all 'n.a.'	9	5	17	16	3	50
4.2 **How do the evaluators assess the impact** of the intervention in the evaluation report?	n.a., no impact at all (1), some impact (2), moderate impact (3), high impact (4)	n.a. report is not analysing this aspect, (1) the intervention has no impacts at all, (2) the intervention has mostly no impact, (3) the intervention has some impact, (4) the intervention has a high impact PLEASE ONLY REFER TO THE REPORT NOT TO THE EXECUTIVE SUMMARY, PLEASE ONLY CODE THE OVERALL ASSESSMENTS PROVIDED IF ANY. THIS IS NOT THE PLACE TO CODE ALL DETAILS; THEY ARE CAPTURED BELOW.	22	1	10	12	5	50
4.3 According to the evaluators, did the **intervention contribute to its overall objective/reach its intended impact**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention did not contribute, (2) the intervention did contribute very little, (3) the intervention did contribute moderately, (4) the intervention did contribute highly PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A. THIS IS THE PLACE TO LOOK AT HIGHER LEVEL IMPACTS, THERE MIGHT BE OVERLAPS TO LONGTERM OUTCOMES, THIS IS OKAY, HOWEVER DO NOT RATE ANY OUTPUTLEVEL ASSESSMENTS HERE. HERE WE DO NOT ASK FOR REASONS AS THEY ARE CAPTURED IN THE THEMATIC SUB-SECTIONS BELOW.	26		7	12	5	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
4.4a According to the evaluators, does the **intervention have any unintended positive impacts**?	n.a., no (1), yes (2)	n.a. report is not analysing this aspect, (1) the intervention did not have positive unintended impacts, (4) the intervention did have positive unintended impacts PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	42				8	50
4.4b If any, please specify								
4.4c What reasons are provided?		If applicable, list all explanatory factors provided by the evaluators, if ambiguous please specify in key words						
4.5a According to the evaluators, does **the intervention have any unintended negative impacts**?	n.a., no (1), yes (2)	n.a. report is not analysing this aspect, (1) the intervention did not have negative unintended impacts, (4) the intervention did have negative unintended impacts PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	45	3			2	50
4.5b If any, please specify								
4.5c What reasons are provided?		If applicable, list all explanatory factors provided by the evaluators, if ambiguous please specify in key words						
4.6a According to the evaluators, does the **intervention contribute to enhance the quality of life of the final beneficiaries**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention did not contribute, (2) the intervention did contribute very little, (3) the intervention did contribute moderately, (4) the intervention did contribute highly PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A., PLEASE DO NOT JUDGE WHETHER THE ASSESSMENT OF THE EVALUATOR IS VALID FROM YOUR PERSPECTIVE, RATHER CAPTURE THE ANSWER. ONLY IF SOMETHING SEEMS VERY SUSPICIOUS, USE THE COMMENT FIELD.	38		1	5	6	50
4.6b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
4.6c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
4.7a According to the evaluators, does the **intervention contribute to enhance institutional quality** (i.e. institutions/services in the partner country/region have been improved)?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention did not contribute, (2) the intervention did contribute very little, (3) the intervention did contribute moderately, (4) the intervention did contribute highly PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	33	1	3	9	4	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
4.7b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
4.7c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) negative?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
4.8a According to the evaluators, has the **intervention contributed to changes in the partner country's/ region's policies**/ to sector reforms?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) the intervention did not contribute, (2) the intervention did contribute very little, (3) the intervention did contribute moderately, (4) the intervention did contribute highly PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	36	4	3	6	1	50
4.8b If any, please specify								
4.8c What reasons are provided?		If applicable, list all explanatory factors provided by the evaluators, if ambiguous please specify in key words						
4.9 Is this section of the report a success story?	no (1), yes (4)	Do you have the impression that this section is a very good example for a very successful project? Then select yes. PLEASE FOCUS ON EXTRAORDINARY WORK.		48			2	50
5. Sustainability								
5.1a Sustainability is discussed.	no (1), yes (4)	Transferred from quality tool		5			45	50
5.1b Sustainability is appropriately captured.	n.a., no (1), rather no (2), rather yes (3), yes (4)	Transferred from quality tool, if 'no' or 'n.a.' no analysis of this subsection possible, all 'n.a.'	5	2	15	19	9	50
5.2 **How do the evaluators assess the sustainability** of the intervention in the evaluation report?	n.a., not sustainable at all (1), somewhat sustainable (2), moderately sustainable (3), highly sustainable (4)	n.a. report is not analysing this aspect, (1) all aspects analysed in sustainability are assessed negatively, (3) most aspects analysed in sustainability are assessed negatively, (3) most aspects analysed in sustainability are assessed positively, (4) all aspects analysed in sustainability are assessed positively PLEASE ONLY REFER TO THE REPORT NOT TO THE EXECUTIVE SUMMARY, PLEASE ONLY CODE THE OVERALL ASSESSMENTS PROVIDED IF ANY. THIS IS NOT THE PLACE TO CODE ALL DETAILS; THEY ARE CAPTURED BELOW.	11	2	17	16	4	50
5.3a According to the evaluators, are **benefits of the intervention likely to continue after the completion** of the intervention? (i.e. Do the final beneficiaries further benefit after the intervention ends?)	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) benefits are not at all likely to continue, (2) benefits are rather not likely to continue, (3) benefits are rather likely to continue, (4) benefits are likely to continue PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	20	2	10	15	3	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
5.3b What reasons are provided for the positive assessment? (i.e. Why did the evaluators assess it (rather) positive?)		If applicable, list all positive explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
5.3c What reasons are provided for the negative assessment? (i.e. Why did the evaluators assess it (rather) not relevant?)		If applicable, list all negative explanatory factors provided by the evaluators, if ambiguous please specify in key words PLEASE KEEP IN MIND THAT IT HAS TO BE ON A GENERAL LEVEL AS THIS IS FOR A SUMMATIVE ANALYSIS OF FINNISH DEVCO, AT A LATER STAGE WE HAVE TO BE ABLE TO UNDERSTAND THEM FROM THIS EXCEL.						
5.4 According to the evaluators, does the **target group have the capacity to make the intervention sustainable**? (i.e. knowledge, know-how)	n.a., no (1), rather no (2), rather yes (3), yes (4)	"n.a. report is not analysing this aspect, (1) beneficiaries do not have the capacity at all, (2) beneficiaries do rather not have the capacity, (3) beneficiaries rather have the capacity, (4) beneficiaries have the capacity, Consider capacity as comprehensive concept, not only human but also institutional capacity PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A. IN CASE THE TARGET GROUP IS AT THE SAME TIME THE IMPLEMENTING ORGANISATION PLEASE COPY YOUR RATINGS TO 5.6."	23	1	8	15	3	50
5.5 According to the evaluators, does the target group have the financial means to make the intervention sustainable?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) beneficiaries do not have the financial means, (2) beneficiaries do rather not have the financial means, (3) beneficiaries rather have the financial means, (4) beneficiaries have the financial means PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A. IN CASE THE TARGET GROUP IS AT THE SAME TIME THE IMPLEMENTING ORGANISATION PLEASE COPY YOUR RATINGS TO 5.7.	36	3	6	2	3	50
5.6 According to the evaluators, do the **implementing partner organisations have the institutional capacity to make the intervention sustainable**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a.report is not analysing this aspect, (1) partners do not have the institutional capacity, (2) partners do rather not have the institutional capacity, (3) partners rather have the institutional capacity, (4) partners have the institutional capacity PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	26	3	7	11	3	50
5.7 According to the evaluators, do the **implementing partner organisations have the financial means to make the intervention sustainable**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) partners do not have the financial means, (2) partners do rather not have the financial means, (3) partners rather have the financial means, (4) partners have the financial means PLEASE LOOK AT EXPLICIT ANSWERS PROVIDED, IF THERE IS ONLY INDIRECT REFERENCE, THEN RATE THIS SUBQUESTION WITH N.A.	26	1	7	5	1	40
5.8 According to the evaluators, does the **intervention have an exit strategy**?	n.a., no (1), yes (4)	n.a. report is not analysing this aspect, (1) the intervention does not have an exit strategy, (4) the intervention has an exit strategy.	27	19			4	50
5.9 Is this section of the report a success story?	no (1), yes (4)	Do you have the impression that this section is a very good example for a very successful project? Then select yes. PLEASE FOCUS ON EXTRAORDINARY WORK.		48			2	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
6. Aid Effectiveness and triple C from a Finnish perspective								
Aid effectiveness								
6.1a Has aid effectiveness been assessed in the report?	no (1), yes (4)	If only implicitly, please specify in comment		38			12	50
6.1b **How do the evaluators assess the aid effectiveness** of the intervention in the evaluation report?	n.a., not effective at all (1), somewhat effective (2), moderately effective (3), highly effective (4)	n.a. report is not analysing this aspect, (1) all aspects analysed regarding aid effectiveness are assessed negatively, (2) most aspects analysed regarding aid effectiveness are assessed negatively, (3) most aspects analysed regarding aid effectiveness are assessed positively, (4) all aspects analysed regarding aid effectiveness are assessed positively	46	1			3	50
6.1c According to the evaluators, has the **intervention promoted ownership**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention has not promoted ownership, (2) intervention has rather not promoted ownership, (3) intervention has rather promoted ownership, (4) intervention has promoted ownership PLEASE LOOK AT SUPPORT OF LOCAL STRATEGIES e.g. Did this project support the implementation of a local strategy?	13	3	4	14	16	50
6.1d According to the evaluators, has the **intervention promoted alignment of priorities with national/ regional priorities**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention has not promoted alignment of priorities, (2) intervention has rather not promoted alignment of priorities, (3) intervention has rather promoted alignment of priorities, (4) intervention has promoted alignment of priorities PLEASE LOOK AT USE OF INSTITUTIONS IN PARTNER COUNTRY/REGION e.g. Did the project use a local institution and procedures to manage the intervention?	21	1	2	13	13	50
6.1e According to the evaluators, is the **intervention embedded in activities by Finland to harmonise aid**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention is not embedded, (2) intervention is rather not embedded in activities of harmonisation of aid , (3) intervention is rather embedded in activities of harmonisation of aid, (4) intervention is completely embedded in activities of harmonisation of aid (complete strategy or several activities are mentioned) E.g. avoiding of duplication of activities, streamling of activities	36		3	2	9	50
6.1f According to the evaluators, has the **intervention promoted management for development results**? (i.e. Did the intervention work and report towards outcomes and impacts?)	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention has not promoted management for development results, (2) intervention has rather not promoted management for development results, (3) intervention has rather promoted management for development results, (4) intervention has promoted management for development results	14	9	10	10	7	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
6.1g According to the evaluator has the **intervention promoted mutual accountability for outcomes**?	n.a., no (1), rather no (2), rather yes (3), yes (4)	n.a. report is not analysing this aspect, (1) intervention has not promoted mutual accountability for outcomes, (2) intervention has rather not promoted mutual accountability for outcomes, (3) intervention has rather promoted mutual accountability for outcomes, (4) intervention has promoted mutual accountability for outcomes PLEASE LOOK AT TRANSPARENCY OF THE INTERVENTION. (e.g. Was transparency promoted? Was information on the intervention's results publicly available and discussed?)	34		5	6	5	50
Complementarity								
6.4a Is the complementarity of the intervention with other donor's activities assessed in the report? (i.e. Did the intervention support thematically/ policy-wise other interventions funded by other EU member states (or other donors)?)	no (1), yes (4)	PLEASE FOCUS ON WHETHER EU ACTOR'S POLICIES ARE IN LINE WITH EACH OTHER. E.G. IF THIS PROJECT WAS ABOUT AGRIBUSINESS PROMOTING DEFORESTATION AND THE NEARBY GERMAN PROJECT SUPPORTED FOREST CONSERVATION, THERE WAS LACK OF COMPLEMENTARITY.		39			11	50
6.4b **How do the evaluators assess the complementarity** of the intervention?	n.a., no complementarity at all (1), somewhat complementarity (2), moderate complementarity (3), high complementarity (4)	n.a. report is not analysing this aspect, (1) all aspects analysed with regard to complementarity are assessed negatively, (2) most aspects analysed with regard to complementarity are assessed negatively, (3) most aspects analysed with regard to complementarity are assessed positively, (4) all aspects analysed with regard to complementarity are assessed positively	39	3		3	5	50
Coordination								
6.5a Are coordinating activities in the intervention assessed in the report? (i.e. Was the intervention coordinated with other initiatives implemented by the same organisation by other donors?)	n.a., no (1), yes (4)	PLEASE FOCUS ON EXCHANGE OF INFORMATION WITH OTHER INTERVENTIONS.		18			32	50
6.5b **How do the evaluators assess the coordinating activities** connected to the intervention?	n.a., no coordination at all (1), somewhat coordination (2), moderate coordination (3), high coordination (4)	n.a. report is not analysing this aspect, (1) all aspects analysed with regard to coordination are assessed negatively, (2) most aspects analysed with regard to coordination are assessed negatively, (3) most aspects analysed with regard to coordination are assessed positively, (4) all aspects analysed with regard to coordination are assessed positively	18	1	8	16	7	50

Specification	Rating 1-4	Guidance	Missing	1	2	3	4	Total
Coherence								
6.6a Is the coherence of the intervention with other Finnish policies assessed in the report? (i.e. Did the intervention support or hamer thematically/ policy-wise other interventions funded by Finalnd in other sectors like education or trade?)	n.a., no (1), yes (4)	PLEASE FOCUS ON WHETHER FINLAND'S POLICIES ARE IN LINE WITH EACH OTHER. THIS IS EVALUATING ISSUES BEYOND DEVELOPMENT COOPERATION.		42			8	50
6.6b **How do the evaluators assess the coherence** of the intervention in the evaluation report?	n.a., no coherence at all (1), somewhat coherence (2), moderate coherence (3), high coherence (4)	n.a. report is not analysing this aspect, (1) all aspects analysed with regard to coherence are assessed negatively, (2) most aspects analysed with regard to coherence are assessed negatively, (3) most aspects analysed with regard to coherence are assessed positively, (4) all aspects analysed with regard to coherence are assessed positively	42	1	2		5	50

ANNEX 9: LIST OF EVALUATION REPORTS RECEIVED AND USED

N°	Year of Report	Report title	MFA Unit	Region	Sector	Project Budget (EUR) by Finland	Project Budget (EUR) Overall	Evaluation Budget (EUR), VAT excluded	ToR available	Used for meta-evaluation
1	2015	FE AU Mediation Support Capacity	ALI-30	Africa	Conflict prevention, resolution, peace and security	3.000.000	3.000.000		Yes	Yes
2	2015	MTR SIP Zambia		Africa	Water and sanitation	9.300.000	18.600.000	22.600	Yes	Yes
3	2015	FE EIBAMAZ Andean Region	ASA-30	Latin America	Education	6.820.000		340.000	Yes	Yes
4	2015	FE EMSP Laos	ASA-10	Asia	Environment/ Climate	9.500.000	9.960.000	90.000	Yes	Yes
5	2015	FE STIFIMO Mozambique	ALI-30	Africa	Communications/ ICT	22.000.000	22.000.000	70.000	Yes	Yes
6	2015	FE MSIA Afghanistan	ASA-40	Eastern Europe and Central Asia	Reproductive Healthcare				No	Yes
7	2015	MTE COWASH Ethiopia	ALI-20	Africa	Water and sanitation	22.000.000	50.000.000	115.000	Yes	Yes
8	2015	MTE Regional Afghanistan	ASA-40	Eastern Europe and Central Asia	Other social services		27.053.564		Yes	Yes
9	2015	MTE REILA Ethiopia	ALI-20	Africa	Agriculture	12.800.000	12.800.000	90.000	Yes	Yes
10	2015	MTR Environment Programme Mekong	ASA	Asia	Environment/ Climate	12.624.996	26.873.207		Yes	Yes
11	2015	MTR Forestry Nepal	ASA-40	Asia	Forestry		135.267.818		Yes	Yes
12	2015	MTR MENA MDTF	ALI-10	MENA	Conflict prevention, resolution, peace and security		6.558.127	27.054	Yes	Yes

N°	Year of Report	Report title	MFA Unit	Region	Sector	Project Budget (EUR) by Finland	Project Budget (EUR) Overall	Evaluation Budget (EUR), VAT excluded	ToR available	Used for meta-evaluation
13	2015	MTR TEVT Nepal	ASA-40	Asia	Education	1.600.000			Yes	Yes
14	2015	Norad's support to UNIDO Trade Capacity Building		Global	Trade policy and regulation				No	No
15	2015	OIOS Eval ITC		Global	Trade policy and regulation				No	No
16	2015	FE Partnership for Market Readiness	KEO-60	Global	Environment/ Climate	4.829.061	102.158.765		Yes	Yes
17	2015	MTE AGRO-BIG Ethiopia	ALI-20	Africa	Agriculture	9.300.000	10.400.000	100.000	Yes	Yes
18	2015	FE BIOCAN		Latin America	Environment/ Climate	4.115.284	6.275.000	180.000	Yes	Yes
19	2015	MTE EEP Southern and Eastern Africa II	ALI-30	Africa	Energy		35.000.000	120.000	Yes	Yes
20	2015	FE FAO-FIN Sustainable Management Forests	KEO-40	Global	Forestry	15.250.000			Yes	Yes
21	2015	MTE FORMIS II Vietnam	ASA-10	Asia	Communications/ICT	9.700.000	10.137.530	75.000	Yes	Yes
22	2015	MTR Sustainable Aquaculture Kyrgyz	ITÄ-20	Eastern Europe and Central Asia	Water and sanitation		1.718.093		Yes	No
23	2015	FE Sustainable Livelihoods Georgia	ITÄ-20	Eastern Europe and Central Asia	Environment/ Climate	1.179.677	1.179.677	12.652	Yes	Yes
24	2015	MTE SWIPSAN	ASA-40	Asia	Conflict prevention, resolution, peace and security	1.556.868	1.556.868		Yes	Yes
25	2015	MTR TA Education Ethiopia		Africa	Education	19.800.000	495.982.000		Yes	Yes
26	2016	FE BIODEV West Africa	ALI-20	Africa	Environment/ Climate	10.000.000	10.000.000	29.900	Yes	Yes
27	2016	FE SWIPSAN	ASA-40	Asia	Conflict prevention, resolution, peace and security	1.541.738	1.541.738		Yes	Yes
28	2016	FE Combat Desertification	ALI-10	MENA	Environment/ Climate	3.000.000	1.188.102		Yes	Yes
29	2016	FE UNESCO Freedom of Expression (Sida-Finida)	ALI-10	MENA	Government and civil society				Yes	Yes

N°	Year of Report	Report title	MFA Unit	Region	Sector	Project Budget (EUR) by Finland	Project Budget (EUR) Overall	Evaluation Budget (EUR), VAT excluded	ToR available	Used for meta-evaluation
30	2016	FE PRESANCA Central America	ASA-30	Latin America	Other social services				Yes	Yes
31	2016	FE Prevention of Violence Central America	ASA-30	Latin America	Conflict prevention, resolution, peace and security				Yes	Yes
32	2016	FE PSDRP Zambia	ALI-30	Africa	Business support services	8.000.000	32.464.276		Yes	Yes
33	2016	FE SEAN Climate Change II Asia	KEO-60	Asia	Environment/ Climate	2.474.034	3.168.409		Yes	Yes
34	2016	MTE Decentralized Forest Zambia	ALI-30	Africa	Forestry	4.384.732	4.384.732	29.000	Yes	Yes
35	2016	MTE Forest and Farm Facility	KEO-60	Global	Agriculture				Yes	Yes
36	2016	MTE Innovation Partnership Programme Vietnam	ASA-10	Asia	Business support services	16.250.000	16.250.000	85.000	Yes	Yes
37	2016	Post-Eval SEAM Nepal	ASA-40	Asia	Environment/ Climate	11.600.000			Yes	Yes
38	2016	FE Policy Dialogue Crimea	ITÄ-20	Eastern Europe and Central Asia	Conflict prevention, resolution, peace and security	1.500.000			Yes	For quality assessment only
39	2016	FE EEP Indonesia	ASA-10	Asia	Energy	4.108.208	4.108.208	80.000	Yes	Yes
40	2016	FE EQUIP Afghanistan	ASA-40	Eastern Europe and Central Asia	Education		466.223.080		No	Yes
41	2016	FE MICCA FAO	KEO-40	Global	Environment/ Climate	6.231.000	> 60.000.000		Yes	Yes
42	2016	MTR Rule of Law Human Rights UNDP Nepal	ASA-40	Asia	Government and civil society		21.642.851		No	Yes
43	2016	MTE Rural Water Supply Sanitation Nepal	ASA-40	Asia	Water and sanitation	13.700.000	21.900.000		Yes	Yes
44	2016	Joint FE School Sector Reform Nepal	ASA-40	Asia	Education		751.187.284	155.000	Yes	Yes

N°	Year of Report	Report title	MFA Unit	Region	Sector	Project Budget (EUR) by Finland	Project Budget (EUR) Overall	Evaluation Budget (EUR), VAT excluded	ToR available	Used for meta-evaluation
45	2016	FE Syrian Voices	ALI-10	MENA	Conflict prevention, resolution, peace and security	400.000		10.000	Yes	Yes
46	2016	FE Green Diplomacy UNEP	KEO-60	Global	Conflict prevention, resolution, peace and security	7.302.195			Yes	Yes
47	2016	MTR WSSCC	KEO-60	Global	Water and sanitation	4.517.139	217.330.295	315.625	Yes	Yes
48	2016	MTE Multi-donor Trust Fund Palestine	ALI-10	MENA	Other social services	8.650.000	97.392.829		Yes	No
49	2016	Summary Report Water Convention		Eastern Europe and Central Asia	Water and sanitation			100.000	No	No
50	2017	MTE ILO Kyrgyzstan Tajikistan	ITÄ-20	Eastern Europe and Central Asia	Other social services		4.000.000	18.937	Yes	Yes
51	2017	MTE ADPP Mosambique	ALI-30	Africa	Business support services	8.000.000	8.800.000	80.000	Yes	Yes
52	2017	MTE Multi-donor Trust Fund Palestine	ALI-10	MENA	Other social services		97.392.829		Yes	Yes
53	2017	MTE PAGE UN	KEO-60	Global	Environment/ Climate	1.383.714	12.011.216		Yes	Yes
54	2017	MTR Recovery Trust Fund Syria	ALI-10	MENA	Conflict prevention, resolution, peace and security		146.500.000		Yes	Yes
55	2017	FE UNWomen Leadership Participation Tanzania	ALI-20	Africa	Government and civil society				No	Yes
56	2017	FE WASH Schools Afghanistan	ASA-40	Eastern Europe and Central Asia	Water and sanitation				Yes	Yes

ANNEX 10: OVERVIEW OF GENERALISED RECOMMENDATIONS PER MAIN TOPIC

Recommendation	N° of reports
Relevance	
To adjust the intervention stronger to the needs of the beneficiaries	5
To adjust the intervention towards enhanced consistency with and/or support of the partner government policies	4
To address international conventions, policies, strategies or goals	3
To continue support largely "as is" because it is deemed very relevant	3
Effectiveness	
To expand the project's activities, change their focus or introduce new ones	11
To improve the work with partners or beneficiaries (via better participation, coordination or capacity building)	6
To focus on consolidating achievements and complete planned activities rather than expanding the project's scope	4
To adjust the results model for the intervention (i.e. programme theory, ToC) and/or improve the quality of baselines and reporting	4
To take into account the context and enabling environment during planning and implementation	1
Efficiency	
To ensure a more adequate and efficient distribution of resources (human, financial, time) through monitoring, visits and assessments	6
To enhance coordination with partners	4
To ensure a better time management	3
Recommendations on the numbers, quality and use of staff	3
Recommendations on management, tools and mechanisms	2
Impact	
Consolidate achievements via scaling-up of pilots, via a second phase/extension of the project or by carrying out additional activities to increase/create impact	4
Adjust the support or management to achieve impact	1
Building trust with beneficiaries and stakeholders	1
Sustainability (Reports containing recommendations on exit strategies are considered in this count and analysis due to the similarity of the two subjects.)	
To develop a sustainability or exit strategy	19
To enhance capacity of the final beneficiaries or implementing partners	14
To ensure financial sustainability either by identifying new sources of funding or by supporting the creation of revenues/development of own financial resources to sustain activities	8
To extend support beyond the initial period, at least in a minor form	6
To ensure that technical issues threatening sustainability are either resolved during the support period or that sufficient capacity for maintenance and repairs is created	5
To support the creation of an enabling environment for sustainability by taking into account the political developments and context, i.e. factors beyond control of the project	4
To disseminate and communicate success stories	1
Aid effectiveness	
To promote ownership by the partner country	6

Recommendation	N° of reports
To promote harmonisation of aid	6
To promote management for development results	5
To promote mutual accountability for outcomes	4
To promote alignment of priorities	1
Complementarity	
To enhance complementarity to other policies, strategies or programmes of the Finnish Government	2
To enhance complementarity to other policies, strategies or programmes of the international community	1
Coordination	
To harmonise policies or programmes by development partners in conjunction with the intervention	16
To harmonize the intervention with input by and interests of local stakeholders	14
Coherence	
Enhance internal coherence for a project or a donor organization	3
Gender	
Improvements to existing gender approaches (strategies, awareness raising, capacity building, recruitment and promotion of women) should be undertaken	12
Gender should be systematically addressed in all project activities	11
Monitoring & Evaluation	
To improve the M&E system	28
To institutionalize monitoring and evaluation	17
To improve the M&E system in terms of data collection	11
To improve the M&E system in terms of efficiency	8
To introduce a results-oriented M&E system	7
To make use of external M&E services	7
To improve the M&E system in terms of data sharing	3
Planning	
Project planning shall be improved in terms of project design and Theory of Change	15
Raise awareness for the importance of project planning, institutionalize planning processes and support implementing institutions in planning	8
Project planning shall be based on a situational analysis and include risk assessment	5
Planning shall be realistic and efficient	4
Provide planning for the remaining period of the project	2
Management	
To change the organisational structure of the project, e.g. by creating new positions or merging/splitting units or shifting responsibilities and tasks	15
To improve the functioning of specific bodies within the project	8
To improve the planning of resources and targets of the project as well as the definition of roles and responsibilities	5
To improve knowledge management within the project	4
To improve the procurement processes and selection of projects	3
To make changes to general approach or modality	2
To implement recommendations (or develop plans how to implement) as soon as possible	2
Scope	2
To extend the scope of activities	16
To narrow or maintain the current geographical scope and scope of activities	10
To extend the geographical scope	7
To assess whether a change in scope makes sense	5

Recommendation	N° of reports
To extend activities to other target groups/beneficiaries	3
Include measures for scaling up already in the project design	1
Time	
Establish a reflected timeline and make use of it efficiently	5
A longer duration of a current/next phase	4
A longer duration for a future intervention	2
Financial	
Improve financial planning/controlling/reporting	9
Improve financing model of the intervention	7
Inform donors about spending decisions	3
Disburse funds on time	2
Concerning wages (Pay higher/equal wages)	2
Mobilize funds for an extension period/ a future period	1
MFA should approve the use of funds for specific purposes	1
Set a spending limit for implementing partners	1
Personnel	
Ensure/improve adequate staffing of institutions of the intervention	9
Create and fill specific key positions	7
Train staff	6
Improve/Adapt recruitment	3
Invest more/adequately in staff	2
Work jointly with Finish Embassy/ MFA staff	2
Limit the time staff has to spend working in dangerous/critical stations	1
Equipment	
Replace old equipment	2
Ensure equipment needs are identified at project start/design	2
Improved technical equipment	1
Ensure availability of supplies for implementing partners	1
Capacity	
To improve the capacity of implementing partners	12
To improve or further develop the quality of capacity building and training activities	8
To conduct further or better assessments of context, needs and stakeholders' capacity	6
To improve the capacity of the beneficiaries to make better use of the services delivered	4
To empower beneficiaries and raise awareness for specific issues related to vulnerable groups	3
Participation	
To enhance participation of stakeholders during project design and management	9
To increase consultation and dialogue with stakeholders for needs assessment and learning purposes	4
Communication	
To improve existing communication methods and channels in terms of frequency and quality (e.g. increase digitalisation and use of web-based media)	14
To start or improve dissemination activities of project knowledge, lessons and results	10
Others	
Liaise with partners to discuss measures and realistic objectives	1
Recognise context and adopt adequate positions and measures	1
Enhance advocacy and support for international best practice examples	1

Recommendation	N° of reports
Conduct further thematic, legal or policy studies	1
Take measures regarding the do-no-harm principle	1
Change the name of the project	1
Follow-up on a human-rights-based approach	1
Ensure the mainstreaming of climate change within the intervention's activities	1

ANNEX 11: STATISTICAL TESTS

Table 16: Project data analysis: Mann-Whitney test for differences between groups

		No. of reports	Means (Euro)	Significance level
Finlands' Project Budget	MFA commissioned	20	9,892,613	0.0199**
	Non-MFA commissioned	14	5,172,025	
Finlands' Project Budget	Individual/independent consultant	12	5,564,218	0.0495**
	Evaluation companies or institutes	22	9,249,545	
Evaluation budget	Individual/independent consultant	4	23,051.25	0.0198**
	Evaluation companies or institutes	17	114,915.4	

Note: *** means significant at the 1% level, ** means significant at the 5% level, * means significant at the 1% level.

Table 17: Report ratings analysis: Mann-Whitney test for differences between groups

		No. of reports	Means (1-4)	Significance level
Rating on Sampling	MFA commissioned	24	1.63	0.0437**
	Non-MFA commissioned	27	2.31	
Rating on methodology	Individual/independent consultant	14	2.19	0.0313**
	Evaluation companies or institutes	37	2.57	
Relevance chapter rating	Final Evaluations	26	2.65	0.0216**
	Mid-term evaluations	23	3.17	
Completeness of Summary	Individual/independent consultants	14	2.12	0.0123**
	Evaluation companies or institutes	35	2.67	

Note: *** means significant at the 1% level, ** means significant at the 5% level, * means significant at the 1% level.

Table 18: ToR ratings analysis: Mann Whitney test for differences between groups

		No. of reports	Means (Euro)	Significance level
Overall ToR Rating	MFA commissioned	22	2.64	0.0117**
	Non-MFA commissioned	23	2.37	
Overall ToR Rating	Individual/independent consultants	13	2.37	0.0657*
	Evaluation companies or institutes	32	2.56	
ToR intervention description	MFA commissioned	22	2.81	0.0087***
	Non-MFA commissioned	23	2.26	
ToR evaluation criteria	MFA commissioned	22	3.00	0.0088***
	Non-MFA commissioned	23	2.67	
ToR evaluation criteria "impact"	MFA commissioned	22	3.86	0.0247**
	Non-MFA commissioned	23	3.09	
ToR methodology	MFA commissioned	22	2.04	0.0324**
	Non-MFA commissioned	23	2.50	
ToR cross cutting topics	MFA commissioned	22	2.50	0.0008***
	Non-MFA commissioned	23	1.67	

Note: *** means significant at the 1% level, ** means significant at the 5% level, * means significant at the 1% level.

Table 19: Spearman Correlation

	No. of reports	Coefficient	Significance level
Overall intervention budget and evaluation budget	19	0.5935	0.0074***
Overall report rating and overall ToR rating	45	0.3044	0.0421**
Overall report rating and ToR: purpose/objectives of evaluation	45	0.4186	0.0041***
Overall report rating and ToR methodology	45	0.3504	0.0183**
Overall report rating and ToR evaluation process	45	0.3438	0.0207**

Note: *** means significant at the 1% level, ** means significant at the 5% level, * means significant at the 1% level.

Table 20: Report summative analysis: Mann-Whitney test for differences between groups

		No. of reports	Means (1-4)	Significance level
Relevance	National intervention	28	3.32	0.0482**
	Regional/global intervention	19	3.79	

Note: *** means significant at the 1% level, ** means significant at the 5% level, * means significant at the 1% level.

META-EVALUATION OF PROJECT AND PROGRAMME EVALUATIONS IN 2015-2017

Ministry for Foreign Affairs of Finland