



EVALUATION

Meta-evaluation of Project and Programme
Evaluations in 2014–2015



Evaluation on Finland's Development Policy and Cooperation

2016/5

EVALUATION

META-EVALUATION OF PROJECT AND PROGRAMME EVALUATIONS IN 2014-2015

Robert LeBlanc (Team Leader)

Max Hennion

Maaria Seppänen



2016/5

This evaluation was commissioned by the Ministry for Foreign Affairs of Finland to Danish Management and Eco-Consult. This report is the product of the authors, and responsibility for the accuracy of the data included in this 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.

© Ministry for Foreign Affairs of Finland 2016

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 978-952-281-494-4 (pdf)

ISSN 2342-8341

Cover design and layout: Innocorp Oy/Milla Toro

CONTENTS

ACRONYMS AND ABBREVIATIONS.....	IX
TIIVISTELMÄ.....	1
REFERAT	2
ABSTRACT	3
YHTEENVETO.....	4
SAMMANFATTNING	12
EXECUTIVE SUMMARY	20
1 INTRODUCTION	28
1.1 Context	28
1.1.1 Background to the meta-evaluation	28
1.1.2 Purposes and objectives of the meta-evaluation.....	29
1.2 Scope	29
1.3 Final report structure	30
2 METHODOLOGICAL CONSIDERATIONS.....	31
2.1 Approach	31
2.2 Methodology overview	31
2.3 Examples of consequences of methodological choices made	32
2.4 Limitations and risk mitigation	33
3 PORTFOLIO ANALYSIS AND QUALITY OF EVALUATION REPORTS.....	37
3.1 Portfolio overview	37
3.2 Portfolio of reports, with an answer to EQ 5 concerning the use of Management Reviews instead of MTE or evaluation	41
4 RESPONSE TO EQ 1: ASSESSMENT OF THE QUALITY OF EVALUATION-RELATED DOCUMENTS	45
4.1 Overview	45
4.2 Answering EQ 6: Analysis of the Quality of Evaluation TOR and Reports – Major Issues Identified	48

5	ANSWERING EQ 7: ASSESSMENT OF THE QUALITY OF APPRAISAL-RELATED DOCUMENTS.....	54
5.1	Overview	54
5.2	Quality of ToR/ITT	55
5.3	Quality of Appraisal Reports.....	57
5.4	Correlation of results between ToR and Reports	60
5.5	Hypotheses concerning reasons for the results obtained.....	61
6	ANSWERING EQ 4: ASSESSMENT OF THE QUALITY OF FINNISH DEVELOPMENT COOPERATION.....	63
6.1	Findings related to 2014-2015 meta-evaluation with respect to the Quality of Finnish Development Cooperation	63
6.1.1	Evaluation Report-based Findings	64
6.1.2	Appraisal report-based Findings.....	81
6.1.3	Induction-based analysis	82
6.1.4	Evaluations not commissioned by MFA	82
6.2	Comparison of 2012–2014 to 2014–2015 meta-evaluations	87
7	CONCLUSIONS	89
7.1	Answering EQ 6 by focussing on Conclusions on the Quality of MFA decentralised evaluation reports and ToR/ITT	89
7.3	Conclusions concerning EQ 4: What evaluation reports reveal about Finnish development cooperation	102
7.4	Concerning avenues to examine for evaluation and appraisal-related capacity improvement	106
8	RECOMMENDATIONS	108
9	DEALING WITH EQ 2: “WHAT IS MFA’S EVALUATION COVERAGE (COMPARISON OF EVALUATION PLANS AND REALIZED EVALUATIONS)?” – CONCERNING THE ESTABLISHMENT OF AN EVALUATION COVERAGE SYSTEM.....	114
	THE EVALUATION TEAM	116
Annex 1	Terms of Reference	118
Annex 2	Key Documents Consulted	124
Annex 3	Detailed Methodology	125
Annex 4	Justifying the use of a mixed Deductive-Inductive Research Approach in Phase 2	143
Annex 5	Portfolio Matrix	146

Annex 6	Quality Assessment Tool for Evaluation ToR/ITT.....	151
Annex 7	Quality Assessment Tool for Evaluation Reports.....	157
Annex 8	Quality Assessment Tool for Appraisal ToR/ITT.....	163
Annex 9	Quality Assessment Tool for Appraisal Reports	168
Annex 10	Quality Assessment Tool for Summary of Finnish Development Cooperation	172
Annex 11	Key Findings, Conclusions and Recommendations (Finnish).....	188
Annex 12	Key Findings, Conclusions and Recommendations (English).....	207

TABLES

Table 1	Ratings given to evaluation reports for the entire n=26 sample of documents studied in Phase One.....	46
Table 2	Ratings given to evaluation ToR for the entire n=26 sample of documents studied in Phase One.....	47
Table 3	Distribution of ratings for the quality assessment grid – evaluation reports	48
Table 4	Lowest and highest scores for ToR and reports according to commissioning agency	49
Table 5	Top-ten evaluation reports by commissioning agency (out of 28 reports; n=28).....	50
Table 6	Reports with scores written by Finnish consultancy companies according to evaluation budget and type	51
Table 7	Average ToR and report scores according to MFA unit (n=26)	51
Table 8	Distribution of ratings for the quality assessment grid – evaluation ToR.....	52
Table 9	Distribution of ratings for the quality assessment grid – appraisal ToR	57
Table 10	Weaknesses in appraisal reports	58
Table 11	Distribution of ratings for the quality assessment grid – appraisal reports	59
Table 12	Correlation between ToR issues, report findings, and report answers	61
Table 13	Ratings given to various analysis areas for Phase Two	77
Table 14	Distribution of ratings for the quality assessment tool – summary of Finnish development cooperation	80
Table 15	Average scores of the evaluation reports per section.....	90
Table 16	Overall scores of appraisal reports.....	97
Table 17	Breakdown of scores in each appraisal report	98
Table 18	Breakdown of scores in each ToR for appraisal.....	100
Table 19	An attribute differentiation table for the Meta-evaluation, based on the complementarity of an inductive research approach to a deductive approach.	143

FIGURES

Figure 1	Evaluation/appraisal reports to be reviewed by regional distribution of projects evaluated/appraised (n=35)	37
Figure 2	Sector distribution (according to OECD CRS codes) of evaluated/appraised projects (n=35)	38
Figure 3	Distribution of projects by geographical scope (n=35)	39
Figure 4	Distribution per project budget (in allocations from Finland) (n=35)	39
Figure 5	Distribution of project budget allocations from Finland between long-term partner countries and other beneficiaries (n=35)	40
Figure 6	Modality of implementation of projects in the portfolio (excluding appraisals) (n=25)	41
Figure 7	Reports in the portfolio by commissioning agency (n=36)	42
Figure 8	Distribution of reports by evaluation budget (n=36)	42
Figure 9	Evaluation conducted by (n=36)	43
Figure 10	Evaluation reports and appraisals (n=36) according to MFA unit	44
Figure 11	Assessment of the TOR for appraisal per area	56
Figure 12	Gap of appraisal reports against the maximum score per assessment area	59
Figure 13	Relation between ToR issues, findings, and answers	60
Figure 14	Score of evaluation reports in the OECD/DAC criteria	64
Figure 15	Ratings used in analysis of aid effectiveness	72
Figure 16	Scores of HRBA and selected cross-cutting issues	75
Figure 17	Comparison of ratings of the OECD/DAC criteria, aid effectiveness, cross-cutting themes and HRBA, and risk management in non-MFA commissioned reports	83
Figure 18	Distribution of Scores given to Evaluation Reports, (per section used in meta-evaluation assessment grid)	90
Figure 19	Comparison of the scores given in percentage to different sections	99
Figure 20	Diagram of the structure of grids showing the difference between headline standards and characteristics for scoring	130
Figure 21	Overview on the deductive process in Phase Two	136
Figure 22	Process through which induction is used as a complementary research approach in Phase Two	140

ACRONYMS AND ABBREVIATIONS

A&M	Approach and Methodology
ALI-10	Unit for the Middle East and North Africa
ALI-20	Unit for Eastern and Western Africa
ALI-30	Unit for Southern Africa
ASA-10	Unit for Eastern Asia and Oceania
ASA-30	Unit of Latin America and the Caribbean
ASA-40	Unit for South Asia
CCO	Cross-cutting objectives
CD	Capacity Development
CIDA	Canadian International Development Agency (now part of Ministry of Foreign Affairs and International Trade)
DAC	Development Assistance Committee of the OECD
Danida	Danish International Development Agency
DEVCO	EC's Directorate-General for International cooperation and Development (DG DEVCO)
DfID	Department for International Development
EC	European Commission
EIA	Environmental Impact Assessment
EQ	Evaluation Question
EU	European Union
EUR-40	Unit for South-Eastern Europe
EVA-11	The Development Evaluation Unit of the Ministry for Foreign Affairs of Finland
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit
GoF	Government of Finland
HRBA	Human rights based approach
IADB	Inter-American Development Bank
ICT	Information and communications technology
IFI	International Financial Institution
INGO	International non-governmental organisations
IR	Inception Report

ITÄ-20	Unit for Eastern Europe and Central Asia
ITT	Instructions to Tenderers
JC	Judgement criteria
KEO	Department of Development Policy
KEO-60	Unit for International Environmental Policy
LFA	Logical Framework Approach
Manual	MFA Evaluation Manual
MFA	Ministry for Foreign Affairs of Finland
MR	Management review
MTE	Mid-term evaluation
MTR	Mid-term review
NA/ND	Not available/not addressed
NGO	Non-governmental organisations
NSA	Non-state actors
OECD	Organisation for Economic Cooperation and Development
PD	Programme Documents
PIU	Project Implementation Unit
QA	Quality assurance
RBM	Results based management
Team	The team of experts performing the present meta-evaluation
ToC	Theory of Change
ToR	Terms of Reference
TUO-10	Unit for Trade Policy
UN	United Nations
UNFPA	United Nations Population Fund
USAID	United States Agency for International Development
WWF	World Wide Fund for Nature

TIIVISTELMÄ

Metaevaluointi arvioi evaluointi- ja ennakkoarvointiraporttien sekä niiden tehtävänkuvauksien laatua. Arviointi perustui ulkoministeriön eri käsikirjoissa laajasti esiteltyihin laatuvaatimuksiin. Toisessa vaiheessa analysoitiin sitä, mitä raportit paljastivat Suomen kehitysyhteistyöstä käyttäen innovatiivista, yhdistettyä deduktiivis-induktiivista lähestymistapaa.

Evaluointiraporttien ja tehtävänkuvausten laatua ei arvioitu kovinkaan korkeaksi lukuun ottamatta tiettyjä vähemmän ratkaisevia ominaisuuksia. Kummissakin havaittiin riittämättömällä todistusaineistolla perusteltuja johtopäätöksiä, puutteellisia vastauksia evaluointikysymyksiin sekä analyysin puutetta sellaisissa aiheissa kuten avun tuloksellisuus, läpileikkaavat tavoitteet ja Suomen ihmisoikeusperustainen kehityspolitiikka. Tehtävänkuvaukset eivät ohjanneet tehtäviä riittävän tarkasti eivätkä vaatineet evaluointien tekijöiltä raportointia useista oleellisista politiikkakysymyksistä. Ennakkoarvointiraporttien laatu arvioitiin erittäin heikoksi erityisesti siksi, että ne oli toteutettu etukäteisevaluointeina hankesuunnittelun vaiheessa, jossa ehdotettu interventio ei ollut vielä riittävän tarkasti muotoiltu.

Suomen kehitysyhteistyön arviointi perustui pienemmälle otokselle sellaisia evaluointiraportteja, jotka olivat saavuttaneet laadullisen vähimmäistason. Loppupäätelmiä varten induktiivisen analyysin tuottamiin näkemyksiin yhdistettiin tiukasti luokiteltu deduktiivinen analyysi.

Suomen kehitysyhteistyön havaittiin olevan erittäin **relevanttia** sikäli, että se on yhdenmukaista sekä Suomen että edunsaajaorganisaatioiden politiikan ja strategioiden kanssa. Se myös vastaa yleisesti hyvin edunsaajien tarpeisiin. Raportit eivät kuitenkaan tarjoa paljon tietoa siitä, missä määrin aloitteet ovat relevantteja tietyille erityisille Suomen kehitysyhteistyön toimintalinjoille. Suomen kehitysyhteistyö näyttäytyi raporttien perusteella vain kohtalaisen **tuloksellisena** korkeamman tason tavoitteidensa saavuttamisessa. **Tehokkuus** sai raporteissa alhaisia pisteitä, ja monet raportit eivät edes maininneet sitä. **Kestävyyden** arvosana oli melko alhainen, sillä raporteista ei ilmennyt, miksi niissä todettiin kehitysinterventioiden olevan kestäviä tuloksiltaan. Erittäin alhainen **vaikutuksellisuuden** arvosana heijastaa sitä, että vain erittäin harvat raportit pystyivät osoittamaan, mikä vaikutus hankkeilla tulisi olemaan. Alhaiseen arvosanaan vaikuttaa suuresti vaikutuksen arviointiin vaadittavan tiedon keräämiseen käytettävien seuranta- ja tietojärjestelmien kertakaikkisen puuttuminen arvioiduissa hankkeissa.

Suosituksat koskevat sekä strategista toimintatasoa että operationaalisen tason johtopäätelmiä. Raportissa esitetään ehdotuksia ulkoministeriön organisaation ja henkilökunnan valmiuksien ja kykyjen kehittämiseksi.

Avainsanat: metaevaluointi, hajautettu evaluointi, ulkoasiainministeriö, kehitysevaluoinnin käytännöt.

REFERAT

Metautvärderingen analyserade kvaliteten hos utvärderingsrapporter och rapporter upprättade som en del av ex-ante bedömningar och deras motsvarande uppdragsbeskrivningar och anbudsinstruktioner. Analysen baserades på kraven som finns utförligt beskrivna i många av UM:s manualer. I en andra fas analyserades vilka indikationer rapporterna ger om det finska utvecklingssamarbetet med hjälp av en innovativ metod där både induktiva och deduktiva verktyg ingick.

Överlag graderades inte kvaliteten på utvärderingsrapporterna och uppdragsbeskrivningarna speciellt högt förutom för ett antal specifika och icke-kritiska faktorer. Återkommande problem var brist på bevis, otillfredsställande svar på utvärderingsfrågorna, frånvaro av analys för ämnen såsom biståndseffektivitet, tvärspektoriella frågor och mänskliga rättigheter och rättighetsperspektiv i det finska utvecklingssamarbetet. Uppdragsbeskrivningarna var inte tillräckligt specifika i sin inriktning och inkluderade inte instruktioner runt behoven av att rapportera ett antal olika policyfrågor. Förhandsbedömningars rapporter graderades överlag väldigt låg, framför allt därför att de behandlades som ex-ante bedömningar i en kontext då den föreslagna insatsen inte ännu var tillräckligt väldefinierad för att en sådan analys skulle kunna vara genomförbar.

Analysen av det finska utvecklingssamarbetet baserades på ett mindre urval av utvärderingsrapporter som levde upp till en miniminivå av kvalitet. Till induktiv analys bifogades också en rigorös deduktiv analys för att slutföra arbetet.

Analysen kom fram till att det finska utvecklingssamarbetet är **relevant** så till vida att den ligger i linje med både finska och mottagande organisationers policyer och strategier. Den är också välanpassad till behoven hos målgruppen. Däremot innehåller inte rapporterna mycket information om den utsträckning till vilken insatserna är relevanta för specifika finska mål inom utvecklingssamarbetet. Det finska utvecklingssamarbetet har endast begränsat **effektivt** när det gällde att nå de övergripande målen. **Kostnadseffektiviteten** fick låga poäng och många av rapporterna som analyserades gick inte in på det ämnet. Betyget för **hållbarheten** var överlag lågt därför att rapporterna inte tydligt förklarade varför de trodde att insatserna skulle vara hållbara. Det mycket låga betyget för **långsiktigseffekten** reflekterar det faktum att endast ett fåtal av rapporterna kunde ange vilka effekterna skulle bli. En stor del av den låga poängen är ett resultat av frånvaron av informationssystem för att samla in data som krävs för att uppskatta effekter.

Rekommendationerna relaterar till de slutsatser som dras för den strategiska såväl som den operativa nivån och förslag ges för hur kapaciteten hos UM och dess personal kan stärkas.

Nyckelord: metastudie, decentraliserad utvärdering, Finska Utrikesministeriet, ledning av utvärderingsarbete.

ABSTRACT

The meta-evaluation assessed the quality of evaluation and appraisal reports and their corresponding ToR and ITT. Assessments were based on the requirements expanded upon in various MFA manuals. A second phase analysed what the reports revealed about Finnish Development cooperation, using an innovative mixed deductive-inductive approach.

The quality of evaluation reports and ToR was not rated very highly except for certain, and non-critical, characteristics. Both presented lack of evidence, insufficient answers to the evaluations questions, absence of analysis on topics including aid effectiveness, cross-cutting issues and Finland's Human Rights Based Approach policy. ToR were not sufficiently specific in their direction and did not include instructions on the need to report on a number of policy issues. Appraisal reports were rated very low, particularly because they were treated as ex-ante evaluations in a context where the proposed intervention was not sufficiently defined to undertake that kind of analysis.

The analysis of Finnish development cooperation was based on a smaller sample of evaluation reports that had met a minimum level of quality. Appraisals and inductive analysis insights were added to a rigorous and rated deductive analysis in order to conclude.

Finnish development cooperation was found to be very **relevant** in that it is aligned to both Finnish and beneficiary organisation policies and strategies. It also responds generally to the needs of the targeted beneficiaries. Reports do not provide much information on the extent to which initiatives are relevant to specific Finnish development cooperation policies, however. It was only moderately **effective** in meeting its higher-level objectives. **Efficiency** was awarded a low score, and many of the reports did not report on it. The **sustainability** rating was rather low because the reports did not show why they believed that interventions would be sustainable. The very low rating given to **impact** reflects the fact that very few reports were able to indicate what the impact would be. An across-the-board absence of information systems to gather required data on impact accounts for a large part of the low score.

Recommendations deal with strategic level and operational level conclusions, as well as suggested avenues for the further development of capabilities of MFA and the abilities of MFA staff.

Key Words: meta-evaluation, decentralised evaluation, Ministry of Foreign Affairs Finland, evaluation management practices.

YHTEENVETO

Tausta, tarkoitus ja tavoite

Suomen ulkoasianministeriön (UM) kehitysevaluoinnin yksikkö (EVA-11) on tilannut tämän metaevaluoinnin UM:n toimeenpanoyksiköiden suorittamista hankkeiden ja ohjelmien evaluoinneista (mukaan lukien etukäteisarvioinnit). Tämä asiakirja on kyseisen metaevaluoinnin loppuraportti.

Tarkoitus

Tämän metaevaluoinnin tehtävänmäärittely (ToR) (katso Liite 2) tunnistaa metaevaluoinnin tarkoituksiksi seuraavat:

ENSIKSI: “alkuvaiheessa UM:n auttaminen evaluoinnin, sen hallintokäytäntöjen ja koko evaluointikapasiteetin kehityksen edistämiseksi. Se tarjoaa myös kokonaiskuvan nykyisestä arviointiportfoliosta, mikä auttaa UM:tä tunnistamaan mahdolliset puutteet”.

TOISEKSI: “seuraavassa vaiheessa sellaisten asioiden ja opetusten esiin tuominen jotka ovat nousseet esiin arviointiraporteista, ja sellaisten suositusten tekeminen jotka auttavat UM:tä parantamaan suomalaista kehitysyhteistyötä. Tähän pyritään arvioimalla ensimmäisen vaiheen arviointiraporteissa tunnistetut suomalaisen kehitysyhteistyöhön liittyvät vahvuudet ja haasteet”.

Tavoitteet

Tämän meta-arvioinnin tavoitteet ovat myös kahdenlaiset, kuten metaevaluoinnin tehtävänmäärittelyssä on mainittu:

ENSIKSI: “metaevaluointi arvioi eri hajautettujen evaluointiraporttien ja niihin liittyvien suunnitteluasiakirjojen laatua. Se myös laatii kokonaiskuvan arviointiportfoliosta vuosina 2014-2015, ja arvioi evaluoinnin kattavuutta vuosina 2013-2015”.

TOISEKSI: “metaevaluointi kokoaa yhteen luotettavat arviointien tulokset ja arviointiraporteista esiin nousevat ongelmat koskien Suomen kehitysyhteistyötä”.

Meta-arvioiden vertailu

Metaevaluointeja on tehty aiemmin vuosina 2007, 2009, 2012 ja 2014. Näissä aiemmissa arvioinneissa käytetyt työkalut ja menetelmät ovat kehittyneet huomattavasti ajan saatossa ja UM:n vakavana tarkoituksena on vakiinnuttaa vertailumenetelmät, jotta ajassa tapahtuvien muutosten vertailu tulisi mahdolliseksi. *Tämä metaevaluointi eroaa* muista merkittävillä tavoilla, joista yksi on se, että metaevaluoinnin ensimmäisessä vaiheessa (eli tehtävänmäärittelyjen ja evaluointi- ja etukäteisarviointiraporttien laatuanalyysissä) käytetyt analyysitaulukot perustuvat nyt täysin erilaiselle logiikalle kuin aikaisemmin. Tämän lisäksi myös metaevaluoinnin toisen vaiheen (eli Suomen kehitysyhteistyön arviointi evaluointiraporttien antaman kuvan pohjalta) analyysin viitekehys

perustuu täysin erilaiselle loogiselle perustalle kuin aikaisemmissa metaevaluoinneissa. UM/EVA-11 pyysi näitä muutoksia keskusteltuaan meta-arviointitiimin kanssa. Näin ollen on oltava varovainen pitkäaikaisia trendejä tai muutoksia tunnistettaessa. Yksi tämän raportin kappaleista vertailee edellisen (2014) ja nykyisen metaevaluoinnin strategisen tason johtopäätöksiä. Johtopäätökset ovat pääasiassa samankaltaisia ja samat ongelmat nousevat usein esiin molemmissa.

Aineiston kuvaus

Metaevaluoinnissa tutkittiin kolmekymmentäkuusi eri evaluointi- (n=26) ja etukäteisarviointi- (n=10) raporttia. Kuusikymmentäkolme prosenttia näistä oli Aasiasta tai Afrikasta, ja 17 % oli globaaleja projekteja. Vain 12 evaluointiraporttia (26:sta) koski Suomen kehitysyhteistyön virallisia ja perinteisiä bilateraalaisia kumppanimaita. Edustettuina oli 13 sektoria, joista neljä muodosti 50 % kokonaismäärästä. Ympäristö ja kolme muuta luonnonvaroihin liittyvää toimialaa edustivat 40 % kokonaismäärästä, eli vuoden 2007 kehitysyhteistyöpoliittisen toimintaohjelman mukanaan tuoma suunnanmuutos näkyy vasta nyt evaluointien otoksessa. Neljäkymmentäkuusi prosenttia raporteista oli maakohtaisia, kun taas 48 % oli joko alueellisia/usean maan alueella tai maailmanlaajuisia. On tärkeää huomata, että 51 %:lla suomalaisista projekteista oli budjetti (Suomen osuus), jonka koko oli alle 5 miljoonaa euroa. Vain 29 %:lla ohjelmista oli yli 10 miljoonan budjetti. Portfolio koostui siis suomalaisen rahoituksen suhteen pääasiassa suuresta määrästä suhteellisen pieniä projekteja ja harvasta suuresta projektista. Vain 56 % suomalaisesta kokonaisrahoituksesta meni bilateraalille kumppanimaille. Tämä antaa ymmärtää, että suomalainen apu on pirstoutunut, mikä tarkoittaa lisääntyneitä hallintokustannuksia sekä toimintojen päällekkäisyyttä ja potentiaalisen vaikutuksen ohentumista.

Metodologia ja riskit

Metaevaluointi vertasi ensivaiheessa eri raporttien sisällön laatua suhteessa useissa UM:n toiminta- ja ohjausasiakirjoissa esiteltyihin vaatimuksiin. Toisessa vaiheessa käytettiin toista arviointikehikkoa tunnistamaan raporttien tuottamat oivallukset liittyen suomalaisen kehitysyhteistyön toimeenpanoon.

Tätä metaevaluointia varten kehitettiin kompleksinen metodologia ja tutkimusprotokolla. Siihen sisältyi innovatiivinen, deduktiivista ja induktiivista lähestymistapaa yhdistävä suomalaisen kehitysyhteistyön analyysi, joka on erittäin epätavallinen tutkimusstrategia. Se esiteltiin työn aloitusvaiheen raportissa (*inception report*) ja perustellaan ja kuvaillaan tämän raportin liitteessä. Kaikki tutkimustiimin jäsenten tekemät arvioinnit tarkistettiin muiden jäsenten toimesta ja tiimi pyrki kaikin keinoin varmistamaan, että tiimin jäsenet arvioivat raportteja ja tehtävänkuvauksia samalla tavalla. Tämä raportin liitteessä kuvataan yksityiskohtaisesti metaevaluoinnin prosessi ja metodi.

Tiettyjä riskejä tunnistettiin jo varhain. Yksi niistä oli se periaatepäätös, ettei tiimi olettaisi mitään asiaa raportoiduksi, ellei sitä nimenomaan mainittaisi analysoidussa raportissa. Jos esimerkiksi avun tuloksellisuudesta ei ollut mitään nimenomaista kuvausta, sellaista ei myöskään oletettu muun tekstin

perusteella. Toinen riski liittyy aineiston edustavuuteen, mutta tiimi uskoo tulosten olevan sekä päteviä että toistettavissa olevia.

Havaintojen, johtopäätösten ja suositusten yhteenveto

Seuraavilla sivuilla esitetään metaevaluoinnissa esiin nousseet oleellimmat havainnot, johtopäätelmät ja suositukset järjestyksessä siten, että ensin tulevat havainnot, sen jälkeen johtopäätökset ja lopulta tärkeimmät suositukset. Aiheesta on tuotettu myös paljon yksityiskohtaisempi, näyttöön perustuvia päätelmiä ja suosituksia sisältävä versio taulukkomuodossa, joka on raportin liitteenä. Taulukko voi toimia johdon vastineen pohjana, mutta se on liian pitkä (yli 20 sivua) sisällytettäväksi lyhennelmään.

Tämän mandaatin tehtävänmäärittely sisälsi seitsemän evaluointikysymystä (EK), joiden mukaan havainnot esitellään.

EK 1: Mikä on UM:n hajautetun evaluointiportfolion (evaluointiraportit ja niiden tehtävänkuvaukset) laatu OECD:n kehitysapukomitean (DAC) evaluointikriteerien perusteella vuosina 2014-2015 sekä suhteessa UM:n evaluointioppaassa annettuun ohjaukseen luokiteltuna maiden, sektorien, budjettien, evaluointityyppien, UM:n hallinnollisten yksiköiden, toimeksiantajan, konsulttiyritysten jne. mukaan? Onko UM:n ja toisaalta UM:n kumppanien tilaamien evaluointien laadun välillä eroja?

Evaluointien tehtävänkuvausten yleisarvosana oli 64,3/100. Metaevaluoinnissa havaittiin heikkouksia tehtävänkuvauksissa useilla avainalueilla, kuten sellaisissa ydinkohdissa joissa tarvittaisiin evaluoinnin tarkkaa ohjausta: evaluointikysymysten muotoilussa, ohjeissa avun tuloksellisuuteen sitoutumisen arvioinnissa, evaluointimetodia/-metodeja koskevassa ohjeistuksessa ja arvioitavan hankkeen kontekstin kuvauksessa. Tehtävänkuvaukset saivat hyvät pisteet tilatun evaluoinnin (ml. etukäteisarvioinnit) taustan, tarkoituksen ja tavoitteiden kuvauksesta sekä evaluointiin käytettävissä olevien resurssien ja evaluointiprosessin kuvauksesta. Voidaan todeta, että UM:n toimeenpanevien yksiköiden ja osastojen kirjoittamien tehtävänkuvausten laatu on suunnilleen kansainvälistä tasoa siinä määrin kuin voidaan olettaa metaevaluoinnin otoksen olevan edustava. Voidaan kuitenkin myös todeta, että tehtävänkuvausten laatu on paljon heikompi kuin sen pitäisi olla, kun otetaan huomioon UM:n virkamiesten velvollisuus varmistaa lopputuotteen laatu ja tehtävänkuvausten merkittävä rooli evaluointipalvelujen hankintaprosessissa.

Evaluointiraporttien yleisarvosana on 64,4, eli sama kuin tehtävänkuvauksilla. Näin ollen keskiarvojen tasolla tämän otoksen raportit eivät ole täysin tyydyttävää laatua. UM:n alueellisten ja temaattisten yksiköiden tilaamat raportit saavat keskimäärin 56,95 pistettä, kun taas muiden avunantajien tilaamat raportit saavat 69,09 pistettä, eli eroa on lähes kolmesta pistestä.

EK 2: Mikä on UM:n evaluointien kattavuus (suunniteltujen ja toteutuneiden evaluointien vertailu)?

Metaevaluoinnissa ei löytynyt riittävästi luotettavaa tietoa tähän kysymykseen vastaamiseksi. UM oli samaa mieltä siitä, että informaatiota ei ollut saatavilla sellaisessa muodossa, joka olisi tehnyt mahdolliseksi kattavuusanalyysin tekemisen.

EK 3: Mikä on etukäteisarviointiraporttien ja niiden vastaavien tehtävänkuvausten laatu?

Metaevaluoinnissa havaittiin, että **etukäteisarviointien tehtävänkuvaukset** eivät olleet tarkkoja ohjeistuksessaan. Havaittiin myös, että kuvaukset olivat heikkoja muotoilemaan arviointikysymykset selkeästi, ja tavallisesti ne esitivät pitkän listan tutkittavia asioita ilman, että näitä olisi kytketty ennalta määritettyihin analyysikriteereihin (kuten UM:n ohjeistuksessa edellytetään). Kun otetaan huomioon kaikki arviointitaulukon osat, joilla etukäteisarviointien tehtävänkuvauksia (ja/tai hankinta-asiakirjoissa esitettyjä ohjeita) analysoitiin, saatiin keskimääräiseksi pistemääräksi 64,78/100. Tätä keskimääräistä arvosanaa on pidettävä hyvin alhaisena, kun otetaan huomioon, että a) nämä asiakirjat on tuotettu UM:n sisällä virkatyönä ja että asiakirjat itsessään ovat UM:n laadunvalvonnan alaisia, ja että b) jotkin näistä asiakirjoista ovat saattaneet olla myös kehitysyhteistyökumppanien, vastaanottajaorganisaatioiden tai toimeenpanevien tahoja laadunvarmistuksen kohteena. Alhaisia pisteitä saivat erityisesti arviointikysymysten tarkka muotoilu, ohjeet avun tuloksellisuuden analysoimiseksi ja arviointien lähestymistavan määrittely. Tehtävänkuvaukset saivat sen sijaan hyviä pisteitä etukäteisarviointien taustan, tarkoituksen ja tavoitteiden sekä arviointiprosessin kuvauksesta.

Etukäteisarviointiraportit olivat yleisesti huonolaatuisia (keskiarvo 46,5 pistettä), mikä saattaa tiettyssä määrin johtua niistä ohjanneiden tehtävänkuvausten puutteista. Erityisesti ne saivat alhaisia pisteitä, koska niissä ei esitetty vankkaa näyttöä havaintojen tueksi sekä koska tehtävänkuvauksissa esitettyihin kysymyksiin vastaamisessa oli puutteita (molemmat ovat minkä tahansa evaluoinnin/arvioinnin ydinelementtejä).

EK 4: Mitä voidaan sanoa suomalaisen kehitysyhteistyön laadusta OECD:n kehitysapukomitean (DAC) evaluointikriteerien mukaan luotettaviksi arvioitujen hajautettujen evaluointiraporttien sekä niihin liittyvien suunnitteluasiakirjojen perusteella?

Suomalainen kehitysyhteistyö havaittiin **tarkoituksenmukaiseksi** sikäli, että se on yhdensuuntainen sekä Suomen että edunsaajaorganisaatioiden politiikkalinjausten ja strategioiden kanssa. Se saavutti ylemmän tason tavoitteensa vain kohtalaisen **tuloksellisesti**. (Osa alhaisesta pistemäärästä johtuu selvästi siitä, että monissa hankkeissa ei ole tavoitteiden saavuttamisen seurantarjestelmiä, joten tuloksellisuudesta ei voitu raportoida). **Tehokkuus** sai alhaisen pistemäärän, pääasiassa koska raportit eivät maininneet siitä mitään. Joissain raporteissa kyllä arvioitiin budjettia ja kulurakennetta, mutta ei tehokkuutta sinänsä; sen sijaan useissa raporteissa todettiin, että byrokratia ja monimutkaiset hankintamenettelyt hidastivat hankkeiden toimeenpanoa huomattavasti. **Kestävyys** sai melko alhaiset pisteet, sillä raporteissa ei perusteltu, miksi hankkeiden väitettiin olevan kestäviä, tai niissä a) puhuttiin “mahdollisesta kestävydestä” tai b) oletettiin hankkeiden vaikutusten/tulosten olevan kestäviä, vaikka samalla todettiin tärkeiden hankekomponenttien jäävän tavoitteistaan. **Vaikuttavuudelle** annettu erittäin alhainen pistemäärä heijastaa sitä, että hyvin harvassa raportissa pystyttiin osoittamaan, mikä vaikutus tulisi olemaan. Vaikuttavuudesta dataa keräävien seurantarjestelmien puute sekä

selkeästi ylioptimistiset väitteet vaikuttavuudesta selittävät suuren osan alhaisesta pistemäärästä.

EK 5: Mitkä syyt selittävät sen, että evaluoinnin sijaan tilataan hallinnollinen katsaus (management review) (mikäli mahdollista selvittää)?

Yksikään analysoiduista raporteista ei tuonut valaistusta tähän asiaan.

EK 6: Mitkä ovat merkittävimmät hajautetuista evaluointiraporteista esiin nousevat asiat? Millaisia menestystarinoita, hyviä käytäntöjä ja haasteita ne tuovat esiin?

Seuraavassa on pieni osa metaevaluoinnin toisessa vaiheessa yksittäisistä raporteista poimituista avainkohdista:

Tarkoituksenmukaisuus:

- Suomalaiset projektit ovat yleensä erittäin hyviä vastaamaan kohderyhmien tarpeisiin.
- Yleisesti raporteissa puhutaan yhdenmukaisuudesta ja -sopivuudesta suomalaisen kehityspolitiikan kanssa, mutta vain erittäin abstraktilla tasolla, jolloin tietoa on vaikea käyttää kehityspolitiikan jatkokehittämisen apuvälineenä.

Tuloksellisuus:

- Induktiivinen analyysi paljasti, että monissa hankkeissa koettiin kohdittuun voimakasta turhautumista niiden kohtaamiin haasteisiin, mutta raporteissa kerrottiin myös useista innovatiivisista keinoista, joita hankkeissa suunniteltiin ja käytettiin kontekstiin liittyvien ja teknisten ongelmien ratkaisemiseksi. Hankkeiden tuotteiden (output) muuttaminen tuloksiksi (outcome) onkin nimenomaan se monitahoisia haasteita sisältävä prosessi, ja näistä useimmat jäävät ilmeisesti näkemättä suunnitteluvaiheessa.
- Suomalaiset hankkeet yleensä eivät saavuta ylemmän hierarkiatason tavoitteitaan; sen sijaan ne yleensä ovat tuloksellisia niin että ne saavuttavat useimmat odotetuista alemman hierarkiatason tuotteistaan eli niistä, jotka seuraavat välittömästi output-tason tuloksista. Jostakin syystä nämä välittömät tulokset epäonnistuvat muuttamaan saavutukset ylemmän tason tavoitteiksi. Vakavia haasteita hankkeissa myös on; näiden joukossa on lähes poikkeuksetta ylimitoitettuja tavoitteita, hankintaprosessien pituuden epäsuhta suunniteltuihin aikatauluihin nähden (varsinkin kansainvälisten järjestöjen tapauksessa), ja haparoi-va hankkeen johtaminen, joka usein aiheutui huonosti määritellyistä tuloksista, ja lista luettelee vain tärkeimmät.
- Monissa teknistä (henkilö-)apua sisältäneissä hankkeissa havaittiin, että kansainväliset hankekonsultit ja heidän paikalliset vastapuolensa tuottivat lakiluonnoksia ja säädös- ja muita ehdotuksia, mutta näitä ei koskaan viety eteenpäin lopullisesti hyväksyttäviksi. Syynä tähän voidaan olettaa olevan se, että henkilöapu (tekninen apu, TA) joko ei tehnyt aloitteita aihepiireistä jotka hyödynsaajaorganisaatio olisi pitänyt relevantteina, teknistä apua ei käytetty tehokkaasti hyödyksi, tai hankkeen tuottamia ehdotuksia ei pidetty haluttuina tai sopivina.

Tehokkuus:

- Interventiot eivät olleet tehokkaita ajankäytön suhteen, ja avainongelmiksi mainittiin pitkät viiveet hankinnassa ja päätöksentekoprosesseissa.
- Suomalainen kehitysyhteistyö sai erityisen kiitosmaininnan joustavuudesta. Kansalliset hallitukset ja kansainväliset järjestöt sen sijaan nähtiin erittäin jäykkinä ja byrokraattisina.
- Hankkeet eivät yleensä sinänsä pyri kustannustehokkuuteen. Niitä kiinnostaa enemmän “tehdä se mitä suunniteltiin sillä tavalla kuin suunniteltiin”, sekä hallinnoida budjettia ja kuluja hyväksytyn maksatussuunnitelman mukaisesti.

Kestävyys:

- Suomalaisissa hankkeissa käytetään yleensä edunsaajien tarpeisiin ja valmiuksiin sovitettuja teknisiä ratkaisuja. Edunsaajat ottavat ne helposti ”omikseen”.
- Taloudellinen kestävyys on harvoin taattu, edes hankkeen päätös-vaiheessa.
- Vaikka valmiuksien kehitys olisi osa interventiota, lopputuloksen saavuttamisen edellyttämä hyödynsaajaorganisaation valmiuksien kestävyys on kovin alhainen.

Vaikuttavuus:

- Metaevaluoinnin perusteella havaitaan, että suomalaisessa kehitysyhteistyössä ei ole selvää käsitystä siitä, missä määrin sen puitteissa toteutetut hankkeet saavat aikaan odotettuja vaikutuksia. Tämän arvioimiseksi vaadittavaa seurantainformaatiota ei kerätä systemaattisesti, ja odotetut vaikutukset tai ylimmät kehitystavoitteet kirjataan niin yleisluontoisin käsittein, että niitä on vaikea evaluoida.

Kehitysyhteistyön tuloksellisuus (Aid Effectiveness):

- Metaevaluointitiimi havaitsi, että useimmat raportit eivät varsinaisesti käsittele avun tuloksellisuutta erillisenä käsitteenä.
- Suomalaisen kehitysyhteistyön yhteensopivuus vastaanottajatahon politiikkalinjausten kanssa on erityisen vahva, varsinkin yleisen tason kehitystavoitteiden kohdalla. Sen sijaan evaluointiraporteissa ei yleensä koskaan analysoida suomalaisen hankkeen yhteensopivuutta ja johdonmukaisuutta alemman tason strategioiden tai tarkkojen kansallisten suunnitelmien kanssa.
- Avunantajien toimintojen keskinäisestä yhteensovittamisesta (harmonisaatiosta) maatasolla evaluointiraportit kertovat varsin harvoin, joskin useissa niissä listataan lyhyesti muita sellaisia rahoittajatahoja, joiden kanssa hanke on tekemisissä.

Ihmisoikeusperustaisuus ja läpileikkaavat tavoitteet

- Tämän analyysitaulukon osalle saatu pistemäärä osoittaa selkeästi, että ihmisoikeusperustaisuus ei toteudu Suomen kehitysyhteistyössä tämän metaevaluoinnin perusteella, tai ainakaan siitä ei raportoida riittävästi.

- Metaevaluointitiimi havaitsi, että vaikka termi ihmisoikeusperustaisuus mainittiin lähes kaikissa vuoden 2012 kehityspoliittisen ohjelman puitteissa kirjoitetuissa raporteissa, ne eivät koskaan arvioineet, miten hankkeet kokonaisuudessaan toteuttivat ihmisoikeusperustaisuutta.
- Sukupuolten välisen tasa-arvon edistämistä hankkeissa käsitellään asi-ana, jota hanke joko edistää tai sitten ei. Suuri osa raporteista mainitsi, että joissain toimissa naiset olivat osallisina toiminnan kohteina, kuten koulutuskurssien ja seminaarien osanottajina, mutta niissä raportoitii myös, että naiset eivät olleet osa päätöksentekoa tai oli tehty tietoinen päätös, eivät naiset olleet toiminnan kohde- tai hyödynsaajaryhmä. Vain kourallisessa hankkeita oli minkäänlainen sukupuolten tasa-arvoon liit-tyvä seurantajärjestelmä.
- Evaluointiraportit eivät käsittele epätasa-arvoa ja sen vähentämistä suo-raan omana alueenaan. Itse asiassa koko termiä käytetään harvoin.
- Monet raportit mainitsivat ilmaston/ilmastollisen kestävyvyyden, mutta lähes poikkeuksetta vain pinnallisina viitteinä.

EK 7: Mitä voidaan oppia etukäteisarviointiraporteista ja niiden tehtävän-kuvauksista koskien suomalaisten kehitysyhteistyöhankkeiden alkusuunnitel-mien laatua?

Yleisesti ottaen ohjelma-asiakirjojen luonnokset eivät ole valmiita arviointiin, sillä niistä puuttuu usein monia avainkohtia, kuten kehityshankkeen tavoite-puu, muutosteoria, tulosten viitekehys, yksityiskohtainen toimeenpanostrategia, keskipitkän aikavälin tulosten ja vaikutusten erittely sekä analyysi tieto-kantojen ja lähtökohtatilannetta koskevan tiedon saatavuudesta.

Pienessä osassa etukäteisarvioita huomautettiin siitä, kuinka vähän hanke-suunnittelua oli arvioinnin toteuttamiseen mennessä tehty, ja koska etukätei-sarvointien tekijöiltä ei edellytetty ohjelma-asiakirjojen luonnoksen muokkaamista, niihin liittyvät suositukset olivat hyvin yleisluontoisia.

Mielenkiintoista kyllä joissain evaluointiraporteissa tunnistettiin, että “heidän” hankkeensa kohtaamat ongelmat olivat seurausta huonosta suunnittelusta.

Suosituks

Metaevaluoinnin tärkeimmät suositukset ovat seuraavat:

A) Strateginen taso

1. Ulkoasiainministeriön tulisi rakentaa mekanismeja koskien kahdenvä-lisen kehitysyhteistyön hallintoa, mukaan lukien seuranta ja laadun var-mennus, jotta se voisi paremmin toimeenpanna omia kehityspoliittisia linjauksiaan.
2. Tulisi suorittaa selvitys kehitysyhteistyön hallinnosta (eli analyyttisellä tavalla toteutettu yksityiskohtainen selvitys pohjautuen hallinnon eri osien vastuualueisiin), jolla kartoitettaisiin vuoden 2016 ja sen jälkeises-sä kontekstissa se hyöty, jonka UM:n toimeenpaneavat tahot kokevat mah-dolliseksi ja tarpeelliseksi saada kehitysyhteistyön evaluointitoimesta.

3. Etukäteisarviointiin liittyvien asiakirjojen huonon yleisarvosanan perusteella suositellaan, että ulkoministeriön tulisi muuttaa etukäteisarviointien roolia siten, että ne tehtäisiin huomattavasti myöhemmässä vaiheessa hankesykliä. Ohjelma-asiakirjojen luonnosten tulisi olla lähes valmiita ja täyttää sisällön ja suunnittelun vähimmäisvaatimukset ennen kuin ne altistetaan sellaiselle kritiikille, jota etukäteisarvioinneilta voidaan edellyttää.
4. Perustuen kappaleen 7.3 johtopäätöksiin (eli Mitä evaluointiraportit paljastavat Suomen kehitysyhteistyöstä), UM:n toimeenpanevien osastojen ja yksiköiden tulisi kriittisesti *pyrkä ymmärtämään syyt niille heikkouksille*, joita löydettiin niiden hallinnoimien kehityshankkeiden relevanssissa, tehokkuudessa, tuloksellisuudessa, vaikuttavuudessa ja kestävydessä. Osana tätä suositusta UM:n tulisi sisällyttää tehtäväkuvauksiinsa viittaus arvioijien velvollisuuteen kytkeä raporteissaan hankkeet Suomen kehitysyhteistyöpolitiikan tavoitteisiin.
5. Perustuen erittäin epätasaiseen ihmisoikeusperustaisuuden toteutumiseen Suomen kehityshankkeissa UM:n tulisi suorittaa sisäinen arviointi (kenties hallinnontarkastuksen muodossa) kyseiseen ihmisoikeusperustaisuuteen liittyvistä käytännöistä ja sille asetetuista tavoitteista.

B) Toiminnallinen taso

6. Evaluointisuunnitelmien metodologiavaatimuksia on kiristettävä huomattavasti (toimeksiantajalle tulisi aina esittää yksityiskohtainen metodologia, johon sisältyy tietolähteet, indikaattorit, työkalut ja menetelmät tiedon keräämiseen ja analysointiin, otantamenetelmät, ja suunnitellut haastattelulomakkeet.
7. Näyttöä on vaadittava kaikkien löydösten ja havaintojen tueksi.
8. Evaluointien ja etukäteisarviointien odotukset on määritettävä paremmin suhteessa kolmeen kriteeriin, eli politiikkajohdonmukaisuuteen, suomalaiseen lisäarvoon ja avun tuloksellisuuteen.
9. On kehitettävä erillinen ohjeistusasiakirja, joka käsittelee erityisesti raporttien hyväksyttävää sisältöä ja tarjoaa niille normit ja standardit.
10. Tätä metaevaluointia varten valmisteltuja arviointitaulukkoja voi muokata hieman ja vaatia, että virkamiehet käyttävät niitä vastaanottamiensa tuotteiden (raporttien) laadun arvottamiseen. Tehtäväkuvausten arviointitaulukkoja virkamiehet voivat käyttää tarkistamaan niiden rakenteen, sisällön ja laadun.

C) Valmiuksien kehittämistä koskevat suositukset

11. UM:n virkamiehien tulisi pystyä arvioimaan sellaisten asiakirjojen laatua, jotka integroivat ihmisoikeusperustaisuuden ja läpileikkaavat tavoitteet arviointikriteereihin (OECD/DAC ja tietyt omat).
12. UM:n virkailijoiden kykyä ymmärtää ja kritisoida evaluointien ja ennakkoarviointien löydöksiä ja johtopäätöksiä sekä seuranta- ja muita raportteja tulisi parantaa merkittävästi suhteessa evaluoitavien hankkeiden tavoitteiden logiikkaan (esimerkiksi loogisen viitekehyksen tai muutosteorian kautta).

SAMMANFATTNING

Bakgrund, Syfte och Målsättning

Enheten för utvärdering av utvecklingssamarbetet (EVA-11) vid det finska utrikesministeriet (UM) har beställt denna metautvärdering av de utvärderingar för olika projekt och program (inklusive ex-ante bedömningar) som utförs av olika enheter på UM. Detta dokument utgör slutrapporten för denna metautvärdering.

Syfte

Uppdragsbeskrivning för denna metautvärdering (se bilaga 2) definierar dess huvudsyfte som följande:

FÖR DET FÖRSTA “att i ett initialskede hjälpa UM med att förbättra sina utvärderingars kvalitet, dess ledning samt den övergripande utvecklingen av kapaciteten för utvärderingar. Studien kommer också att ge en övergripande bild av den nuvarande kapaciteten för att hjälpa UM att identifiera eventuella luckor.”

FÖR DET ANDRA: “i den efterföljande fasen lyfta fram frågor och lärdomar som framkommit i utvärderingsrapporterna och ge rekommendationer avsedda att hjälpa UM att förbättra det finska utvecklingssamarbetet. Detta kommer att ske genom att rapporten analyserar de typer av styrkor och utmaningar som idag förknippas med det finska utvecklingssamarbetet och som har identifierats under analysens första fas.”

Målsättning

Målsättningen med denna metautvärdering som fastställs i uppdragsbeskrivningen är:

FÖR DE FÖRSTA: “metautvärdering kommer att utvärdera kvaliteten hos de olika decentraliserade utvärderingsrapporterna och relaterade planeringsdokument. Den kommer också att teckna en övergripande bild över hela utvärderingsportföljen under 2014-2015 och bedöma utvärderingstäckningen under 2013-2015.”

FÖR DET ANDRA: “metautvärdering kommer att summera tillförlitliga resultat som framkommit i utvärderingarna som gjorts av det finska utvecklingssamarbetet.”

Jämförelser mellan metastudier

Metastudier har tidigare genomförts 2007, 2009, 2012 och 2014. De verktyg och den metodologi som tidigare använts har utvecklats signifikant över tid, och UM är mån om att stabilisera dem för att i ett senare skede möjliggöra jämförelser över tid. *Denna metastudie skiljer sig* från de tidigare på ett betydande sätt, inte minst genom det faktum att bedömningstabellerna som används för den första fasen av studien (d.v.s. kvalitetsanalys av uppdragsbeskrivningen och rapporter av utvärderingar och förhandsbedömningar) nu helt och hållet

är baserade på logisk grund. Dessutom är analysramen för metastudiens andra fas (d.v.s. analys av det finska utvecklingssamarbetet på basis av utvärderingsrapporterna) **också** baserade på en helt annan logisk grund än vad som tidigare har varit fallet. Båda dessa förändringar har tillkommit på initiativ av UM och är ett resultat av dialogen med metautvärderingsteamet. Därmed måste försiktighet iakttas vid identifiering av eventuella långsiktiga trender och förändringar. Ett avsnitt av rapporten är tillägnad jämförelser av slutsatser på en mer strategisk nivå som dras i denna rapport med de som framkommit i tidigare studier. I stort sett är slutsatserna som dras den samma, och samma faktorer lyfts ofta fram.

Beskrivning av urval

Trettiosex olika utvärderingsrapporter (n=26) och förhandsbedömningar (n=10) studerades. Sextiotre procent av dessa hade anknytning till Asien eller Afrika, och ytterligare 17 % var globala projekt. Endast 12 (av 26) av projekten ägde rum inom ramen för det officiella och traditionella bilaterala samarbetet runt finskt utvecklingsarbete. Totalt fanns 13 olika sektorer representerade, varav 4 tillsammans utgjorde 50 % av helheten. Miljö och tre andra naturresursrelaterade sektorer utgjorde 40 % av helheten, vilket innebär att den omprövning för inriktningen av finskt biståndarbete som har skett sedan 2007 nu har fått genomslag i urvalet av rapporter. Fyrtiosex procent av rapporter var landspecifika, medan 48 % antingen var "regionala/multinationella" eller "världsvida". Vad som också är viktigt att notera är att 51 % av de finska projekten hade budgetar (d.v.s. finsk del av budgeten) omfattandes mindre än 5 miljoner euro (MEUR). Endast 29 % av projekten hade budgetar som överskred 10 miljoner. Portföljen utgjordes därför av en relativt omfattande mängd mindre projekt vad avser finansiering från Finland, samt ett mindre antal mycket stora projekt. Endast 56 % av den finska finansieringen gick till bilaterala länder. Detta innebär att finskt utvecklingsbistånd är fragmentiserad med allt vad det innebär av extra administration och styrning.

Metod och risker

På det övergripande planet analyserade metautvärderingen kvaliteten i innehållen i de olika rapporterna på basis av de instruktioner och krav som finns definierade i olika vägledande dokument som tagits fram av UM. I en andra fas utfördes en ytterligare analys syftande till att utröna vilka typer av slutsatser som rapporterna drar som har anknytning till hur finskt utvecklingssamarbete bedrivs.

En komplex metodik och ett analytiskt ramverk användes för denna metautvärdering. Den inkluderade en innovativ integrering av både deduktiva och induktiva tillvägagångssätt i analysen av finskt utvecklingssamarbete, en ovanlig typ av forskningsstrategi som finns närmare beskriven och motiverad i en bilaga till den här rapporten. All analys som utfördes av en teammedlem dubbelkontrollerades av de andra teammedlemmar, och betydande ansträngningar lades ner på att tillse att alla teammedlemmar var till fullo anförtrögn med hur olika typer av faktorer skulle graderas på samma sätt. En mycket utförlig och detaljerad bilaga beskriver detta tillvägagångssätt i detalj.

Ett antal olika risker identifierades i ett tidigt skede, inklusive konsekvensen av ett ställningstagande av teamet att inte förutsätta att något hade rapporterats om det inte fanns explicit antecknat. Om det till exempel inte fanns någon tydlig beskrivning av ansträngningarna som gjorts inom biståndseffektivitet fick man inte själv bedöma detta på basis av annan skriftlig information. En annan risk är förknippat med urvalets representativitet, men teamet är övertygad om att resultaten är valida och reproducerbara.

Summering av resultat, slutsatser och rekommendationer

I bilagan till rapporten finns det en tabell med resultat, slutsatser och rekommendationer, men den är för lång (över 20 sidor) för en sammanfattning. Den presenterar ungefär de samma saker som presenteras här men med mer empiriskt prov. Uppdragsbeskrivningen för denna studie omfattade sju olika forskningsfrågor (FF) som organiserar presentationen.

FF 1: Vilken kvalitet höll UM:s decentraliserade utvärderingsportfölj (utvärderingsrapporter och deras motsvarande uppdragsbeskrivningar) på basis av OECD/DAC:s standarder för utvärderingar 2014-2015 och den vägledning som ges i Utvärderingsmanualen, klassificerad enligt länder, sektorer, budgetar, utvärderingstyper, ansvariga enheter vid UM, utförare, konsultbyråer etcetera? Finns det en skillnad i kvalitet mellan de utvärderingar som utförs direkt av UM och de som genomförs av UM:s partners?

Den övergripande graderingen för **uppdragsbeskrivningar för utvärderingar** var 64,3 av 100. Metautvärdering fastslår att uppdragsbeskrivningarna uppvisar brister inom ett antal viktiga områden, inklusive "kärnområden" där specifik vägledning för insatsen krävs: utformningen av forskningsfrågor, instruktioner gällande åtaganden relaterade till biståndseffektivitet, rekommendationer gällande metodik samt kontext. Graderingen var positiv inom området "logik, syfte och mål", "resurser" samt beskrivning av utvärderingsprocessen. Det är möjligt att fastslå att kvaliteten på de uppdragsbeskrivningar som tas fram av de ansvariga enheterna och departement på UM i huvudsak har liknande kvalitet som de som tas fram av sina internationella motsvarigheter, förutsatt att metastudiens portfölj är representativ. Samtidigt råder det inga tvivel om att kvaliteten fortfarande är lägre än vad den borde vara, givet uppdragsbeskrivningarnas roll i upphandlingsprocessen och den kvalitetssäkring som skall genomföras av UM:s tjänstemän.

Den övergripande graderingen för **utvärderingsrapporterna** är 64,4, samma som för uppdragsbeskrivningarna. Överlag är genomsnittsnivåerna för utvärderingsrapporterna som omfattas av studien inte helt och hållet av tillfredställande kvalitet. Rapporter som tagits fram av regionala och tematiska enheter på UM fick i genomsnitt 56,95 poäng, medan rapporter som tagits fram av andra aktörer i genomsnitt fick 69,09, vilket alltså innebär en skillnad på nästan 13 poäng.

FF 2: Vilken omfattning har UM:s utvärderingar (jämförelser mellan planerade utvärderingar och de som faktiskt genomförts)?

Metautvärdering lyckades inte få fram data som krävs för att svara på denna fråga. UM höll med om att det inte fanns tillräckligt med information tillgängligt som skulle kunna användas för att genomföra en analys över "täckningen".

FF 3: Vilken kvalitet höll förhandsbedömningarna och deras motsvarande uppdragsbeskrivningar?

Metautvärdering visade att uppdragsbeskrivningarna för ex-ante bedömningarna inte var tillräckligt specifika när det gällde att peka ut riktningen. Den visade också att uppdragsbeskrivningarna hade brister när det gällde att definiera tydliga frågor, och att de presenterade regelbundet ett stort antal forskningsfrågor utan att samla dem runt fördefinierade analyskriterier (vilket begärs av UM:s standarder). När alla delar av förhandsbedömningarna och uppdragsbeskrivningarna/instruktioner till anbudsgivarna togs i beaktning så var det genomsnittliga resultatet 64,78 av 100. Givet att a) personerna bakom dessa dokument är personal på UM och dokumenten genomgår kvalitetsanalys av handledare på UM, och b) vissa av dokumenten även kan ha varit föremål för granskning av utvecklingspartners, mottagande organisationer eller verkställande organ så är det genomsnittliga resultatet mycket lågt. Låga poäng gavs specifikt i de frågor som skulle besvaras, instruktionerna runt biståndseffektivitet och metodologiskt tillvägagångssätt. Uppdragsbeskrivningarna fick däremot höga poäng i "logik, syfte och mål" och i beskrivningen av bedömningsprocessen.

Rapporterna som producerades inom ramen för ex-ante bedömningarna var generellt av låg kvalitet (genomsnittligt resultat 46,5) vilket kan vara en återspeglning av bristerna i de uppdragsbeskrivningar som låg till grund för dem. Viktigt att notera här är att de uppvisade låga poäng när det gällde att presentera evidensbaserade resultat och att ge tillfredsställande svar på frågor som ställs uppdragsbeskrivningen (båda dessa faktorer är centrala när det gäller bedömningar).

FF 4: Vad kan sägas om kvaliteten på det finska utvecklingssamarbetet baserat på de pålitliga decentraliserade utvecklingsrapporterna, och relaterade planeringsdokument, utifrån OECD/DAC:s kriterier?

Det finska utvecklingssamarbetet bedömdes vara **relevant** så till vida att den ligger i linje med både finska och mottagande organisationers policyer och strategier. Den var endast måttligt **effektiv** när det gäller den övergripande måluppfyllelsen (en del av de låga poäng är utan tvekan ett resultat av avsaknaden i många projekt av informationssystem för att uppfölja måluppfyllelsen vilket resulterade i att rapporterna inte kunde analysera den). **Kostnadseffektiviteten** fick låga poäng, till stor del på grund av att rapporterna inte tog denna faktor i beaktning. Vissa rapporter omfattade budgetar och utgifter men inte kostnadseffektiviteten *per se*, förutom att antyda att byråkrati och komplexa upphandlingsprocesser väsentligt saktade ner genomförandet. **Hållbarheten** fick relativt låga poäng då rapporterna inte lyckades kommunicera varför man borde anse att insatserna var hållbara på sikt, alternativt att de a) nämnde "potentiell hållbarhet", eller b) förutsatte hållbarhet även fast de samtidigt fastslog att viktiga komponenter inte skulle leva upp till målen. Den mycket låga poängen som sattes på **effekten på lång sikt** reflekterar det faktum att väldigt få av rapporterna var kapabla att ange vad effekterna skulle bli. En övergripande frånvaro av informationssystem för insamling av data runt effekter tillsammans med "svepande" bedömningar runt effekterna ligger bakom en stor del av de låga resultaten.

FF 5: Vilka är anledningarna till att man utför en exekutiv översyn (management review) snarare än en utvärdering (om möjligt)?

Inga av de rapporter som analyserades bidrog med några trovärdiga insikter i denna fråga.

FF 6: Vilka är de huvudsakliga slutsatserna som dras av utvärderingsrapporterna? Vilka är framgångshistorierna, de goda exemplen och utmaningarna?

Följande faktorer utgör endast ett litet urval av de slutsatser som dras av metautvärderingen som kommer från de slutsatser som dras av de olika enskilda rapporterna:

Relevans:

- De finska projekten lyckas i huvudsak bra med att tillgodose behoven hos målgrupperna.
- På ett övergripande plan tenderar rapporterna att nämna anpassning till finsk policy, men endast på högsta graden av abstraktionsnivå vilket gör det svårt att använda informationen för policyutveckling.

Resultat:

- Den induktiva analysen pekar på en relativt hög grad av frustration vad avser de utmaningar som varje insats ställs inför. Samtidigt noterades också många innovativa åtgärder som utvecklades och implementerades för att lösa kontextuella och tekniska problem. Det är transformeringen av output till resultat som står inför mångfacetterade utmaningar, varav de flesta uppenbarligen inte förutses i planeringsstadiet.
- Det finska utvecklingssamarbetet kan inte ses att nå sina övre utvecklingsmål. På andra hand, de finska insatserna tenderar att producera de flesta av de direkta effekter som de förväntas producera och huvuddelen av de mellanliggande effekterna men de lyckas inte att transformera dessa till övre utvecklingsmål. De är också inte fria från allvarliga utmaningar; dessa inkluderar nästan alltid överoptimistisk målsättning, mismatchning av upphandlingar och den tid de behövde (särskilt med multilaterala organ), en brist på ledningsfokus som ofta berodde på dåliga resultatdefinitioner.
- För många projekt som involverade tekniskt stöd noterades att utföraren och deras lokala motsvarigheter tog fram utkast på lagar och andra dokument som sedan aldrig behandlades.

Effektivitet:

- Åtgärderna var inte tidseffektiva och uppvisade omfattande förseningar där upphandlingar och beslutsfattande pekades ut som huvudansvariga.
- Finskt bistånd kännetecknades av sin flexibilitet. Nationella regeringar och multilaterala organ kännetecknades av att vara alltför rigida.
- Projekt strävar inte efter kostnadseffektivitet per se. De är i betydligt högre utsträckning inriktade mot att "genomföra det som planerades på det sätt som planerades" samt att hantera budgeten och utgifterna inom den upprättade utbetalningsplanen.

Hållbarhet:

- Finska insatser använder generellt tekniska lösningar som är anpassade till behoven och kapaciteterna hos målgruppen vilket gör att denna lätt tar till sig dem.
- Ekonomisk hållbarhet är sällan garanterad, inklusive vid projektets slut.
- Även om kapacitetsutveckling är en del av insatsen så är den "organisatoriska hållbarheten" som krävs för att åstadkomma tillfredställande resultat i slutändan mycket låg.

Inverkan:

- Det är uppenbart att det inte är helt klart i vilken utsträckning insatserna inom ramen för det finska utvecklingssamarbetet faktiskt ger den långsiktseffekten som är tänkt. Den information som skulle krävas för att kunna svara på detta samlas inte in på ett systematiskt sätt och utlåtanden runt inverkan och till och med resultat på ett mer övergripande plan beskrivs i väldigt konceptuella termer och är inte enkelt att utvärdera.

Biståndseffektivitet:

- Teamet som utförde meta-analysen fastslog att de flesta rapporter inte behandlar området biståndseffektivitet som ett separat koncept.
- Inriktning med mottagarlandets policyer är mycket stark, framför allt på högre nivåer. Den används aldrig i rapporterna som ett koncept för att indikera anpassning till delstrategier eller detaljerade nationella handlingsplaner.
- Harmonisering redovisas mycket sällan som ett problem, även fast rapporterna ibland räknar upp andra donatorer med vilka åtgärderna samspelar.

Mänskliga rättigheter, rättighetsperspektiv (HRBA) och genomgående målsättningar

- Den gradering som gjorts ger UM tydliga indikationer på att dess policy om rättigheter och genomgående målsättningar inte implementeras alternativt inte rapporteras.
- Teamet bakom metastudien fann att även fast termen rättsperspektiv (HRBA) nästan alltid nämndes i rapporterna som skrevs inom ramen för 2012 års utvecklingspolitisk riktlinje så utvärderade de aldrig helt detta tillvägagångssätt.
- Jämlikhet mellan könen behandlas som en "gör eller gör inte"-fråga. En stor del av rapporterna noterade att vissa av aktiviteterna inkluderade kvinnor som "mål", såsom att inkludera kvinnor i kurser som deltagare, men noterade samtidigt att de inte var involverade i beslutsfattande eller inte var de främsta mottagarna som ett resultat av ett samvetet beslut. Endast en handfull av insatserna berördes av övervakningssystem som tog med kön i beaktande.

- Utvärderingsrapporterna behandlar inte "ojämlikhet" som ett specifikt område. Faktum är att termen mycket sällan används.
- Många rapporter nämnde klimatet, men i nästan samtliga fall rörde det sig om ytliga referenser.

FF 7: Vilka lärdomar kan dras av förhandsbedömningarna (och deras uppdragsbeskrivningar) vad avser kvaliteten på den inledande utformningen av insatser inom ramen för det finska utvecklingssamarbetet?

På det stora hela så är inte programdokumenten kapabla till bedömningar eftersom viktiga delar av utformningen ofta saknas, inklusive logiken bakom en eventuell insats, resultatramverket, detaljerad implementeringsstrategi, fastställandet av mellanliggande resultat och output samt analys av den utsträckning till vilken informationsdatabaser och data för baslinje finns tillgängliga.

Ett antal förhandsbedömningar noterade hur litet som gjorts och då de inte hade mandat att ändra i programdokumenten var deras rekommendationer mycket breda och "svepande".

Intressant nog så identifierade vissa utvärderingsrapporter den utsträckning till vilken de problem som "deras" insatser ställdes inför var ett resultat av dålig utformning.

Rekommendationer

Metastudiens huvudsakliga rekommendationer är som följer:

A) Strategisk nivå:

1. UM borde etablera mekanismer, inklusive mekanismer för övervakning och kvalitetssäkring, för att understödja dess policyer runt ledningen av bilaterala samarbeten.
2. En inledande analys bör genomföras på ledningsnivå (d.v.s. en noggrann analys och lämpligt analytiskt tillvägagångssätt med bakgrund av ansvarsområdena för UM:s tjänstemän), för att på så sätt identifiera vad UM:s projektledare känner att de kan och borde få ut av utvärderingsfunktionen under 2016 och därefter.
3. Mot bakgrund av de slutsatser som dras gällande den relativt låga kvaliteten på dokument relaterade till ex-ante bedömningar borde UM ändra dessa dokumenters funktion så att de upprättas betydligt senare under projektcykeln. Utkast till programdokument bör vara i de närmaste slutförda och leva upp till minimistandard vad avser innehåll och utformning innan de blir föremål för den typ av granskning som genomförs i en förhandsbedömning.

4. Mot bakgrund av de slutsatser som dras i kapitel 7.3 (m.a.o. vad de olika utvärderingsrapporterna säger om det finska utvecklingssamarbetet) så bör UM:s operativa enheter med ett kritiskt förhållningssätt **sträva efter att förstå vad som ligger bakom svagheterna** som konstateras inom områden som relevans, effekt, effektivitet, resultat och hållbarhet för dess insatser. Som ett led i denna rekommendation borde UM inkludera i uppdragsbeskrivningar plikten hos utvärderare och bedömare av projekt för att direkt knyta dessa till det finska utvecklingssamarbetets riktlinjer.
5. På grund av slutsatserna som relaterar till en ojämn tillämpning av rättighetsperspektivet och mänskliga rättigheter i det finska utvecklingssamarbetet borde UM genomföra en intern utvärdering (eventuellt i form av en förvaltningsrevision) runt de metoder som är förknippade med ett sådant perspektiv och de mål och utfall som sattes upp för det.

B) Operativ nivå

6. Signifikant strama upp de metodologiska kraven för startrapporter (kunden bör godkänna en detaljerad metodologi som innefattar dataresurser, indikatorer, verktyg för datainsamling och analys, urvalsmetoder, intervjuguider och intervjuanteckningar).
7. Insistera att bevis presenteras som stöd för alla påstådda observationer.
8. Bättre definiera förväntningarna som föreligger för utvärderingar och bedömningar vad avser kriterierna för koherens, finskt mervärde och biståndseffektivitet.
9. Utveckla ett separat vägledande dokument som specifikt behandlar vad som utgör ett acceptabelt innehåll i rapporter och fastställa normer och standarder för dem.
10. Lätt justera bedömningstabellerna som togs fram för denna metautvärdering och insistera på att berörda personer använder sig av dessa för att bedöma kvaliteten i de rapporter de mottar. Internt så kan tjänstemän använda sig av bedömningstabellerna för uppdragsbeskrivningar för att kontrollera deras struktur, innehåll och kvalitet.

C) Rekommendationer med anknytning till kapacitets- och förmågeutveckling

11. Berörda personer på UM borde kunna bedöma kvaliteten i dokumenten som integrerar rättighetsperspektivet och genomgående målsättningar (inklusive OECD/DAC och specifik utrikesministerium).
12. Kapaciteten hos berörda personer på UM att verkligen förstå och granska utvärderingars och förhandsbedömningars iakttagelser och slutsatser, såväl som uppföljnings och andra rapporter, mot bakgrund av logiken runt särskilda insatser (genom till exempel enlogg eller *Theory of Change*) bör stärkas avsevärt.

EXECUTIVE SUMMARY

Background, Purpose and Objective

The Development Evaluation Unit (EVA-11) of the Ministry for Foreign Affairs of Finland (MFA) has commissioned this meta-evaluation of the evaluations of projects and programmes (including appraisals) performed by implementing units of the MFA. This document represents the Final Report for that meta-evaluation.

Purpose

The Terms of Reference (ToR) for this meta-evaluation (see Annex 1) identify the purposes of the meta-evaluation as:

FIRST: “in an initial phase, to help the MFA improve the quality of evaluations, the evaluation management practices and the overall evaluation capacity development. It will also provide an overall picture of the current evaluation portfolio which helps the MFA to identify possible gaps”.

SECOND: “in a subsequent phase, to bring forward issues and lessons learned that emerge from the evaluation reports and to give recommendations which will help the MFA to improve Finnish development cooperation. It will do this by assessing the kinds of strengths and challenges facing Finnish development cooperation that are identified in the different evaluation reports studied in the first phase”.

Objectives

The objectives of this meta-evaluation, as stated in the ToRs, are also twofold:

FIRST: “the meta-evaluation will assess the quality of different decentralized evaluation reports and related planning documents. It will also draw an overall picture of the evaluation portfolio in 2014-2015 and assess the evaluation coverage in 2013-2015”.

SECOND: “the meta-evaluation will synthesize reliable evaluation findings and issues rising from the evaluation reports on Finnish development cooperation”.

Comparison between meta-evaluations

Previous meta-evaluations were conducted in 2007, 2009, 2012 and 2014. The tools and methodologies of these past efforts have evolved considerably over time, and MFA is anxious to stabilize them so that longitudinal studies can begin in earnest. This *meta-evaluation differs* from the others in significant ways, not the least of which is that the assessment grids for the first phase of the meta-evaluation (i.e. the quality analysis of the Terms of Reference and reports of appraisals and evaluations) are now based on an entirely different logical foundation. In addition, the analysis framework for the second phase of the meta-evaluation (i.e. the analysis of Finnish development cooperation from the perspective of evaluation reports) is **also** based on entirely different

logical foundations than it was the case in the past. Both of these changes were requested by MFA as a result of interchange with the Meta-evaluation team. So care must be exercised when trying to identify long-term trends and changes. A section in this report compares more strategic-level conclusions from the last meta-evaluation with those developed for this one. By and large, the conclusions are similar, and the same issues are often highlighted.

Description of sample

Thirty-six different evaluation (n=26) and appraisal (n=10) reports were studied. Sixty-three percent were from Asia or Africa, and a further 17% were global projects. Only 12 (out of 26) of the evaluation projects took place in the official and traditional bilateral partner countries of Finnish development cooperation. There were 13 sectors represented, with four accounting for 50% of the total. Environment and three other natural resources-related sectors presented 40% of the total, meaning that the change of direction given to Finnish development cooperation by the 2007 Development Policy Programme is only now visible in the sample of evaluations. Forty-six percent were country-specific while 48% were either “Regional/multi-country” or “worldwide”. Importantly, 51% of Finnish projects had budgets (Finnish part) of less than 5 Million Euros (MEUR). Only 29% of projects had budgets greater than 10 million. The portfolio therefore consisted of a relatively large number of small projects in terms of funding from Finland, and few very large projects. Only 56% of total Finnish funding went to bilateral partner countries. The implication of all this is that Finnish aid is fragmented, with the added overhead and function duplication/thinning that that entails.

Methodology and risks

The meta-evaluation essentially compared the quality of the contents of the various reports to the requirements that are spelled out in various MFA policy and guidance documents. In a second phase, another assessment identified what insights could be brought forward by the evaluation reports on the extent to which Finnish development cooperation policy was being implemented.

A complex methodology and research protocol was set up for this meta-evaluation. It included an innovative integration of both a deductive and inductive approach in the analysis of Finnish cooperation, a highly unusual research strategy that is justified and described in an annex to this report and in the inception report. All assessments done by one team member were cross-checked by others, and a significant effort was invested into ensuring that team members understood how to rate each characteristic in the same manner. A very detailed Annex describes the process and methodology.

A number of risks were identified early on, including the effect of a position taken by the team to not assume that something had been reported on unless it was actually specifically written. If there was no description of efforts in aid effectiveness, for example, none was assumed from what else may have been written. Another risk is associated with representability of the sample, but the team believes that the results are valid, and replicable.

Summary of Findings and Conclusions

The Terms of Reference for this mandate included seven evaluation questions (EQ).

EQ 1: What is the quality of MFA's decentralized evaluation portfolio (evaluation reports and their corresponding ToRs) based on the OECD/DAC evaluation standards in 2014-2015 and the guidance given in the Evaluation Manual and the requirements classified by countries, sectors, budgets, evaluation types, managing units of MFA, commissioner, consultant companies etc.? Is there a difference between the quality of MFA commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?

The overall rating for **evaluation ToRs** was 64.3 out of 100. The meta-evaluation concluded that the ToRs were weak in a number of important areas, including the “core” sections where specific direction on the intervention is required; the statements of EQ, instructions on aid effectiveness commitments, recommendations concerning methodology and context. The ToRs scored positively for “rationale, purpose and objectives”, “resources” and describing the evaluation process. It can be concluded that the quality of TOR written by the implementing units and departments of MFA is more or less at the level of their international peers, as far as it can be assumed that the meta-evaluation portfolio is representative of that universe. It can also be concluded that the quality of ToR is much lower than it should be, considering its role in the contracting process and the quality assurance that must be exercised by MFA officials.

The overall rating for **evaluation reports** is 64.4, the same as for the TORs. As such, on the level of averages, the reports in this portfolio are of a not totally satisfactory quality. Reports commissioned by the regional and thematic units of MFA score, on the average, 56.95, while the reports commissioned by some other agency score 69.09, almost thirteen points of difference.

EQ 2: What is MFA's evaluation coverage (comparison of evaluation plans and realized evaluations)?

The meta-evaluation was not able to find the data required to answer this question. MFA agreed that the information was not there in a form that would have enabled the team to help develop a “coverage” analysis.

EQ 3: What is the quality of the appraisal reports and their corresponding ToRs?

The meta-evaluation found that **appraisal ToRs** were not specific in their direction. It also found that the ToR were weak in developing clearly defined issues and regularly identified a large number of issues to study without (as requested by MFA standards) clustering them around pre-defined analysis criteria. When all parts of the assessment grid for appraisal ToR/ITT are taken into account, the average number of points given is 64.78 out of a possible 100. Given that a) the authors of these documents are internal to MFA and that the documents themselves are subject to QA by MFA supervisors and b) some of the documents may have been subjected to review by development partners, recipient organisations or implementing agencies, the average score is very low. Low ratings were given for the specificity of issues to appraise, the instructions for aid effectiveness and the approach. ToRs scored highly in “rationale, purpose and objectives” and the description of the appraisal process.

Appraisal reports generally were of poor quality (average rating of 46.5) which may reflect, as it were, the ToR that generated and guided them. Importantly, they scored poorly on presenting evidence-based findings and on providing satisfactory answers to issues identified in the ToR (both are core elements of any appraisal).

EQ 4: What can be said about the quality of Finnish development cooperation based on the reliable decentralized evaluation reports, and related planning documents, by each OECD/DAC criteria?

Finnish development cooperation was found to be **relevant** in that it is aligned to both Finnish and beneficiary organisation policies and strategies. It was only moderately **effective** in meeting its higher-level objectives (part of the low score is clearly attributable to the fact that many interventions do not have information systems that monitor these objectives so the reports could not report on them). **Efficiency** was awarded a low score, largely because the reports did not report on it. Some reports measured budget and expenditures but not efficiency per say except to indicate that bureaucracy and complex procurement procedures slowed down execution considerably. The **sustainability** rating was rather low because the reports did not show why they believed that interventions would be sustainable, or they a) spoke of “potential sustainability” or b) assumed sustainability even if major components were reported as not going to meet objectives. The very low rating given to **impact** reflects the fact that very few reports were able to indicate what the impact would be. An across-the-board absence of information systems to gather required data on impact, coupled with what were very clearly “lofty” expressions of impact together account for a large part of the low score.

EQ 5: What are the reasons to commission a management review instead of an evaluation (if possible)?

None of the reports analysed provided any research-quality insights into this question.

EQ 6: What are the major issues emerging from the decentralized evaluation reports? What are success stories, good practices and challenges?

The following are but a small sample of the key points identified in the second phase of the Meta-evaluation as identified in the individual reports:

Relevance:

- Finnish projects tend to be very good in specifically addressing the needs of targeted groups.
- Generally, reports tend to speak of alignment with Finnish policy, but at the highest levels of abstraction only, making the information difficult to use for policy development.

Effectiveness:

- The inductive analysis indicted a relatively high degree of frustration with the challenges facing any intervention, but noted many innovative measures that were designed and implemented to resolve context and technical problems. It is the transformation of outputs into outcomes

that faces multifaceted challenges, most of which were apparently not foreseen in the design stage.

- While Finnish **cooperation** overall was not shown to be effective at meeting higher-end objectives (see page 4 above and the relevant sections of this report), Finnish **interventions** tend to achieve a majority of the (lower-lever) effects that were identified in the results chain analysis (i.e. those that are directly generated by the outputs). These lower-end effects somehow do not get transformed into higher-end effects; the analysis that would shed light on the reasons for this are way beyond the scope of this meta-evaluation. The meta-evaluation analysis does show that interventions have many serious challenges however, including over-scoping, mismatching of procurement deliveries and the time they were needed (especially with multilateral agencies, a lack of management focus that often resulted from poor result definition (an illustrative and partial list only).
- Many projects that involved Technical assistance noted that the TA and their counterparts produced final drafts of proposed laws, regulations and other documents that were never brought forward for adoption. The hypothesis that can be drawn here is that either the TA were working on tasks that were not seen as relevant to the “client” and they (i.e. the TA) were not efficiently used, or that the solutions proposed were not seen as appropriate or wanted.

Efficiency:

- Interventions were not time efficient, with long delays for procurement and decision-making noted as key problems.
- Finland aid was noted for its ability to provide flexibility. National governments and most multilateral agencies were specifically identified as being overly rigorous.
- Interventions do not generally manage efficiency per se. They are much more concerned about “doing what was planned the way it was planned”, and managing the budget and the expenditures within an approved disbursement plan.

Sustainability:

- Finnish interventions generally use technical solutions that are adapted to the needs and capabilities of the target beneficiaries. Beneficiaries easily adopt and “own” them
- Financial sustainability is rarely assured, even at project end.
- Even if a capacity development component is part of the intervention, the “organisational sustainability” required to continue towards outcome achievement is very low.

Impact:

It is clear that Finnish development cooperation does not have a handle on the extent to which its interventions contribute to expected impact. The information required is not gathered systematically and the state-

ments of impact or even of higher-level outcomes are written in high-level conceptual terms and are not readily “evaluable”.

Aid Effectiveness:

- The meta-evaluation Team found that most reports do not specifically address the issue of aid effectiveness as a separate concept.
- Alignment is particularly strong, especially at higher levels. It is never used as a concept to indicate alignment with sub-strategies or detailed national plans.
- Harmonisation is very rarely reported against as such, although reports briefly list other donors with which the intervention interfaces.

Human Rights Based Approach (HRBA) and Cross-cutting Objectives (CCO)

- The rating given provides MFA with a clear indication that its HRBA policy is not being implemented or is not being reported upon as such.
- The Meta-evaluation Team found that while the term “HRBA” was almost always mentioned in the reports that were written under the 2012 policy umbrella, the reports never evaluated such an “approach”.
- Gender equality is treated either as a “do-or-do-not” issue. A large proportion of reports noted that some activities involved women as “targets”, such as including women in training course, but also noted that they were not involved in decision-making or were not the direct beneficiaries as the result of an overt decision. Only a handful of interventions had monitoring systems concerned with gender at all.
- Evaluation reports do not deal specifically with “inequality” as a specific domain. In fact, the term is rarely used.
- Many reports did in fact mention climate but almost all were superficial references.

EQ 7: What can be learned from appraisal reports (and their ToRs) on the quality of the initial design of Finnish development cooperation interventions?

Overall, the draft Programme Documents are not ready for appraisals because key parts of the design are most often missing, including the development intervention logic, the results framework, the detailed implementation strategy, the statement of intermediate results and outcomes and the analysis of the extent to which information database and baselines are available.

A small number of appraisals noted how little had been done and, since they were not mandated to change the draft PD, their recommendations were rather broad and all-inclusive.

Interestingly, some evaluation reports identified the extent to which the problems “their” interventions faced were the result of poor design.

Recommendations

The key recommendations of the meta-analysis are:

A) Strategic level

1. MFA should put in place mechanisms, including those for monitoring and quality controls, to help it better enforce its own policies concerning the management of bilateral cooperation.
2. An “uptake” analysis should be done on a managerial research basis (i.e. with rigorous analysis and an appropriate analytical approach based on the accountability framework of MFA managers), in order to identify, within the 2016-and-beyond context, the benefits that MFA managers feel they could and should extract from the evaluation function.
3. Based on the conclusion dealing with the poor overall ratings given to appraisal-related documents, MFA should change the role of appraisals so that they take place considerably later on in the project cycle. Draft PDs should be in a near-complete state and meet minimum content and design standards before being subjected to the critique that can only be rendered through an appraisal.
4. Based on the conclusions in chapter 7.3 (i.e. What evaluation and appraisal reports reveal about Finnish development cooperation), MFA’s operating divisions should critically **seek to understand the causes for the weaknesses** found in the relevance, effectiveness, efficiency, impact and sustainability of all of its interventions. As part of this recommendation, MFA should include in its ToRs a reference to the obligation of evaluators and appraisers to specifically link the interventions to Finnish development cooperation policy.
5. Based on the conclusions related to the very uneven application of the HRBA policies of the Government of Finland, MFA should undertake an internal assessment (perhaps in the form of a management audit) of the practices associated with that HRBA policy and the objectives and outcomes that were set for it.

B) Operations level

6. Significantly tighten methodology requirements for inception reports (the client should approve a detailed methodology that included the data sources, indicators, tools for data collection and analysis, sampling methods, interview guides and interview notes).
7. Insist that evidence be specifically provided to support all findings.
8. Better define the expectations of evaluations and appraisals with respect to the three Finnish criteria coherence, Finnish value-added and aid effectiveness criteria.
9. Develop a separate guidance document that specifically addresses the acceptable content of reports, and provides norms and standards for them.

10. Modify slightly the assessment grids prepared for this meta-evaluation and insist that officials use them to judge the quality of the deliverables (reports) they receive. Internally, officials and supervisors can use the ToR assessment grids to double check the structure, content and quality of TOR.

C) Recommendations dealing with capability and ability development

11. MFA officials should be enabled to assess the quality of assurance-related documents that integrate HRBA and CCO into the management criteria (including OECD/DAC and specific MFA). This is fundamentally a question of design policy.
12. The ability of MFA officers to truly understand and critique evaluation and appraisal (ex-ante evaluation) findings and conclusions, as well as monitoring and other reports, in the light of the centrality of the logic of specific intervention (through a log frame or Theory of Change, for example) should be significantly improved.

The assessment grids for the first phase and the analysis framework for the second phase are based on entirely different logical foundations than those used in past meta-evaluations.

1 INTRODUCTION

1.1 Context

1.1.1 Background to the meta-evaluation

The Development Evaluation Unit (EVA-11) of the Ministry for Foreign Affairs of Finland (MFA) has commissioned this meta-evaluation of the evaluations of projects and programmes (including appraisals) performed by implementing units of the MFA. This document represents the Final Report (FR) for that meta-evaluation.

Decentralised evaluations and appraisals are the responsibility of the MFA departments and units that are charged with the development cooperation programmes in specific countries, regions or with international organisations. Decentralised evaluations include appraisals (ex-ante evaluations), mid-term evaluations, and final, or ex-post evaluations. These are clearly defined in the Bilateral Project/Programme Manual and in the Evaluation Manual of the MFA.

Previous meta-evaluations were conducted in 2007, 2009, 2012 and 2014. The tools and methodologies of these past efforts have evolved considerably over time, and EVA-11 is anxious to stabilize them so that longitudinal studies can begin in earnest. This meta-evaluation differs from the others in significant ways, not the least of which is that the assessment grids for the first phase of the meta-evaluation (i.e. the quality analysis of the Terms of Reference (ToR) and reports of appraisals and evaluations) are now based on an entirely different foundation. In addition, the analysis framework for the second phase of the meta-evaluation (i.e. the analysis of Finnish development cooperation from the perspective of evaluation reports) is also based on entirely different logical foundations than it was the case in the past, as explained in later sections of this Final Report.

Meta-evaluations are useful in a number of ways, not the least of which are:

- As a means of implementing the GoF policy on transparency wherein each ministry must report on its accountability framework to Parliament every four years;
- As a means of helping the management teams of both EVA-11 and the MFA to prepare annual and mid-term plans;
- As a means of independently reporting on the quality of the work done by external contractors;
- As a means of data-mining and consolidating lessons learned from a wide variety of evaluative research evaluations so as to inform policy making within the MFA;
- As a means of identifying gaps and opportunities for improving the execution of evaluation and project/programme management cycles within the Ministry;

- As a means of identifying possible areas where capability gaps and capacity weaknesses may exist within MFA and its key partners (i.e. what has to be done to improve the capability of MFA staff to manage the evaluation process and individual evaluation actions);
- As a means of analysing coverage (extent to which those parts of the Finnish development cooperation that is supposed to be evaluated is actually evaluated);
- As a means of identifying and analysing long-term trends.

1.1.2 Purposes and objectives of the meta-evaluation

Purposes:

The Terms of Reference for this meta-evaluation (see Annex 1) identify the purposes of the meta-evaluation as:

FIRST: “in an initial phase, to help the MFA improve the quality of evaluations, the evaluation management practices and the overall evaluation capacity development. It will also provide an overall picture of the current evaluation portfolio which helps the MFA to identify possible gaps”.

SECOND: “in a subsequent phase, to bring forward issues and lessons learned that emerge from the evaluation reports and to give recommendations which will help the MFA to improve Finnish development cooperation. It will do this by assessing the kinds of strengths and challenges facing Finnish development cooperation that are identified in the different evaluation reports studied in the first phase”.

Objectives:

The objectives of this meta-evaluation, as stated in the ToR, are also twofold:

FIRST: the meta-evaluation will assess the quality of different decentralized evaluation reports and related planning documents. It will also draw an overall picture of the evaluation portfolio in 2014-2015 and assess the evaluation coverage in 2013-2015.

SECOND: the meta-evaluation will synthesize reliable evaluation findings and issues rising from the evaluation reports on Finnish development cooperation.

The results of this meta-evaluation were compared to the results of the Meta-evaluation of Project and Programme evaluations 2012-2014 in order to find trends, patterns and changes. Because of the revised assessment tools and the use of an additional inductive approach during Phase 2, there are many constraints to develop and then interpret this comparison. These constraints will be spelled out in the methodological section.

1.2 Scope

The scope of the meta-evaluation is clearly defined in the ToR provided (refer to ToR, pp. 2 and 3):

The research domains of the meta-evaluation are the MFA's decentralised evaluations and appraisals, and specifically their ToR and reports. The number of reports assessed is 36.

The quality of the evaluations conducted on Finnish development cooperation initiatives and programmes was compared to those of its key partners.

- The first part of the evaluation requires an assessment of the quality of the appraisals, evaluation reports and their ToRs and Instructions to Tenderers (ITT). The evaluation ToRs require that the assessment tools developed in the 2012–2014 meta-evaluation be improved so that they may become standardized over time. In fact, they have been significantly modified.
- The temporal scope for evaluations of all types spans the period 09/2014–08/2015. Appraisal reports that were approved from Jan 2013 to Aug 2015 are part of the sample.
- The geographical and institutional domains of the meta-evaluation are the decentralised evaluations and appraisals, and specifically their ToR and various reports (evaluation and appraisal reports and related ToRs). The final total number of the reports assessed is 36. The 36 reports derive from 35 projects/interventions, one of which has two evaluation reports. Countries, sectors, budgets, evaluation types, managing units of MFA, consultant companies of the evaluations and appraisals have been described by the team based on the various reports, ToRs, and ITT.
- The first phase was to include an (annual) systematic assessment of MFA's evaluation coverage. This has proven to be impossible given the way information required is stored and classified within MFA. A section on this issue is included in this report.
- The evaluation also required that the quality of the evaluations conducted on Finnish development cooperation initiatives and programmes be compared with those of its key partners. This is mostly done in the portfolio analysis and in the analysis of the quality of deliverables and products.

1.3 Final report structure

The Final Report contains seven sections:

1. An introduction (background, purpose, objectives, scope)
2. The methodological considerations, including limitations
3. The portfolio analysis concerning the documents analysed
4. The assessment of the quality of evaluation-related documents
5. The assessment of the quality of appraisal-related documents
6. Issues and lessons learned from the evaluation reports concerning Finnish development cooperation
7. Conclusions
8. Recommendations
9. The establishment of an evaluation coverage system within MFA

A series of technical annexes follow, including a detailed methodology and the assessment grids prepared for the meta-evaluation.

2 METHODOLOGICAL CONSIDERATIONS

2.1 Approach

The meta-evaluation's approach directly reflect the instructions laid down in the Terms of Reference and the proposed response to that document found in the mini-tender of Danish Management /Eco Consult and the Inception Report prepared by the meta-evaluation Team. The **key** elements of that approach were:

- A two phase approach where Phase One was an assessment of the quality of documentation used for evaluations and appraisals; in this case the ITT/ToR and the appraisal or evaluation reports. Phase Two was an assessment of the “quality” (the term used in the ToR) of Finnish cooperation, based on an analysis of the evaluation reports that had received the highest scoring in Phase One.
- The development of a set of analysis grids that were applied to a set of evaluation and appraisal ToR documents and reports.
- The use of deductive and inductive reasoning in Phase Two.
- An assessment of quality based on the requirements spelled out in MFA's own manuals and policy documents.
- The complete cross-checking of all assessments done so that three Team members look at each document as a mitigation strategy against analysis bias.
- A quality comparison between MFA commissioned documents and those commissioned by others.
- A portfolio analysis of all the documents retained for analysis.

2.2 Methodology overview

Annex 3 contains a comprehensive description of the methodology used in this mandate.

The mandate's execution began with an analysis of past meta-evaluations and the development of a set of talking points that were discussed with EVA-11 at a start-up meeting. The Team then developed the Inception Report as required and in so doing commented on the consequences of using the OECD/EU Quality Grid as the baseline for a quality assessment of documents when the MFA should, in the opinion of the Team, be comparing the documents it generated (or received as deliverables) against what it specified it wanted or needed (ex. within its evaluation and other manuals).

In Phase One, Team members often had to judge whether a deliverable met the MFA's standards of CONTENT, a function that one could argue should have been done by the official who accepted the deliverable in the first place.

The MFA responded with a request to change the baseline standard, requiring a new IR, a completely new set of analysis grids and other analysis tools. Versions of a revised IR were prepared until the MFA was satisfied with both the application of the approach that would be required, and the structure and content of the assessment tools that would be used. Ranking systems and weighting protocols were defined with MFA guidance.

Reports and ToR/ITT that were to be used as the sample of evaluation as well as appraisal ToR and reports were divided among the three meta-evaluation team members for a comprehensive analysis and then cross-checked by both other team members. Results were posted and analysed resulting in the analysis found later in this report.

At the same time, a table containing all the identifiers of the set of documents that the team analysed was prepared and cross-checked. This “portfolio analysis” was discussed with the MFA and the structure and content were agreed to. The portfolio analysis was later used to help in the research of both Phase One and Phase two.

A list of evaluation documents that had received a minimum score (n=18 reports that had a minimum of 60 points) during Phase One was drawn up and communicated to the MFA as the sample for Phase Two. The analysis of these reports (requiring both deductive and inductive analysis) was followed by the writing of this report.

Data validity was addressed by applying a rigorous QA process throughout the meta-evaluation process; external experts reviewed key documents and worked with the Team to find solutions to the many issues brought forward through the mandate. Triangulation per se was not always possible but a rigorous cross-checking of analysis results was done.

2.3 Examples of consequences of methodological choices made

- The methodology selected places a great deal of importance on what could be termed as “compliance audit” during Phase One. The Team members often had to judge whether a deliverable (or parts thereof) met the standards of CONTENT, a function that one could argue should have been done by the official who accepted the deliverable in the first place and authorized the payment of the invoice for services rendered. Meta-evaluations should be in the future oriented more towards the quality of the content, and less of its existence, even if the Team undertook such analysis in the course of its work. In this meta-evaluation the Team integrated an element of “quality of deliverable” by having each Team member provide an overall rating of multiple characteristics at a headline standard level and not through the simple process of calculating averages. The Methodology annex provides further insight into how this was done.
- Any evaluation report that did not achieve a 60% overall score and therefore did not form part of Phase Two could have provided insight into possible issues to correct in the future (there were 10 of these representing almost a third of all possible reports).

- The evaluation Team made it clear in its mini-tender and its Inception Report that it would not “assume” that something was included in a report unless that “thing” was specifically spelled out. For example, if the report did not contain “answers” but had a findings section that spoke of higher-level judgements on findings, then the Team considered that “answers were not there”. Not presenting any specific link to findings through evidence also was treated as a weakness and “not met” ratings were given. Conclusions had to be specifically based on findings, and recommendations had to be linked to findings and conclusions, or poor ratings were given, for example. The scores may be lower than would have been the case has another set of protocols been used.
- The meta-evaluators chose to place emphasis on the evidence that would demonstrate whether a document “mainstreamed” approaches such as “HRBA” and cross-cutting objectives such as environmental sustainability. If these issues were not specifically addressed, were not supported by evidence or were not mainstreamed, the meta-evaluators assigned lower ratings. In fact, as noted in this report, very few reports even mention HRBA at all and barely recognise CCOs, or if they do they do not relate the approach or CCOs to the intervention’s contribution to the achievement of outcomes.
- The referential approach in Phase One discriminates against sector, or content-focused documents; if a sector expert was retained to write a report, he/she might not be aware of more “political or policy” considerations. As a result, the content part of the documents may satisfy the needs of content managers but not those of development cooperation managers or their partners. In effect, the documents they wrote would not necessarily rate highly.

2.4 Limitations and risk mitigation

There are important epistemological premises that are at play in this meta-evaluation; hypotheses have been laid down and assumptions made concerning the relationship between the raw observations (in the documents of the sample) and the nature of the possible conclusions that might be derived from the sample. Some of these hypotheses need to be examined:

- a) **The evaluation reports and the appraisals can be assessed in an objective manner using a comparison approach with standards and norms.** Reference-based assessments are particularly effective when the norm is not open to interpretation (ex. the height of children must be at least 140 cm if they are allowed to travel alone on an airplane) and when the norm is closed and self-contained (ex. the electrical code for residential housing for the elderly in Canada requires a fire alarm for every kitchen area, a norm for which very little room for interpretation is allowed). That is not the case in this evaluation. Standards are often open to interpretation (ex. words such as “adequately” or “improved” are common and undefined), and norms are open to considerable interpretation (ex. the quality of the indicators or the structure of the Logical Framework). The Team

The Team did not “register” that something was included in a report unless that “thing” was specifically spelled out.

The referential approach in Phase One may discriminate against sector or content-focused documents. In those cases, good sector-focussed reports would not necessarily rate highly.

Reference-based assessments are particularly effective when the standards are both not open to interpretation and are self-contained. That is not the case in this evaluation. Standards were often open to interpretation (ex. words such as “adequately” or “improved” are common and undefined), and are open to considerable interpretation.

As early as 2009, meta-evaluations have identified the level of capacity of MFA personnel as a constraint, and have noted that turn-over and a lack of time and interest are factors affecting quality.

reacted to this situation by i) writing down, within the assessment tools, the interpretations that should be given or a description of how to treat the characteristic to be assessed and ii) cross-checking reports so that similar interpretations are given to any particular report and any possibility of analysis or subjective bias is mitigated against to the extent possible.

- b) **Cross-checking eliminates analysis bias between researchers.** Although the mini-tender clearly referred to the cross-analysis being done on a “sample” basis (refer to Step 4a in the mini-tender), the MFA preferred a 100% cross-checking process. Guaranteeing replicability through cross-checking is not necessarily possible, especially when the objects of the analysis are different. Some level of bias is therefore represented in the analysis in this meta-evaluation, but steps have been taken to try to minimise that risk. To that end, two complementary strategies were used, one based on comparison analysis of responses and the other was a formal triangulation process where feasible. **One hundred percent of the reports have been cross-analysed in both Phase One and Phase Two.**
- c) **The quality and scope of the reports are sufficient to allow for the assessment required in Phase 2.** The ToR indicated that the second phase should reflect on Finnish development cooperation based on the documentation. While it may be possible to gather and sort findings and conclusions, the relatively small sample used in Phase Two, when one considers the complexity and scope of Finnish external policies including development assistance and international relations (for all possible ramifications), it is a giant leap to suggest that Phase Two should be anything more than a contribution to a wider and more comprehensive analysis. A much more comprehensive, systematic and real-time meta-evaluation system would have to be developed for that objective to be realised.
- d) **The staff of MFA (who manage evaluations and appraisals) are able to manage the quality of the evaluation and appraisal products.** As early as in the 2009 meta-evaluation, and as recently as the last a scant year ago, regular meta-evaluations have identified the capacity of the MFA personnel as a constraint, and have noted turn-over and a lack of time and interest as factors affecting quality. This meta-evaluation certainly assessed the quality of reports that should have been higher if MFA managers were better qualified in this area or supported (see sections on the overall quality of documents analysed during Phase One), including through training but also through relevant systems and other elements that address capability and not only capacity. This line of thinking is particularly important when it comes to Phase 2 because there are no means at our disposal to judge the extent to which the deliverables were fully comprehensive and insightful *in the context of the Phase 2 objective* (i.e. the development cooperation of the GoF). For example, the Team was not in a position to understand the country strategies involved, or the context involving the relationship between MFA and the development partners/recipient country organisations.

- e) **While all the previous meta-evaluations have adopted different rating scales and methodologies, it cannot be assumed that there is a seamless longitudinal logic (i.e., spanning all meta-evaluations) that may be derived from this evaluation.** Since the baseline has shifted from OECD/DAC Quality Grids to MFA manuals and their inherent policies, standards and norms, care should be taken when comparing meta-evaluation reports. On the other hand, this meta-evaluation is based on the internal requirements of MFA and should therefore become the baseline for future meta-evaluations.
- f) **It is assumed that the analyses conducted in Phase 1 and Phase 2 will enable MFA to specifically identify pockets of capacity and capability gaps and address them.** Compared to larger donors, the divisions and departments of the MFA that deal with development cooperation are relatively small organisations with a wide policy spectrum and an imposing geographical coverage. Evaluation is a management assurance mechanism and requires early stage evaluability frameworks and clear definitions of the results that need to be generated to achieve (or to have achieved) expected outcomes. Relatively speaking, it is easy to evaluate or assess inputs and outputs (the data is generally readily available in monitoring reports or periodic contracting reports), but it takes systems, standards and skill (not to mention resources) to evaluate effectiveness, sustainability impact, value-added, aid effectiveness and policy/execution coherency. The Team has mitigated against this risk by qualifying its conclusions (arrived at both deductively and inductively) within a capacity-development paradigm.
- g) **Rating systems and dealing with the interpretations that arise from them.** The team has tried to take the preoccupations of EVA-11 concerning rating systems and their consequences into account. It understands that a rating system based on relative concepts (ex. excellent, very good, good), such as that used in the OECD/EU Quality Grid, is difficult to manage because it is open to much interpretation and personal preferences and the results of that type of “quality” assessment is very hard to communicate. Studies show, for example, that the EU does not use its Quality Grids for lessons learned. Assessment systems based on steps or levels of excellence (such as the one in this meta-evaluation) are much more useful as management assurance tools because they are based on transparent and communicable levels of performance for deliverables (they are based on known norms or standards). In the case where standards have not been complied with by report authors (such as consistently not including data on total budgets) then there are bound to be an important proportion of the reports that get low ratings in one category or other. This is not really a problem of rating but one of quality control and critique (i.e. compliance with norms) on the part of the authors and MFA managers.

Some appraisal reports clearly show that the draft PD analysed was not at a sufficient stage of design to be subjected to that level of ex ante analysis. Intervention performance suffered.

- h) The inductive approach in the meta-evaluation has been linked to each part of the Analysis Grid for Phase 2, a strategy that enables the evaluators to link the results of a deductive analysis directly to the results of an inductive analysis. The **depth of that inductive analysis is limited** however, to the expression (statement) given to the deductive parts of the grid. It is clear that the extent of the usefulness of the inductive research in this meta-evaluation is limited to the contents of the documents and their subsequent deductive approached-based analysis.
- i) The portfolio of reports and ToR for both appraisals and evaluations is not homogeneous and **care must be taken when interpreting the results of their analysis**. Some are much larger than others in terms of budgets; some are another phase of a project and have had years of experience to build on; some are not implemented by MFA at all and so the intervention is, by its very nature, different from an MFA-implemented or managed intervention. And so on. Extrapolating should be done with great care.
- j) The people involved in authoring the documents may not all be fully knowledgeable about the interventions. Some appraisals, for example, have been done on documents that were prepared by those responsible for previous phases of an intervention, and some reports clearly show that the draft PD never should have been subjected to the appraisal because it was not at a stage of design that would allow it to be considered for management approval as a PD. In this way, the MFA may have been using appraisals as a substitute for local efforts to design good PD. The appraisal, in effect, may have been commissioned too early. It is interesting to note that the Team was not made aware of an appraisal that was done after a previous appraisal on the same intervention that had identified major weaknesses.

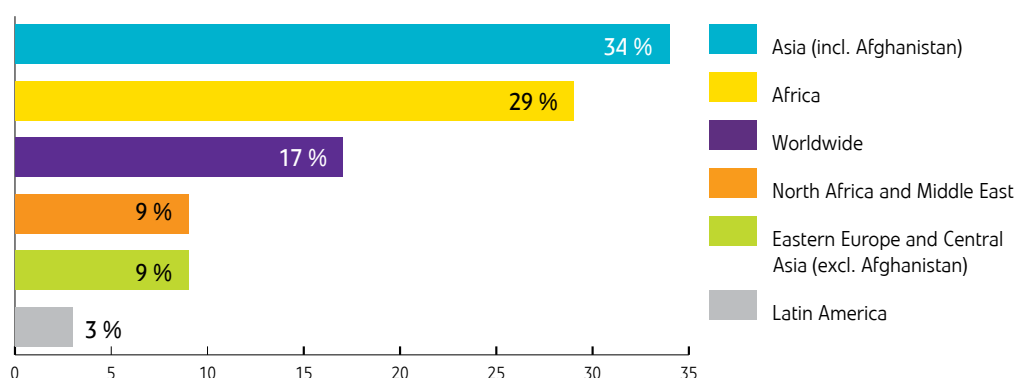
3 PORTFOLIO ANALYSIS AND QUALITY OF EVALUATION REPORTS

3.1 Portfolio overview

The original meta-evaluation portfolio delivered to the Team by EVA-11 consisted of 52 reports; however, by common understanding, many of those original reports were taken out, essentially management reviews, concessional credit scheme appraisals and a self-evaluation. Thus the population of reports retained for the current meta-evaluation is 36, corresponding to 35 projects/interventions with Finnish funding of which one project with two evaluation reports submitted for the meta-evaluation. All percentages in the graphs and statistics in this section have been calculated on the basis of 35 projects (n=35) and 36 reports (n=36).

The **regional distribution** of projects in this portfolio is shown below. In this regional classification, Asia is the largest region in terms of projects evaluated (34 percent) in this portfolio. Over one fourth (29 percent, ten projects) of the projects in our portfolio were in sub-Saharan Africa, mainly in Eastern and Southern Africa. African, Asian, North African (Southern shore of the Mediterranean) and the Middle Eastern projects represent 72 percent of the portfolio. Only 12 (slightly over one third) of the evaluated projects took place in the official and traditional bilateral partner countries of Finnish development cooperation (Ethiopia, Kenya, Mozambique, Nepal, Tanzania, Vietnam and Zambia), despite the firm decision taken in the 2004 Development Policy Programme to concentrate on fewer countries, and fewer sectors in those countries.

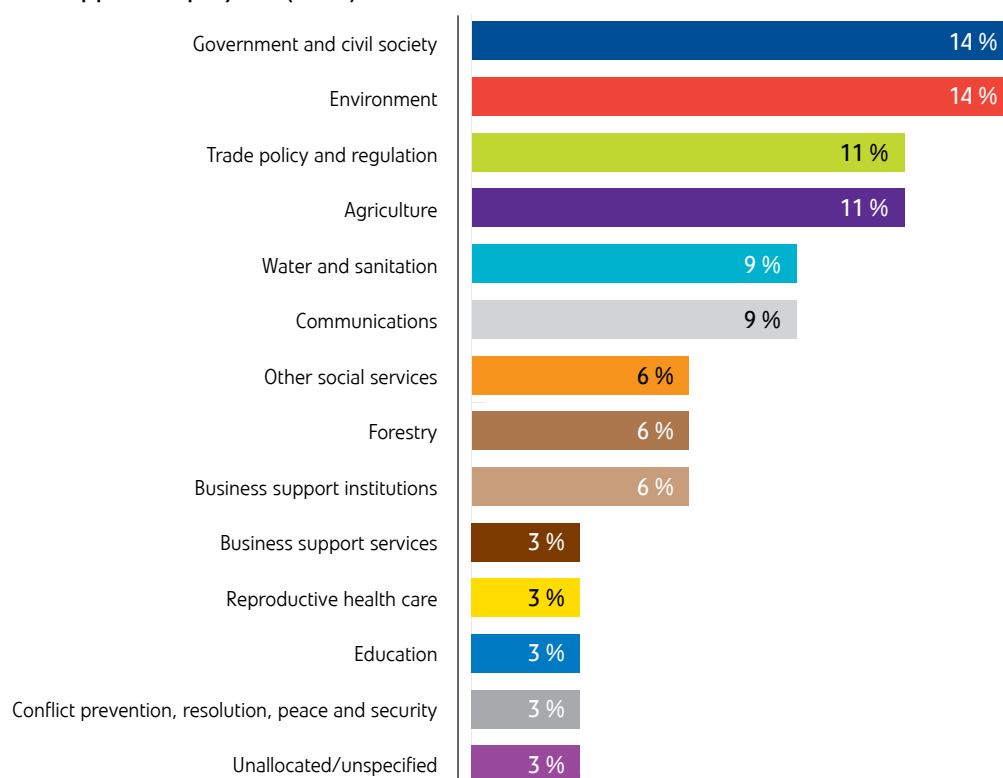
Figure 1: Evaluation/appraisal reports to be reviewed by regional distribution of projects evaluated/appraised (n=35)



Source: Meta-evaluation team

Concerning the development cooperation **sectors** that are covered by the portfolio of projects, the reports can be classified as per the typology below. The sector classification is taken from the OECD CRS codes (in three digits) as they are indicated in statistical summaries of projects/funding decisions of MFA. The largest number of projects was in the environment and government and civil society sectors (five projects in each). One of these latter, however, is very close to the environment sector, i.e. the project that provides support to cartographic services for natural resources and land use mapping (in Lao PDR). Health, a traditional priority sector for Finnish development cooperation, and education (that has made Finland famous internationally through the PISA rankings) together only represent six percent of the projects, one project (3%) for each sector. Compared to the 2012-2014 meta-evaluation, this is a big change as environment was not among sectors named in the distribution (meaning that less than 5% of the projects were in that sector; p. 104). In concluding, it can be said that the orientation of Finnish development cooperation introduced by the 2007 Development Policy Programme, with an important emphasis on environment, agriculture and business, and trade, only now is strongly visible in the portfolio of projects and evaluation reports.

Figure 2: Sector distribution (according to OECD CRS codes) of evaluated / appraised projects (n=35)



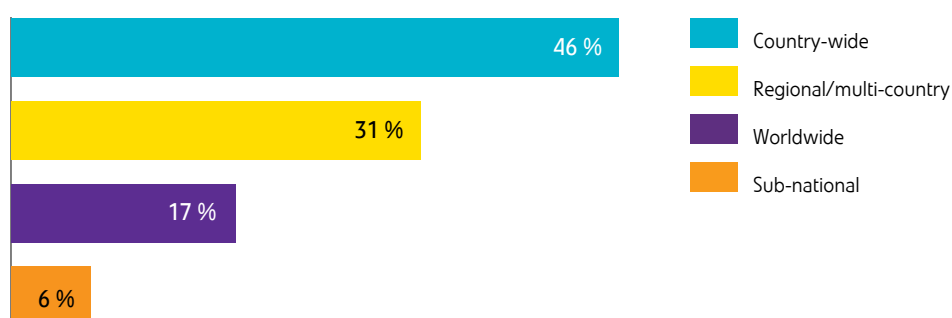
Source: Meta-evaluation team

Moreover, when the larger objective of the 2007 Development Policy Programme, “ecologically sustainable development” and the use of natural resources is considered, the concentration of Finnish aid on the “larger” environment sector becomes even more salient. By clustering environment proper (including climate), agriculture, water and sanitation and forestry together, the

meta-evaluation portfolio represents 40 percent of projects directly related to natural resources. If representative, this portfolio then seems to confirm the observation expressed in 2009 by the Development Policy Committee, an advisory board for the Government, that “Finland’s development cooperation [...] appears to be shifting under the [2007] Development Policy Programme from country-specific to sector- or theme-specific cooperation” (The State of Finland’s Development Policy 2009, p. 20).

With regards to the geo-focus of the projects evaluated in this portfolio, the distribution is presented below. Country-wide projects dominate, and when added to the category of sub-national (one or several regions in a country), the overall one-country projects represent about one half of the projects evaluated (52%). Yet, worldwide projects/programmes make almost one fifth of the portfolio, and when added to the category of regional or multi-country projects, they make up the other half of evaluated projects. One can thus say that the sample of projects with Finnish funding represented in this meta-evaluation portfolio consists of one half of projects in one country, and the other half of projects in a larger number of countries (3-6 countries in the category of regional/multi-country projects).

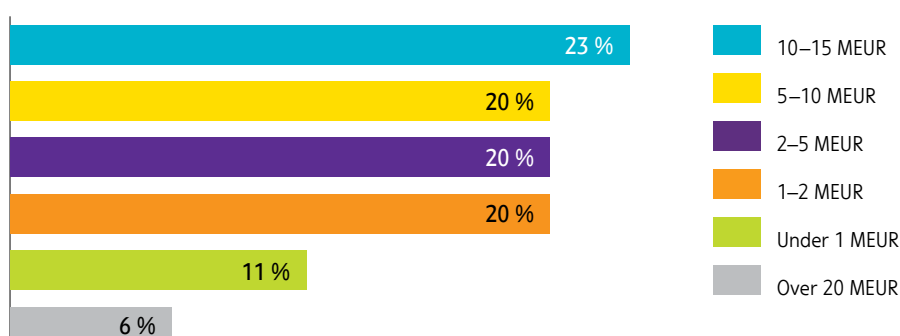
Figure 3: Distribution of projects by geographical scope (n=35)



Source: Meta-evaluation team

The distribution of projects according to **budget allocation** from Finland is presented below in Figure 4. This Figure shows a high level of fragmentation of the portfolio of projects evaluated in this meta-evaluation, perhaps even more than in the case of sector and geographical distribution. Over one half of the projects are small (51%) with a budget of max. 5 MEUR, and six percent (two projects) benefitted from a budget of over 20 MEUR.

Figure 4: Distribution per project budget (in allocations from Finland) (n=35)



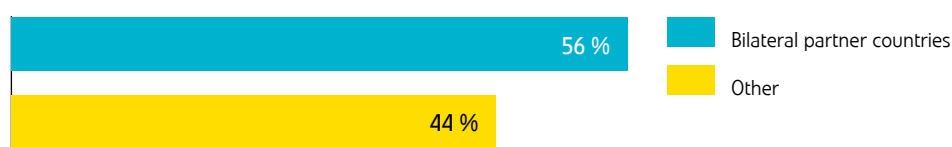
Source: Meta-evaluation team

The portfolio consists of a large number of small projects, and few large projects in terms of funding from Finland.

When analysing more carefully the distribution of projects according to budget, the meta-evaluation found that four projects (11%) had a budget of one million Euros or less, seven projects (20%) a budget between one and two millions; and seven projects (20%) between two and five millions. When taking into account that the two largest projects budget-wise represented 20 percent (n=35) of the sum of all budget allocations (49 MEUR, or 28% when excluding appraisals, n=26), we can conclude that the portfolio consists of a large number of small projects in terms of funding from Finland, and few very large projects.

The meta-evaluation calculated the percentage of budget allocations destined to bilateral long-term partner countries. When including **and** excluding appraisals, the percentage of budget allocations to bilateral, “traditional” partner countries is 56% in both cases. The implication of this is that there is constancy and continuity in shares to bilateral long-term countries and non-traditional beneficiaries, respectively, between past or current evaluated projects forming the portfolio of this meta-evaluation, and future projects in the sample of appraisals.

Figure 5: Distribution of project budget allocations from Finland between long-term partner countries and other beneficiaries (n=35)



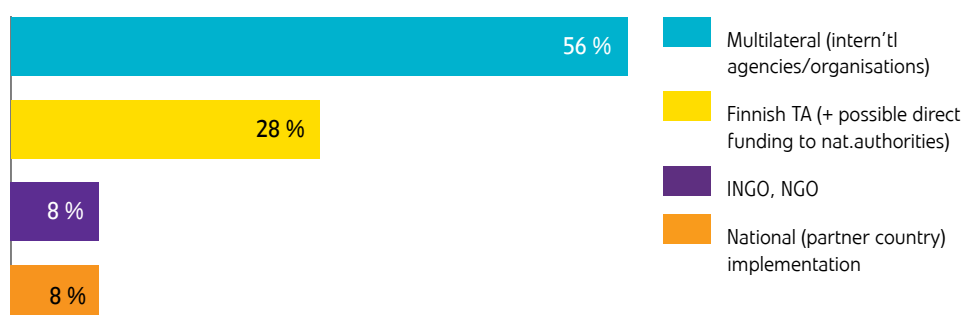
Source: Meta-evaluation team

Fragmentation of aid has long been an issue in Finnish development cooperation. This was already raised by the OECD-DAC peer review of 2003 which observed that allocable bilateral aid to top-ten partner countries had dropped to 53% in 2000/2001, and that Finland, the sixth smallest DAC donor member, had 96 ODA recipient countries in 2003. The peer review recommended that Finland focus its aid to a maximum of 10 partner countries, in order to have “*more cumulative impact, enhanced effectiveness, and increased ability to influence other donors and partner policy dialogue*” (DAC Peer Review 2003, 253-255). The next year, Finland’s new development policy set, as its objective, to increase to 60% of aid the share of Finland aid going to long-term partner countries (2004 Government Resolution). The meta-evaluation cannot compare the budget percentage of 56 to partner countries, as it presents itself in this portfolio, to any other number because the total universe of Finnish aid allocations is not known, and the OECD-DAC peer reviews manage the total expenditures shown in official statistics. However, once again the 2012 OECD-DAC peer review observed that aid was not concentrated in key partner countries (p. 47-48). Assuming that the portfolio in our meta-evaluation is representative, we can conclude that this practice (i.e. fragmentation) has not changed¹.

¹ This assumption has been discussed with EVA-11 and is included in the risk section of the report. Without knowing what the “coverage” is of MFA appraisals and evaluations, the meta-evaluation cannot determine the extent to which the portfolio is “representative”.

The way (modality) the projects represented in this meta-evaluation are implemented is presented below. Appraisals have been left out because it is not known if the proposed projects were effectively initiated after appraisal. Almost six projects out of ten (56%) are multilateral projects implemented by international organisations, development banks or other multilateral agencies, and 28% only are implemented in the “traditional” Finnish way through consultancy firms offering TA services.

Figure 6: Modality of implementation of projects in the portfolio (excluding appraisals) (n=25)



Source: Meta-evaluation team

This information is closely connected to the above-mentioned fragmentation of Finnish development aid budgets, which becomes clearly visible with some additional calculations. Out of the nine projects the budget of which is 10 MEUR or over (including appraisals, n=36), six are bilateral, implemented by Finnish TA, and three are multilateral - where Finnish TA also may be involved but is not the main implementation vehicle. Further, out of the fourteen multilateral projects (excluding appraisals, n=25), ten have benefitted from funding decisions of a maximum 5 MEUR (c. 2/3 of them). There seems to be a correlation [bilateral implementation \approx -> large funding decisions], and on the other hand [multilateral implementation \approx -> small budget allocations]. Unfortunately this cannot be calculated in exact correlations because qualitative denominations cannot be crossed with quantitative elements. It would be a useful avenue of research for MFA in the future, however.

3.2 Portfolio of reports, with an answer to EQ 5 concerning the use of Management Reviews instead of MTE or evaluation

After a thorough analysis of the documents that were retained for this meta-evaluation, the meta-evaluation did not find a single reference to why a choice was made to call a report a management review instead of an MTE. In fact, this EQ was not possible to answer because management reviews were taken out of the portfolio in common agreement with EVA-11. The Meta-evaluation team defined the difference between evaluation (ex-ante, ex-post or other) and management review in the following way: evaluations assess in the first place a project's achievements, while management reviews study and analyse the way a project works, not its progress and outcomes.

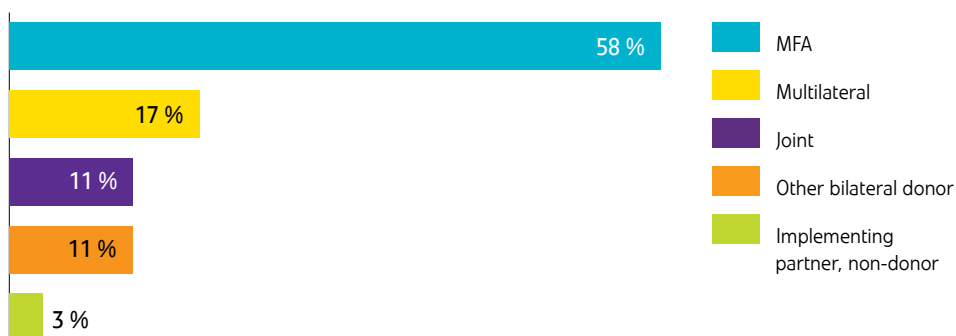
Almost six projects out of ten (56%) analysed are multilateral projects implemented by international organisations, development banks or other multilateral agencies, and only 28% are implemented in the “traditional” Finnish way through consultancy firms and TA services.

Including appraisals, 58 percent of the reports were commissioned by the MFA; the rest by other agencies such as bilateral donors and multilateral organisations.

Sizes of evaluation budgets are likely an important factor in the quality of all reports.

The portfolio of *reports* is described in this section. Including appraisals, 58 percent of the reports have been commissioned by the MFA (n=36); the rest by other agencies such as bilateral donors and multilateral organisations. Figure 7 below visualises the distribution.

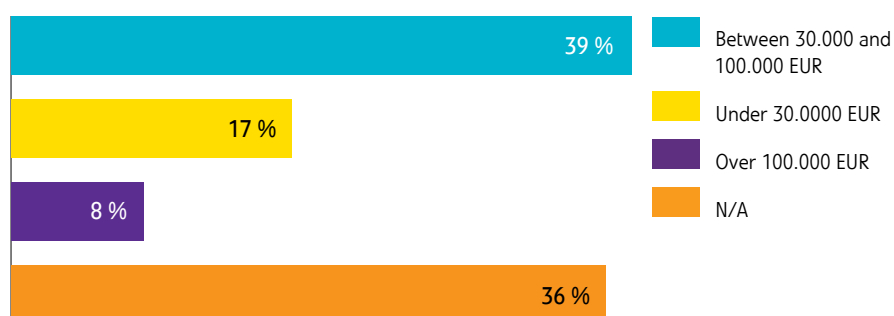
Figure 7: Reports in the portfolio by commissioning agency (n=36)



Source: Meta-evaluation team

Concerning the evaluation budgets, the largest group of reports are those where the budget was between 30,000 EUR and 99,999 EUR (14 reports), and the second largest those with no information concerning the evaluation budget (13 reports). Six reports (17%) were the result of very small evaluation budgets; mainly carried out by individual consultants. The Meta-evaluation team tried to establish a correlation between the evaluation budget and the quality of report as scored by the assessment tool. However, no strong correlation was found, although there was a tendency to have lower scores for reports with small evaluation budgets. The Team cannot prove causality; that is, it cannot say if the small budget causes the report to have lower scores, or if lower scores are shown for smaller evaluation budgets because appraisals have a high frequency of both low evaluation budgets and low scores. It does, however, hypothesize that the evaluation budget is likely to be a factor in the quality of reports.

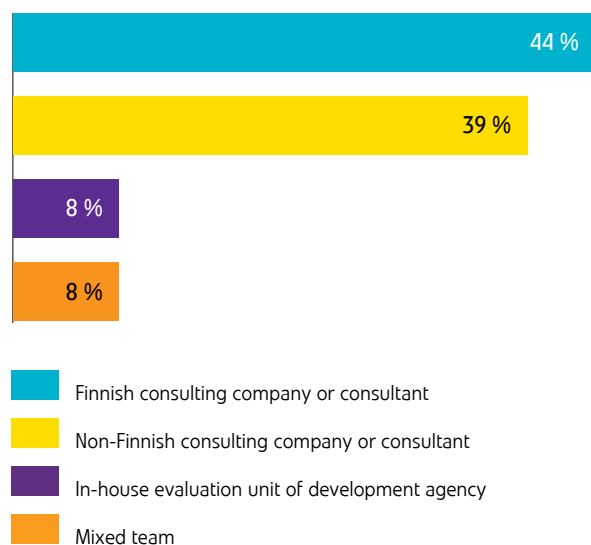
Figure 8: Distribution of reports by evaluation budget (n=36)



Source: Meta-evaluation team

The evaluations and appraisals in the portfolio have been conducted in 44% of the cases by Finnish consultancy companies or Finnish individual consultants. The non-Finnish consulting companies represent almost four out of ten (39%) of the reports. In three cases the report have been conducted and commissioned by international organisations through their in-house evaluation units, and three reports were written by mixed team, with one Finnish consultant joining evaluators of other nationalities.

Figure 9: Evaluation conducted by (n=36)

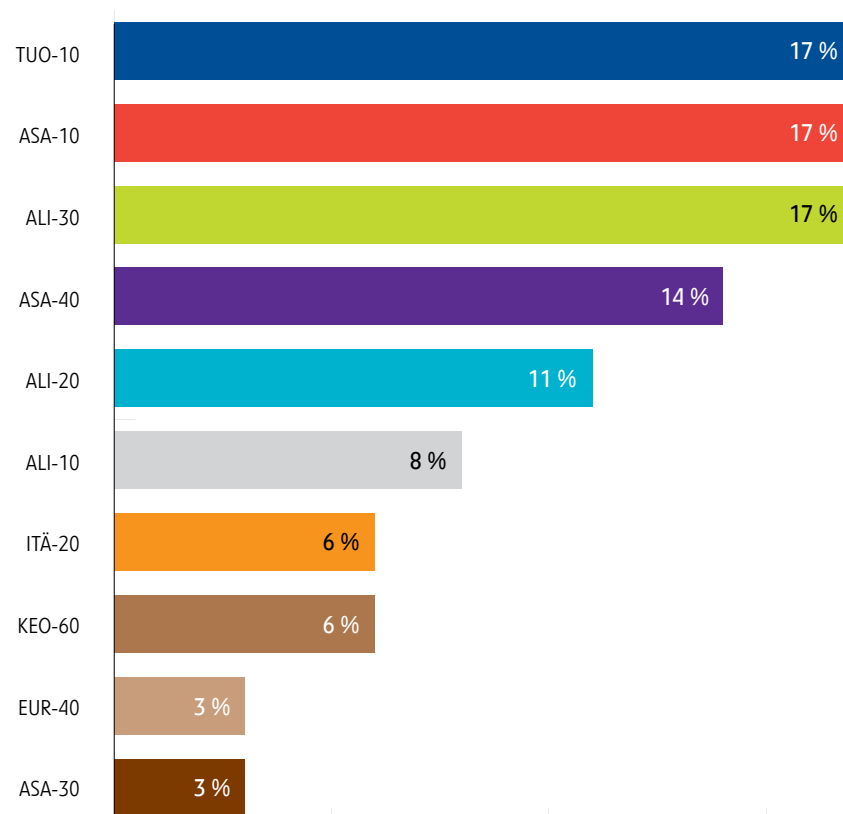


Source: Meta-evaluation team

Concerning the MFA unit which delivered reports for meta-evaluation, the distribution shows below in *Figure 10*. Three Units (External economic relations, Unit of Trade policy (TUO-10), Eastern Asia and Oceania (ASA-10) and the Unit of Southern Africa (ALI-30) delivered seven reports each (17%), while the Unit of Latin America and the Caribbean (ASA-30) and the Unit for South-Eastern Europe (EUR-40) delivered one report. In the following subchapter below the quality of evaluation reports is assessed according to MFA unit. All reports delivered for meta-evaluation by TUO-10 were multilateral evaluations commissioned and paid for by other donors than Finland.

The evaluations and appraisals in the portfolio have been conducted by Finnish consultants in 44% of the cases.

Figure 10: Evaluation reports and appraisals (n=36) according to MFA unit



Source: Meta-evaluation team

4 RESPONSE TO EQ 1: ASSESSMENT OF THE QUALITY OF EVALUATION- RELATED DOCUMENTS

4.1 Overview

This section deals with (n=26) evaluation reports and n=23 TOR. The conclusions of the analytical narrative in this chapter may have a limitation: because the assessment grids of TOR and evaluation reports followed the standards and norms of Finland's manuals and guidelines, the scores given to deliverables commissioned by other than MFA may suffer from a bias, as naturally they are not bound by the instructions of the development evaluation function of Finland.

The table below is a complete rendering of the ratings given to the various **evaluation reports** studied in this meta-evaluation.

Table 1: Ratings given to evaluation reports for the entire n=26 sample of documents studied in Phase One

	Preliminaries	Main text	1. Introduction	2. Context	3. Description of programme	4. Approach, methodology and limitations	5. Evidence-based findings	6. Answers to Evaluation Questions	7. Conclusions	8. Recommendations	9. Lessons learned	Annexes	Non-content issues	Total
Zambia small scale irrigation	7	53	1	2	2	7	12	8	6	10	5	3	3	66
COWASH Ethiopia MTE	6	55	1	2	4	10	8	8	10	12	0	4	3	68
Mékong core environment pg	5	52	1	4	4	6	8	8	10	12	0	2	1	60
Marie Stopes Afghanistan	6	70	1	1	2	15	14	9	12	12	4	5	2	83
Afghanistan and neighbours, drug prevention	5	67	1	3	5	14	10	6	10	13	5	5	3	80
Afghanistan SNGP II	6	65	1	4	4	12	13	9	12	10	0	4	3	78
IGAD	5	27	0	1	2	3	5	4	5	7	0	1	2	35
AU mediation capacity	4	53		4	4	10	8	5	12	10	0	3	3	63
Laos EMSP eval	7	52	1	3	4	7	9	6	7	12	3	2	3	64
Andean BioCAN evaluation	6	65	1	4	5	10	9	9	12	10	5	0	3	74
UNCTAD academies	4	65	1	3	4	12	12	10	12	8	3	3	3	78
TEVT Nepal Soft Skills MTE	6	56	1	4	3	8	12	6	10	10	2	3	1	66
North Africa & Middle East Trust Fund MTR	3	49	1	3	3	5	7	8	8	10	5	2	1	55
Land Administration, Palestinian Territories	7	49	0	2	5	6	6	6	6	13	5	2	3	60
Land administration (REILA), Ethiopie	6	75	1	3	5	14	15	10	14	13	0	4	3	88
PALWECO (Kenya)	7	49	1	2	2	6	12	5	9	7	5	2	3	61
Nepal Forestry Programme	6	80	0	3	5	14	15	9	15	14	5	4	3	93
AFT Kosovo	3	33	0	0	0	5	6	0	5	12	5	1	1	38
Advisory Centre on WTO Law	5	42	1	2	3	7	10	5	5	10	0	2	3	52
STIFIMO (Mozambique)	5	25	0	2	2	0	4	4	4	6	3	1	0	31
OCSE Framework Agreement	3	38	0	2	2	7	7	7	4	6	3	5	0	46
Partnership for Market Readiness	4	62	1	3	4	13	12	7	10	8	4	5	3	74
Enhanced Integrated Framework	3	43	1	3	3	8	7	5	5	10	2	5	3	54
International Trade Centre (1)	7	70	1	3	5	15	15	10	2	14	5	3	3	83
International Trade Centre (2)	6	43	1	1	2	8	5	5	8	13	0	5	1	55
UNIDO's Trade Capacity	4	64	1	2	4	14	9	10	10	12	2	3	2	73
Average	5,2	53,8	0,8	2,5	3,3	9,1	9,6	6,9	8,6	10,5	2,7	3,0	2,3	64,4
Max	7,0	85,0	1,0	4,0	5,0	15,0	15,0	10,0	15,0	15,0	5,0	5,0	5,0	100,0
Score on 100	74,7	63,3	84,0	62,0	66,5	60,4	64,1	68,8	57,2	70,0	54,2	60,4	45,4	64,4

Source: Meta-evaluation team

Table 2: Ratings given to evaluation ToR for the entire n=26 sample of documents studied in Phase One

	1. Background information	2. Rationale, purpose and objectives	3. Scope of the evaluation	4. Relevant and specific evaluation questions	5. Aid effectiveness commitments	6. Methodology	7. Evaluation process and management structure	8. Resources	9. Annexes and structure of the ToR	Total
Zambia small scale irrigation	2	12	1	16	1	2	3	18	2	57
COWASH Ethiopia MTE	3	15	1	30	1	1	3	19	2	75
Palestine Land Administration	2	8	2	22	0	2	4	8	2	50
Marie Stopes Afghanistan	3	8	2	25	0	2	2	16	2	60
Afghanistan and neighbours, drug prevention	4	15	2	30	1	4	5	16	1	78
Afghanistan SNGP II	4	14	3	32	3	2	5	19	2	84
IGAD	2	5	1	15	0	1	2	15	1	42
AU mediation capacity	3	12	2	29	0	2	5	16	2	71
Laos EMSP eval	4	13	2	27	0	2,5	3	22	2	75,5
Andean BioCAN evaluation	5	10	2	30	1	3	2	20	2	75
UNCTAD academies	3	15	3	30	0	4	5	20	2	82
TEVT Nepal Soft Skills MTE	3	15	1	27	0	1	3	10	1	61
North Africa & Middle East Trust Fund MTR	2,5	8	1,5	5	0	2	0	4	1	24
Land administration (REILA), Ethiopie	4	12	2	28	0	0	3	13	0	62
PALWECO (Kenya)	3	14	2	30	0	4,5	3	18	2	77
Nepal Forestry Programme	5	15	3	33	0	5	4	21	2	88
AFT Kosovo	4	15	2	30	0	3	3	17	1	75
Advisory Centre on WTO Law	2	10	2	18	0	2	3	8	1	46
OCSE Framework Agreement	3	12	2	17,5	0	3,5	2,5	5	2	47,5
Partnership for Market Readiness	2	15	2	28	0	5	5	20	2	79
Enhanced Integrated Framework	4	10	2	28	0	4	5	10	2	65
International Trade Centre (2)	2	11	2	16	0	2	4	0	1	40
Average	3,2	12,0	1,9	24,8	0,3	2,6	3,0	14,4	1,6	64,3
Max	5	15	3	35	5	5	5	25	2	100
On 100	63,2	80,0	64,4	71,0	6,4	52,3	60,4	57,6	79,5	64,3

Source: Meta-evaluation team

With respect to evaluation reports, there are serious deficiencies with the CORE requirements of a) rigorous approach and methodology, b) identifying evidence-based findings, c) providing answers to evaluation questions and d) providing legitimate conclusions.

With respect to Terms of Reference, Table 2 presents the same type of information in the same way as for the report table just above. There is a smaller number of projects in Table 2 than in Table 1 because some projects did not have any ToR associated with them.

In support of the above observations, the following table indicates the distribution of ratings that were given to the various elements of the Quality Assessment Grids for Evaluation Reports. It should be remembered that it requires as rating of “4” or better to meet standards and expected quality norms. **Interpreting the following table, it is very evident that there are serious deficiencies with the CORE categories of 4, 5, 6, 7 and 8. This is expanded upon in section 4.2.**

Table 3: Distribution of ratings for the quality assessment grid – evaluation reports

Main category	How many reports received a rating of:					In how many reports this information was missing	Total no. of assessed reports
	1	2	3	4	5		
Table of Contents and Acronyms	0	0	0	25	0	1	26
Executive Summary	0	5	9	10	1	1	26
1. Introduction	2	1	8	11	0	4	26
2. Context	2	6	6	10	0	2	26
3. Description of programme or project	1	2	12	9	2	0	26
4. Approach, methodology and limitations	4	6	8	5	3	0	26
5. Evidence-based findings	0	5	11	6	4	0	26
6. Answers to Evaluation Questions	1	4	10	5	5	1	26
7. Conclusions	1	9	10	5	1	0	26
8. Recommendations	0	4	11	10	1	0	26
9. Lessons learned	1	2	7	5	1	10	26
Annexes	4	6	6	4	6	0	26
Non-content issues	2	3	4	12	4	1	26

Source: Meta-evaluation team

4.2 Answering EQ 6: Analysis of the Quality of Evaluation TOR and Reports – Major Issues Identified

EQ 6 is actually dealt with in two separate parts of his report. This chapter presents findings developed as a result of the analyses of the quality of evaluation reports and their corresponding TOR. Another section (7.1) deals with the same question by focussing on Conclusions, rather than findings.

Concerning the quality of ToR in our portfolio, the average score for all ToR (n=23) is 64.3. When considering only the ToR drafted by MFA, thus commissioned by Finland, the average score of TOR is practically identical with the overall ToR score. For ToR drafted by some other commissioning agency (other

bilateral donor, international organisation or bank etc.), the average score of ToR is 63.9. In fact, all three average ToR scores are within 0.9 points out of 100 from each other. It can therefore be concluded that the quality of ToR written by the implementing units and departments of MFA is more or less at the level of their international peers, as far as we can assume that the meta-evaluation portfolio is representative.

Unfortunately the findings do not suggest as comforting a conclusion for the quality of evaluation reports (MTE and evaluations). The average score for reports (n=26) is 64.4, which is almost the same as the average ToR score (64.3). This suggests some kind of correlation between the overall quality of ToR and reports; an issue to which we will return later. As such, on the level of averages, the reports in this portfolio are of a minimum acceptable, although not totally satisfactory quality. However, when split into two categories according to the commissioning agency, the picture changes. The reports commissioned by the regional and thematic implementing units of MFA score, on the average, 56.95, while the reports commissioned by some other agency score 69.09. This is about thirteen points of difference with other agencies' reports and almost five points of difference with the average score for ToR drafted by MFA. (Were appraisals included, the average for Finnish reports would be 51.6; fifteen points below the average non-Finnish report score.) This difference between Finnish commissioned and those managed by other donors, and between the quality of ToR produced by MFA and the reports approved, suggests serious deficiencies in evaluation management in implementing units' evaluations: reports scoring over 20 points below the ToR score were approved.

There is another way of looking at the issue of scores. When comparing the highest and the lowest scores of ToR and reports commissioned by Finland and those commissioned by other agencies, we come to the following figures:

Table 4: Lowest and highest scores for ToR and reports according to commissioning agency

Commissioned by	Two lowest		Two highest		Average
	Lowest ToR score	Second lowest ToR score	Second highest ToR	Highest ToR	
Finland MFA	42	47.5	75.5	77	64.6
Non-Finnish agency	24	40	84	88	63.9
	Lowest report score	Second lowest report score	Second highest report score	Highest report score	
Finland MFA	30.5	35	74	87.5	57
Non-Finnish agency	51.5	54	83 (2 reports)	92	69

Source: Meta-evaluation team

A specific sub-question that is part of EQ 1 concerns the difference between the quality of MFA-commissioned evaluations and those of MFA's partners. The ToR commissioned by other agencies than MFA have a larger scatter of points

The quality of ToR prepared by MFA is more or less at the level of its international peers.

Reports commissioned by units of MFA score an (quite low) average of 56.95, while reports commissioned by some other agency score an average of 69.09.

between poor and good, 64 points, while in the case of Finland the scatter does not exceed 47 points. Overall, the scores of non-Finnish ToR suffer from the fact that in that category there was one ToR particularly poorly assessed which lowers the average significantly.

Unfortunately it is not possible to make reliable average scores by type of commissioning agency; in most cases the number of individual reports per agency other than MFA-Finland is too low. But this topic can also be approached by listing the highest 10 reports by commissioning agency:

Table 5: Top-ten evaluation reports by commissioning agency (out of 28 reports; n=28)

No.	Score of report	Commissioning agency; type	Carried out by; type
1.	92	Bilateral non-Finnish donor	Non-Finnish consultancy company
2.	87.5	MFA	Finnish consultancy company
3.	83	Implementing agency/ INGO	Non-Finnish consultancy company
4.	83	Multilateral organisation/ UN system	In-house evaluation
5.	80	Multilateral organisation/ UN system	In-house evaluation with collaboration with consultants
6.	78	Bilateral non-Finnish donor	Non-Finnish consultancy company
7.	78	Multilateral organisation/ UN system	Non-Finnish think tank/ individual consultant
8.	74	MFA	Finnish consultancy company
9.	74	International financial institution / multilateral	Non-Finnish consultancy company
10.	73	Bilateral non-Finnish donor	Non-Finnish consultancy company

Source: Meta-evaluation team

Two reports commissioned by MFA are on the top-ten list, while three non-Finnish bilateral donors get to this list, too; however no conclusions should be drawn based on this fact as only a very limited sample of other donor's reports make their way to a meta-evaluation commissioned by Finland. An important category of agencies on this top-10 list is multilateral organisations, which should be no surprise as they are the ones who set the standards in evaluation.

Due to the fact that only one Finnish consultancy company had conducted more than one evaluation or MTE in this portfolio, it is not possible to calculate averages across consulting or executing "organisations". Instead, the Meta-evaluation team listed the reports delivered by consultancy companies according to the score of the report in Table 6. The two evaluation reports with the lowest scores were conducted by individual consultants with low evaluation budgets, again suggesting that there may be a correlation between available resources and the quality of the reports. The highest score belongs to a report written by an evaluation budget of 90,000 EUR.

Table 6: Reports with scores written by Finnish consultancy companies according to evaluation budget and type

Score report	Evaluation budget	Type of Evaluation
87.5	90,000 €	Mid-term evaluation
74	180,000 €	Evaluation
68	115,000 €	Mid-term evaluation
66	74,000 €	Mid-term evaluation
64	90,000 €	Evaluation
61	80,000 €	Evaluation
45.5	50,638 €	Evaluation
38	Not known (assumed under 30,000 €)	Mid-term evaluation
35	20,000 €	Evaluation

Source: Meta-evaluation team

The average scores of reports according to MFA units are presented below:

Table 7: Average ToR and report scores according to MFA unit (n=26)

Score TOR	Score report	MFA Unit	Number of reports
37	57.3	ALI-10	2
64	62.9	ALI-20	4
64	53.2	ALI-30	3
75.5	62	ASA-10	2
75	74	ASA-30	1
77.5	83.3	ASA-40	4
75	38	EUR-40	1
47.5	45.5	ITÄ-20	1
79	74	KEO-60	1
58.8	65.8	TUO-10	7

Source: Meta-evaluation team

An important factor to consider is the extent to which the generators of the ToRs provided a quality product, in keeping with the norms and standards of the MFA. The following table indicates the dispersion of ratings given by the meta-evaluation team to the evaluation TORs it studied. (A total of 26 evaluation reports were assessed, but 4 of them were not accompanied by ToR therefore, the number of ToR assessed is 22).

Table 8: Distribution of ratings for the quality assessment grid – evaluation ToR

Main category	How many ToR received a rating of:					In how many ToR this information was missing	Total no. of assessed ToR
	1	2	3	4	5		
1. Sufficient background information to the evaluation/ review provided	0	6	8	6	2	0	22
2. Rationale, purpose and objectives of the evaluation are clearly described	0	1	6	5	10	0	22
3. Appropriate and sufficiently detailed description of the scope of the evaluation	0	3	12	6	1	0	22
4. Evaluation objectives are translated into relevant and specific evaluation questions	1	1	5	6	9	0	22
5. Implementation of aid effectiveness commitments is described	5	0	1	0	0	16	22
6. Proposed methodology is appropriate and capable of addressing the evaluation questions	3	10	2	3	3	1	22
7. Evaluation process and management structure are adequately described	1	3	9	3	6	0	22
8. Resources required for this evaluation are sufficiently described	3	4	6	8	1	0	22
9. Annexes and structure of the TOR	1	1	5	10	5	0	22

Source: Meta-evaluation team

Those categories where ratings were high (the meta-evaluation team considered that a rating of 4 or 5 would be needed to indicate a “good or very good” rating, and that a combined total of 17 (out of 22) would be required for the overall performance of the MFA (TOR writers) to have been considered as good (i.e. 75% of 22). No categories met this requirement. The highest was 65% (i.e. 15 out of 22). When these are the corporate instructions to individuals and firms

for the execution of the strategically-important evaluation function, this overall performance is poor. This, combined with the logical (and obviously simplistic) observation that it is not possible to prepare good evaluations from poor instructions, provides a possible roadmap for future internal capability development.

Of note is the observation that 15 out of 22 ToRs scored 3 or less (i.e. poorly) for “methodology” proposals.

The table above also provides an indication of the extent to which the standards of the MFA are adhered to with respect to inclusiveness (i.e. are all the parts that are supposed to be there actually included?). Almost all categories scored very highly in this regard with the exception of “aid effectiveness” where 16 ToRs did not even address the category at all.

Contrary to the previous meta-evaluation which did not find a significant correlation between the quality of ToR and the quality of reports, the current meta-evaluation found a statistically significant, although not robust, correlation (0.58) between those two. In other words, if the ToRs are of good quality, there is a 58% probability that the report, too, is good. On the other hand the table above indicates that the correlation is far from total: there are cases of very good ToR and poor reports, and on the other hand, reports that score significantly higher than their ToR.

Considering that evaluation TORs are the corporate instructions to individuals and firms for the execution of the strategically-important evaluation function, the overall performance is poor.

Sixteen ToRs did not address “aid effectiveness” at all.

The meta-evaluation found that appraisal ToR's were not specific in their direction and were weak in presenting clearly defined issues specific to the PD. Resources allocated to appraisals were often inadequate to undertake an evidence-based and triangulated study that would have critically examined all the issues in the ToRs.

5 ANSWERING EQ 7: ASSESSMENT OF THE QUALITY OF APPRAISAL-RELATED DOCUMENTS

5.1 Overview

Great care was taken during the inception phase to fully understand what should be the standards and norms against which to assess appraisal ToR/ITT and Reports. The concept paper prepared in early 2015 on ex-ante evaluation within MFA pointed to the fact that the present use and scope of MFA appraisals are relatively constrained when compared to the practices of other donors. GoF MFA uses appraisals to examine and “critique” whatever documentation is proposed for use as a Programme Document (PD) before it is approved; the appraiser is not required to develop another version of the PD but must offer suggestions and recommendations that would allow the existing draft PD to meet the standards and norms that such documents must meet. Nevertheless, the appraisal must be based on evidence and must provide “answers” (or a “learned judgement”) to a set of issues that are pre-defined by the client. The findings, specifically, must be evidence-based and must be the cornerstone for conclusions and recommendations.

The meta-evaluation found that appraisal ToR's were not specific in their direction (for example, indicating that the priority for an appraisal is the “feasibility” of the intervention is a tautology and does not serve to focus or scope the mandate). It also found that the ToR were weak in developing clearly defined issues and regularly identified a large number of issues to study without (as requested by MFA standards) clustering them around pre-defined analysis criteria. By comparing the requirements (mainly EQ) from the ToRs for any particular appraisal to the budgets allocated for the appraisal, the Team was able to identify whether resources allocated might be appropriate. By first deducting the approximate costs of travel, allowances, and other “reimbursables”, the Team was left with an approximation of the number of days that would be allocated to international and local consultants. By eliminating the effort required for travel, in-bound and out-bound briefings, report writing and adjusting reports, the remainder was an approximation of the effort that could be devoted to research. In fact, the team found that many international researches only had 7-8 field days to manage the appraisal, even though their reports spoke of major issues, gaps and weaknesses. The Team has therefore found that resources allocated to appraisals were quite often totally inadequate to undertake an evidence-based and triangulated study that would have critically examined

all the issues presented in the ToR. The ToR's therefore clearly seemed to be based on the assumption that the available documentation (including the all-important draft PD) was on the right track (in terms of project justification and design) and that resources were not required for the appraiser to undertake critique, grounded investigation, independent observation of assumptions presented, the development of alternative scenarios or the analysis of the possible effects of the proposed investments. The absence of any of these types of requirements was evident in all the ToR's.

The ratings showed that appraisal reports generally were of poor quality and may reflect, as it were, the ToR that generated and guided them. The team also hypothesises that the inability of officials to quality control the appraisal reports due to a possible lack of detailed knowledge concerning the intervention may have been a defining factor in the poor quality of the appraisals. There are many reasons why this absence of detailed knowledge could be a factor, but MFA's managerial requirements for basic information and analysis required before appraisals can be initiated, and a project cycle description that is clear concerning what constitutes a Draft PD would help frame the timing and usefulness of appraisals so that the implementation of interventions does not suffer from errors and oversights during the pre-approval stages of projects (as has been reported upon in evaluation reports). The team found that the reports themselves were often written in a "sector" perspective and were not preoccupied with the broader policy and contextual issues that are real and important parts of PDs, such as the integration of HRBA, the analysis of coherency and aid effectiveness or the focus on intervention impact as a factor of the Country Strategy. Much of the analysis was superficial and not supported by evidence (or at least the evidence was not presented to support the arguments and findings).

5.2 Quality of ToR/ITT

When all parts of the assessment grid for appraisal ToR/ITT are taken into account, the average number of points given is 64.78 out of a possible 100. Given that a) the authors of these documents are internal to MFA and that the documents themselves are subject to QA by MFA supervisors and b) some of the documents may have been subjected to review by development partners, recipient organisations or implementing agencies, the average score is very low.

The range of scores for the ToR/ITT shows that 5 out of the 10 documents had scores that were between 50 and 69 points, while 6 had scores between 60 and 79. One report had a very low score and the highest was graded at only 80.5 points. No document was graded at over 81 points which indicates that top-quality ToR documents are not being prepared.

Areas where the documents were particularly weak include:

a) translating appraisal objectives (within the specific contexts of the proposed intervention) into relevant and specific issues to be appraised (half of the 10 ToR were scored at 20 points or less out of a possible 35); b) The integration of aid effectiveness commitments into the appraisal; c) Providing some minimum amount of direction and guidance on the most appropriate approach and

Ratings show that appraisal reports generally were of poor quality.

Much of the analysis found in appraisal reports was superficial and not supported by evidence.

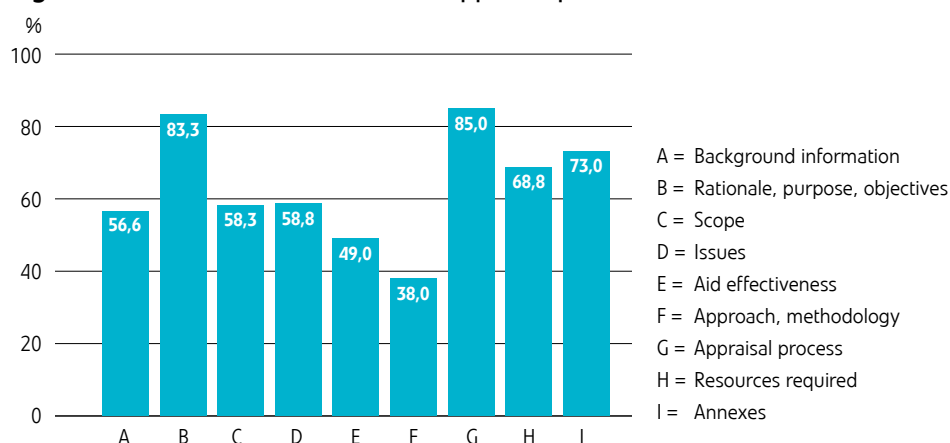
Given that the authors of appraisal ToRs are internal to MFA and that the documents themselves are subject to QA by MFA supervisors, the average score (64.8) is very low.

No areas of appraisal ToRs were well rated except for the description of the appraisal management process by MFA, which was mostly cut and pasted from MFA documents.

methodology given the author's privileged knowledge of context and available information sources, and d) Insufficient levels of resources and inappropriate/unclear match of mandate and expertise.

There were no areas where the ToRs were generally well rated (over 80% of possible points on average) except for the description of the appraisal management process by MFA.

Figure 11: Assessment of the TOR for appraisal per area



Source: Meta-evaluation team

The ToR Assessment -% score chart is a visual depiction of the analysis in this section. It shows the score per “category”. It is particularly interesting to consider the scores for what could be called the “core” of the appraisal: i.e. issues; aid effectiveness; approach (all of the foregoing rated less than 60%) and resources (rated at 70%). In the view of the meta-evaluation team, these scores warrant serious consideration by the MFA.

The following table provides an overview of the assessments given within the various categories that were identified for the assessment of the appraisal reports.

Table 9: Distribution of ratings for the quality assessment grid – appraisal ToR

Main category	How many ToR received a rating of:					In how many ToR this information was missing	Total no. of assessed ToR
	1	2	3	4	5		
1. Sufficient background information to the appraisal provided	0	4	3	2	1	0	10
2. Rationale, purpose and objectives of the appraisal are clearly described	0	0	3	2	5	0	10
3. Appropriate and sufficiently detailed description of the scope of the appraisal	1	1	5	2	1	0	10
4. Appraisal objectives are translated into relevant and specific appraisal issues	1	2	3	3	1	0	10
5. Implementation of aid effectiveness commitments is described	2	3	4	1	0	0	10
6. Proposed methodology is appropriate and capable of addressing the appraisal questions	4	4	1	1	0	0	10
7. Appraisal process and management structure are adequately described	1	0	1	4	4	0	10
8. Resources required for this evaluation are sufficiently described	0	1	4	4	1	0	10
9. Annexes and structure of the TOR	0	1	3	6	0	0	10

Source: Meta-evaluation team

The following are noted from an analysis of this table:

- All required parts of the ToR were present for all categories.
- Three important categories were not well developed: Methodology (8 out of 10 ToR were unacceptable); aid effectiveness (5 out of 10 were unacceptable), and background (4 out of 10 were unacceptable).
- The following categories scored relatively highly: Rationale and purpose; appraisal process, and resources required

5.3 Quality of Appraisal Reports

The meta-evaluation found that appraisal reports were of poor quality overall with an average scoring of only 46.35 points out of a possible 100 points. Particular weaknesses were shown in the following areas, with an indication of the extent to which average scores reflect the maximum scores available.

Appraisal reports were of rather poor quality overall with an average scoring of only 46.35. Particular weaknesses were in: a) descriptions of methodology, b) evidence-based findings, c) legitimate conclusions, d) lessons-learned and e) risks.

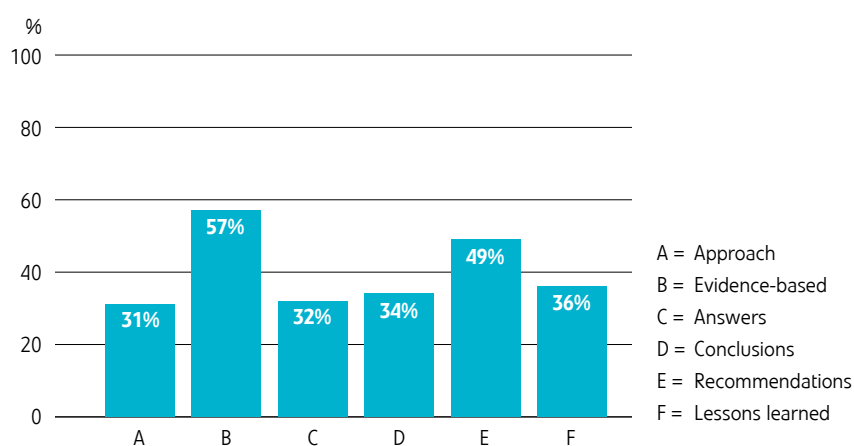
Table 10: Weaknesses in appraisal reports

Category of assessment	Percentage of maximum actually realised
Providing the reader (client) with a description of an analytical approach and methodology that would reflect the needs of MFA and the contexts of the eventual intervention. This section also should have shown how the data collection and analysis methods were appropriate and, in so doing, it should have assessed the validity and reliability of the data and the analysis performed upon it.	31%
Presenting evidence-based findings: the section should have presented empirical data, facts and other evidence relevant to the indicators that were used to analyse the appraisal issues. Findings were to be clustered by OECD/DAC evaluation criteria and those of the MFA, and should have supported (or not) the proposed intervention logic. Finally, the findings should have included evidence related to the MFA's policies on HRBA and cross-cutting objectives	57%
The "answers" section in the report should have provided strategic analysis of the issues (provided in the ToR). They should have been demonstrably based on findings	32%
The "conclusions" section should have contained the assessment of the likely and probable performance of the project/programme based on the findings in relation to the set evaluation criteria, performance standards or policy issues.	34%
The "recommendations" section should have included proposed improvements, changes, and actions to improve the project design or to capitalise on strengths identified. Recommendations must be based on the findings and conclusions. There should be a clear indication of: <ul style="list-style-type: none"> • to whom is the recommendation directed • who is responsible for implementing the recommendation, and • when the recommendation should be implemented 	49%
The "lessons learned" section should have contained any general conclusions that are likely to have the potential for wider application and use across the programmes or across MFA generally	36%

Source: Meta-evaluation team

The gap between the percentage reached and the maximum can best be appreciated with a visualisation that shows the gap between the maximum (represented below by a line at the 100% level) and the actual. The Figure 12 shows these gaps and a detailed analysis of the completed assessment grids offers clues as to the possible reasons for the gaps. For example, the "evidence-based" category reaches almost 60% largely because of the dominance of sector-related descriptions. It should be recognised that the gap between each of the characteristics and the 100% baseline is very large and should be of concern, especially since the quality of appraisal reports has a direct effect on the quality of Programme Documents and therefore affects the extent to which Finnish development cooperation policy is achieved.

Figure 12: Gap of appraisal reports against the maximum score per assessment area



Source: Meta-evaluation team

The following table provides an overview of the assessments given within the various categories that were the core of the assessment of the appraisal reports.

Table 11: Distribution of ratings for the quality assessment grid – appraisal reports

Main category	How many reports received a rating of:					In how many reports this information was missing	Total no. of assessed reports
	1	2	3	4	5		
Table of Contents and Acronyms	0	0	0	0	10	0	10
Executive Summary	1	1	5	0	2	1	10
1. Introduction	2	0	3	3	0	2	10
2. Context	0	0	3	2	4	1	10
3. Description of programme or intervention being appraised	0	2	3	2	2	1	10
4. Approach, methodology and limitations	3	4	2	0	0	1	10
5. Evidence-based findings	1	2	6	0	1	0	10
6. Answers or strategic analysis of issues based on findings	2	0	5	0	0	3	10
7. Conclusions	3	0	2	0	2	3	10
8. Recommendations	2	1	4	1	1	1	10
9. Lessons learned	0	4	0	0	2	4	10
Annexes	0	0	0	2	6	2	10
Non-content issues	0	4	2	2	1	1	10

Source: Meta-evaluation team

A significant number of appraisal reports did not contain important sections required by the MFA (or if they did, the content was judged to have been superficial).

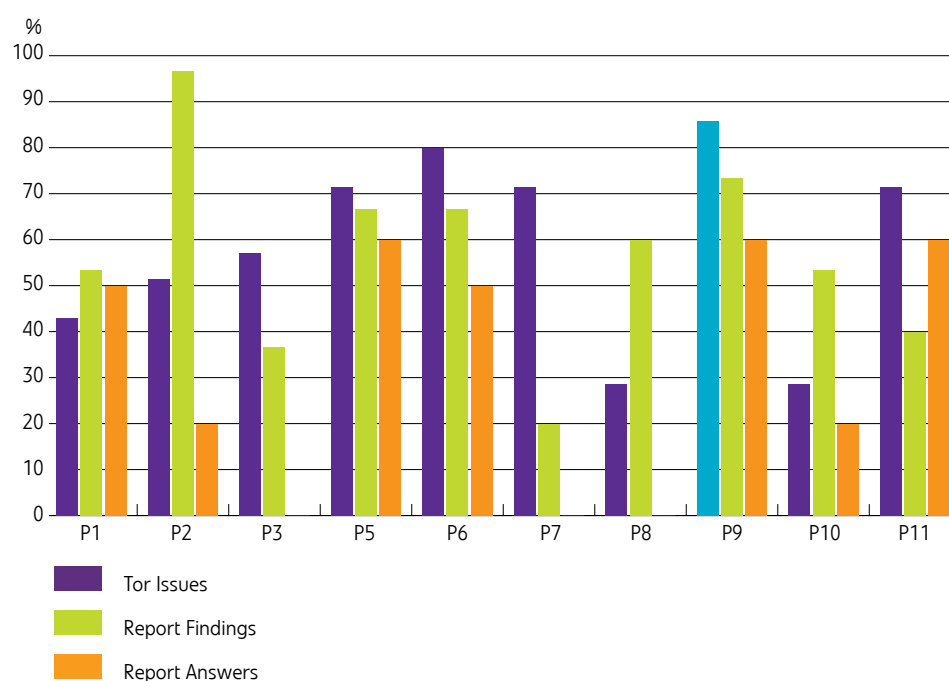
It should be noted that in order to generate this table, a transformation algorithm had to be developed in order to base the values on the same 5-point rating scale as was used in evaluation ToR and evaluation reports tables. Using the same logic as in section 4.2, a combined score of 7.5 or better would have to be obtained in the ratings 4 and 5 in order for the quality of the appraisal reports to be good or better. This score is only obtained with the following categories (combined scores in brackets): Table of Content (10) and Annexes (8). This lends quantitative support to the overall findings above.

The table draws attention to the finding that a significant number of reports did not include important sections required by the MFA (or if they did the content was judged to have been superficial). The following are of note, with a few summary observations concerning the distribution:

- Answers or strategic analysis of issues based on findings (3 missing out of 10). It should be noted here that five out of 10 appraisals only scored a barely passable “3” and two of the 10 scored only a “1”. These observations are noteworthy because these are the “Core” of an appraisal report.
- Conclusions (3 missing out of 10). Three out of 10 scored only a “1” while 2 out of 10 scored only “3”.
- Lessons learned (4 missing out of 10). Four out of 10 scored a very low “2”

5.4 Correlation of results between ToR and Reports

Figure 13: Relation between ToR issues, findings, and answers



Source: Meta-evaluation team

In examining the scores given the various sections of the documents for 10 projects studied in phase I of the evaluation process, interesting correlations (or lack thereof) stand out. There is no strong correlation between the individual scores given the Approach and Methodology (A&M) section in the ToRs and the A&M section of the Reports. On average, both scored poorly; however, that section in the higher scoring ToRs did not necessarily result in good quality A&Ms sections in the reports (or vice versa). Moreover, relatively good report A&M sections were produced in a few reports, even when the related ToRs scored very poorly there. That would imply that the impetus for producing good A&M sections was in the hands of the report authors. It should be noted that the ToR (or any other document) do not spell out what kind of A&M section is required.

Also compared were the scores given to the Issues section of the ToRs and two sections of the Reports related to ‘issues’- the Findings and Answers sections. These are the “core” sections of the appraisal related documents. The Answers section scored very poorly throughout the reports with an average of only 32%, and only once scoring higher than the Issues section (which is also poor at an average of 58.86%). Answers also only surpassed the “Findings” section once. That being said, there is no strong statistical correlation between the quality of the answers and the findings and issues.

Table 12 below shows the correlation coefficient between the elements in the above diagramme (Figure 13):

Table 12: Correlation between ToR issues, report findings, and report answers


Relationship Between	Correlation Coefficient
TOR Issues and Report Findings	-0.024
TOR Issues and Report Answers	0.534
Report Findings and Report Answers	0.325

Source: Meta-evaluation team

5.5 Hypotheses concerning reasons for the results obtained

The meta-evaluation hypothesises that there are essentially three reasons why the appraisal reports have such poor overall ratings:

- The standards and norms of the MFA concerning the content and structure of reports may not be well understood, especially the internal logic of the reports and the link between appraisals and the entire policy framework concerning evaluation. Since appraisals are considered to be ex-ante evaluations, they fall under the evaluation manual’s guidance, for example. This meta-evaluation applied the letter and intent of both the Bilateral and Evaluation manuals, so authors who were not as familiar with their requirements for content will not have been judged to have provided “quality” work.

- 
- b) The direction provided by the corresponding Terms of Reference is weak. As shown above, the average score for the ToRs assessed is only 64%, but the range of points is significant in the sense that it shows a wide range (from 42.5 to 80.5, with a median of 66). Overall, the performance of MFA officials and their supervisors that have prepared these TORs is not encouraging; fully half of the dossiers assessed having a score in the 65-72 range when it would seem logical to expect that those officials responsible for the dossiers should be in a position to accurately prepare ToR based on the MFA Bilateral Manual, the Evaluation manual and other MFA policy documents.
- c) The content of the assessed appraisal reports clearly indicates that the draft PD that were “appraised” were clearly not “Ex-ante evaluation-ready”. Evidence shows that a significant portion did not have an intervention logic on which to judge the appropriateness of any implementation strategy or the sustainability of the intervention; they also had a poor monitoring framework and significant problems with management and oversight. They also poorly addressed HRBA and cross-cutting issues, both of which are of key importance to Finland, a fact that should have been known to the authors of the draft PD (to name only a few issues).

6 ANSWERING EQ 4: ASSESSMENT OF THE QUALITY OF FINNISH DEVELOPMENT COOPERATION

Please note: also refer to section 7.3 for conclusions related to EQ 4

6.1 Findings related to 2014-2015 meta-evaluation with respect to the Quality of Finnish Development Cooperation

The meta-evaluation administered a mixed deductive-inductive methodology to n=18 reports to assess the extent to which evaluation reports (and to a lesser extent appraisals) could provide insights into the quality of Finnish development cooperation. The aspects subjected to deductive logic were rated on a five-point scale so all ratings quoted in this section are based on a maximum of five points (e.g. the lowest score is a “1”). Where a particular topic was not covered, it was so indicated, so care must be exercised when analysing the tables and values in this section because the value may represent an average that has fewer than n=18 sample points; in that case, one needs to ask why the report did not contain the information. As explained in Annex 3, there are a significant number of standards and sub-standards used in the analysis. The following analysis refers to aggregated levels of analysis (at the standard level) and sub-aggregated levels of analysis for the sub-standard or sub-issue level.

The aspects subjected to inductive logic were integrated into the analysis grids so that ideas and reflections that were not picked up through the rated systems could be picked up and later analysed. They are not rated per se as explained in Annex 3 Detailed Methodology, The Analysis Grid can be found as an Annex to this report. Based on the many contributions of the inductive analysis, a sample has been introduced into the following sections. Where the contribution has been essentially “cut and pasted” into this report, it is treated as an “observation”; where the inductive analysis showed that a number of reports repeated the same message (usually where at least five reports had the same input), the meta-evaluation team transformed them into a “finding”. *The latter are indicated by the use of the letter (F) at the end of the statement of finding.*

In the context of the OECD-DAC, reports rated well (4.1) on the extent to which they reflected the MFA policies on relevance. However, ratings on effectiveness (2.6), efficiency (2.6) and sustainability (2.3) were very poor, and impact was rated extremely weak at 1.3.

The structure of the evaluation-based analysis grid was such that there were four main parts:

- a) an analysis of the content of reports on the basis of the five OECD/DAC evaluation criteria,
- b) an analysis of the content of the reports with regard to their treatment of Finnish policy on aid effectiveness , and
- c) an analysis of the way the reports dealt with Finnish policy dealing with HRBA and CCOs
- d) an analysis of the use made of risk analysis within interventions

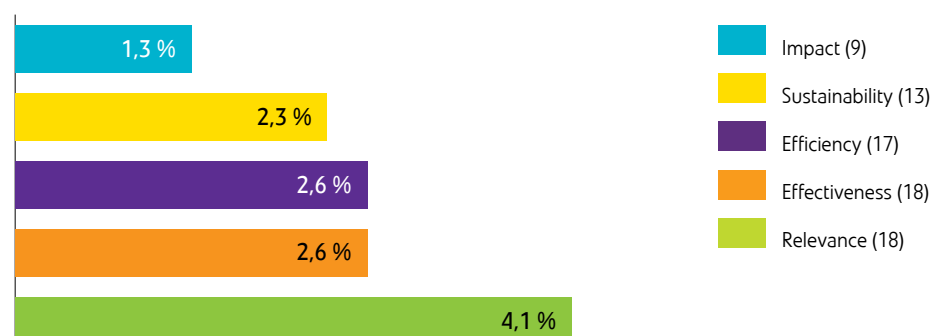
This section deals with each of these in turn. It is important to note that the analysis below deals with the way the evaluation reports took Finnish policy into account. A separate section at the end of this chapter deals with a few insights on the quality of Finnish development cooperation that can be extracted from **appraisal** reports.

6.1.1 Evaluation Report-based Findings

6.1.1.1 Findings based on the OECD/DAC evaluation criteria

Overall, the aggregated level analysis indicates that reports rated well on the extent to which they reflected the MFA policies on **relevance** (4.1). **Effectiveness** (2.6), **efficiency** (2.6) and **sustainability** (2.3) were not rated highly, but **impact** was rated very low at 1.3. The following diagram illustrates the aggregated level analysis, and brings out, visually, the considerable difference in the ratings. The number in brackets indicates the number of reports that dealt with the criteria; for example, in the case of impact, the average value is based on nine reports and not eighteen. It is interesting to note that part of the reason for these absences is because there is no obligation (or no perception of obligation) to report against impact where MTE are concerned.

Figure 14: Score of evaluation reports in the OECD/DAC criteria



Source: Meta-evaluation team analysis

On closer inspection, it is noted that the “Degree of **Relevance** to the intervention” was well rated (4.1 at the aggregate level) because it scored highly in three sub-aggregated levels.

- Consistency with the needs of the target group
- Alignment with national development goals and strategies
- The extent to which the report linked the intervention with Finland’s aid priorities and policies

A fourth area, namely the “extent to which the report explicitly deals and analyses the reference through the prism of HRBA and CCO”, received a much lower score of 2.3.

The induction-based analysis indicated that:

- a) relevance can sometimes disappear rapidly if organisational legitimacy disappears or if motivation falters; (F)
- b) relevance should be measured in response to concrete actions, not promises or broad concepts; (F)
- c) relevance should be defined, in part, by impact. Projects that are really very marginal in the scheme of things should be qualified somehow; (F)
- d) “the “needs” of the target group should be made clear, not the (partial) responses given to some of the symptoms of that need”;
- e) “the deliverables or results should be directly linked to the need”;
- f) something can only be relevant if it addresses the pursuit of another objective. Just because an intervention in trade expansion may generate some employment does not mean that it is directly relevant to the climate change strategy of the country; (F)
- g) ownership is not only a concept but a real and palpable thing. Inclusion and ownership are not necessarily the same thing; (F)
- h) alignment should not be defined in relation to to-level generic plans but to the strategies for delivering outcomes; (F)
- i) “the appearance of relevance is enhanced through a strong and participative design process”;
- j) the extent to which an issue is a “threat” (ex. flooding) increases the perception of relevance and focusses “needs”; (F)
- k) “poor design and scoping, weak or vague objectives and results and poor execution transform relevance into dis-motivation;
- l) “where all possible priorities are addressed by the programme under the umbrella of one particular sector, the result becomes somehow artificial, patching together activities and targeting each objectives without an integrated approach. Relevance becomes meaningless except at conceptual level.”

There is an urgent and significant need for MFA to clarify what information it wants to receive insofar as the assessment of relevance in project and programme evaluations is concerned.

MFA is likely finding it difficult to report on the extent to which interventions are effective, particularly at the outcomes level and specifically in terms of CCOs or HRBA.

In consequence, there is an urgent and significant need to clarify what content MFA wants to be informed of insofar as the assessment of relevance in project and programme evaluations is concerned.

The “degree of achievement of stated objectives, or likelihood to do so (**Effectiveness**)” criteria was analysed deductively using two sub-aggregated levels in four areas. The overall rating at the aggregated level for effectiveness is 2.6. The sub-levels were:

- The achievement of outcomes
- The extent to which the report linked and analysed effectiveness through the prism of HRBA and CCOs.

These results show that MFA’s development cooperation programme is likely finding it difficult to report on the extent to which interventions are effective, particularly at the outcomes level. It also shows that the interventions do not reflect effectiveness as being supported by CCOs or HRBA. The average rating for the latter (HRBA/CCO) was very low, with only 12 of the 18 documents dealing with the topic at all.

The inductive analysis on effectiveness supported the results of the deductive analysis. Only a few of the more salient points raised were:

- a) “Effectiveness was significantly reduced due to overly bureaucratic processes and long decision-making processes of organisations involved. Should have had more delegation to one body for action and decision-making”;
- b) “The complexity of development problems faced by communities cannot be solved by isolated interventions - a more holistic approach is needed, where technical, financial and social aspects are tackled together”;
- c) All the items in the project’s results framework are activities of the project, not actual indicators or results which would measure the outcomes of these activities; (F)
- d) “At the output level, achievements are relatively poor as per the evaluation report. Laws and decrees were drafted but none passed in Parliament. Trainings and other capacity development initiatives were organised but had positive results only for individuals, hardly for municipalities. Even staff hired by the programme to implement activities are of limited qualification”;
- e) “The time span was too short to accomplish the objectives. It was hard enough to develop the tools but the project was internally complicated and decisions hard to make efficiently. There was a one-year gap in the project but that was not the main stumbling block”;
- f) No real outcome level results that are significant except at a very local level; (F)

- g) “The level of programme flexibility when re-focusing the LFA in 2012 has been a major factor in the programme’s success. Administrative staff may be transferred without too much impact on the programme but highly technical skills like monitoring technicians, laboratory technicians, EIA reviewers, environmental auditors cannot”; (F)
- h) “The project was over scoped for the time and implementation strategy selected. This should have been evident to the designers. Too many stakeholders with the management structure and implementation programme. Much community-based and ownership-participation involved which is fine but not with the timeframe and resources allocated. Lack of synergy between components”. On-the-roll design without fixing the performance parameters of systems and organisations means that no one is happy with products because they are never finished and adapted.”

Overall, the inductive analysis indicted a relatively high degree of frustration with the challenges facing any intervention, but mentioned many innovative measures that were designed and implemented to resolve context and technical problems at the output level. The level of success in meeting expected performance targets is fairly high in most components of most interventions. It is the transformation of outputs into outcomes that faces multifaceted challenges, most of which were apparently not foreseen in the design stage.

The overall aggregated average rating for “Degree of performance of the intervention oversight and management (**Efficiency**)” was 2.6. Five sub-level standards were used:

- Extent to which outputs were achieved as planned
- Extent of transformational efficiency
- Extent of time efficiency
- Degree of quality of the oversight, decision-making and management reactivity
- Extent to which the report explicitly deals with, and analyses the efficiency as being supported through CCO and HRBA

Results show that the first (outputs) was rated at 3.3 out of five, with only one report not providing details at all. The next three sub-standards rated from 2.4 to 3.0, but the last rated a mere 0.5 (with only four reports even mentioning the topic, and not well done at that). As an overall finding, the MFA is not in a solid position to report on the all the elements of the efficiency of its programmes. The term efficiency was broadly interpreted in this context, and included not only financial or cost efficiency, but the efficiency of the chain of strategies chosen for transforming inputs into impacts. Efficiency also includes “the extent to which outputs were generated”. MFA can note that it registers much better with transformational efficiencies than it does with time of reaction and change, which continues to be a problem that many interventions have had: the time required to react to a change is too great when dealing with donor processes and multi-decision-maker structures.

The Team considered the efficiency of the entire chain of strategies chosen for transforming inputs into impacts. MFA does not receive solid data to enable it to report comprehensively on the efficiency of its programmes.

The inductive-based analysis provided much insight including:

- a) “The overall efficiency rating is less than efficient trending towards inefficient”;
- b) “Relatively efficient overall except for inefficiencies caused by process control and lack of decentralisation of decision making and authority. Overall the resources were adequate and there was no other way of doing this except through GoE”;
- c) The report deals with efficiency as if it equated with disbursement plans and reality; (F)
- d) “Project management performance would normally be assessed based on level of success in achieving the expected results. As mentioned in section 3.1.1 however, this is difficult in the case of SIP since there is no clear overall results framework that has guided the project throughout the implementation phase. The PIU has generally performed well in terms of work planning, reporting and accounting, although work plans and budgets have been consistently too optimistic”
- e) Being input-based, the MFA-finance TA was very efficient in transforming money in experts’ working days and delivering draft reports. Only few expected deliverables were still in progress at the time the MTE was undertaken; (F)
- f) The report does not say much about transformation efficiency. Incidentally, HR qualification and dedication are praised but counter-balanced by organisational issues, lack of presence on the field. The absence of data linking costs with outputs is a major impediment to come to an objective assessment; (F)
- g) “Share of transformation costs by project is slightly higher (16% vs 13%) for NORAD projects that for other donors’ project implemented by UNIDO. UNIDO staff is often assesses by the report as efficient and dedicated.”
- h) “While it remains difficult to compare different type of schemes, SIP’s cost per hectare are definitely on the high side”.
- i) The programme is not really time-bound, even is country programmes are so. Achievements to date are presented as significant in their context; (F)
- j) Delays in implementing the project are not pinpointed specifically by the report; (F)
- k) “Most projects are delayed. Much of the delay is attributed to the partner countries”
- l) The report emphasises the role of weak oversight, lack of decision-making and poor management as one of the key issues of the programme(F)

- m) "A major historical weakness of UNIDO identified by the report is the lack of a functional RBM system capable of capturing data at all levels of the results chain and systematically reporting the achievement of a project towards outcomes. Other aspects of oversight and decision-making were outside the scope of the evaluators".
- n) MFA financed TA component performed poorly. The TA delivered a number of reports, guidelines, etc. but failed to get them really approved and owned by the project; (F) management. Though quality of the reports is not neatly assessed as poor by the evaluator.
- o) "At the output level, achievements are relatively poor as per the evaluation report. Laws and decrees were drafted but none passed in Parliament. Trainings and other capacity development initiatives were organised but had positive results only for individuals, hardly for municipalities. Even staff hired by the programme to implement activities are of limited qualification".

At an aggregated level, the "Degree or prospects for **sustainability**" standard was given an overall rating of 2.3. It had six sub-level areas of analysis:

- Degree of, or prospects for social sustainability
- Degree of, or prospects for financial/economic sustainability
- Degree of, or prospects for environmental sustainability
- Findings related to technical sustainability
- Degree of, or prospects for organisational sustainability
- Extent to which the report explicitly deals with, and analyses the sustainability as being supported through CCO and HRBA

The first five sub-levels registered from 1.5 to 2.4, all in all a relatively poor performance with between 9 and 14 reports dealing with the topic at all. The last issue only received a rating of 0.5 (with nine reports mentioning the topic). It is clear from the above that MFA may have cause to be preoccupied with the sustainability of its development cooperation initiatives. The overall ratings for financial/economic sustainability reflect the often-written sections in the reports indicating that there are still (at the time of writing of the report) no concrete steps taken to ensure that the programme or objective to which the intervention has contributed will receive the ongoing financial support required from the governments. The reports also provide information showing that the organisations have important gaps in capacity and capability as well as in organisational stability. The issue of how the governments will be able to fill the shoes left when the PIUs are disbanded is very rarely discussed in the reports. The reports do not refer to "social sustainability" in those terms but the meta-evaluation has used the contents of the report to provide ratings: social sustainability is directly dependent upon the existence of an enabling/organisational environment and the financial/technical resources to service the needs of communities. If these are seen as sustainable (even in the mid-term), then the communities are "motivated" or feel they have "ownership", to name a few indicators. That is possibly the reason (hypothesis) why the rating for social sustainability reflects the ratings of these variables.

MFA has many reasons to be preoccupied with the sustainability of its development cooperation initiatives.

Although many items were identified in the **inductive analysis** dealing with sustainability, most were already identified through the deductive analysis. The weight of mid-term reports in the sample of 18 reports weaken somehow the scope for inductive analysis of reporting on sustainability because those reports are expected to focus on short-term decision making, thus mainly effectiveness and efficiency. A hint across the sample is that sustainability is positively assessed (with caution) for the few projects involved in some sort of HRBA.

A few of the more interesting inputs from that inductive analysis include:

- a) Rated “likely” by the report, but very unlikely by the meta-evaluator. The issues faced are deeply rooted and recommendations issued quite far-reaching; (F)
- b) The report indicates that it is far too early to assess sustainability; (F)
- c) “CDF/CMP implemented schemes are highly sustainable compared to other approaches. Ownership and commitment of the community and WASHCOs for supervision of construction quality and for O&M and protection of the scheme contributes to sustainability”;
- d) “To bolster sustainability prospects, a comprehensive capacity development strategy is needed, with a 2-year no-cost extension proposed to allow the programme to recover time lost in the first years and to spread the still considerable financial resources over a longer period”;
- e) “Significant challenges remain, For example, running a monitoring systems costs a lot of money for a poor country. Political commitment is voiced but not done, as guidelines, rules, etc. are developed by never put into law or effect”;
- f) “By lack of indication of an HRBA or another approach of activating a social demand, social sustainability is unlikely altogether against economic forces”;
- g) “Successes include support to participatory development of CIDP and good awareness creation on land rights issues. High success claimed on improved extension services and new businesses for youth but to be corroborated. Capacity building suffers from limited focus on mostly technical training”;
- h) “An important hindrance to substantial impact is the small scale of the activities undertaken and the fact that substantial replication of activities is not yet taking place (except on a limited scale for rocket stoves)”;
- i) “Where the project is managed as a cooperative the prospects look good, but that is only a small part of the total acreage, No mention of financial sustainability for the latter. These projects are historically very difficult to finance on the long term (operations and maintenance) and the report does not deal with these aspects sufficiently”;
- j) No CC/environment component in the programme; (F)
- k) “The problem is not the impact of the project on climate but the other way around”. This is made clear in the text; (F)

- l) This appears to be the main benefit of the programme through tools such as strategic evaluations. The programme however does not involve much in capacity development and training; (F)
- m) There is no technological transfer whatsoever in this project; (F)
- n) Much of the report is about changing the administration of the programme. As it stood when under scrutiny, the organisational sustainability was rate poor; (F)
- o) "The management model developed by SIP promises to balance farmer ownership with professional and cost-efficient scheme management. The business plans show good commercial viability, and initial financial needs have been covered through loans negotiated with FNB (a commercial bank) and through facilitation by Zambia Sugar. Management support provided to the irrigation companies by AMSCO through separate funding from Finland will help ensure institutional sustainability".

At an aggregated level, the "Contribution to the achievement of **impact** (intermediate) level results (even if not an "impact evaluation" per se)" standard was given an overall rating of 1.3 with nine reports commenting on the topic (none of the MTE reported on impact). It was assessed using two sub-level criteria:

- Degree of achievement of main intended intermediate impacts
- Extent to which the report explicitly deals with and analyses the extent to which "IMPACT " is supported through CCO and HRBA

Fundamentally, the reason for the low ratings is that evaluators apparently did not feel they had the information or analysis required to judge the extent to which the intended impacts would be attained. In fact, many reports note that they were unable, for a variety of reasons, to judge on outcomes.

The inductive analysis indicated that:

- a) the statement of impact was too lofty for the scope of the project and the programme it contributed to: (F)
- b) monitoring tools did not include the measurement of impact indicators even if these sometimes existed: (F)
- c) the evaluators thought that it was too early to judge on impact, all the while not specifying when it would be a good time. Very few report described any negative impacts; one noted "reduced social inequity could be an impact of such a land registration process but intended or unintended achievements in this respect are nowhere indicated". (F)
- d) "market distortions are presented as potential unintended effect (if the project design process is poor). That hardly apply to norms and standards supported (by the donor)".

6.1.1.2 Findings based on Aid Effectiveness

The analysis the extent to which evaluation reports reflect on the quality of Finnish development cooperation in terms of Aid Effectiveness was done using six independent analysis areas. There is no aggregated level. The ratings are given in brackets:

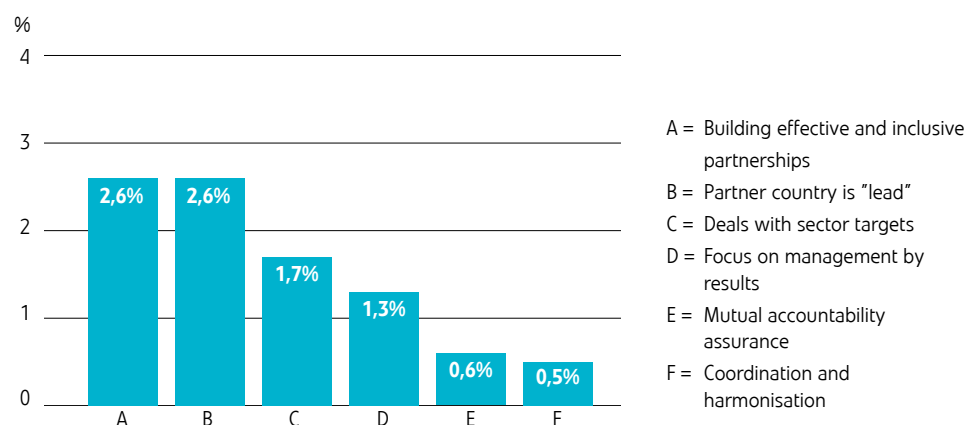
"Contribution to the achievement of impact (intermediate) level results" was given an overall rating of 1.3 with only nine reports (out of 18) commenting on the topic at all.

Most reports do not specifically address the issue of aid effectiveness as a separate concept.

- Degree of coordination and harmonisation with other donors, host country organisations, civil society, NSA etc. (2.6)
- Extent to which mutual accountability is assured (0.5)
- Extent to which the partner country is in the lead (i.e. formulate and implement their own national development plans, according to their own national priorities, using, wherever possible, their own planning and implementation systems (harmonisation)) (2.6)
- Extent to which the intervention actually focusses on the management for results: I.e. focus on the result of aid, the tangible difference it makes in poor people's lives. Develop better tools and systems to measure this impact (1.3)
- Extent to which the intervention contributes to building more effective and inclusive partnerships (1.7)
- Extent to which the report deals with the sector targets and indicators for aid effectiveness (0.6)

Visually the scores show the extent to which there is a considerable level of variation between these areas. Some areas score twice as high (and more) as others:

Figure 15: Ratings used in analysis of aid effectiveness



Source: Meta-evaluation team

The meta-evaluation Team found that most reports do not specifically address the issue of aid effectiveness as a separate concept. The Team has therefore had to "data mine" the reports to some extent to be able to identify relevant information to use in the analysis of these six areas.

The issues related to the Paris, Accra and Busan agreements (especially Paris) are interesting to study as a “cluster”: in the relevance section above, “alignment” has an overall rating of 4.6. Mutual accountability in this section is 0.5, and coordination/harmonisation in this section is rated at 2.6. Overall, these reflect the experience of most donors (see the OECD’s evaluation of the Paris Declaration on Aid Effectiveness): alignment is rather straightforward and easy to do at the strategic level (since it is a comparison of higher-level objectives versus other higher-level objectives), while harmonisation is much more standards-based and norms-based and so requires “compliance” mechanisms and “competency assurance” (to name a few) that needs to be developed and managed on a daily basis in order to “evolve”. Mutual accountability is rarely, if ever, discussed or monitored or managed using that specific term (even the OECD evaluation used a very indirect proxy measure to study a small part of the complex issue of “accountability”). As a cluster, it would appear that MFA is reported as having been successful at alignment, as having some difficulty but evident successes in harmonisation, but has not been seen as being successful in managing its mutual accountability commitments. It is recognised that the concept of “mutual” requires at least two parties; the evaluation reports never discuss how the recipient country tries to execute its own mutual accountability commitments).

As noted above, the reports also very rarely reflect on aid effectiveness as a concept in and of itself. That explains the very low score given to the sixth area. It is clear that if MFA wants to use evaluations as a means of gathering information on aid effectiveness, it should begin to include that in the Terms of Reference.

Finally, the meta-evaluation team found it interesting that the evaluation reports rarely spoke of concerted or planned efforts to “develop partnerships” or any other related concept. An exception to this observation is when such an objective is explicitly stated as part of the intervention’s log frame or component-based structure. In some cases a dozen or more organisations may be involved, but the most the reports will do is to speak of “coordination” for the purposes of the intervention’s own objectives. They have not evaluated any form of partnership value-added, nor have they discussed the benefits to the intervention of “twinning”, working with highly-qualified TA or consulting firms. In fact, the TA used in the interventions is very rarely evaluated at all, whereas other forms of resources and inputs are noted and contextualised.

6.1.1.3 Findings based on HRBA and CCOs

Five main analysis areas were studied using specific quality criteria. The description and rating given are followed with an analysis of key findings:

1. “The evaluation report indicates that the project or programme supported by the MFA effectively addresses the crosscutting objective of HRBA” (0.8).

Only six reports deal with HRBA or CCO in a meaningful way. There are many tangential references but little analysis, thus explaining the low rating given.

With respect to aid effectiveness, MFA has been successful at “alignment” and to a much lesser extent, “harmonisation”. It has not been successful in managing mutual accountability.

Reports do not deal with partnership value-added. In fact, the TA strategy used is rarely evaluated, whereas other forms of resources and inputs are noted and contextualised.

MFA's HRBA policy is not being implemented or is not being reported upon.

Except for two projects, reports did not present disaggregated data or conclusions based on gender (or any other similar variable). MFA is not being provided with data on those policies.

The 2012 GoF policy context emphasises the Human Rights Based Approach, but the 2007 policy context also contains references to human rights, but as a cross-cutting issue. The Meta-evaluation Team found that while the term “HRBA” was almost always mentioned in the reports that were written under the 2012 policy umbrella, the reports never evaluated such an “approach”. In fact, they almost always noted that the intervention was not yet (in the case of MTE) or had not (final reports) put in place the means to plan, monitor, execute, manage etc. an approach that was founded on human rights. That being said, the implementation itself may (or may not) have dealt with issues and problems of human rights (access to food and health, education, water etc.). In this case the Team specifically rated the reports in the light of their reference to a concerted and planned approach. The rating given provides MFA with a clear indication that its HRBA policy is not being implemented or is not being reported upon as such.

2. “The evaluation report indicates that the project or programme supported by the MFA effectively contributes to the cross-cutting objective of gender equality (planned higher-level results in this area are realised” (2.1). Surprisingly, only thirteen reports analysed the issue.

The team was rather taken aback with the ratings given for this analysis area. It observed three phenomenon: a) the ratings were either very high (4-5) or very low (1). This implies that gender equality is treated either as a “do-or do not” issue; b) a large proportion of reports noted that some activities involved women as “targets”, such as including women in training course, but also noted that they were not involved in decision-making or were not the direct beneficiaries as the result of an overt decision, and c) only a handful of interventions had monitoring systems concerned with gender at all. Except for two projects, no other evaluation reports presented, in the main part of the report, disaggregated data based on gender (or any other similar variable for that matter) and only a few reports presented conclusions and recommendations based on the intervention’s experience with gender.

The rating given should provide MFA with an avenue to explore further. The problem may not be that gender is not taken into account; it may not be “structured” appropriately or it may not be reported against adequately.

3. “The evaluation report indicates that the project or programme supported by the MFA effectively addresses the cross-cutting objective of reduction of inequality (planned higher-level results in this area are realized)”. (1.4). Only nine reports analysed this issue.

Overall, it is clear that evaluation reports do not deal specifically with “inequality” as a specific domain. In fact, the term is rarely used. Only one project scored highly (5) in this area in spite of its centrality in Finland’s development policy framework. It was also observed that the expression “reduction of inequality” was not used in the description of the interventions’ objectives or components.

4. “The evaluation report indicates that the project or programme supported by the MFA effectively addresses the crosscutting objective of climate sustainability (planned higher-level results in this area are realized” (2.5). Eleven reports commented on this issue.

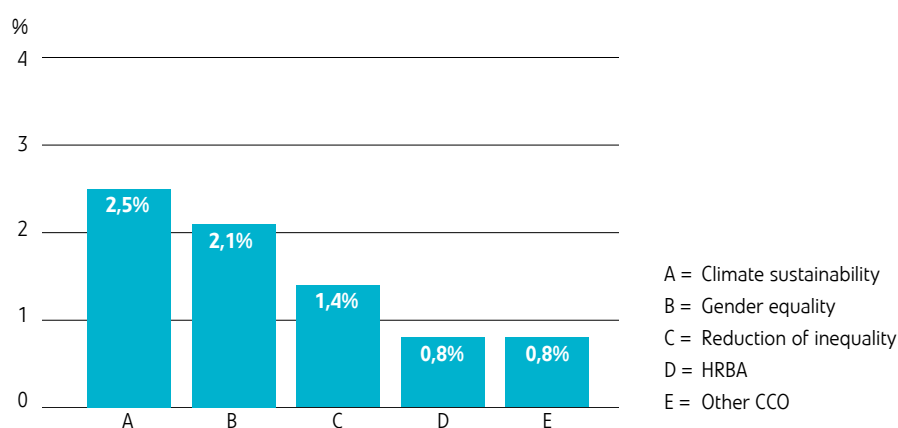
In contrast to the analysis area dealing with “environment” within the “sustainability” criteria above, the Meta-evaluation team was specifically looking for the treatment of “climate”. It was noted that many reports did in fact mention climate but almost all were superficial references. References to concepts of greenhouse gasses, or carbon sequestration, for example (to name only two climate change related concepts) were not present in the sample. A few projects reported that climate change had more of an effect or impact on the project than the project had on climate change (ex. irrigation schemes). The rating is a little higher than it perhaps should be: the meta-evaluators gave positive responses where the issue was raised at all, even if it was superficial.

Overall, it is clear that the MFA cannot use the meta-evaluation to report on the extent to which its development cooperation activities are supporting the GoF commitments on climate change. To do that would require a much more rigorous monitoring or reporting than what was presented in the reports. What it can report on is the fact that “climate change management within interventions” is a variable that is starting to rise to the surface, but which has not yet found successful models.

5. “The evaluation report indicates that the project or programme supported by the MFA effectively addresses another cross-cutting objectives or emerging themes.” (0.8). Only four reports actually provided substantial analysis on this topic.

A visual representation of the entire set of five areas shows the relative ratings obtained and the extent to which some of MFA’s flagship policy objectives (HRBA for example) are rated just above the lowest possible rating (of “1”).

Figure 16: Scores of HRBA and selected cross-cutting issues



Source: Meta-evaluation team

Overall, it is clear that the MFA cannot use evaluations to report on the extent to which its development cooperation activities are supporting the GoF’s commitments on climate change.

MFA may have policies on risk management but its interventions are not implementing them. Either that, or the evaluation reports are systematically not reporting on them.

Of note is the large number of issues (239) that were not dealt with within the entire set of 18 projects; No report spoke of risk management and only 11 spoke to RBM.

6.1.1.4 Findings Related to Risk Management

At an aggregate level, the “Degree of implementation and success of the risks management strategy” standard was rated at a very low 0.3, with only 2 reports actually providing some form of detailed analysis on the topic. The two sub-levels were:

- Existence of an articulated risks management strategy (i.e. a thorough and reasoned analysis with identified risks, their chance of occurrence, the impact of occurrence and the mitigation strategy)
- Findings demonstrating a gain in utilising the risks management strategy

Not a single report dealt with the issues of the contribution of risk management strategies to the realisation of planned results.

The results noted above paint a clear picture: MFA may have policies on risk management but the projects are not implementing them. Either that, or the evaluation reports are systematically not reporting on them. It was also observed that the only project to score well (3.0) on risk management also scored well (3.5) on the use of RBM. That, in the opinion of the meta-evaluation team, is not likely to be a coincidence.

6.1.1.5 Overall picture of the entire set of Phase 2 results

The following table contains the complete set of Phase 2 results in each of the aggregated (orange) and un-aggregated (orange) levels. It also shows the arithmetic average of all the scores given to the 18 projects selected for the Phase 2 analysis and, importantly, the “size of “n”, or the number of projects (out of the 18) that **dealt with each issue at all** at the un-aggregated levels. The “Average of values” column (second from right) is always equal to the “Total of all values” divided by the number of projects in the sample, (i.e. 18).

The column entitled “size of “n”” provides an interesting perspective of the degree to which reports **comprehensively** dealt with the quality of Finnish development cooperation. Where the “n” is close to 18, (such as 4.1 “transformation efficiency”) the reader is informed that all, or almost all, of the reports deal with that issue. Of note is the large number of issues (239) that were not dealt with within the entire set of 18 projects; some issues were not dealt with at all by any project (i.e. risk) while others were dealt with by a fraction of the 18 projects (ex. 6.4, “management for results” where only 11 projects dealt with the issue). This provides some insight for MFA executives as it indicates the extent to which the reports as a whole provide them assurance information.

When considering the hypothetical question: “How well did issues get dealt with by those reports that actually dealt with them?”, the last column provides some insight. In this case the “Total of all values” column was divided by the number of projects that registered some positive rating. In other words, the effect of project reports that were silent on an issue is not taken into account. As expected mathematically, the effect is larger as the number of “absent projects” increases. For example, note that the average for issue 3.1, dealing with intermediate impacts, goes from 1.3 to 3.0, indicating that the quality of the reporting was fair. The same level of effect is not always present, however: consider issue 6.2 dealing with mutual accountability. The revised average goes from 0.5 to 2.3, but even that revised score is a poor performance for the combined set of reports that dealt with the issue.

Table 13: Ratings given to various analysis areas for Phase Two

	Total of all values	Size of "n" (n=# projects that dealt with issue)	Average of values of all 18 projects	Average of values, only projects that dealt with issue
RELEVANCE				
1. Degree of relevance of the intervention	74,5	18	4,1	n.a
1.1 Consistency with the needs of the target group	79	18	4,4	4,4
1.2 Alignment with national development goals, policies and plans	83	17	4,6	4,9
1.6 The report explicitly deals with and analyses the extent to which "relevance" is supported through CCO and HRBA	37	13	2,1	2,8
1.7 Link of objectives with Finland's aid priorities (refer to note in upper right- hand quadrant)	70	16	3,9	4,4
EFFECTIVENESS				
2. Degree of achievement of stated objectives (or likelihood to do so) during the implementation period. Not only "during" but within planning horizon. Relevance needs to cover the future	47	18	2,6	n.a
2.2 Degree of achievement of the outcome level	50	18	2,8	2,8
2.7 The report explicitly deals with and analyses the extent to which "EFFECTIVENESS" is supported through CCO and HRBA	29	12	1,6	2,4
IMPACT				
3. Contribution to the achievement of impact (intermediary) level results (even if not an "impact evaluation" per se)	23	9	1,3	n.a
3.1 Degree of achievement of main intended intermediary impacts	24	8	1,3	3,0
3.4 Extent to which the report explicitly deals with and analyses the extent to which "IMPACT" is supported through CCO and HRBA	11	4	0,6	2,8
EFFICIENCY				
4. Degree of performance of the intervention oversight and management	46	17	2,6	n.a
4.1 Extent of transformation efficiency (outputs)	54	17	3,0	3,2
4.2 Extent of time efficiency	46	16	2,6	2,9
4.3 Degree of quality of the oversight , decision-making and management reactivity	43	15	2,4	2,9
2.1 Were the outputs achieved as planned in the implementation period?	60	17	3,3	3,5
4.6 Extent to which the report explicitly deals with and analyses the extent to which "EFFICIENCY" is supported through CCO and HRBA	9	4	0,5	2,3
SUSTAINABILITY				
5. Degree or prospects for sustainability	42	13	2,3	n.a
5.1 Degree or prospects for social sustainability	40	11	2,2	3,6

	Total of all values	Size of "n" (n=# projects that dealt with issue)	Average of values of all 18 projects	Average of values, only projects that dealt with issue
5.2 Degree or prospects for financial/economic sustainability	33	10	1,8	3,3
5.3 Degree or prospects for environmental sustainability	27	9	1,5	3,0
5.4 Findings related to technical sustainability	43	14	2,4	3,1
5.5 Degree or prospects for organizational sustainability	39	13	2,2	3,0
5.8 The report explicitly deals with and analyses the extent to which "SUSTAINABILITY" is supported through CCO and HRBA	9	5	0,5	1,8
AID EFFECTIVENESS				
Average of all values for each project divided by n=6	28	18	1,6	n.a
6.1 Degree of coordination and harmonisation with other donors, host country organisations, civil society, NSA etc.	47	14	2,6	3,4
6.2 Extent to which mutual accountability is assured	9	4	0,5	2,3
6.3 Extent to which the partner country is in the lead (i.e. formulate and implement their own national development plans, according to their own national priorities, using, wherever possible, their own planning and implementation systems)	47	11	2,6	4,3
6.4 Extent to which the intervention actually focusses on the management for results : i.e. focus on the result of aid, the tangible difference it makes in poor people's lives. Develop better tools and systems to measure this impact	23	11	1,3	2,1
6.5 Extent to which the intervention contributes to building more effective and inclusive partnerships .	31	10	1,7	3,1
6.6 Extent to which the report deals with the sector targets and indicators for aid effectiveness	11	5	0,6	2,2
OVERALL MANAGEMENT AND CONSEQUENCES OF HRBA and CROSS-CUTTING OBJECTIVES				
Human Rights-Based Approach				
7. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the crosscutting objective of HRBA.	15	6	0,8	2,5
Gender Equality				
8. The evaluation report indicates that the project or programme supported by the MFA effectively contributes to the cross-cutting objective of gender equality (planned higher-level results in this area are realized).	38	13	2,1	2,9
REDUCTION OF INEQUALITY				
9. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the cross-cutting objective of reduction of inequality. (planned higher-level results in this area are realized).	25	9	1,4	2,8

	Total of all values	Size of "n" (n=# projects that dealt with issue)	Average of values of all 18 projects	Average of values, only projects that dealt with issue
CLIMATE SUSTAINABILITY				
10. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the crosscutting objective of climate sustainability. (planned higher-level results in this area are realized).	45	11	2,5	4,1
Other Cross-cutting Objectives or Emerging Themes				
11 The evaluation report indicates that the project or programme supported by the MFA effectively addresses another cross-cutting objectives or emerging themes.	15	4	0,8	3,8
RISK MANAGEMENT				
12. Degree of implementation and success of the risks management strategy.	6	2	0,3	n.a
12.1 Existence of an articulated risks management strategy (i.e. a thorough and reasoned analysis with identified risks, their chance of occurrence, the impact of occurrence and the mitigation strategy).	6	2	0,3	3,0
12.2 Findings demonstrating a gain in utilising the risks management strategy.	0	0	0,0	0,0

Source: Meta-evaluation team

In support of the findings outlined in the table above, the following table provides an overview of the distribution of the ratings that were allocated for the 18 projects retained for Phase 2. It should be noted that there is no single rating for "Aid Effectiveness" since it is a concept that is composed of many component parts that cannot be artificially aggregated. Attention is drawn to the number of reports that did not deal with specific elements (see 6.2, 6.6, HRBA, impact, inequality, emerging themes and risk management).

Table 14: Distribution of ratings for the quality assessment tool – summary of Finnish development cooperation

Main category	How many reports received a rating of:					In how many projects was at least part of this information missing	Total number of assessed reports
	1	2	3	4	5		
Relevance	0	1	3	6	8	0	18
Effectiveness	1	4	5	6	2	0	18
Impact	1	3	3	2	0	9	18
Efficiency	1	7	4	5	0	1	18
Sustainability	0	4	1	5	2	6	18
Aid effectiveness							
Degree of coordination and harmonisation	3	3	2	2	5	3	18
Extent to which mutual accountability is assured	1	1	2	0	0	14	18
Extent to which the partner country is in the lead	0	1	1	3	6	7	18
Extent to which the intervention actually focusses on the management for results	5	3	2	1	1	6	18
Extent to which the intervention contributes to building more effective and inclusive partnerships.	2	3	0	2	3	8	18
Extent to which the report deals with the sector targets and indicators for aid effectiveness	2	0	3	0	0	13	18
Human rights-based approach	2	1	4	0	0	11	18
Gender equality	1	4	4	3	1	5	18
Reduction of inequality	2	2	2	2	1	9	18
Climate sustainability	0	2	1	2	6	7	18
Emerging themes	0	0	2	1	1	14	18
Risk management	0	0	2	0	0	16	18

Source: Meta-evaluation team

6.1.2 Appraisal report-based Findings

Appraisals, by their very nature, are specifically interested in feasibility. In the case of the MFA, that term is used in a broad enough sense to mean that appraisals should provide analysis and recommendations that will help to ensure that draft Programme Documents are (or will become) congruent (i.e. comply) with the standards, norms and policies of the MFA.

Without repeating the analysis and findings above, the team found, among other findings, that:

- a) The results-chain logic on which MFA policy is structured is rarely prepared at the time of the appraisal, or whatever was done has important weaknesses. In short, that means that a preliminary intervention design was generated without a logical framework in spite of MFA guidance on that topic.
- b) Appraisals (confirmed by evaluations) indicate that RBM is not applied in project design, contrary to MFA instructions.
- c) The draft PDs are not based on HRBA and only deal superficially with CCOs. Targets and indicators are very rarely available. This is clearly not in line with MFA policy.
- d) The draft PDs rarely explicitly and comprehensively deal with efficiency, sustainability or effectiveness, but they do focus on relevance and impact. This, eventually, will cause problems with the approval process and will constrain the policy on “evaluability”. MFA will find it hard to report on the basis of the OECD criteria.
- e) Appraisals consistently identify that the management systems for interventions are weak, including those for monitoring, supervision, and oversight. This finding is important because it may indicate that no matter what the success (or weakness) of an intervention may be, the MFA will not have the data for early-warning and change management, or for reporting and transparency management.
- f) The topic of aid effectiveness is not well treated in appraisals (given a rating of only 50%), indicating perhaps that the appraisers were either not instructed on expectations in that regard in the ToR or were not made aware of the MFA’s requirements made explicit in other documents. MFA must report nationally and internationally on aid effectiveness, but does not necessarily have the information it needs to deal with the issue in detail.
- g) Overall, the appraisal reports are not structured along the lines of the OECD/DAC criteria or the MFA policy domains. Evaluating policy/guidance then becomes very difficult without information.
- h) Appraisals (and later evaluations) rarely provide MFA with lessons learned. This is important in the context where MFA sees itself as a knowledge-based organisation

Many preliminary intervention designs were generated without a logical framework or without an RBM-based approach, in spite of MFA guidance on that topic.

It is clear that draft PDs are not based on HRBA; rarely comprehensively deal with efficiency, sustainability or effectiveness, and only deal superficially with CCOs. Targets and indicators are very often unavailable.

Appraisals and evaluations rarely provide MFA with solid lessons learned. This is important in the context where MFA sees itself as a knowledge-based organisation.

While appraisals often indicate that risk management needs to be included in the programme documents assessed, evaluations point out that this is not followed even if it is MFA policy.

- i) Appraisals often indicate that some form of risk management needs to be included in programme documents. MFA guidance includes the management of risk.

The nine appraisal-related key statements complement the evaluation-based analysis presented above.

Many of the findings from appraisal and evaluation reports that deal with intervention weaknesses or difficulties in meeting MFA's policies, standards and norms are clearly systemic; moreover they are often identified during the initial planning stages of the MFA'S project cycle. One report may have provided very wise advice when it noted that: **"the problems encountered in the project would not have occurred if solid front-end analysis would have taken place"** (the quotation has been synthesized by the Meta-evaluation team).

6.1.3 Induction-based analysis

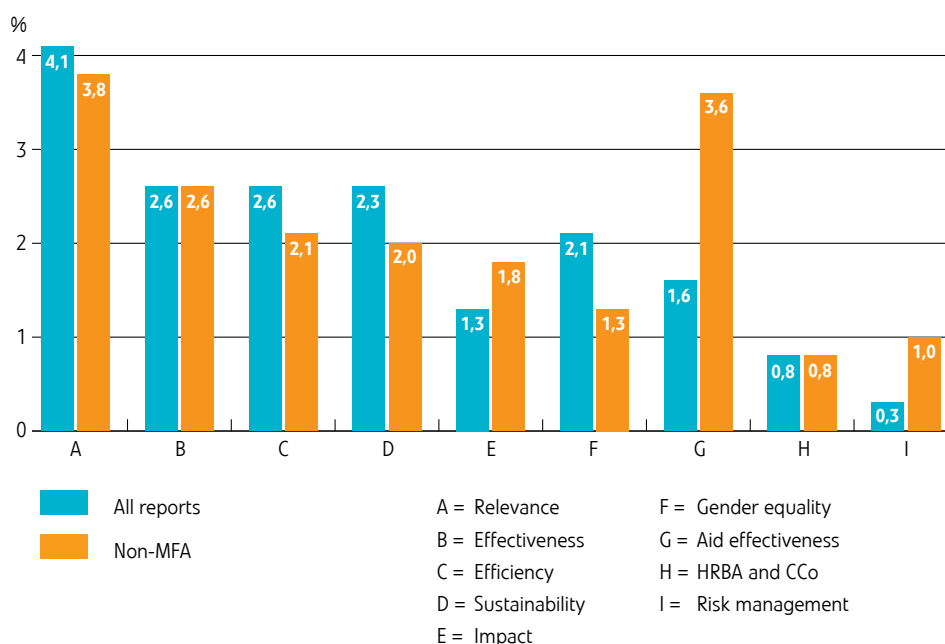
The meta-evaluation team has prepared an annex that gathers and structures all of the "inductive-based" information that was generated during the meta-evaluation. In its raw form it is rather long (over 90 pages) and contains repetitions and duplications. It has been shared with EVA-11 as an electronic document for future reference. In that annex, the information is structured along the same lines as the Phase 2 Analysis grid. The data gathered in that Annex represents the meta-evaluation team's findings and insights as well as selections from specific reports. The structure of the annex enables the analyst to see all the relevant observations that were made on any one topic on the same "page". The results have been integrated into the different sections as relevant (ex. 6.1.1.1).

6.1.4 Evaluations not commissioned by MFA

The analysis above was based on the entire set of n=18 evaluation reports used in Phase Two and the n=10 appraisal reports assessed in Phase One. Of interest to MFA is the extent to which the reports commissioned by non-MFA agencies (on projects partially or wholly funded by Finland) offer insights into the quality of Finnish development cooperation.

It is risky to compare the ratings of MFA and non-MFA commissioned reports one-on-one, especially if the object is to see the congruity of those reports with Finnish development cooperation. But with all due care being exercised, the ratings compare as follows:

Figure 17: Comparison of ratings of the OECD/DAC criteria, aid effectiveness, cross-cutting themes and HRBA, and risk management in non-MFA commissioned reports



Source: Meta-evaluation team

Comparison of the RELEVANCE criteria in MFA and Non-MFA commissioned reports

The overall rating for relevance for these documents is 3.8, which is the highest score among the 5 OECD evaluation criteria. Alignment with national development frameworks (4.5) and consistency with needs of the target group (4.4) rate significantly higher than most other standards in these reports. The overall rating of 3.8 is lowered because of the inability to assess (positively or negatively) the role and positioning of HRBA and CCOs, and to a lesser extent, by the underlining of the coherence with one or several Finland aid priorities (3.3).

The inductive analysis casts some doubts whether a common understanding of relevance exists amongst these commissioners, particularly regarding the problems associated with the needs they are trying to fill. Each and every report may have its own interpretation and indicators (or indications, arguments) allowing it to assess the extent to which an intervention answers the needs of target groups in varying ways. The reasoning or meaning of the terms becomes even more imaginative when the project (report) deals with global or regional collective actions. The difference in the underlying linguistic and management rationales may be that there is always some need to satisfy through aid interventions, but some intervention designs do not specifically define their relevance within the paradigm of the “needs of beneficiaries”.

Another striking result is that while all UN projects reviewed (albeit a small sample in this meta-evaluation), were assessed positively for relevance, none had formally adopted the HRBA, even though their mandate fundamentally stemmed from the Declaration of Human Rights. The only NGO project in the

sample is closer to implementing an HRB approach than are other commissioners; it also is based on more grounded analysis and a particularly strong emphasis on sustainability. In a similar line of thinking, the results of the analysis calls for the question “Can relevance can be assessed positively if the designers were not engaged in a participatory and non-discriminatory process that identified and prioritized specific needs of target groups?”. Otherwise, at collective levels (global, national, local) levels, consistency with needs duplicates with alignment, which may partly explain why relevance was systematically positively assessed.

Comparison of the EFFECTIVENESS criteria in MFA and Non-MFA commissioned reports

The global **effectiveness** rating of non-MFA projects is a relatively low (2.6). Here again, the absence of adoption of an HRBA and the uneven coverage of CCOs (1.8) have contributed to reduce the rating although but not so extensively as was the case for relevance. The extent of achievement of outputs² is only (3.6), as is the degree of achievement of outcomes (2.6)³. The underlying negative statements are compounded by the fact that all reports within this group (i.e. non-MFA) are mid-term evaluations or reviews. Issues in achieving outputs are naturally impacting the ability to perform at outcome level. The only recurrent feature in this respect is that projects’ performance at output level are systematically uneven between components of the same project, which questions project design either on the nature of the prioritised activities, and/or the existence (or quality of) of risk mitigation strategies (which were found to be systematically missing in all reports).

The relatively low rating (2.6) on **outcomes** reported in non- MFA “commissioned” evaluation is biased upwards by two projects rated at 5.0: a) the only NGO project to have a focus on health and to integrate a gender and human rights approach, and b) an academic project. Symptomatically, only the reports dealing with the NGO project and a separate sector project identify that they have or will achieve “outstanding” outcome. The rule is more that shortcomings are more frequently quoted (in relation with political, institutional or contractual blockages).

The low rating (1.8) for the role of HRBA & CCOs in achieving outcomes is linked to the quasi systematic absence of an HRB Approach; but it is also directly related to a “check mark in the box” approach to reporting on CCOs. In all cases save the only NGO project, the reports underline that CCOs are not mainstreamed but developed within a silo approach, whether it is the main focus of the project (environment) or affixed to a component or other of the project.

² Most donors consider “outputs” as part of an effectiveness analysis.

³ It is understood that MFA treats the generation of outputs as part of “efficiency”. In this paragraph, outputs are referred to simply because they are part of the logic chain that leads to outcomes.

Comparison of the IMPACT criteria in MFA and Non-MFA commissioned reports

The global rating is as low as 1.8 (out of a population of 3 reports that provided some tentative assessment of impact). It is not possible to analyse this result further; it should be underlined however, that this rating is consistent with the score awarded the achievement of the outcomes (2.6). As will be further developed in the section of this report dealing with Phase 2 results, poorly conceived impacts that rely on lofty and non-measurable outcomes within an overly-scoped (and under-budgeted) intervention were too-often reported on in the documents examined.

Comparison of the SUSTAINABILITY criteria in MFA and Non-MFA commissioned reports

Sustainability was only assessed within half the sample of non-MFA commissioned reports (five out of ten reports), the other half not having reported on that topic. The average rate is only 2.0, meaning that sustainability is almost certainly not being achieved or is not reported on. Reports show that the best achievements or prospects for sustainability are linked to technical sustainability (ex. delivery of tools or technology/knowledge transfers), rated at a good 3.5; and environmental sustainability (3.0).

Organisational sustainability appears as a relatively frequent problem (rated at 2.3) within evaluation reports. Social and financial/economic sustainability are rarely predicted to occur (respectively 1.6 and 1.7). Strikingly, but consistent with the above-noted findings of the meta-evaluation, HRBA and CCOs are never identified as driving factors for sustainability. Evaluators are observed to not report on HRBA or CCOs, and (hypothesis) they may not be familiar with the details of the HRBA or of CCOs. They have never reported on the potential role of HRBA and CCOs in developing ownership and demand-driven sustainability.

Issues identified by the inductive analysis are similar to those of MFA-driven projects: uncertainties dealing with the actual adaptation to changes in contexts or the regulatory framework and its subsequent implementation; political blockages at one level or another; staff or champions instability; isolation of rights-holders in patronage relationships etc.

Comparison of the EFFICIENCY CRITERIA in MFA and Non-MFA commissioned reports

The average overall rating is 2.1 (n=7), downgraded (by 0.8) by the lack of identification of the role of HRBA/CCOs in increasing performance for all reports. A sub-criteria, transformation efficiency, ranks above average (3.1, n=9), as do two other sub-criteria (time efficiency 2.9, and quality of oversight 2.6; n=8).

According to inductive analysis findings, those rates are artificially high due to the confusion between evaluators when reporting on the efficiency criteria. There are as many types of analyses as evaluators (some reports *combine* 2-4 approaches), covering all possible understandings of value-for-money and outcome attainment strategies, to name a few. Most merely deal with budget and expenditure management or cash flow, and do not present evidence to justify broad claims and obvious conclusions (ex. the procurement system took three

years to buy a key asset, so the project is inefficient) to come to a defensible assessment.

Without ever having reported on that basis, the language used in efficiency sections of reports is often much more guarded and neutral than it is in other sections, hence the hypothesis can be developed that the evaluators seize the opportunity to be lenient on an issue for which clients and management generally demonstrate a high sensitivity.

It is recognised that the structure of evaluation reports promoted in the MFA evaluation manual is not followed in non-MFA commissioned evaluations; the meta-evaluation took that reality into account in its analysis.

Comparison of AID EFFECTIVENESS in MFA and Non-MFA commissioned reports

As expected due to the nature and value sets of non-MFA commissioners, coordination and harmonisation rated relatively high at 3.6 (n=7). This result should come as no surprise since any form of co-financing implies a minimum of coordination under the Paris Declaration and subsequent international guidelines. On the other side of the spectrum, mutual accountability is never assessed in reports (n=5); nor is the extent of use of sector targets and indicators for aid effectiveness.

Interestingly, the leadership of the partner country is not very high when one recollects the nature of the executing agencies (3.0, n=5). Even if RBM has been implemented in more than half of the reports (2.7, n=7) an analysis shows that it may have been superficial with poor results frameworks. Result-based monitoring by these commissioners is a recurrent feature in all projects, demonstrating a trend with high potential for development aid. With respect to results frameworks, the reports indicate that UN agencies often define results-chains in lofty terms and then face hurdles translating high level commitments to grounded implementation.

The aid effectiveness agenda comes out unevenly in non-MFA evaluations, as shown by the diverse values of n in the above. Information gathered in grids is more often incidental than specific analyses. This again is likely related to the over-representation of UN agencies' programmes in the subsample.

Comparison of RISK MANAGEMENT in MFA and Non-MFA commissioned reports

None of the 10 evaluations identified that a comprehensive risk management strategy had been integrated into a project.

Comparison of HRBA/CCO in MFA and Non-MFA commissioned reports

Save for the one NGO project referred to above, HRBA is altogether absent in this cluster of 10 projects. Even for that project, HRBA is not a core approach but one of the several approaches that were adopted to contribute to sustainability and ownership, and eventual approval of the project as a whole (in parallel with advocacy targeted on religious leaders, for example).

As noted earlier, CCOs are rarely mainstreamed and used as a tool to contribute to achieving outcomes and impacts. They are included in projects to develop a compatibility with donors' priorities, not to ensure coherence with their development strategies and the principles that are underlying them.

6.2 Comparison of 2012–2014 to 2014–2015 meta-evaluations

Since they are not based on the same indicators nor have a common frame of reference, it is not advisable to directly compare the results of these two meta-evaluations as if the second was an update of the first. It should be remembered that the two reports are not based on the same structure either, so a side-by-side comparison is impossible. Another key difference is that appraisals were not part of the 2012–2014 meta-evaluation. A detailed description of the key methodological and epistemological differences between the two meta-evaluations is found in Annex 3 Detailed Methodology.

What **is** feasible (and epistemologically justifiable) is to take notice of what each had to say about specific topics that are common to both within Evaluation Reports. With respect to the OECD/DAC criteria:

Effectiveness

- The present meta-evaluation is not as positive as the previous one about the extent to which evaluated projects were achieving their expected results. Many projects reported that the results frameworks were weak to non-existent; that monitoring and evaluation systems in place could not identify when (or whether) outcomes would be attained, and that some components were not going to perform adequately to reach expected targets. Contrary to the previous meta-evaluation, this Team did not automatically consider a “partial success” as meeting objectives, regardless of the reason.

Impact

- As with the previous meta-evaluation, this one found that most of the projects evaluated (both MTE and final) would fall short of being able to prove that they achieved their intended impact objectives. For the most part monitoring and evaluation systems that provide relevant impact information are still not in place, and the impact statements used in the interventions are lofty enough to require a significant investment in ex post impact evaluations to be able to judge in impact. Nothing of consequence has changed since 2014.

Relevance

- as with the past meta-evaluation, this one found that that the finalities for the interventions were largely aligned with national plans and priorities; they addressed real needs of the target beneficiaries, and they represented the majority (but not all) of the policies of the GoF.

It is not epistemologically correct to compare results of the 2012–2014 meta-evaluation to this one. But many strategic conclusions are in the same vein.

Efficiency

- As with the previous meta-evaluation, this one found that most projects faced challenges with efficiency. The same problems were present, including inadequate design, lack of risk management strategies and plans, and the absence of baselines on most topics. This meta-evaluation broke down the concept of efficiency more than did the previous one, and was less fundamentally preoccupied with timeline challenges or the cost of developing outputs; it stressed outcomes and intermediate results. What the two meta-evaluations do agree on, however, is that the processes, procedures and oversight/control frameworks applied to most development cooperation initiatives are factors that seriously imperil the achievement of results, especially when mutual accountability and oversight are lacking.

Sustainability

- As with the previous Meta-evaluation, most evaluation reports placed serious doubt on the likelihood that outcomes and intermediate results will be sustainable. Some of the same reasons are still valid including a lack of concrete action (in deference to commitment or promises) to guarantee or at least enable sustainability to take hold. Contrary to the previous meta-evaluation, this one considered interventions in a comprehensive and holistic manner: “partial sustainability”, or “sustainability in one variable without a corresponding sustainability in another (enabling or dependent) variable was not deemed to render a development action “sustainable”.

7 CONCLUSIONS

7.1 Answering EQ 6 by focussing on Conclusions on the Quality of MFA decentralised evaluation reports and ToR/ITT

Conclusion 1: As a strategic-level conclusion, it is clear that there are important content and structural weaknesses in evaluation reports and their ToR.

The quality of MFA decentralised evaluation reports, while assessed against the conformity of the reports to guidelines which were explicitly referred to in TORs and were available to all commissioned evaluators, shows evidence of important content and structural weaknesses. These are noted as separate conclusions below. One critical weakness that can be observed in almost all reports is the absence of evidence to support or qualify the report's findings, and the failure to logically link conclusions to evidence-based findings and (then) to recommendations.

Conclusion 2: Regarding the conformity of the evaluation reports with the requirements of the MFA Evaluation Manual, the **core** parts of reports (i.e. findings, EQ answers and conclusions) are homogeneously weak in that they score only between 57-69 points out of 100.

Regarding conformity with the requirements of the MFA Evaluation Manual, the meta-evaluation team calculated "Average" scores for each level, based on the assessment grids. It then calculated the extent to which that average represented the maximum possible score for those levels. For example, if the average score for all documents for an issue was three and the maximum was five, the extent of what might be called "perfect performance" was 60%. These calculations are important because they show which parts of the reports were well done or not, (i.e. compared to the requirements of MFA manuals and guidelines).

The performance scores (in percentages) attributed to the various sections of the assessment grids are relatively homogeneous, ranging roughly between 50-70 points. The scores are reproduced below.

Table 15: Average scores of the evaluation reports per section

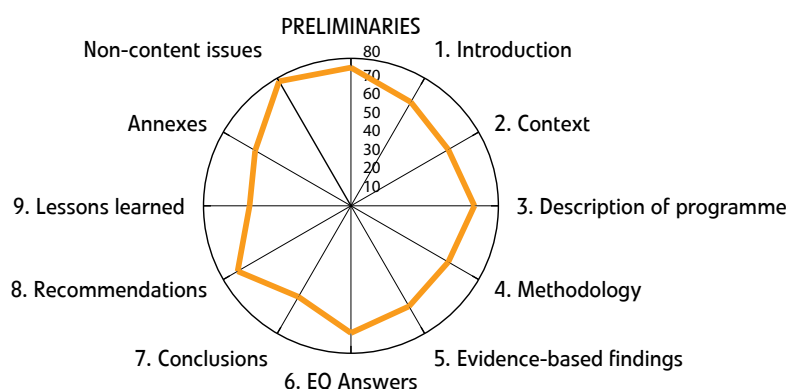
Section	Average score
Preliminaries	75%
Introduction	65%
Context	61%
Description of programme	67%
Methodology	61%
Evidence-based findings	63%
EQ Answers	69%
Conclusions	57%
Recommendations	71%
Lessons learned	55%
Annexes	60%

Source: Meta-evaluation team

The core sections were not rated highly, and the meta-evaluation team hypothesises that this must have an impact on the overall quality of the implementation of Finnish initiatives or programmes.

The points awarded to the recommendations section are the highest in all of the “content” areas, while “lessons learnt” is the weakest section because many reports did not address them or if they did, it was in a very superficial manner. The meta-evaluation team noted that the tendency to “synthesize” rather than “critique” was very real and palpable. Even the conclusions were often more related to “meta-findings” than they were of systemic rendering of a “judgment or decision reached by reasoning”. The conclusions were thus weakened and, importantly, were not often a strong basis on which to develop strategic or operational recommendations.

The table below illustrates the percentages in a “radar” chart. The concentric circles (refer to the point scales from 0.0 to 100.0 in increments of 20 points) indicate the range of possible percentages, and the blue line drawing represents the percentages. It should be noted that overall the performances are not very high for an evaluation function.

Figure 18: Distribution of Scores given to Evaluation Reports, (per section used in meta-evaluation assessment grid)

Source: Meta-evaluation team

The meta-evaluation found that the standard template for structuring the report is used in most cases, usually starting with an introduction, then a methodology and then a successive presentation of context, findings, conclusion and recommendations (and in rare cases lessons learnt). This structure is appropriate for evaluation reports and should lead the authors to base judgments on the logical progression from observation, to findings, to conclusions and then to recommendations. Unfortunately, most of these sections are dealt with in silos; for example, the conclusions are very seldom qualified by referring to the context, and the limitations in the validity of findings/conclusions are seldom referred to as the consequences of the “problems” of methodology. The recommendations do not always follow the same logic as the conclusions.

Annexes are not perceived as the place where detailed evidence should be presented. The meta-evaluation did not find an annex dealing with the evidence that supports the findings through indicators, for example. In fact, the logic that should be there concerning the formal link between the evaluation matrix and the indicators for the mandate is not found either in the main report or in the Annexes. Evidence on how the methodology was executed (i.e. in the form of a detailed methodology) is not found in most reports, and the meta-evaluation showed that there were few annexes dealing with any “quantitative” data that should have been, or was, used in the analysis.

The meta-evaluation also found that the reports do not comply with the MFA standards and norms when it comes to structuring the report on the basis of the MFA’s evaluation criteria (MFA has accepted the five OECD/DAC criteria as its base). Often only a sub-set of these are used as a means to design the overall structure of the report (generally relevance, effectiveness, sustainability and efficiency). The analysis shows that:

- a) impact is generally only referred to marginally without assumptions or hypotheses or statements of enabling conditions that are necessary for the impacts to be achieved. Time-to-impact is never discussed;
- b) the three Finnish criteria are very rarely ever referred to;
- c) the reports do not comprehensively examine how relevance was evaluated. Stating that an intervention is aligned to a poverty-reduction strategy or a sector strategy, for example, is not a precise enough analysis to claim that it is relevant. If that were the case then every possible activity could be “relevant”; the meta-evaluation found few cases where relevance was determined as a result of a contribution of the expected outcomes to some larger impact, or where it was related to the resolution of problems and policies so that a desirable change occurs (the expected impact). The meta-evaluation found many examples of what could be called “statements of the obvious”: saying that education is required in a county and so the intervention was relevant because it worked in that domain, for example, is stating the obvious.
- d) contexts were not developed sufficiently and most always took the form of a listing of relevant policy and strategy documents. Organisational and social interfaces or lessons learnt from previous development efforts, for example, were not described.

- e) efficiency was rarely described in terms of the selection of the appropriate strategies for the intervention. Very few reports had any details of costs and where they were noted it was generally to compare actual disbursements to planned budgets.
- f) the reports only dealt with alignment to Finnish policies in a very cursory way, generally by **listing** relevant high-level policies but not explaining the specific reason why they are aligned.
- g) only a small proportion of reports (and generally those that received a higher overall rating) are structured on the basis of the evaluation questions. Where one would expect to find an “answer” to the EQ, there is often only a diffused presentation of findings and it is left to the reader to decipher the answer.
- h) Many reports introduced the concept of “key” findings and then jumped directly to recommendations without generating conclusions.
- i) In very rare cases, activity clusters are used to provide evidence findings. While this may be useful in some areas (ex. lower-level activity-based outputs such as holding training sessions), activities are generally observations, not findings; it is the **result of the activity** that should have been presented in the report.

Conclusion 3: Those parts of the ToR that can be essentially “copied” from MFA manuals (ex. rationale, purpose, generic evaluation questions, evaluation process) scored higher than core content or policy parts (ex. aid effectiveness, appropriate methodology) which require detailed knowledge of the intervention.

Even a cursory analysis of Tables 1 and 2 in this report will point out that sections such as the “purpose and objectives”, “evaluation process” or “annexes” scored relatively higher than most other areas. These are parts of ToR that can essentially be “cut and pasted” from manuals or existing ToR. Scores are lower where thought and creativity were required in order to make the ToR become specific to the intervention.

Conclusion 4: The quality and substance of facts, figures and documentary references presented to sustain evidence-based findings is poor overall.

The overall volume and quality of facts, figures and documentary references presented to sustain evidence-based findings is poor overall. In (many) worst cases, mere enunciation (ex. “experts say”) appears sufficient to the evaluator (and thus to the evaluation manager), when it is clear that the “independent and objective research” required for evaluation judgement making cannot be satisfied using only those types of data from those sources. The best performers in evidence-based findings were found within the reports generated through UN specialised agencies and in some atypical reports (ex. Nepal forestry programme, and the Mary Stropes International initiative in Afghanistan).

In these cases, several methods for collecting data were utilised, justified and presented and information was checked to be congruent (even if not triangulated as such). If it were not possible to validate then a reference was placed in the report to that effect.

In most cases, the authors of the reports have collected a significant amount of reasonably cross-checked information (mostly through interviews and not triangulated through various forms of collection) to enable them to describe the intervention in a narrative and synthesis form. Only in rare cases do they succeed to convince the reader that they did so by following an epistemologically-sound methodology and that their conclusions and recommendations are truthful and logical consequences of their key findings. Since it is clear from ToRs that the evaluations need to be evidence-based, this is more than a simple oversight; moreover, it is systemic across reports. In fact, it is often very difficult, if not impossible, to identify whether a “finding” is derived from “observation” or “personal opinion”.

Conclusion 5: With the exception of sustainability, the concepts of, and use of evaluation criteria by different authors varies to the point where making cross-report analyses becomes risky.

There is a great deal of variation between evaluation reports on the meaning given to each evaluation criteria; save of impact and sustainability where lesser variations were noted, the other six criteria (three OECD/DAC and three MFA) were used quite differently by different authors. None of the reports refers to the MFA manual to clarify how OECD criteria or MFA criteria were to be understood and analysed. The evaluation questions in TORs are rarely an added-value and a guide for framing the analysis and targeting of evaluation on key issues specific to the project or programme. When taken into consideration in evaluation reports, they become a rigid framework rather than a guide for enlightening the analysis of the evaluation criteria. Being not required by the Evaluation Manual to be the vehicle for structuring the analysis and evidencing findings (with further judgment criteria and indicators), they are most often understood as a substitute to evaluation criteria.

Conclusion 6: Reports are typically not structured to compare the evaluation findings and conclusions to the logic of the intervention; the logical framework or theory of change are seriously underutilised and where used are reported on superficially.

Only rare reports present an evaluation matrix elaborated from the project logical framework. Among those few, again a high diversity can be found in the methodology to be applied to elaborate the matrix and ways to use it. With an elusive analysis of achievements on key outcomes of the intervention logic or the theory of change, it is difficult for the evaluation report to provide recommendations that can alleviate underperformance or resolve blockages.

Conclusion 7: HRBA and CCO are marginalised in reports and are not well analysed with respect to the evaluation criteria; they are typically addressed as a separate section rather than being integrated.

The Human-Rights Based Approach is altogether ignored by almost all reports (even those who should have used the 2012 policy framework as a starting point). However, inductively, it is clear that this key approach for Finnish aid is not mastered by evaluators who have, for example, absorbed the HRBA into one or another cross-cutting objective, rather than seeing HRBA as a deep transformation of the approach to development projects.

Cross-cutting objectives are systematically reported on in the evaluation reports as if they were an add-on, separate from the main thrusts of the evaluation itself. They are almost always reported on in a separate sub-section as if they were silos; unfortunately, almost all reports identify that CCOs are not mainstreamed in the field and are not integrated into the project. Reports present CCOs in a brief and superficial section and almost never refer to them for conclusions and recommendations. Most report that the monitoring and evaluation systems do not capture appropriate and sufficient data on CCO's, a systemic weakness that MFA will need to address if its policies are to be executed.

Conclusion 8: Reports do not clearly link interventions to GoF development policy context.

Evaluation reports are difficult to use in the manner explicitly specified in ToRs, to improve congruence of projects with principles and priorities of Finnish aid. They refer to global policy documents without specifying the part of the context to which they are referring.

Conclusion 9: Reports do not enable the reader to judge the value/validity/replicability of the contents of the reports (especially its analysis).

Methodology sections are often little more than basic references to very mundane activities (ex. a document-based search was followed by a field mission, followed by a field briefing, etc.). Only the best reports present a basic stakeholders mapping and/or an approach to sampling respondents or sub-projects or activities (most convincing is the funnel method). Much of the methodology is based on what MIGHT be found in the brief field visit (which is dependent upon respondents' availability, logistical constraints, access to key people who may have moved on, etc.) This approach is indeed pragmatic and reflects widely acknowledged field experience in MFA, but it almost always has a severe incidence on the reliability of findings; these "incidences" are very rarely brought out in the report.

Conclusion 10: Executive Summaries are designed to offer information on the intervention itself, not the evidence to support analysis or the links to GoF policy contexts.

Executive summaries rarely include a table presentation of key findings with related conclusions and recommendations, even if the MFA standards require it. They also do not have a template that would help decision-makers to quickly capture the key elements. Key information (recommendations) is lost amongst very operational data. However, Executive Summaries succeed in most cases to present the main conclusions and recommendations of the evaluation. Although they are required to identify the organisation that should be primarily concerned with the recommendations, Executive Summaries very rarely contain this information.

Even where a table format was used in the Executive Summaries (the case in only a handful of the “highest scoring” reports), very few tables were drafted in a way that presented the links between findings (even key findings) and conclusions, and then conclusions and recommendations. It is therefore impossible for the evaluation manager to check if all key issues or main strengths are fully covered by conclusions and recommendations.

Conclusion 11: The ToRs do not always provide sufficiently clear and unambiguous direction to evaluators, and MFA officials and managers have not provided a means to ensure that ToRs are of sufficient quality (i.e. meet MFA standards and norms) before being published.

The high diversity of ways to respond to TORs and the sometimes vague guidance on specific evaluation concepts and expectations (to name a few) provided by the Evaluation Manual and the Bilateral Programme Manual have failed to some extent to provide clear direction to evaluators on a) what is wanted, b) what the performance expectations of the evaluation are, c) what the standards are against which the deliverable will be judged and d) to what extent key concepts of interest to MFA (ex. HRBA) are to be studied. But it is clear that the officials who prepare ToR, and their supervisors, also are having a great deal of problem in directing the mandates, specifying what they want and how they want it, and then controlling the quality of the deliverables when they are sent to them. The fact that many reports do not provide the answers to the questions stated in the ToR is a sign that MFA officials are likely approving sub-quality deliverables. There does not appear to be any reason to believe that evaluation managers (in the decentralised system) are providing a common framework of understanding of the evaluation guidelines or ensuring a minimum level of conformity with the Evaluation Manual and the many MFA policies and guidelines.

As noted above, the overall quality of ToRs is low, even if one considers that the authors are ministry officials and should be the “masters” of MFA’s processes and standards, with access to their mentors and supervisors as well as unlimited access to documentation and past examples. Beyond their low average score, key issues identified for ToR quality are the relatively poor added-value of the evaluation questions (reflecting many MFA policy preoccupations but less so the specificities of the initiative itself). The Meta-evaluation also found that the ToRs most often did not offer an alignment between the work to be done within the context of the intervention, and the allocation of human resources (profiles, team) to the mandate. In only too few cases, evaluation questions in the ToR conveyed a direction and identified the specific factors (evaluation issues) that truly influenced performance and impacts. Profiles of the expertise required gen-

erally reflect an appropriate mix of experience and professional backgrounds, although often are insufficiently specific. Local consultants are rarely required, losing the opportunity to familiarise further with the context and socioeconomic environment. There was not one example where the evaluation was to have taken place in partnership with the key recipient country executing agency or with another agency of that government, representing a missed opportunity to pass on the capacity for evaluation to recipient countries.

Conclusion 12: Peer review or quality assurance is not perceptible, nor presented.

No reference to peer review or quality assurance was ever stated as being required on the part of the contractor.

Conclusion 13: Resources allocated for the execution of evaluations were often inadequate to enable the evaluator to perform objective and triangulated analysis or to independently develop observations.

The meta-evaluation team always analysed how many days of professional effort could be devoted to the evaluation given the indicated budget. It assumed costs and resource-days for a kick-off meeting where required, travel to and from the country, travel inside the country, research locally, time for in-bound and out-bound debriefings, and time for writing the report, among others. It divided what remained into payment for international and local consultants, and then figured out approximately how many days of effort would be spent on research, at home and in the field.

What it found was that many of the ToRs dealt with complex and multi-faceted interventions with problems, constraints and demonstrated successes, but rarely allowed the evaluators enough time to develop observations and findings. Some reports had annexes that indicated that evaluators interviewed many people, but when the team compares the number of people “met” with the time in the field, it has to conclude that many people were met in a “meeting” mode with its own dynamics and space for freely expressing ideas.

Overall, the level of effort required to “EVALUATE” (and not just describe) what is defined in the ToR, to a depth that is adequate for an evaluation, is not always appropriate.

7.2 Dealing with EQ 3: On the Quality of appraisal reports and ToR/ITT

Conclusion 1: Most draft PDs were not ready to be subjected to an ex-ante evaluation (appraisal).

Based on comments in both the appraisal and evaluation reports, it is clear that a significant proportion of appraisals are, in fact, project design activities with major pieces of intervention design not done. In some cases the evaluation reports identify how the “Appraisal” was wrong in its recommended strategies

and made proposals based on incomplete data and weak analysis. This conclusion is also based on the findings reported on in Section 5 of this document. Appraisal reports identified that the intervention logic, results chain logic, results framework and other core design components were not ready and, of course, many appraisal reports recommend that these be done.

Conclusion 2: The overall quality of appraisal reports, in terms of their compliance with MFA standards and norms, is very low. Out of a possible score of 100 points, the average score given to appraisal reports is only 46.4.

The points awarded and the maximum weighting possible for a section are:

Table 16: Overall scores of appraisal reports

Section	Max score	Score awarded
Preliminaries	7	4.1
Introduction	1	0.45
Context	4	2.5
Description of intervention	5	3.1
Approach, methodology, limitations	15	4.7
Evidence-based findings	15	8.5
"Answers" to issues presented in ToR	10	3.2
Conclusions	15	5.0
Recommendations	15	7.4
Lessons learnt	5	1.8
Annexes	5	3.8
Non-content quality not in evaluation manual	3	1.45
Total	100	46.35

Source: Meta-evaluation team

The table below is the dashboard representation of the ratings give at the main levels of the analysis of appraisal reports. Attention is drawn to the fact that the meta-evaluation created four sections, one of which being the "Main Text" which contains the core analytical elements of the appraisal reports. Significantly low ratings were awarded in the parts of the main text dealing evidence-based findings; answers to issues; conclusions and recommendations; these were the parts of the report what were assigned the greatest weighting, so the overall average score for the "Main Text" is only 36.95 out of a possible 85. Closer inspection reveals that conclusions were particularly rated low, as were answers to issues.

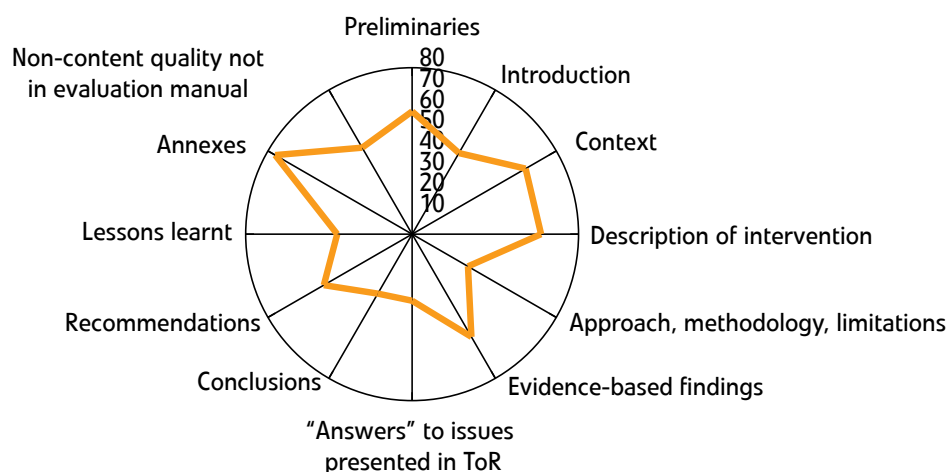
Table 17: Breakdown of scores in each appraisal report

	Preliminaries		Introduction Chapter contains:	Context. This chapter contains:	Based on the existing version of the PD, description of the programme or intervention being appraised, including	Approach, methodology and limitations. The rating should be based on the following:	Evidence-based Findings:	"Answers" or strategic analysis of issues based on findings	Conclusions. The assessment of the performance of the project/programme based on the findings in relation to the set evaluation criteria, performance standards or policy issues	Recommendations. Proposed improvements, changes, action to improve the project design or to capitalise on strengths.	Lessons learned	Annexes. The following annexes, as a minimum, are included	NON-CONTENT QUALITY ISSUES NOT SPECIFICALLY MENTIONED IN EVAL MANUAL BUT REQUIRED FOR META_EVAL	
	A	B												
01 - DDC	5	24,5	0,5	2	2	3	8	5	1	3	0	0	0	29,5
02 - ADPP	7	47,5	0,5	3,5	4	3	14,5	2	2	13	5	5	1	60,5
03 - AWEPA	4	24,5	0	2	3	8	5,5	0	3	3	0	5	1	34,5
05 - BIO FISA	1	58	0	2	3	10	10	6	15	10	2	5	1	65
06 - LAO SNGS	4	51	1	3	5	5	10	5	10	10	2	4	1,5	60,5
07 - LAOS Forest Sector	2	12	0	4	3	0	3	0	0	0	2	0	2	16
08 - Vietnam WSPST 3	4	46,5	0,5	4	4	5	9	0	10	12	2	4,5	2,5	57,5
09 - Vietnam IPP 2	5	56,5	1	4	5	5	11	6	9,5	10	5	5	3	69,5
10 - Nepal RVWRMP 3	6	26	0	3	2	3	8	2	0	8	0	5	1	38
11 - Aid for trade - Central Asia	3	23	1	0	0	5	6	6	0	5	0	5	1,5	32,5
Average Score	4,00	36,95	0,45	2,75	3,10	4,70	8,50	3,20	5,05	7,40	1,80	3,85	1,45	46,35
Max. score	7	85	1	4	5	15	15	10	15	15	5	5	3	100
Score on 100 (%)	57,14	43,47	45,00	68,75	62,00	31,33	56,67	32,00	33,67	49,33	36,00	77,00	48,33	46,35

Source: Meta-evaluation team

This information can be visualised as the “percentage of the points awarded as a function of the maximum that could have been awarded”. That information is represented in the radar graph below. Note that only the ‘annexes’ go over 70%.

Figure 19: Comparison of the scores given in percentage to different sections



Source: Meta-evaluation team

All the sections do not have the same weighting. The majority of the weight (70 out of 100) has been linked to five core sections. The “percentage of maximum” for what are arguably the most important parts of appraisal reports, range from a cluster around 32 (31, 32 and 33) to a high of only 57. These are very low scores. What appears to be “good news” in the radar graph above concerning Annexes is somewhat misleading because “Annexes” only represent 5% of the weighting overall.

Conclusion 3: The ToRs for appraisals are not specific enough to direct the appraisal towards specific requirements of the decentralised manager.

The meta-evaluation found that ToRs tended to use relatively generic terms to indicate what was required, such as “feasibility” (i.e. “is the project feasible?”). Further, the sections of ToR that deal with core content to be developed (on the part of the appraiser) are not contextualised and do not refer to the specific needs and concerns of the MFA; performance standards are not defined insofar as actual levels of detail that should be in the appraisal report (ex: “is logic framework adequate”). It is noted that the appraiser is not mandated to actually refine the draft PD. Those sections that can almost be cut and pasted from MFA manuals scored highest (ex. rationale, purpose and appraisal process). Both of these cored over 80 out of 100 points, while appraisals overall scored rather low at 4.6.

Table 18: Breakdown of scores in each ToR for appraisal

	1. There is sufficient background information to the appraisal provided in the TOR or ITT	2. The rationale, purpose and objectives of the appraisal are clearly described in the TOR or ITT	3. There is an appropriate and sufficiently detailed description of the scope of the appraisal	4. The appraisal objectives are translated into relevant and specific appraisal issues	5. The implementation of aid effectiveness commitments is described	6. The proposed methodology is appropriate and capable of addressing the appraisal questions	7. The appraisal process and management structure are adequately described	8. The resources required for this evaluation are sufficiently described	9. Annexes and structure of the TOR	Total
01 - DDC	3	15	1	15	3	4	5	20	0	66
02 - ADPP	6	15	3	18	3	2	5	15	1	68
03 - AWEPA	2	10	1,5	20	3,5	1	4	15	0	57
05 - BIO FISA	5	12	1,5	25	2	1	4,5	20	1	72
06 - LAO SNGS	3	13	1	28	2,5	1	5	18	1	72,5
07 - Lao Forest	4	12	3	25	1	2	5	12	1	65
08 - Vietnam WSPST 3	2	1	2	10	3	2,5	4,5	17	0,8	51,8
09 - Vietnam IPP2	5	13	2	30	4	2	3,5	20	1	80,5
10 - Nepal RVWRMP3	2	10	0,5	10	0,5	2,5	1,5	10	0,5	37,5
11 - Aid for trade Central Asia	2	15	2	25	2	1	4,5	25	1	77,5
AVG	3,40	12,50	1,75	20,60	2,45	1,90	4,25	17,20	0,73	64,78
MAX	6	15	3	35	5	5	5	25	1	100
Percent	56,67	83,33	58,33	58,86	49,00	38,00	85,00	68,80	73,00	64,78

Source: Meta-evaluation team

The table above is the dashboard representation of all the ratings given to appraisal ToRs.

The table indicates that ratings were high when dealing with material that could easily be extracted from manuals or previous ToRs, such as “rationale, purpose and objectives” (83% of maximum) or the “description of the appraisal process”, rated as 85 percent. Much lower average ratings were given to areas that are particularly specific to the intervention itself or the expression of the needs of MFA. For example, “appraisal issues” was rated at 59 percent and the “implementation of aid effectiveness” was rated at only 49 percent. Particular note should be taken of the very low rating given to methodology and approach (38 percent).

Conclusion 4: The quality of appraisal reports does not directly correlate with the quality of ToRs.

The scores awarded for the core sections of appraisal reports, when compared to the same sections in the corresponding ToRs, do not show a strong positive or negative correlation. Both scored poorly overall, but those section in the higher scoring ToRs did not necessarily result in good quality corresponding sections in the reports (or vice versa). Moreover, relatively good report core sections were produced in a few reports, even when the related ToRs sections scored poorly. One could hypothesize that the key quality variable at play here is the “deliverer”, or appraiser, who may know “what to do and how to do it”, even if it is not spelled out in an instruction (i.e. in the ToR).

Conclusion 5: Appraisals are not being executed as ex-ante evaluations except in the broadest sense.

A concept paper prepared by one of the team members of the meta-evaluation in early 2015 indicated that the way that MFA used ex-ante evaluation was quite different than the practices of other donors. It was noted that the appraisers in the Finnish context actually very often generate the design they are being asked to “evaluate”. As a result, the appraisers are often asked to generate (i.e. plan, design) the very core parts of interventions. It is not clear why this does not take place in the field by other, more “project-cycle” or programme development and management logics, rather than the “assurance” logic which is the core *raison d’être* of appraisal and ex-ante evaluation.

Conclusion 6: Although MFA maintains that appraisals are, in fact, ex-ante evaluations and are subject to the evaluation policies of MFA, the authors of ToR and reports do not adhere to the required processes and structures of evaluation management.

There is a significant difference in terminology and construction of ToR and reports between the two: ToRs speak of “issues” and evaluations speak of “questions”. The evaluation rigour, research approaches and epistemological constructs are not applied in appraisals. Even if they look similar on the surface, the analysis and reporting requirements are very different.

The purpose of Phase Two was to identify any facts, tendencies and variables that could assist MFA in understanding the “Quality” of its development cooperation as a whole.

Conclusion 7: Even if the appraisal is based on the 2012 policy construct, they only deal superficially with HRBA, aid effectiveness, RBM and CCOs, if at all.

Section 1.2 (e) in the assessment grid for appraisal ToRs, looks at the information provided to appraisers in terms of HRBS, gender and other CCOs. Of the 10 appraisals, none had provided any information, and four asked the appraisal to examine the topic. Section 5.4 in the assessment grid for appraisal reports refers to an assessment of the quality of findings related to HRBA and CCO. Of the ten projects, only three described such findings in some detail. None of the others did. Some of the “others” indicated that the design preparedness at the time of the appraisal did not incorporate either a strategy or indicators for monitoring and managing HRBA or CCO. It should be noted that only a few of the 10 projects were appraised using the 2012 policy context, but that does not change the conclusion, especially since even the 2007 context referred to human rights and CCOs.

7.3 Conclusions concerning EQ 4: What evaluation reports reveal about Finnish development cooperation

The following are the key conclusions that can be drawn on the “quality of Finnish development cooperation” from the evaluation reports that were the subject of the first phase of the meta-evaluation. It should be noted that the purpose of this second phase was to identify, from the evaluation reports (but not the evaluation ToR or any appraisal reports or ToR), facts, tendencies and variables that could assist MFA in understanding the “Quality” of its development cooperation. Phase 2 is NOT about evaluating the reports per se. The members of the meta-evaluation team adjusted their analysis to take this in mind; for example, if a report did not have a section entitled “HRBA” but noted that the intervention dealt with human rights under a variety of forms, the ratings were still positive.

Of the original 28 Evaluations and mid-term evaluations, only 18 were considered for the second phase of the meta-evaluation. Of the 18 reports, only eight were commissioned directly by MFA (code 1 in the portfolio analysis) and 10 were commissioned by others (code 2, 3, 4 and 5). All issues were assessed using a grid that used a 5-point rating system. Since the objective is to understand the extent to which one object (the reports) is congruent with the expectations of another object (I.e. Government of Finland policy), no system of weighing between sections was introduced. Each section (ex. Relevance) is given a score of one-to-five, corresponding to the specific rating scale developed for that purpose.

To recover the maximum amount of benefit from Phase 2 analysis, the reader would have to examine in detail the summary grids that were produced. The details contained in a document that was produced by the meta-evaluation team wherein the various comments generated through an “inductive” process would also lead an interested analyst to discover avenues for further research.

The conclusions presented in this section are those that most closely respond to the question: “what do evaluation reports teach MFA about the quality of Finnish development cooperation?”. This question is answered by using the MFA’s own evaluation criteria.

It has proven difficult to identify “best practices” in the sample of 18 projects covered by Phase 2, e.g. projects showing the way ahead, successfully combining HRBA, RBM and dividends of the long-term partnership leaving the partner country in the driver’s seat in order to deliver expected outcomes. These common principles of MFA’s overall policy context were hard to isolate and discern in the projects evaluated. This could, it is hypothesized, indicate that there are systemic forces at play to prevent their integration in project design, implementation and oversight. This meta-evaluation was not mandated to follow-up on that hypothesis.

Conclusion 1: With respect to “relevance”, the evaluation reports indicate that MFA interventions are designed to meet the identified needs of targeted beneficiaries; are very much aligned with national goals and strategies (at least at strategic levels), and have proposed objectives that reflect Finnish aid priorities and policies (at least at strategic levels).

Inductive analysis underlined the criticality of the design phase to the relevance and effectiveness of any intervention, including the initial selection of projects or partners, and then the quality of the project logic through a logical framework or theory of change. In terms of the strategic level of oversight (ensuring relevance, effectiveness and impact specifically, and other criteria generally), the poor quality of the oversight function was often cited (rated 2.5).

Conclusion 2: With respect to “effectiveness” the evaluation reports indicate that expected outcomes from MFA interventions were only partly achieved (or their achievement could not be assessed from the report).

It is interesting to note that evaluation reports tend to consider the effectiveness of an intervention through a component-by-component analysis; if a key component is not going to meet its objectives, or is several years behind schedule, the overall assessment may nevertheless remain positive; what is important here is that a comprehensive perspective is often lost with the compartmentalisation that occurs. What is clear is that the simple-to-execute activities are essentially all completed during the intervention, including things such as training sessions, TA-on-the-ground, and setting up PIUs. But a significant portion of interventions do not have the indicators and monitoring systems in place to allow the evaluation of outcomes (or CCO performance) to take place, so most reports stay at the activity narrative level. Most reports also do not deal explicitly with the extent to which effectiveness is supported through HRBA and CCOs. The meta-evaluation also noted that “effectiveness” is most often judged on the basis of activities/outputs⁴ or intermediate outcomes rather than on higher-level outcomes or impact. There is not one example of where a contribution analysis was used to justify the effectiveness of an intervention. Part of

⁴ In spite of Finland’s interpretation of the term “effectiveness”.

Many interventions do not have the indicators and monitoring systems in place to allow the evaluation of outcomes (or CCO performance) to take place.

The meta-evaluation also noted that “effectiveness” is most often judged on the basis of activities/outputs or intermediate outcomes rather than on higher-level outcomes or impact.

There is not one example of where a contribution analysis was used to justify the effectiveness of an intervention. Results frameworks are stated in lofty terms and indicators, outcomes and impacts are not clearly stated in measurable terms.

There is not one record of an analysis of an oversight committee's specific actions to promote and leverage higher-level outcomes and impacts; what is reported on is the logistics.

No report compares the efficiencies of the resources applied with other possible resourcing strategies.

the problem appears to be that log frames and results frameworks are stated in lofty terms and indicators, outcomes and impacts are not clearly stated in measurable terms; this hypotheses needs to be checked out further with documentation that the meta-evaluation team does not have. Indeed if outcomes are not achieved, the whole rationale and process of design and implementation of the projects must be questioned.

The inductive analysis brought forward a number of interesting avenues to explore further: the MFA's practices concerning aid effectiveness are seen as positive, particularly for leaving the partner country direct its own development (rated 3.1), and for attempts at coordination and harmonisation (rated 2.9). Coordination's relatively good rating may be more related to the frequent recourse to co-financing than operational coordination with other donors intervening in the same field or area. Only few examples of harmonisation (in the sense of a strict use of national management systems) are at hand.

Conclusion 3: When dealing with impact, the Government of Finland cannot rely on evaluation reports to provide it with information on, or the potential for, impact. Evaluation reports just don't deal adequately with impact, even though there are sporadic examples of anecdotal justification provided.

It is understandable that Mid-term evaluations may find it difficult to judge on impact (potential or otherwise) but MTE and final reports fail to analyse the constraints to impact and the actions would be needed (if any) to facilitate the movement from outputs to outcomes and then to impact. There is not one record of an analysis of oversight committee's actions to promote and leverage higher-level outcomes and impacts; what is reported on is the logistics (i.e. how often they met, who sits on the committee, and who chairs it). There is not one example where a report has analysed impact in the light of HRBA or CCOs. The meta-evaluation team recognises that mid-term revaluations do not typically cover impact, and so the comments above take that into account.

Conclusion 4: Efficiency is not well analysed in reports, so the Government of Finland cannot judge the extent to which its development cooperation programme is efficient, other than at the level of a comparison between budgets and disbursements over time.

Analysis is not offered in the reports of the extent to which the strategies selected for the intervention were the most efficient, or the extent to which efficiency was influenced by the use of a particular mix of instruments or strategies, for example. Most reports do not distinguish between the various types of efficiencies (ex. transformational, financial, economic, time, resource, etc.). Where there is reporting, it tends to deal with budgets and disbursements. There are only a few reports that mention that other intervention strategies would have worked better. No report compares the efficiencies of the resources applied: the use of TA, for example, is always assumed to be an efficient strategy to achieve innovation objectives, and training is always assumed to lead directly to an

increase in the intervention's ability to generate outcomes. There is no prima facie reason to believe either hypothesis in any given context.

Conclusion 5: With respect to sustainability, the Government of Finland can generally only count on anecdotal evidence to judge the sustainability of its interventions. The reports suggest that most MFA interventions are either wholly or partially sustainable.

Overall, the prospect for sustainability in MFA interventions cannot be defined except in the specific context of the interventions themselves. Terms such as “the project should become sustainable over time”, or “there is potential for sustainability” are commonly used but the meta-evaluation Team never saw a business case, a detailed sustainability analysis, a cost-benefit or a cost per unit analysis. It never saw a multi-year financial analysis or a scenario-by-scenario analysis with different assumptions concerning sustainability. If an output was going to be managed through a participatory or collectivity-based process, it was almost always assumed in the reports that it would be sustainable. Overall, each report dealt with sustainability, but generally in a less-than-thorough manner. Reports have not analysed sustainability from the perspective of HRBA or CCOs.

Conclusion 6: Aid effectiveness is not a key factor in intervention design or management, either through mainstreaming or other management strategies.

The vast majority of reports assessed do not deal with the issue directly; overall, and the meta-evaluation had to dig into the text to extract information on ownership, mutual accountability, partnerships, alignment and harmonisation. Monitoring systems do not follow the roll-out of aid effectiveness measures and therefore cannot report on them.

Conclusion 7: The Government of Finland's policy frameworks covering such all-encompassing policies as the application of HRBA or the need to take into account and manage various CCOs are not being respected generally.

Very low scores were given to the extent to which the intervention addressed HRBA and the extent to which the programme or project support addresses inequality. Slightly higher ratings (but still less than 3.0) were given for the extent to which the intervention contributes to gender equality and to the extent to which climate change objectives were taken into account. Coverage of the CCOs is not particularly encouraging (climate sustainability 2.9; gender equality 2.6; reduction of inequality 1.9) and the inductive analysis does not find much mainstreaming in these domains; rather, a silo approach is used almost exclusively.

There were considerable problems encountered in analysing sustainability itself.

There is an urgent need to understand why evaluations are not better managed; there is a need for decentralised officers and their supervisors to better master the logic and research rigour that evaluations need to have.

Conclusion 8: The MFA's efforts to integrate risk management and Results-based Management into its programmes have not been very effective.

Evaluation reports just do not deal with risk mitigation or strategies at all (rating given was 1.4 out of 5); the majority of reports mention RBM but are quick to state that the intervention is not focussed or structured that way (1.8 out of 5). The meta-evaluation discovered that that rating is artificially (and wrongly) inflated by projects managed by UN agencies.

7.4 Concerning avenues to examine for evaluation and appraisal-related capacity improvement

Unlike the 2012–2014 meta-evaluation (refer to Chapter 6 of the previous meta-evaluation report), this one did not have the mandate to interview MFA officials or to engage in questionnaire development and administration with the objective of studying capacity gaps within MFA. The few comments that are offered are based on the observations, findings and analysis carried out within this mandate alone.

It is interesting to note that this meta-evaluation believes that many of the needs identified in the previous meta-evaluation with respect to capability/ability are still relevant. This meta-evaluation observes that:

1. There is an urgent need to understand why evaluations are not better managed; the assumption being that if they were useful to MFA then they would have a much higher level of quality overall.
2. There is clearly a need for decentralised officers and their supervisors to better master the logic and research rigor that evaluations need to have
3. Much better instructions need to be given to appraisers and evaluators (within ToRs but also with much clearer policy, process and functional frameworks that are supported by reference material that is clear and can be used as performance specification references. Thus armed, they become enabled through a performance framework that will define the quality of their deliverables.
4. The evaluation manual is an appropriate and adequate reference for understanding the function of evaluation within MFA; it does not provide a sufficient level of comprehensive guidance.
5. There is an urgent need to significantly raise the level of the QA that is provided internally to officials who are involved in evaluations or appraisals. Supervisors are the frontline resources for doing this and so they need to be able to “supervise” (especially control, direct and assist functions) their employees. If this is not possible, for whatever reason, the assist function can be outsourced and the control function can be tightened up internally (QA panels for example). In this light the help desk function, begun in 2009, can be a model again if there is a requirement to use it as a QA measure.

6. A significant number of evaluation and appraisal reports were accepted even if their quality was clearly below standard. This phenomenon needs to be studied and steps taken to improve the critique and gate-keeping abilities of officials. The meta-evaluation believes that the issue is not only one of training and information provision, but is due to a systemic weakness. The standards for managing evaluation processes (and those of the evaluation function itself) should not be any lower than those for managing invoices or contracts; the processes and functions are either managed as per MFA requirements, or they are not.

Beyond the above, this meta-evaluation has observed that there are fundamental weaknesses in the management and execution of appraisals (ex-ante evaluations) within MFA. The concept paper prepared for EVA-11 in 2005 contains a number of observations and recommendations, but the key ones are a) the positioning of the appraisals within the project cycle, and b) (related to the first), the extent to which Programme Documents contain the core elements of project design before they are approved.

If appraisals are to be conceived as ex-ante evaluations then the direction given to them through the ToRs and the bilateral project and evaluation manuals needs to be significantly improved. At the moment appraisals are better classified as “project design” activities than an assurance strategy.

Those core elements include (partial list) a comprehensive Theory of Change proposal with intermediate results, a set of assumptions for progress and a risk/mitigation analysis. Core elements also include a comprehensive (although not a final detailed version) version of an RBM-based performance framework and implementation plan which integrates HRBA and CCO objectives. An evaluability matrix should be in the PD including indicators and monitoring strategies and plans that cover each part of the logic chain.

8 RECOMMENDATIONS

This meta-evaluation was not mandated to examine the cause of any findings that it might bring to the surface, so it will not be a surprise that the vast majority of recommendations do not propose ways to “fix problems”, especially at strategic levels. They offer avenues to explore in order to uncover the causes of weaknesses that were identified in this report, and they offer rather operational recommendations based on the operational findings. For simplicity and ease of reading, and unless stated otherwise, it is assumed that all of these recommendations deal with decentralised evaluations and appraisals, either commissioned by MFA or not, and not those carried out by EVA-11.

Based on the conclusions found in this report, the following recommendations are proposed. The term “operational” is meant to convey that some units and departments deal directly with the planning, programming and implementation of interventions while other units and departments deal with policy, administrative support and other (non-operational) functions. The term “operational” is directly borrowed from the terminology used in the productive sectors.

A) Strategic Level Recommendations affecting the quality and execution of evaluations and appraisals

Recommendation number	Statement of Recommendation	To whom is recommendation addressed	Relative priority of recommendation
1	MFA should put in place mechanisms, including those for monitoring and quality control, to help better enforce its own policies concerning the management of bilateral cooperation. Specifically, it should the basic constructs of its Bilateral Manuals that require that the core of intervention design be a logic analysis (developed via a results-chain analysis that presents how an intervention will contribute to outcomes that will meet the needs of beneficiaries). This should be done, as per MFA policy, using Logical Framework, Theory of Change or similar approaches; the design should be crafted using RBM and HRB approaches and correspond to all the requirements of the Bilateral programme manual.	MFA executives	1
Justification and expansion: These mechanisms could, for example, be soft (ex. improved management supervision); policy and guidance-based (ex. development of very clear performance requirement statements and the means to access them); systems-based (information systems that support the evaluation function and assist the MFA officers, evaluators or appraisers) and/or assurance-based (control frameworks tied to individual performance appraisal systems. This recommendation should apply to all cooperation mechanisms and instruments including trust funds, budget support and financing global or regional collective actions.			

Recom- mendation number	Statement of Recommendation	To whom is recom- mendation addressed	Relative priority of recommen- dation
2	An “uptake” analysis should be done on a managerial research basis (i.e. with rigorous analysis and an appropriate analytical approach based on the accountability framework of MFA managers), in order to identify, within the 2016-and-beyond context, the benefits that MFA managers feel they could and should extract from the evaluation function. The analysis would also identify if, or how, the evaluation function in MFA should adapt itself in order to provide decision-makers with the information and analysis they feel they need.	MFA executives EVA-11	1
Justification and expansion: “Uptake” analyses have been generated in many donors (including the EU-DEVCO) and are regularly used in large, multi-faceted and complex organisations where they are often integrated in to strategic business case analysis. The analysis does not have to cover all of MFA but could start (as a pilot) with a few priority areas. An MFA operations champion would be very useful to ensure buy-in.			
3	Develop, using MFA policies, norms and standards as a base, a rolling meta-evaluation function that would provide real-time information on the effectiveness, impact and sustainability (at least at first) of Finnish development cooperation through all forms of evaluation-based deliverables.	EVA-11 MFA executives	3
Justification and expansion: This recommendation is an expansion of the “coverage” analysis that was noted in the ToR for this meta-evaluation (see Chapter 9). The key part of this recommendation is that not only is it important to ensure that all evaluations and appraisals take place in a timely manner as prescribed by MFA policy, but it is important that senior managers are constantly aware of the strategic significance of the content of those efforts. By “rolling”, the Team implies that meta-evaluations should not take place on a fixed-period basis but constantly; it is recognised that some version of this form of “developmental evaluation” has been integrated into every major model of knowledge management; strategic management; landscape strategies development, and open organisations/open systems. The Team’s research shows that this will not be simple and will likely require important changes to sub-systems. Without putting into question the independence of the evaluation function in MFA, EVA-11 and KEO should jointly develop the performance requirements of meta-evaluation-based reporting so that the MFA may be in a position to communicate the effectiveness and impact of its development cooperation investments.			
4	Based on the conclusion dealing with the poor overall ratings given to appraisal-related documents, MFA should change the role of appraisals so that they take place considerably later on in the project cycle. Draft PDs should be in a near-complete state and meet minimum content and design standards before being subjected to the critique that can only be rendered through an appraisal. Refer to the concept paper prepared in 2015 on ex-ante evaluation for a clearer distinction between “due diligence” and “ex-ante evaluation”.	MFA executives KEO	2
Justification and expansion: Ex-ante evaluation is a very powerful tool for assurance purposes but it is not particularly effective at ‘programme design, particularly when the ToR and MFA policy imply that the role of the appraisal is not to undertake changes to the draft PD. Appraisals are part of “due diligence” and are there to provide “managerial assurance”.			

Recom- mendation number	Statement of Recommendation	To whom is recom- mendation addressed	Relative priority of recommen- dation
5	Based on the conclusions in chapter 7.3 (i.e. What evaluation and appraisal reports reveal about Finnish development cooperation), MFA's operating divisions should critically seek to understand the causes for the weaknesses found in the relevance, effectiveness, efficiency, impact and sustainability of all of its interventions. The results of this meta-evaluation can help managers to pinpoint areas of research and place those areas within a broader context. It is suggested that this be an internal analysis, and that it should take place using the operational concept of a learning organisation so as to develop internal capacity and ownership. As part of this recommendation, MFA should include in its ToRs a reference to the obligation of evaluators and appraisers to specifically link the interventions to Finnish development cooperation policy.	MFA executives EVA-11 Relevant operational units and departments	1
Justification and expansion: This should be either an ongoing responsibility of the policy divisions of MFA or could become a separate research-based activity.			
6	Based on the conclusions related to the very uneven application of the HRBA policies of the Government of Finland, MFA should undertake an internal assessment (perhaps in the form of a management audit) of the practices associated with that HRBA policy and the objectives and outcomes that were set for it. A major focus of that assessment should be on identifying the causes for the weak implementation; it should also carry out a CAPABILITY GAP analysis to ensure that all the parts of MFA concerned with development cooperation have "WHAT IT TAKES" to implement HRBA policy.	MFA executives EVA-11	1
Justification and expansion: The conclusions arrived at in this meta-evaluation point to a situation where one of MFA's flagship policies is just not being applied. Since Finland in particular (and other Nordic countries generally) have adopted a focus on Human rights at the highest levels, action should be taken now to understand what is not working and put in place "what it takes" to get it to work.			
7	Based on the conclusions dealing with effectiveness and impact, and taking into account the small average budgets allocated to the interventions, MFA should undertake a rigorous analysis of the effect of the fragmentation of Finnish Aid on the level of contribution that it can produce (towards outcomes).	MFA executives EVA-11	1
Justification and expansion: This issue is serious for impact and effectiveness, and as far back as the 2003 DAC peer review, this issue has been on the table. It is acknowledged that there may be political or develop, mental reasons to fund an intervention at all level, but the large number of countries and projects, coupled with the size of the budget for Finnish Aid, warrants a cost-benefit study at the very least.			

B) Operations level recommendations affecting the quality and execution of evaluations and appraisals

Recommendation number	Statement of Recommendation	To whom is recommendation addressed	Relative priority of recommendation
8	Insist on visible evidence that the reports go through a QA process.	EVA-11	2
Justification and expansion: Findings and conclusions show that many reports do not reflect the contents of their ToRs. They also are not based on evidence and respond only weakly to the issues and answers that are found in the ToR. A QA system would help the evaluator or appraiser to identify where there are incoherencies and where the reports have not taken relevant MFA policy and guidance into account.			
9	Significantly tighten methodology requirements for inception reports (the client should approve a detailed methodology that included the data sources, indicators, tools for data collection and analysis, sampling methods, interview guides and interview notes).	EVA-11	2
Justification and expansion: The meta-evaluation team found that there was often no obligation to prepare and present inception reports; many contracts were not based on ITT, TOR or proposals either. Based on those findings, the Team proposes a recommendation that forces someone, at some point in time, to explain to the client (MFA) what they are going to do and how they are going to do it. The inception report need not be more than a few pages long and should be almost ready when the contractors arrive on site to brief the Embassy staff. If there are no proposals or inception reports, it is clear that the MFA officials have no basis on which to question the quality of the deliverables.			
10	Insist that evidence be specifically provided to support all findings. One possible model for doing this is the one used by EU-DEVCO (evaluation matrix in annex); another is to provide a box with key findings at the end of the analysis of each OECD and MFA evaluation criteria.	EVA-11	1
Justification and expansion: The MFA guidelines direct that evaluations and appraisals are to be "evidence-based". They do not clarify what that term means and how MFA wants it applied. This is not a question of freedom of research on the part of contractors; it is a policy requirement. And means must be provided to specify performance expectations.			
11	Better define the expectations of evaluations and appraisals with respect to the three Finnish criteria coherence, Finnish value-added and aid effectiveness criteria.	EVA-11	3
Justification and expansion: The conclusions point to a wide variation in the interpretation of key terms by authors (MFA officials as well as contractors). Where these three criteria are reported against (very small number of reports), they are not well treated as evaluation criteria and more closely resemble CCOs. It is important that some significant level of coherency between documents exist. The M&E systems, for example, will not be able to gather comparable data if different terms are applied.			

Recommendation number	Statement of Recommendation	To whom is recommendation addressed	Relative priority of recommendation
12	Clearly define the expectations for reporting and assurance related to "value for money", as part of the effectiveness thrust of GoF for 2014-2018. This will require a much more robust RBM platform on the ground and a requirement to install adequate M&E systems at the individual project, portfolio and "selected sub-sets" of interventions levels.	EVA-11 Relevant operational units and departments	2
Justification and expansion: None required.			
13	The MFA should provide a tool that officers can use to allocate resources (budgets) in line with the complexity of the work they are asking to be done.	EVA-11 Relevant operational units and departments	3
Justification and expansion: None required.			

C) Recommendations influencing the evaluation function, process of cycle

Recommendation number	Statement of Recommendation	To whom is recommendation addressed	Relative priority of recommendation
14	Develop a guidance document that specifically addresses the acceptable content of reports, and provides norms and standards for them. This document would expand considerably on the evaluation manual and tie-in appraisals as an ex-ante evaluation.	EVA-11	2
Justification and expansion: None required.			
15	Modify slightly the assessment grids prepared for this meta-evaluation and insist that officials use them to judge the quality of the deliverables (reports) they receive. Internally, officials and supervisors can use the ToR assessment grids to double check the structure, content and quality of TOR.	EVA-11 Relevant operational units and departments Embassy staff	1
Justification and expansion: None required.			

D) Capability or capacity-related recommendations

Recommendation number	Statement of Recommendation	To whom is recommendation addressed	Relative priority of recommendation
16	Based on the conclusions dealing with the quality of ToRs for both evaluations and appraisals, specific training should be given to officials and supervisors on the nature and construct of evaluative analysis applied to MFA interventions. The exact needs (gaps) should be based on a capability analysis so that the focus is not only on the individual but on the systems, resources and authorities that are in place (or need to be).	EVA-11 Relevant operational units and departments	1
Justification and expansion: None required.			
17	MFA officials should be enabled to assess the quality of assurance-related documents that integrate HRBA and CCO into the management criteria (including OECD/DAC and specific MFA). This is fundamentally a question of design policy.	EVA-11 Relevant operational units and departments Human relations	2
Justification and expansion: Conclusions and findings point to a very low level of analysis and reporting on HRBA and CCOs. The Team has to hypothesize "why it is that MFA officials would accept the deliverables without these constructs in them?". The ability to critique the deliverables requires not only individual training sessions, but perhaps mentoring, best cases, hands-on experience in drafting these types of contents, systems and tools to help analyse, etc. This should become a factor in employee personal appraisals.			
18	The ability of MFA officers to truly understand and critique evaluation and appraisal (ex-ante evaluation) findings and conclusions, as well as monitoring and other reports, in the light of the centrality of the logic of specific intervention (through a log frame or Theory of Change, for example) should be significantly improved.	EVA-11 Relevant operational units and departments Human relations	2
Justification and expansion: This is an ability that they need in order to carry out their core functions and goes to the heart of strategic management.			

9 DEALING WITH EQ 2: “WHAT IS MFA’S EVALUATION COVERAGE (COMPARISON OF EVALUATION PLANS AND REALIZED EVALUATIONS)?” – CONCERNING THE ESTABLISHMENT OF AN EVALUATION COVERAGE SYSTEM

At the start of the meta-evaluation it was agreed with EVA-11 that the task of evaluation coverage analysis was not expected to provide a full analysis but rather to indicate ways in which evaluation coverage can be analysed and assessed in the future. The background for the need to have some control over what is evaluated and by whom at the MFA is the internal norm of EVA-11 from February 2015 which states that all funding decisions have to be evaluated at one point of time. The norm is not retroactive however; therefore earlier funding decisions are not necessarily bound by it. The meta-evaluation team was provided with evaluation plans of MFA regional departments since 2011, and one list of realized evaluations (2010–2011).

The meta-evaluation team proceeded to produce a suggestion for the format of a database organised by MFA regional and thematic units enabling the comparison between planned and realised evaluations. Besides indicating the country or region where an evaluation was planned to take place, the format considered as possible relevant factors the implementing modality (on the axis from bilateral to multilateral), name of the intervention (project), the sector according to OECD-DAC CRS codes, the type of evaluation (appraisal, MTE, evaluation etc.), the year the evaluation was expected to start and the year the report was delivered. The format was inserted in a calculus spreadsheet (Excel), which would later, once completed with information, make it possible to calculate comparisons and correlations between the units and the different factors. The implementation modality was included as potential factor based on the hypothesis

that in multilateral interventions the organisation of an evaluation could be lengthier and more cumbersome than in bilateral projects.

An intent was made to fill in the spreadsheet on the basis of the information of the evaluation plans starting from those in 2012, which then was contrasted with available information on realised evaluations (those object of the 2012-2014 meta-evaluation and the current one). The earlier evaluation plans were not included because there were contradictions in plans within one single year and the information in evaluation plans had significant gaps. It may be worth noting here the observation that the evaluation plans for 2014 and 2015-2016 were of much precise nature than those from earlier years.

However, the database format is not attached to this report as annex because the available information was too fragmentary. There were several reasons for this. First, the meta-evaluation has not been able to determine the real representativeness of the evaluation plans; that is, with the means available for a meta-evaluation it cannot be known what percentage of total number of projects/funding decisions the evaluation plans represent. Second, particularly in earlier years the names of projects used in the evaluation plans were “nick-names” from units’ internal use, such as, for instance, “Mekong water project” or “Kenya rural development project”, making it impossible to know which of the sometimes several same-sector projects the plan was directed to. Third, the two meta-evaluations had a significant number of evaluation reports which were not in the evaluation plans delivered to the meta-evaluation team, and fourth, a meta-evaluation does not have the means to know beyond its given portfolio which evaluations in the plans effectively have been carried out. Summing up factors 1 and 4: there is no information on what is the mathematical relation between the available sample and the total universe, which is unknown for a meta-evaluation.

THE EVALUATION TEAM

Robert N. LeBlanc

Mr. LeBlanc holds an MBA in International Trade and Commerce. He has over 40 years of experience in development cooperation and has been the director of many complex evaluations, including the evaluation of the 3C provisions Maastricht Treaty, over forty regional or country level evaluations. With these experiences, he is recognized as a research expert in the domain of organisational and multi-organisational capability development in the pursuit of outcomes and strategic results. He has developed private-sector models of capability development for use in international cooperation and has been mentoring national teams on their application. With over 10 years of experience in meta-evaluation and analysis for development cooperation policies and strategies, he has been the Team Leader in this meta-evaluation. Notably, he has led and managed strategic evaluations the European Commission, USAID, Dfid, Sida, the AIDB and the World Bank. He was the only person ever to evaluate, for Cuba, the robustness of its public service. He has also researched and advised development cooperation policies and strategies and assessed evaluation capacity for various donor agencies.

Maaria Seppänen

Ms. Seppänen, PhD, has a background in development geography and development studies. She also holds a European MA in human rights and democratisation. Ms. Seppänen has 30 years of experience in development-related work and has worked in the MFA as advisor from 2002 to 2005, and from 2006 onwards she has been working as a senior consultant on a number of evaluation assignments for the MFA including meta-analysis of Development Evaluations 2007-2008 as well as the EC. Furthermore, in her quality of adjunct professor in development studies, she regularly provides courses at the university level on topics related to development cooperation. Her specialisation is in the larger governance sector and cross-cutting issues (gender and social equality, democracy, fight against corruption, and human rights).

Max Hennion

Mr. Hennion holds the Master of Economics and PhD of Geography. He has over 25 years of development experience. He is highly experienced in evaluation of development cooperation interventions particularly in the transport and environment sectors. During the past 15 years, he has been involved in a number of complex evaluations at the programme, sector and country/regional strategic level commissioned by the African Development Bank, the UNFPA, and the EC etc. Mr. Hennion's is also an expert in formulation of sector and thematic strategic framework and development cooperation policy.

Keitaro Hara

Mr. Hara is a consultant/analyst with 6 years of professional experience in the field of development cooperation. He has been involved in various evaluation-related research assignments, where inter alia he has reviewed and analysed evaluation systems of bilateral and multilateral development organisations. Mr. Hara is well versed in the evaluation principles and standards, their institutional settings, the evaluation practices, and the current evaluation trends.

Annegrete Lausten

Ms. Lausten is an expert in monitoring & evaluation, review, and research and has 15 years of professional experience in these areas. As Chief Consultant for Danish Management, she has designed M&E and survey methodologies, conducted result-based monitoring, and managed development interventions for a number of assignments commissioned by various development organisations such as MFA Finland, the European Commission, Danida, GIZ, and the World Bank etc. Ms. Lausten has also played a key role as Project Director and QA Expert in these projects. Furthermore, she has extensive experience in the energy and ICT sectors and crosscutting issues.

Dietrich Busacker

Mr. Busacker has over 25 years of professional experience in development cooperation and is Managing Director of ECO Consult. He has profound expertise and experience in all aspects of evaluation. He has been involved in different types of evaluation assignments such as strategy and policy evaluations, impact evaluations, and project and programme evaluations commissioned by various development organisations such as, the European Commission, GIZ, WWF and MFA Finland. His typical responsibility covers development of evaluation methodologies and frameworks, implementation of M&E, peer review, quality assurance, institutional analysis and training. Mr. Busacker is particularly specialised in the environment, education, and rural development sectors.

ANNEX 1: TERMS OF REFERENCE



MINISTRY FOR FOREIGN AFFAIRS
OF FINLAND
EVA-11 Mattila Ilona

TERMS OF REFERENCE
EVALUATION
5.10.2015
V 0.1

UH2015-014383

UHA2011-006988, 89887901

Meta-evaluation of Project and Program Evaluations in 2014-2015

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 quality 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 2014 and August 2015. The evaluation will be based on the assessment of the decentralized evaluation reports, appraisal reports and corresponding Terms of References (ToR) documents.

Meta-evaluation can provide a clear account of the evaluation function of Ministry for Foreign Affairs of Finland (MFA) during a certain period of time by classifying decentralized evaluation reports by commissioner, country, sector etc. and by assessing the quality of the reports. Meta-analysis of decentralized evaluations can also bring together otherwise scattered evaluation findings on the results of development cooperation projects and programs 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. PURPOSE AND OBJECTIVE OF THE EVALUATION

The purpose of the meta-evaluation is twofold: first, the meta-evaluation helps the MFA to improve the quality of evaluations, 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 evaluation is expected to 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-evaluation 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 the quality of different decentralized evaluation reports and related planning documents. It will also draw an overall picture of the evaluation portfolio in 2014-2015 and assess the evaluation coverage in 2013-2015. Second, it 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 2012-2014 in order to find trends, patterns and changes.

In order to enhance the long term utility of Meta-evaluations they will be carried out annually and the requisite assessment tools will be institutionalized.

3. SCOPE

The meta-evaluation will be carried out in two phases. The first phase will concentrate on the quality of appraisals, evaluation reports and their corresponding ToRs. The quality assessment tools used in the previous meta-evaluation will be further developed in the beginning of this phase. During the first phase the meta-evaluation will produce an overview of the quality of MFA's decentralized evaluation activities classified by countries, sectors, budgets, evaluation types, managing units of MFA, consultant companies etc.

The meta-evaluation 2014-2015 will also start a systematic assessment of MFA's evaluation coverage, i.e. identifying if there are projects funded by MFA that have never been evaluated, by comparing annual evaluation plans and realized evaluations. This assessment will start from the year 2015 and will be continued in future meta-evaluations. The current evaluation norm obliges all development funding to be evaluated at some point. However, this norm came into effect in early 2015 and does not apply projects or programmes prior to that.

The quality assessment of the evaluation reports (mid-term evaluations, final evaluations, ex-post evaluations and impact evaluations) will include all decentralized evaluation reports conducted between September 2014 and August 2015, their corresponding ToRs, ITTs and Inception Reports if they are available. It will assess the quality of the reports and their ToRs applying the OECD/DAC evaluation criteria. During the quality assessment also a comparison of the quality between MFA commissioned evaluations and evaluations commissioned by MFA's partners will be made. A selection of reliable evaluation reports will be made based on the quality assessment and only selected evaluation reports will be included in the summative meta-analysis carried out during the second phase of the meta-evaluation.

The appraisal reports will be analysed separately from the other evaluations and they will not be included in the summative meta-analysis of Finland's development cooperation. The quality assessment will be made to appraisal reports conducted between January 2013 and August 2015 and their corresponding ToRs and Invitation to Tenderers (ITT). Possible management reviews will be analysed as the appraisal reports. In addition, the reasons to commission a management review instead of an evaluation will be analysed.

The second phase of the meta-evaluation will provide a synthesis, that is a summative meta-analysis of reliable evaluation findings on Finland's development cooperation verified against the OECD/DAC evaluation criteria and demonstrate how Finnish development policy goals have been achieved. The summative meta-analysis will utilize data driven inductive analysis (see grounded theory), i.e. it will sum up the major issues evident in current development cooperation emerging from the decentralized evaluation reports. The synthesis will also conclude what are the main reasons for success or challenges in development cooperation projects and programs and what are the lessons learned.

The second phase will also provide information to the extent possible on the quality of entry of MFA's development cooperation projects based on the appraisal reports.

4. EVALUATION QUESTIONS

Phase 1:

1. What is the quality of MFA's decentralized evaluation portfolio (evaluation reports and their corresponding ToRs) based on the OECD/DAC evaluation criteria in 2014-2015 classified by countries, sectors, budgets, evaluation types, managing units of MFA, commissioner, consultant companies etc.?

- Is there a difference between the quality of MFA commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?
- 2. What is MFA's evaluation coverage (comparison of evaluation plans and realized evaluations)?
- 3. What is the quality of the appraisal reports and their corresponding ToRs?

Phase 2:

1. What can be said about the quality of Finnish development cooperation based on the reliable decentralized evaluation reports, and related planning documents by each OECD/DAC criteria.
2. What is the quality at entry of Finnish development cooperation projects and programs based on the appraisal reports?
3. What are the reasons to commission a management review instead of an evaluation (if possible)?
4. 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. An assessment tool developed during the previous meta-evaluation will be further developed with EVA-11 and used in this meta-evaluation.

The main sources of information will be the evaluation reports (appraisals, mid-term evaluations, final evaluations, ex-post evaluations, impact evaluations, and possible management reviews) and their corresponding ToRs as well as Development Policy Programme documents, guidelines, earlier meta-evaluations and other centralized evaluations, Government Reports to the Parliament and administrative in-house norms.

The evaluation team is expected to cross-analyse the evaluation reports 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.

6. EVALUATION PROCESS AND DELIVERABLES

The evaluation consists of the following phases and will produce the respective deliverables. A new phase is initiated only when all the deliverables of the previous phase have been approved by EVA-11. The reports will be delivered in Word-format (Microsoft Word 2010) including all the tables and pictures. The tables and pictures will also be delivered separately in their original formats.

Phase 1:

I Start-up meeting and a work shop

The purpose of the start-up meeting is to discuss the entire evaluation including evaluation approach, practical issues related to the evaluation, reporting and administrative matters.

The purpose of the work shop is to discuss the methodology of the meta-evaluation and develop the assessment tools further together with EVA-11.

The start-up meeting and the work shop will be organized by EVA-11 after the signing of the contract and they will take two days. The whole evaluation team must be present in person in the start-up and work shop meetings.

Deliverables: Assessment tools

II Overall Description of Evaluations and Test of Assessments Tools

This phase will produce an Inception report which includes the overall description of MFA's evaluation portfolio and finalization of the assessment tools. The assessment tools will be tested on five evaluation reports in order to ensure their usability. The approach, methodology and sources of verification will be explained in detail, including the methods and tools of analyses, scoring or rating systems, example figures and tables, and alike.

The inception report will be kept concise and will not exceed 25 pages (annexes excluded). It will also suggest an outline of the final report.

Deliverables: Inception report.

III Quality Assessment of evaluations

After EVA-11 has approved the inception report the evaluation team will carry out a quality assessment of all evaluation reports and select which reports will be included in the analysis of the development cooperation.

During the quality assessment the evaluation team is expected to compare the quality between different evaluations classified by evaluation type, commissioner etc.

Phase 2:

IV Meta-analysis and Reporting

The meta-analysis will combine statistical and qualitative methods when analysing the selected evaluations. Limitations of statistical analysis must be recognized and explained.

One possible approach to the qualitative analysis of emerging issues is inductive approach:

"The idea for using an inductive approach in meta-analysis is to (a) condense raw textual data into a brief, summary format; (b) establish clear links between the evaluation objectives and the summary findings derived from the raw data; and (c) develop a framework of the underlying structure of experiences or processes that are evident in the raw data." (David R. Thomas, American Journal of Evaluation: <http://aje.sagepub.com/content/27/2/237.abstract>)

The draft final report will be kept clear, concise and consistent (max 40 pages + annexes). The report will contain the evaluation findings, conclusions and recommendations concerning the quality of the evaluation reports and evaluation capacity of the MFA. They should be logical and based on verified evidence. In addition, the draft final report will contain evaluation findings and conclusions concerning the Finnish development cooperation based on the meta-analysis. However, the meta-analysis does not form an adequate basis for recommendations concerning the Finnish development cooperation and therefore such recommendations will not be made. The evaluation team must pay extra attention to visualization of final data and results i.e. the format of statistics etc.

When the draft final report is ready it will be subjected to a round of comments after which a validation seminar/work shop will be held in Helsinki. The purpose of the seminar is to validate the results and discuss the evaluation with relevant stakeholders. The evaluation team must be at present in person and prepare a short presentation of the evaluation for this seminar. The draft final report may also be subjected to an external peer review of internationally recognized experts. The comments and remarks of the peer review will be anonymously made available to the evaluation team.

A public presentation will be held as a Webinar session when the report is finalized and the MFA has prepared a management response for the evaluation.

Deliverables: Draft final report, validation seminar presentation, and a public Webinar session.

The final report will be finalized based on the comments and discussion raised in commenting round and validation seminar. The final report must include abstract and executive summary in Finnish, Swedish and English as well as a summary matrix in Finnish and English. The consultant is responsible for the translations. The layout of the final report must be according to the writing instructions and template provided by EVA-11.

Deliverables: Final report, account of quality assurance and interim evidence documents.

The MFA requires access to the evaluation team's interim evidence documents, e.g. completed matrices, although they are not expected to be of publishable quality. All confidential information will be handled properly.

7. EXPERTISE REQUIRED

The Framework agreement contractors are invited to suggest a team of one KEH-1 level Team leader and 2 KEH-1 or KEH-2 level experts for the meta-evaluation. Successful conduct of the meta-evaluation requires a profound understanding and experience of international development policy and cooperation as well as conducting development policy/cooperation evaluations and knowledge on meta-evaluations and their methodology. Some of the documents are in Finnish and therefore a good command of Finnish language is required from one of the team members. Each team member must have fluency in English and at least Master level education. The minimum requirements and evaluation criteria are indicated in the Invitation to tender letter and the cv-form.

8. BUDGET AND TIMETABLE

The meta-evaluation will not cost more than 200,000€ (VAT excluded). Therefore a price tender is not needed. A detailed budget according to the prices of the framework agreement will be included in the mini tender.

All reports are subject to the approval by EVA-11 and the payments will be made only after the reports have been approved.

The tentative starting time of the evaluation is October 2015. The whole evaluation team must participate in person in the kick off and work shop meeting in Helsinki in November. Preliminary findings of phase 1 must be available no later than in early December and preliminary findings of phase 2 must be available no later than in the end of December. The whole evaluation must be ready no later than the end of January 2016.

9. MANAGEMENT OF THE EVALUATION

Development Evaluation Unit (EVA-11) will be responsible for the management of the evaluation. EVA-11 will work closely with other units and departments of the MFA during the evaluation process.

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. The evaluation team does not represent the MFA of Finland in any capacity.

As part of reporting process, the Consultant will submit a methodological note explaining how the quality control was addressed during the evaluation. The Consultant will also submit the EU Quality Assessment Grid as part of the final reporting.

The consultant will attach Quality Assurance expert(s) comments/notes to the final report, including signed EU Quality Assessment Grid, as well as a table summarizing how the received comments/peer review have been taken into account.

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 the “Creative Commons” license in order to promote openness and public use of evaluation results.

11. AUTHORISATION

Helsinki, 5.10.2015

Jyrki Pulkkinen

Director

Development Evaluation Unit

Ministry for Foreign Affairs of Finland

ANNEX 2: KEY DOCUMENTS CONSULTED

DAC Peer Review Finland 2003. Paris: OECD-DAC, 2003. At: http://www.keepeek.com/Digital-Asset-Management/oecd/development/dac-peer-review-of-finland_journal_dev-v4-art25-en#page27 (visited 9 March 2016).

Development Policy Committee. *The State of Finland's Development Policy in 2009*. Helsinki: Development Policy Committee, 2009. At: <http://www.kehityspoliittinentoimikunta.fi/public/default.aspx?contentId=167463&nodeId=37559&contentlan=2&culture=en-US> (visited 3 March 2016).

Peer Review 2007 Finland. Paris: OECD-DAC, 2007. At: <http://www.oecd.org/dac/peer-reviews/39772751.pdf> (visited 3 March 2016).

Peer Review 2012 Finland. Paris: OECD-DAC, 2012. At: (visited 3 March 2016). At: <http://www.oecd.org/dac/peer-reviews/PRFINLAND2012.pdf> (visited 3 March 2016).

Ministry for Foreign Affairs. *Development Policy 2004. Government Resolution*. Helsinki: MFA. Available at: <http://formin.finland.fi/public/default.aspx?contentid=84290&contentlan=1&culture=fi-FI>.

Ministry for Foreign Affairs. *Development Policy Programme 2007. Towards a Sustainable and Just World Community*. Helsinki: MFA. Available at: <http://formin.finland.fi/public/default.aspx?contentid=103136>.

Ministry for Foreign Affairs. *Development policy programme 2012. Government Decision-in-Principle 16 February 2012*. Helsinki: MFA. Available at: <http://www.formin.fi/public/default.aspx?contentid=251855&nodeid=49559&contentlan=2&culture=en-US>.

ANNEX 3: DETAILED METHODOLOGY

Approach and Methodology for both Phases

The meta-evaluation was carried out by strictly following the instructions laid down in the Terms of Reference and the proposed response to that document found in the mini-tender of Danish Management Group. The **key** elements of that approach were:

- A two phase approach where **Phase One** was an assessment of the quality of documentation used for evaluations and appraisals; in this case the ITT/ToR and the appraisal or evaluation reports. **Phase Two** was an assessment of the “quality” (the term used in the ToR) of Finnish cooperation, based on an analysis of the evaluation reports that had received the highest scoring in Phase One
- The development of a set of analysis grids that were applied to a set of evaluation and appraisal terms of reference and reports
- A quality comparison between MFA commissioned documents and those commissioned by others.
- A portfolio analysis of all the documents retained for analysis

Overview of the steps involved in Phase One: Assessment of the quality of appraisals, evaluation reports and corresponding ToR.

Inception phase

After an initial review of the available material (obtained by EVA-11), the team became better acquainted with the various types of documents used in the Finnish project and programme cycle (specifically from the perspective of the interventions that would be analysed). Based on the 2012-2014 meta-evaluation report, a more detailed draft approach and methodology for the mandate, as well as a list of issues to be discussed during the Start-up meeting, was prepared. A more detailed Meta-evaluation Framework Matrix, similar to Annex 5 in the 2012-2014 report, was prepared, as was a set of first-level suggestions for improving Annexes 7 & 8 (tools for the ToR and evaluation reports). These were discussed by the entire team prior to the Start-up meeting so that the team would become very familiar with the assessment and analysis tools developed for the 2012-2014 report, and would be in a position to identify how those tools could be adapted to better reflect MFA management concerns. It should be noted that the original approach (later changed) of EVA-11 was to minimize any deviations from the approach, standard and norms that had been used in the 2012-2014 meta-evaluation in order to enable a more longitudinal analysis. The Team believed that the key objective of management accountability needed to be much better reflected in the existing tools so that the result of the meta-evaluation could be used as an important base from which MFA could make decisions concerning the implementation of its development cooperation mandate.

A one-day meeting was held in Helsinki where some of the Team’s concerns were aired, but time constraints resulted in a shorter meeting than originally planned, and a number of issues were not discussed. Current concerns with the relative ineffectiveness of capacity development identified globally⁵; the lack of a theory of change within project design plans; the poor level of ownership, and the failure of most interventions to use results-based management approaches in spite of formal guidelines are examples of the managerial concerns that the Team believed should legitimately be included in the meta-evaluation. The extent to which interventions are designed and man-

⁵ See, for example, the Danida-Sida-Norad joint evaluation on Capacity Development (CD) performed in 2014-2015 which identified that most CD is not sustainable and not measurable.

aged to focus on organisational outcomes (and the resulting link to impact) was also a key managerial concern that the team believed should be at the heart of the tools and their focus on analysis.

The team then set about preparing an inception report. The original intent, as reflected by the MFA at the start-up meeting, was to make few adjustments to the grids used in previous meta-evaluations so that a longitudinal analysis could be made over a large number of years. As noted below, the reality was quite different and the assessment tools were radically changed at MFA's request (based partly on suggestions put forward by the Team) following its analysis of the first version of the Inception Report.

In preparation for the first version of the IR, and in line with the discussions during the start-up meeting, the Team adjusted the tools used in the 2012-2014 meta-evaluation and three evaluation reports were selected to test the applicability of the "adjusted tools". They represented a cross-section of the entire evaluation portfolio, insofar as that was possible with such a small sample.

Particular emphasis was placed on two key questions: **first**, were the tools, criteria and rating scales designed clearly enough so that any team member would give the same assessment as any other team member? This implied that the criteria and assessment issues would be clearly enunciated and not inherently confusing. It also implied that the ratings were clear enough to be easily used. For example it was shown that the 2012-2014 meta-evaluation contained ratings that could be interpreted differently by different people: attention is given to the definition of "good" on p. 109 of that report, where a rating was tied to the concept of "mostly met" and could be interpreted differently depending on whether one considers whether it was the number or the importance of the objectives that were "met". **Second**, were the tools so designed that only "excellent" evaluations could be fully assessed? Would an unacceptable amount of "not assessable" or "not determined" answers be generated if the tool were applied the same way to all reports? If so, then serious weaknesses, opportunities, trends and lessons learnt could be slipping through without being detected. The quality assessment process proposed by the 2014-2015 meta-evaluation Team, including the cross-analysis by different team members on the same three reports, served to help identify these potential problems and others; ways and means to mitigate against the effects of those methodology weaknesses were proposed in the IR.

The Team also prepared an overall description of the MFA's evaluation and appraisal portfolio for the 2014-2015 temporal scope. From an initial 52 reports that were sent to the Team by EVA-11, a reduced number (n=38) reports, were selected by eliminating those that were basically credit scheme appraisals and self-evaluations. Of the 38, two projects had two reports each, **so the final result is that the Team assessed 38 reports related to 36 projects**. In terms of describing the population of reports that will eventually be used for the initial quality assessment, the portfolio description included the following characteristics (partial list):

- Geographical coverage (where do the reports refer to geographically?): Regional, national, sub-national
- Sector coverage, using the OECD/DAC sector classification system
- Size of project or programme, in terms of budgets. Value of the evaluation budgets
- Type of implementation partner
- Commissioned by whom? MFA? Implementation partner? Etc.
- Firm or individual that generated the report.

A separate section of this annex describes the methodology that was used for the portfolio analysis.

The Inception Report (IR) was written so as to reflect the requirements set out in the TOR, p. 4, especially: *“The approach, methodology and sources of verification will be explained in detail, including the methods and tools of analyses, scoring or rating systems, example figures and tables... It will also suggest an outline of the final report”*. The IR also presented an Evaluation Framework Matrix and work plan; identified a risk mitigation plan; proposed a communications plan between the TL and the Evaluation Manager; identified missing or required documents, and identified epistemological and logistical limitations to the mandate.

As noted above, a fundamental change in the approach and methodology took place on December 15th when MFA advised the Team that it had “decided that the basis for quality assessment of evaluation reports should be the guidance given in the Evaluation Manual (table 11, page 70 onward) instead of the EU Quality Assessment Grid. This means that the quality assessment tools used in the previous meta-evaluation will not be used in this meta-evaluation.”⁶ The Team then set out to re-design the entire approach and methodology it had presented both in the IR and in its mini-proposal, including all the assessment tools and the standards and norms that would be used in the analysis. What followed was a number of further versions of the IR wherein EVA-11 and the Team developed and commented on grids, approaches and tools to meet the much higher levels of complexity in the quality analysis than would have been the case had the change not occurred. The Team was always pleased, from a professional perspective, that the change had been requested by MFA: the analysis tools are now much more powerful analytically and represent much better the management concerns and standards and norms of the MFA. The team found it challenging to integrate some of the standards and norms into the grids: some standards, for example, were not stated (in policy documents) in terms that would enable different evaluators to come up with the same rating (replicability) because the norms were not sufficiently precise (for example, what constitutes a sufficient and appropriate focus on an integration of HRBA into specific interventions?).

The introduction of the inductive process into the analysis for Phase Two was also challenging because, while it is easy to describe the difference between deductive and inductive research, it is quite another to specifically describe how to analyse documents inductively where many standards and norms are to be taken into account. A concept paper was prepared by the Team and sent to MFA; the IR was subsequently adjusted to reflect a mutual understanding of the application of inductive logic into this meta-evaluation. The inductive logic concept was subsequently integrated into the Phase Two assessment tool. A separate section to this Annex describes the epistemological underpinnings of both the inductive and deductive approaches as they are applied in this meta-evaluation.

The IR being finally accepted, the Team proceeded with the rest of the mandate. In Phase One, the Team members analysed the quality of the evaluation and appraisal TOR and their corresponding reports. The three Team members divided the documents amongst themselves (based on sector knowledge or context familiarity, among other factors) and then undertook an initial comprehensive analysis all the while filling in an assessment grid. Each TOR and report was then cross-checked by the two other members and, where the ratings were significantly different (where the second or third opinion would result in a relative drop or increase in the overall total rating in a way that would represent a point “drop” of ten points or more, or where the baseline 60 points for inclusion within the Phase Two analysis was not going to be reached with revised points), the members communicated their differences and found a compromise. The original member was always the one responsible for accepting or making any changes in ratings.

⁶ Source: email received by the Team from MFA.

Specific Methodology: Data gathering and analysis - Phase One

The basic approach to data gathering was to assemble and collate the various reports that were required to properly (i.e. comprehensively) assess an evaluative event, whether it was a MTE/MTR, a final or ex post evaluation, or an appraisal. As an operating protocol, the Team required the following in order to complete an assessment of a sample in the portfolio: a report and a ToR. If an ITT were available, it was assessed at the same time as the ToR to which it referred, and only one assessment was made on the documents (i.e. ITT and ToR as one unit).

As agreed at the start-up meeting, this meta-evaluation did not require interviews with MFA or other Government of Finland (GoF) officials, as was the case with the previous meta-evaluation. The Team worked closely with EVA-11 which provided invaluable support in the gathering of relevant data documents for Phase One. EVA-11 gathered data on overall and specific budgets and sent them to the Team; this was integrated into the Team's analysis. EVA also identified an official who assisted the Team in gathering reports, ToRs and other relevant documents. One of the Team members was identified as having the responsibility of document gathering on behalf of the team and that person reviewed each document to see if it was complete and in a final form (i.e. not a draft). Discrepancies and questions were dealt with in collaboration with EVA-11 contact persons.

The standards for appraisals require that "The appraisal team should not directly revise the Project Document" (a direct quote from the Bilateral Manual), so the appraisal reports almost always provided **lists** of things that need to be done without providing any suggested content, leaving the rest of the work to be done by others. For example, an appraisal tested noted that a logic chain needed to be done along with a set of results and indicators; no suggestions were offered. The point here is to confirm that the Team did not seek to obtain revisions of Programme Documents or other reports that followed on the heels of the appraisal report. In that case, it was nearly impossible to speak of the "quality upon entry" analysis that was required in the ToR. The MFA agreed that this particular analytical focus would not be included in the mandate any further. The team would, however, be required to bring insights based on Phase One analysis of appraisals to the attention of the MFA as part of Phase Two.

One of the tasks assigned to the meta-evaluation Team was the comparison of the quality between MFA-commissioned evaluations and the evaluations commissioned by MFA's partners (refer to the ToR of the meta-evaluation). This was done by assigning a special code to all the evaluation reports commissioned (by the MFA or by other agencies than the MFA, including joint evaluations where the leading role has been carried out by another donor), and fixing this code as the basis of correlations to highlight possible differences in quality and other variables between those commissioned by MFA and those by others.

Assessment tools described

There are six tools that were developed for this meta-evaluation. The first five were used in Phase One, the last one only in Phase Two:

1. Baseline tool for gathering information on each project in the Meta-evaluation (used for population analysis)
2. Quality Assessment Tool for Evaluation ToR/ITT
3. Quality Assessment Tool for Evaluation Reports
4. Quality Assessment Tool for Appraisal ToR/ITT
5. Quality Assessment Tool for Appraisal Reports
6. Reference frame of the analysis of Finnish Development Cooperation (for Phase Two)


Each of the tools for Phase One is presented in separate worksheets found in other annexes to this report. A large number of important changes were made the original version of the tools⁷ that, according to the ToRs for this meta-evaluation, were to be used, including:




- The number of sub-categories or sub-standards was significantly reduced to focus more on managerial/strategic issues that are at the heart of the assurance function of management within MFA. Issues dealing with very minor process requirements or format have thus been eliminated;
- At the specific request of the EVA-11, following the recommendations of the team, the assessment tools are not based on the OECD/DAC Quality Grid as was the case in previous meta-evaluations but on the standards and norms promulgated by the Finnish development cooperation (MFA) ;
- The tools contain specific instructions to the meta-evaluators on how to interpret the sub-standards and what to look for in their assessments;
- The tools are based on five possible scenarios that are identified by the meta-evaluator. The highest rating is given if the part of the report under consideration: “Exceeds most key quality criteria and standards”. The lowest rating is given to parts of the report that have “serious deficiencies in terms of meeting the standards”. In between the ratings progress in five possible increments. The previously-used concept of “very good” was abandoned altogether, the logic being that something could meet a standard or a norm or exceed it, but it **is not required to exceed**, and extra points should not be allocated to “exceed”. The Team also qualified what may happen if some part of report does not meet or exceed expectations: if the deficiency is minor (i.e. it is not rejected as being poor), the meta-evaluator must make a distinction between the case where the deficiency puts into question the core of the evaluation report (i.e. findings, conclusions, EQ analysis, etc.) or not. If it strikes to the core it receives a lower rating than if the deficiency deals with non-core issues. The definitions for the rankings are clearly spelled in each assessment grid. In practice, the team members found that they almost always would have given the same rating as another member; this occurred because the three Team members very often consulted one another and discussed how to react to particular situations. A “what if” learning approach was used by the team and it learned how to adapt to special or specific situations.

The ratings were applied at the sub-category level but the points (weights) were calculated at a “Headline Standard” level. What this meant in practice is that the meta-evaluators would read and analyse a printed or an electronic copy of the report he/she was going to assess. They made marginal notes (on paper) or comments (electronic format) to show they had analysed the report in detail. They then used the assessment tools (the first for ToR, then for the report) to systematically make notes about what they found. Once each sub-category was rated, the Team member assigned an overall rating to the standard and then used the pre-assigned weighting to translate a rating into a score. For example, if all the categories were rated at 4 (meets all standards) then the headline standard would also receive the same rating of 4. Transposing that rating to a “Score” meant using the weighting protocols that were developed with the EVA-11. In the weighting for a headline standard was to be 25 points and the overall rating was “4” as in the example above, the final score for that category would be 35. The rating system was proposed by the Team and accepted by EVA-11. The diagram below illustrates the principles described in this paragraph.

⁷ Refer to 2012-14 Meta-evaluation

Figure 20: Diagram of the structure of grids showing the difference between headline standards and characteristics for scoring

Headline Standard 

	Max.	Score
5. Evidence-based findings, including:	X	Y
a) An analysis of empirical data, facts proving a sound level of evidence to finding related to:		
b) Overall progress of the implementation (for expected outputs, outcomes and impacts)		
I) Relevance	Characteristics 	
II) Effectiveness		
III) Impact		
IV) Sustainability		
V) Efficiency		
This section must also integrate:		

Source: Meta-evaluation Team

Where weighting is concerned (as different from rating), the EVA-11 indicated its preferences for the relative weights that should be given for the key (i.e. most important) headline standards. The Team had participated in this process by proposing a weighting system and then invited MFA to impose its own relative weights based on its priorities. Here are some of the key algorithms that were used:

- With respect to TOR, the priority issue 1 (i.e. rationale, purpose and objectives) were each given an overall weight of 15 out of 100.
- With respect to priority issue two (i.e. issues and evaluation questions), the weight should be 35 out of 100.
- Issue 3 (i.e. adequate resources for the mandate) should be given a weight of 25 out of 100.
- With respect to the evaluation and appraisal report assessments, a separate section on findings was introduced with a weighting of 15 out of 100 (the same weighting as conclusions and recommendations).
- Finally “Evaluation questions” is a separate section with a weighting of 10 out of 100.

The remainder of the points, as initially proposed by the Team in its IR, were not changed. The complete weighting distribution may be found in the annexes containing the various assessment grids.

It should be noted that:

- The distribution of the weighting for reports gives a strong prominence to the reliability of findings (hence conclusions and recommendations) over more formal aspects of evaluation reporting; Weighting for TORs is more equally distributed among criteria with an extra focus on identifying EQs/issues.
- For each criteria, an option to rate something as Not available/not addressed (NA/ND) has been provided in case the standard does not apply in this particular case or if the report does not deal with that issue or standard at all. This may legitimately occur for evaluation reports that are not expected to follow MFA Evaluation Manual, (for example joint evaluations, trust fund management reviews, etc.), but it also happens when reports just do not deal with the issues or standards at all (ex. there is no discussion of logic or oversight, or there is no analysis of context as it affects effectiveness of efficiency or sustainability). In principle, consistency with ToRs might lead to identify specific expectations or exception to the general rules set by the Manual but this could apply only in the future if ToRs are required to specifically identify exemptions granted to a given assignment against the general rules.

The overall judgement on the headline standard is based on “expert opinion based on the relational principle of research” and is not necessarily a mathematical sum of the ratings provided on the components. This is easily verified by examining any category with three or more sub-categories. Some sub-categories are clearly more important than others.

The various characteristics and factors (the standards and norms) that need to be assessed by the meta-evaluation Team are found in the far left-hand column of each assessment grid. These are, for the most part, extracted directly from the Bilateral Program/Project Manual or the MFA Evaluation Manual, and are exactly as described in the three-layer model that was sent to MFA in December. Each tool has a column on the far right that explains to the evaluator how the document should be assessed against any given standard. Some of the cells are not filled in because no clarification is needed. Illustrations of some of the important methodological specifics that were built into the tools follow. There were many others that could have been in the list.

- The Evaluation report assessment tool contains a number of questions wherein the meta-evaluator is asked to assess the report in areas that are not, strictly speaking “standards”. For example, a question asks whether the reviewer would suggest using the report in Phase Two and whether there is a capacity development/training issue to flag for that function of EVA-11.
- Assessment tools for appraisal ToR/ITT and Reports were created under this evaluation and tested. There were no such tools used in the previous meta-evaluations. The standards and norms are those found in Bilateral Programme and Project Manual and the Evaluation Manual. No new or arbitrary (i.e. on the part of the reviewers) standards have been introduced.
- Each of the appraisal-related tools contains a column wherein specific instructions indicate how to interpret the standard and norms. These interpretations were generated by the Team based on its experience with project cycle management generally and appraisals specifically. The overall intention is to assess the extent to which the appraisal reports enable MFA management to proceed to approval (or to cancel). MFA policy requires that appraisal reports not re-write the draft Programme documents but should, as minimum, be very explicit about what they suggest as changes or improvements. We have interpreted this approach as indicating that reports should not, for example, suggest that indicators need to be generated without indicating what kind of indicators would be appropriate and possibly suggesting how they should enable evaluability and results management to be overriding principles of the intervention.

There has been a concerted effort to harmonize both the structures and the weighting distribution between the assessment grids used for the ToRs (evaluation and appraisal), as well as the reports themselves (evaluation and appraisal). In addition, the grids are designed in such a way as to mainstream the five OECD and the three MFA evaluation criteria into headline standards, and to also mainstream the findings relating to the HRB approach and the Cross-cutting objectives of the MFA.

The set of guidances extracted from the MFA Evaluation Manual was combined for each section of the standard table of contents (introduction, context, description of the project...) as criteria. A three-layer model was developed by the team based on their required content and approved. It integrates the various MFA-defined norms and standards that apply to evaluation and appraisal deliverables.

Specific methodology for the generation of the “specific standards and characteristics” used in the assessment grids for Phase One

There are a number of MFA policy documents that describe what is current policy on specific topics (ex. sustainable environment) but the two key documents that describe what should be included in evaluation and appraisal documents (ToR and reports) are the Bilateral Manual and the Evaluation Manual. The challenge faced by the Team was to specify what should be included within the sections that were

often identified by one or two words. For example, in the main text of evaluation or appraisal reports, the following should be present:

1. Introduction
2. Context
3. Description of the project
4. Purpose and objectives of the mandate
5. Findings
6. Aid effectiveness
7. Answers to the evaluation questions
8. Conclusions
9. Recommendations
10. Lessons learned

But what is meant, specifically, by “context” and, as importantly, what should be included within a section that deals with “context”? In the same line of thought, how should HRBA be mainstreamed into the report and how should the report reflect OECD and MFA evaluation criteria? Through trial and error, much team interplay and the integration of Quality Assurance advisors into the discussion, the Team proposed a number of versions of assessment grids. They can still be improved, but they are now fairly easy to use by trained experts.

Each of those items comes as a separate heading (light orange) in the assessment grid, on which rating and scoring is applied. The key parts of what may be called “preliminaries” (ex. executive summary, table of contents, list of acronyms and abbreviations and annexes) are also separated out in the grid and points are allocated for them. For each of those sections, criteria, norms and standards are extracted from the standards set by MFA Evaluation Manual, the Bilateral Project Manual and various policy documents including those for the HRB approach and any Cross-cutting issues, and not, as previously done, from the OECD-EU Quality grids. A typical example can be taken from section 4 of the MFA Evaluation Manual related to “Approach, methodology and limitations” where criteria for assessing the quality of an evaluation report are essentially the existence of a presentation, in the report, of:

- a) The overall evaluation approach;
- b) The evaluation matrix including the evaluation questions approved by MFA in the inception report, indicators, sources of data and the correspondence between the EQ and the evaluation criteria used by MFA (p.57 of the Manual);
- c) The methodology used and its risks and limitations;
- d) The data collection and analyses techniques used and their limitations;
- e) A description of the sources of information;
- f) The logic for the use of case studies if any;
- g) A critical assessment of the validity and reliability of data and the analysis conducted upon it;
- h) Any limitations on process, methodology or data and how they may affect validity and reliability;
- i) A statement that describes any obstruction to a free and open evaluation process that may have influenced findings;

- j) A statement to the effect that there were no discrepancies between the planned and actual implementation and products of the evaluation.

Characteristics were developed by meta-evaluators with a high degree of consistency with MFA manual(s) standards, combined with a professional assessment of the quality of implementation of the standards. For example, the mere existence, in the section 5 related to “Answers to evaluation questions” of some paragraphs presenting opinions concerning “... An analysis of empirical data, facts, evidence (findings) relevant to the indicators of the evaluation questions” was not judged sufficient to confirm positively that the report provides valuable and reliable evidence. Focus, structure, volume, clarity, consistency with stated methodology, and plausibility were judged to be very important and cannot be assessed in any other way than peer reviewing.

Before proceeding with Phase Two, the following questions need to be answered for each report or ToR/ITT that were assessed in Phase One:

- Would you recommend including this report in Phase Two and why?
- Were the resources allocated sufficient to carry out the evaluation?
- Are there any points that should be pursued in terms of evaluation capacity building?
- Are there points that should be pursued in terms of how MFA manages the evaluation cycle?

The appraisal assessment tools follow the same logic as explained above.

Reliability and replicability testing of Phase One assessment tools

The assessment tools have been tested in three ways.

- First, the tools were developed jointly by all the members of the Team so that each other's experiences could be brought to bear. Where there were differences of opinion, a workable solution was found and an algorithm was devised. For example, it was noted that there was a great deal of latitude that could be given to many standards that involved EQ or issues or results. After various exchanges a decision was made to ensure that the Team stick closely to the letter of the “standard” or “norm”. The result of that “internal decision” was the insertion of that guidance in the instructions of each assessment tool.
- Second, three evaluation reports were selected for testing against the original version of the assessment tools; this number is down from five as required in the ToR, with a reduced number being agreed in the start-up meeting. The projects selected were MTE of PALWECO in Kenya (Agriculture and Livelihoods sector); Final Evaluation of EIBAMAZ, (three Andean countries in inclusive education), and the Final Evaluation of Institutional development to IGAD (as a regional integration organisation). These three were selected for a variety of reasons including the fact that some of the Team Members were aware of the initiatives or their host organisations; it was thought desirable to have a mix of sectors and geographic areas, and all the required documentation was available when needed for the three projects. Each of the three senior members of the Team performed an assessment of three projects and the results of their assessment were recorded in the appropriate tool. The results of those assessments were compared and the following was noted: 1) that the assessments of the ToR/ITT were very similar between Team members. The variations were analysed and clearer instructions were inserted into the assessment tools. 2) The evaluation report assessments identified wider differences in assessment results. The reasons for these differences were analysed and the team members have developed more “common” ways of dealing with the standards. It was agreed among team members that once these adjustments were taken into account similar (but not necessarily identical) responses would be forthcoming. It is interesting to note that team members all had similar approaches to the “heart” issues of

the reports, including the approach to findings, conclusions and recommendations. Part of the variation in ratings was also due to the mechanical process of calculating the overall value at the level of the “headline standard”. This has been rectified. For the information of the EVA-11, a short description of the analysis performed on the three projects is included in Annex 4. Overall, the revisions introduced in the assessment tools improved their capacity to seize the reliability of the findings-conclusions-recommendations and lessons learnt chain, and thus will ensure better quality inputs to Phase Two. The test on the three selected projects, covering the diversity of MFA portfolio, demonstrated that the tools are operative and allowed to fine-tune the instructions to the meta-evaluators, particularly of the judgement criteria to be applied on the structuration of the evaluation reports by EQs and the importance to be given to evidence-based findings.

- Each evaluation and appraisal document (i.e. all ToR and all reports) were assessed through a comprehensive analysis process and the other Team members have cross-checked the results of the assessment. In the vast majority of cases, there was a significantly high level of consensus on the analysis itself as well as on the ratings. The comments of the experts are included on the assessment grids themselves.
- The two QA experts retained to work on this mandate provided invaluable insight and suggestions during the generation of the IR, and especially with the design of the various assessment tools.

Methodology: Data Gathering and Analysis - Phase Two

Overall approach and sources of information

The purpose of Phase Two is quite different from that of Phase One and requires a much higher level of “interpretation”, “synthesis” and “critique”. As specified in the ToR, Phase Two analysis is structured on a combined and complementary deductive and inductive approach. The sample for Phase Two is drawn from the Mid-term evaluations and evaluation reports that were analysed in Phase One, with the following filters and baseline standards:

- Information sources for Phase Two are limited to the reports provided;
- The sample (eligible reports) for Phase Two will be drawn from reports examined in Phase One;
- The strategy for selecting the reports that were used in Phase Two is based on a required minimum baseline score of 60 points TOTAL for the reports. An algorithm was included in the IR in case the total number of reports that reached at least 60 points was not approximately 80% of the number of reports assessed (i.e. 18 out of 28), but it was not necessary to use it. Although 18 are two less than the targeted number of 20 reports, the remaining reports score all at 55 points or lower and would not, from an evaluation view point, provide any real value added to phase two.
- For this meta-evaluation, all evaluation reports, regardless of commissioning agents, or budgets, or any other qualifiers, were eligible for consideration as part of the sample for Phase Two, providing they met the 60% baseline;
- Results were consolidated on a single worksheet and analysis began using an inter-category (how reports deal with a single category) as well as a multiple category approach (is there a pattern between categories such as between low levels of effectiveness and the use of results-based approaches?).
- A particularly challenging issue that arose in this meta-evaluation is the fact that the policy contexts of 2007 to 2012 are quite different from those after 2012. Differences had to be taken into account in the assessments. All the points in the Phase Two analysis grid that refer to the priorities of the CCO or of Finnish development Cooperation were be qualified with the terms “as indicated in the appropriate ToR” (i.e. for that report), as a means of taking those policy contexts into account.

Combining deductive and inductive research approaches in this meta-evaluation

It is important to justify why, from an epistemological perspective, both an inductive and a deductive approach are (or could be) used in a complementary manner within the same meta-evaluation research effort. An annex to the IR and a separate concept Paper were developed by the Team to deal with that issue and sent to the MFA.

To begin with, it is clear that the choice of Induction or Deduction is a key part of the selection of a “RESEARCH APPROACH” in any evaluation, or any research effort for that matter. It is not merely a methodological choice, nor is it merely an analytical option. There are methodological consequences of that choice, however, including an important set of limitations and constraints.

An analysis of the applicability of the inductive approach to this mandate was performed by the Team and the results, in the form of a comparison of the characteristics of this meta-evaluation against a set of epistemological or research criteria. That analysis was sent to MFA. Overall, it showed that the inductive approach was applicable in this meta-evaluation, but given the small number of sample points and the vast differences in the nature and contexts of the reports, it may have been difficult to obtain high levels of correlation in the findings. A combined inductive and deductive approach was therefore recommended.

As noted above, Phase Two relied on a combined deductive and inductive approach. For each report that needed to be analysed, the normative deductive analysis was done first by a Team member through an analysis (i.e. assessment) grid that reflected the OECD/DAC evaluation criteria and key Finnish policy points. Then, the same team member used an inductive approach to identify issues and trends that would not normally be identified through the deductive process. The same analysis grid used for the deductive process enabled the team member to write down these “issues” (called research memos) and trends in cells specifically reserved for that purpose (in other words, the grid is designed to **also** provide for a frame of reference that can “stimulate” the creative process for the researcher).

In fact, the Team members thoroughly re-analyse assigned reports⁸ and then used the Phase Two analysis grid to rate the extent to which the intervention that is represented by the report reflects (or not) the pre-defined criteria and component standards of MFA specifically and GoF generally. A rating is given for each criteria (OECD/DAC) and component/policy level (ex. HRBA) in a manner similar to Phase One. Further, a clear definition of what to look for and how to interpret the standard is included in the grid. The rating scale is somewhat different even if it is still based on a five point system where:

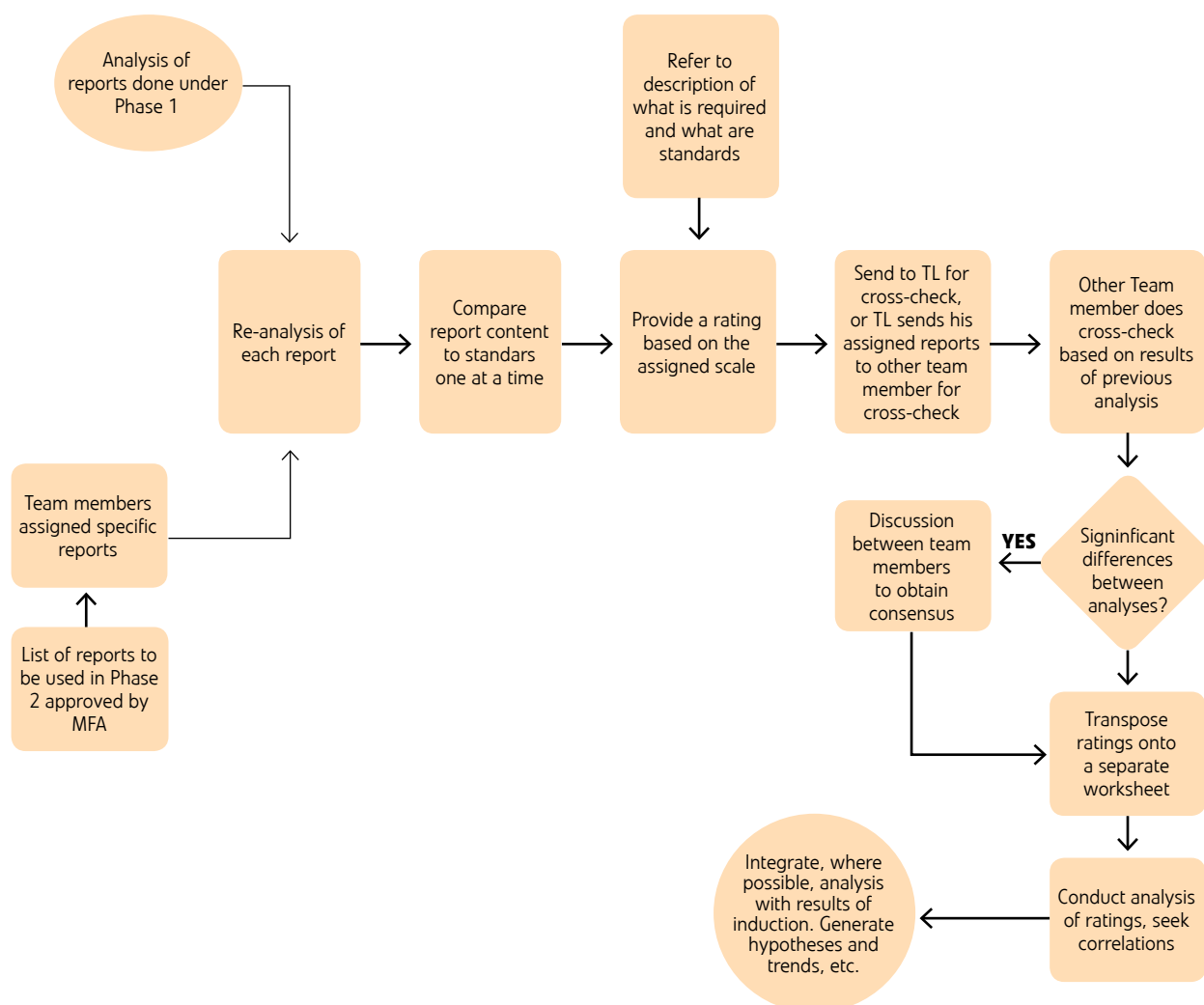
- 5 = Component assessed positively (i.e. the project, as evaluated, met or exceeded its expected outcomes/results under the appropriate evaluation criteria, and therefore contributed to meeting Finnish development objectives).
- 4 = Minor restrictions to a positive assessment of the component/criteria (i.e. no major setbacks in terms of meeting expected outcomes under the evaluation criteria being analysed).
- 3 = No more than one serious restriction to a positive assessment of the component under the evaluation criteria. Intervention must have met sustainability and effectiveness criteria.
- 2 = Major restrictions to more than one positive assessment of the component under the evaluation criteria.
- 1 = Component assessed negatively (i.e. did not meet expectations or was not analysed sufficiently in the report to enable rating to take place).

⁸ They have already been analysed under Phase One

Once the expert completed the assessment and the cross-check process had taken place (on all the reports), all the ratings were transposed to a separate worksheet and the team began to analyse the combined data (from all the reports). Among other things, it looked for domains where there were consistent occurrences of low ratings, and areas where good (or bad) ratings correspond to the same criteria used in the portfolio analysis (Annex 8), such as country or region, commissioning agent, authors, etc.

The following diagram illustrates the essential parts of the deductive process described above:

Figure 21: Overview on the deductive process in Phase Two



Source: Meta-evaluation Team

The following paragraphs will first lay-out the epistemological underpinnings that the Team used for the meta-evaluation and then explain the methodology that it applied. While conceptually enticing, the inductive approach has its own set of conditions and contexts that must be considered if the results of any research effort are to provide valid and useful contributions. Because it is open ended and creative in nature, team members (i.e. in situations where there is more than one researcher) need to be able to find a way to conduct the analysis more or less in the same way. Comparatively, the deductive approach is much more “bordered” and the team members essentially merely need to follow the pre-defined analysis grid provided. For that reason, a more detailed description is required for the inductive approach than for the deductive approach.

Induction is one of the fundamental backbones of qualitative approaches⁹; and a key characteristic of all those approaches is that no hypotheses tests are used, contrary to quantitative approaches that are based on scientific explanation models. Any scientific hypothesis used in deductive approaches is based on a background theory, typically assuming the form of a proposition whose validity depends on empirical confirmation. Otherwise, a hypothesis is nothing but an imaginative conjecture¹⁰.

By contrast, inductive-referenced qualitative researchers generally contend that their work does **not** consist of proposing and testing hypotheses. **Their primary interest is to achieve understanding** of a particular situation, or individuals, or groups of individual, or (sub)cultures, etc., **rather than to explain and predict future behaviours** as do the so-called hard sciences, with their arsenal of laws, theories, and hypotheses employed or rejected on the basis of their predictive value.

In addition to the before mentioned deductive approach, the team therefore used an inductive approach to seek to “achieve understanding” of Finland’s development cooperation experiences in Phase Two. To do that it applied the same level of rigor in its work as would be expected in any research effort.

In terms of how to arrive at trends, premises and conclusions using induction logic, there is today a complex, diversified praxis influenced by a large number of schools, authors, and epistemological perspectives. The academic world adds a very palpable and real layer of analytical rigor to the approach because of its historical “gatekeeping” functions, while “practitioners” (including consultants and programme management officials) constantly attempt to find ways of reducing the level of effort and the impact of methodological constraints that would normally be required to meet “standards” of research.

The Meta-evaluation Team maintained that there is a **minimum level** of analytical rigor that must be introduced, and that the data that will be used (from which observations will be made) has to meet a minimum set of quality characteristics (ex. stability, replicability, accuracy, known relation to context). The Team therefore began by defining the analytic core that is required in any qualitative data analysis method used for meta-evaluation; this core is found in the processes used in the research cycle composed of **data coding, categorizing, and conceptualizing**, followed by measures such as counting, scoring and rating in order to quantify or qualify patterns.

9 Much of the conceptual underpinnings explained in this section is based on the works of Pedro F. Bendassolli, and a good reference is his “Theory Building in Qualitative Research: Reconsidering the Problem of Induction”, Qualitative Social Research Journal Vol 14, No. 1, Article 25, January 2013. <http://nbn-resolving.de/urn:nbn:de:0114-fqs1301258>. Some of his work has been introduced here with editing; all appropriate attribution of rights remains with his works.

10 The preceding should not be considered as an absolute. In specific cases the inductive approach can be adapted to conduct more quantitative research as well (ex. When it is used as a basis for exploratory data analysis).

1. The Team would have already begun by establishing an initial contact with the material by means of an initial analysis performed in Phase One. Once MFA agreed on the selection of cases that were to be used in Phase Two (based either on the 60% baseline proposed in this IR or by the inclusion, through “justified sample definition” of a particularly interesting item), the team would follow-up with a “careful reading” of each piece of information (report) that would form part of Phase Two. As noted above, no documents other than the evaluation reports already studied under Phase One were to form part of Phase Two. The use of the “justified sample definition” algorithm was not necessary since 18 projects met the 60% entry level.

The team member would first use the Phase Two rating grid to rate the components and criteria according to the norms and standards specified in the grid. Then, the team member would re-examine the report to analyse the text for “issues that might arise”. Data/text analysis for that purpose would require the taking of “research notes” in the form of “research memos” (refer to many reference works on this practice¹¹) to record impressions and insights, which were to be used in later stages of the analysis.

Through coordination and in-house training, the team members would already have been prepared to keep in mind and work with **two** different sets of “frames” with which to analyse the text of these reports: the first was non-structured, based on the experience of the researcher and the interface of that experience with the data itself. In this case, the Team member identified what he/she considered to be the data that “stands out”. The second was a general research frame that had been developed by the team itself and was incorporated into the Phase Two grid. It was designed to be dynamic, evolving, and flexible (in the sense that the researcher was free to add to the frame as circumstances required), and reflected the relevant managerial or policy concerns of Finland with respect to its development cooperation. Using the frame, for example, the team would look for any references to local ownership issues, or oversight weaknesses, or impact on poverty (to name a few). Other examples included the identification of best cases, or examples of factors that enabled outcomes to be achieved. None of these examples would have been brought forward using only an “assessment-type” of grid or a rating system applied to criteria or components.

2. Based on the analysis of the evaluation reports carried out under 1) above, it was foreseen that some themes and patterns would start to emerge from that text analysis; that is, that they would inductively reveal themselves to the team in the data’s interaction with the empirical “tools” as given above. Once the themes and patterns started to reveal themselves, the Team attempted to concentrate the “findings and make them more and more “strategic” in nature.

A fundamental question, and one which has a direct impact on the relation between theory and empirical data, is what would be understood as a “theme,” “pattern,” or “category.” After all, what the meta-evaluation Team was being asked to do was to identify factors or meta-level findings concerning the QUALITY of Finnish development cooperation. Using Bendassolli’s¹² models (or similar), the team sought to identify themes that were related to central meanings (factors that define the intent, structure, logic and management of Finland’s development cooperation) that organise experiences (factors that have been identified as a result of the real-world evaluations performed). Examples of “from where” themes can evolve include (but are not limited to...):

¹¹ Strauss, Anselm & Corbin, Juliet M. (1998). Basics of qualitative research. Thousand Oaks, CA: Sage

¹² Pedro F. Bendassolli, “Theory Building in Qualitative Research: Reconsidering the Problem of Induction”, Qualitative Social Research Journal, Volume 14, No. 1, Art. 25 - January 2013

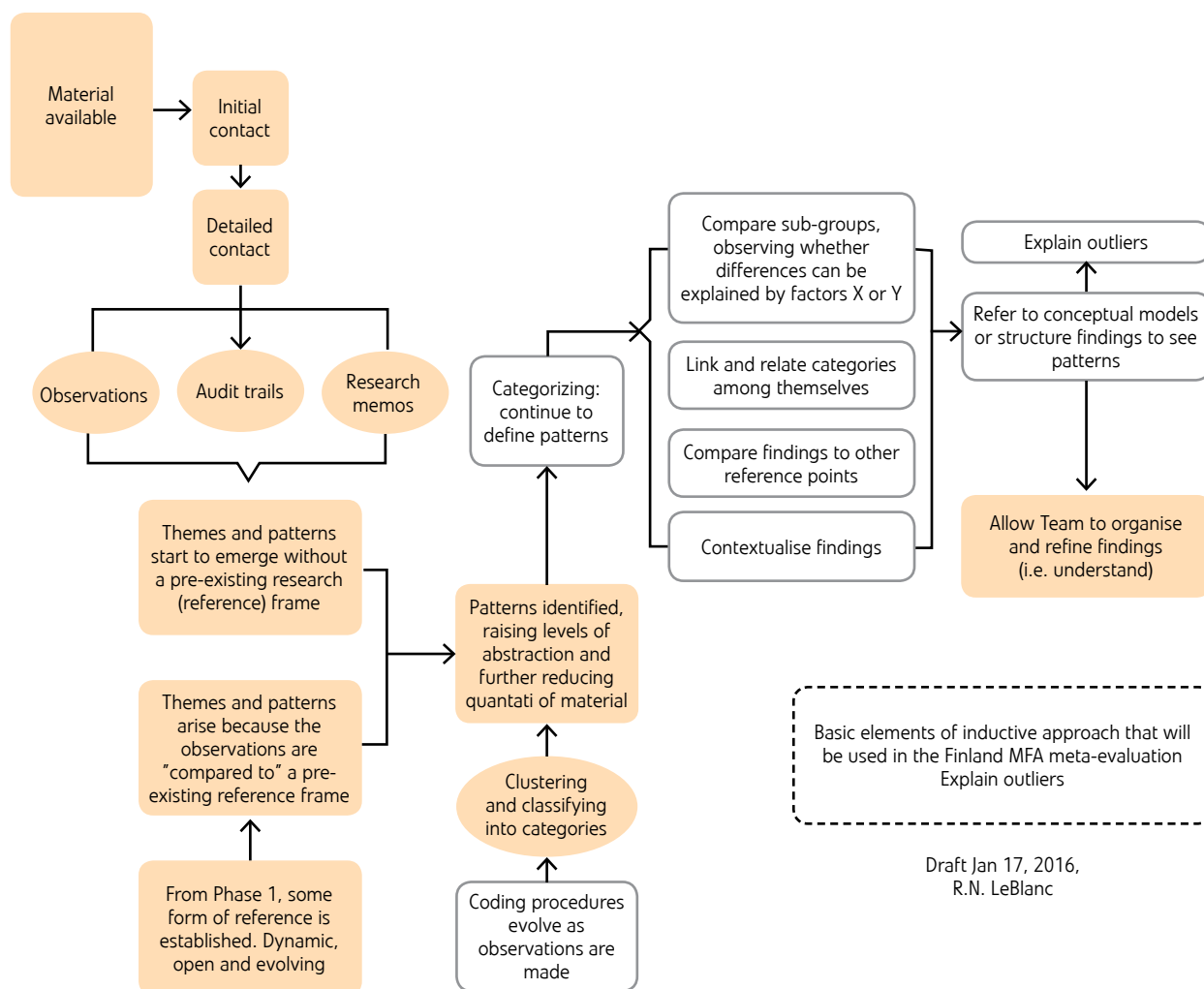
-
- repeated ideas, concepts (ex. similar conclusions and findings in different reports);
 - in similarities among units that make up the analysis material (for example, among different commissioning bodies);
 - in similarities of content, or sector, or other purposeful findings between reports;
 - in the concepts used by evaluators to describe or justify findings (ex. ownership, RBM, HRBA);
 - in the frequency and intensity of repetition in the material under analysis;
 - in the location of the themes in discourse and in its centrality as a cognitive element and effective organiser of experience;
 - in the similarities and differences of ratings given to specific assessments (ex. the rating given to “recommendations”).

In summary, themes were expected to assume both categorical (an instance of the experience, a unit of meaning), and frequential (repetition of themes or their location in networks or schemes) forms.

The Team then began to consolidate a list of findings based on the inductive approach it had used. The specific analysis that was carried out with these findings was framed or depended on many factors, including the strength and validity of the findings themselves, and the extent to which linkages could be made between findings. The final report that was prepared described the findings but did not propose conclusions on the development cooperation programmes of GoF.

The overall structure of the inductive approach as it was applied to the meta-evaluation is illustrated in the following diagram (Figure 22).

Figure 22: Process through which induction is used as a complementary research approach in Phase Two



Source: Meta-evaluation Team

The results of the inductive analysis were reported in the final report (mostly as text concerning findings or examples grouped by “themes” or “patterns”, along with the results of the “deductive” analysis that arose from the use of the analysis grid for Phase Two (mostly structured by OECD/DAC evaluation criteria, supplemented by MFA policy issues such as HRBA). All of these analyses are meant to be complementary in order to meet the objectives for Phase Two.

Portfolio and report quality analysis: methodology

Specific activities

The population of reports was divided in four categories: appraisals, mid-term evaluations (MTE), evaluations and (mid-term) reviews. While appraisals are a clear and well defined category, the Team classified as “MTE” all mid-term reports which use the OECD/DAC evaluation criteria at least to a certain extent independently of what they were called in the report titles. This solution comes from the desire expressed by EVA-11 to start using the term MTE for what used to be called mid-term reviews, and reserve the term “review” for narrower reports carried out for management purposes. The Team categorized as “evaluations” all final, ex-post and other evaluations without qualifiers.

There may be a certain degree of overlapping in MTEs and evaluations: in some cases an end-of-phase assessment report is called MTE, especially when the stated objective is to continue funding a further phase of the same project, while in some other similar cases the report was simply called an evaluation. The Team decided to maintain the original differentiation between MTEs and evaluations because it does not have any possibility to know if the evaluation effort ended up being “final” or whether there was an additional phase. It also wanted to enable a distinction to be made between reports that deal with impact and those that do not typically do so (i.e. MTE)

Of the population of 38 reports, ten are appraisals; another ten could clearly be classified as MTEs and fifteen correspond to the definition of evaluations. The Team found two “reviews” that could not be classified either as appraisals, MTE or evaluations; they did not use the DAC evaluation criteria and concentrated on management structures and administrative problems and/or arrangements. They could best be characterized as a performance audit for one of them, and an internal joint-donor inception review for the other. The common point for what here have been called management reviews is that they analyse how the project works rather than elucidating what it purports to achieve and how.

According to an analysis of the reports and evaluation budgets, the ITT were missing (21 evaluation events were not available, over 55 percent of the total number of cases assessed), mainly for smaller evaluations contracted by direct procurement (no tendering involved) and in cases when another donor commissioned the evaluation, although in some of the latter cases the evaluation budget was indicated in the ToR of the evaluation. The ToR were missing for 5 reports (out of 38), mostly in cases of joint-donor or multilateral evaluations; in 33 cases the ToR were either available separately or annexed to the reports. The budget of an evaluation, if indicated in a currency other than EUR, was converted into Euros according to the average exchange rate of the month and year when the ToR of the evaluation were signed.

For the regional distribution of projects of which reports have been submitted to the present meta-evaluation, the Balkans (one project) were clustered together with Eastern Europe and Central Asia (excluded from the rest of Asia) because both are part of the “larger Europe”, either through (potential or existing) association agreements with the European Union and/or as beneficiaries of the Wider Europe Initiative of Finnish development cooperation. In this classification Afghanistan (three projects) is clustered as part of Asia. North Africa and the Middle East were defined as the Southern and Eastern shores of the Mediterranean. The long-term partner countries of Finnish development cooperation development cooperation are Vietnam, Nepal, Ethiopia, Kenya, Tanzania, Zambia and Mozambique.

The budgets in the portfolio presented a challenge for the meta-evaluation. Only few reports presented the budget of their evaluation objects, even fewer indicated the total budget from all stakeholders and funding agencies (including the beneficiary partner government). Therefore to rely on the reports only would have provided very fragmentary information on the financial scale of the evaluated projects. The Team thus turned to the only available reliable source of information which is the body of funding decisions from Finland, provided by EVA-11 for the meta-evaluation. For each project/programme, thus for each report, there was a table of funding allocations drawn from the administrative system of the MFA. The Team is aware that the funding allocation from Finland is only one part of the total budget of a project, but faced to the scant, fragmentary information, this was the only sensible solution.

Methodology for Analysing the Portfolio

The meta-evaluation took the following path to analyse the “portfolio” of evaluation reports. For each evaluation report, basic information was collected using a template derived from an MS Excel work sheet. The information on the entire set of reports was inserted into another MS Excel spreadsheet in horizontal rows with variables as vertical columns. For each variable, a coding system was designed to enable analysis and cross-referencing.

When all the variables and their codes were entered into the table, graphs based on horizontal (stacked) bars were produced on a number of the most important variables. In some cases the correlation function of MS Excel allowed for calculating correlations between any two of the variables; for instance, between the budget of the evaluation and the quality score of the report, or the commissioning regional unit or agency, consultancy company (Finnish or non-Finnish) and the geographical scope of the evaluated project. The limitation was that it is not possible to reasonably calculate correlations between qualitative information (e.g. the type of commissioning agency) and quantitative (and quantifiable) information (e.g. report scores). In these cases, the meta-evaluation team found other solutions, such as top-ten lists, averages and other similar forms of visualisation in order to make appear tendencies.

ANNEX 4: JUSTIFYING THE USE OF A MIXED DEDUCTIVE-INDUCTIVE RESEARCH APPROACH IN PHASE 2

When undertaking evaluative research it is compulsory to specify which approach one is adopting, and to justify that selection on the basis of the research object or purpose. Researchers often develop and then share the contents of a table comparing how various attributes would be treated using inductive and/or deductive approaches (Table 19). As the nature of what is being sought is clarified (i.e. the fields of interest, domains, or concerns of the meta evaluation), researchers begin to understand the effects of their choice (of using the inductive and/or deductive approach) on such methodology options as those dealing with the selection of primary data collection methods, research processes, validation tools, etc.

Table 19: An attribute differentiation table for the Meta-evaluation, based on the complementarity of an inductive research approach to a deductive approach.

Attribute	Deductive	Inductive	This meta evaluation's specific applications of inductive and deductive approaches (examples)
Direction	"Top-Down"	"Bottom-Up"	The inductive approach can complement the deductive approach. While the latter will use the "theories" established by the norms and standards of the MFA and work its way through a rigorous analysis of reports, the inductive approach will not use pre-set theories. Instead it will be based on (i.e. developed from) a set of observations from the assessments of "better quality" evaluation reports. The two approaches are complementary and will provide rich data for the development of findings and the identification of patterns.
Focus of research	Prediction changes, validating theoretical construct, focus in "mean" behaviour, testing assumptions and hypotheses, constructing most likely future	Understanding dynamics, robustness, emergence, resilience, focus on individual behaviour, constructing alternative futures	At this point, the only focus that has any legitimacy are those specifically spelled out in the ToR. EVA-11 has clearly indicated that it does not wish to define or suggest domains for phase 2; the team will nevertheless have to create boundaries (frame of reference) for its research. These will be based on the ToR, relevant MFA manuals and the most recent 2012-2014 meta evaluation.
Spatial scales	Single (one landscape, one resolution)	Multiple (multiple landscape, one resolution)	All evaluation reports are distinct and cover multiple spatial scales, from national to regional, from one ministry to sector-approaches. An inductive approach is therefore warranted and can complement the more deductive research approach

Attribute	Deductive	Inductive	This meta evaluation's specific applications of inductive and deductive approaches (examples)
Temporal scales	Multiple (deterministic)	Multiple (stochastic)	Each evaluation report covers a different time scale. Much overlap but not necessarily similar temporal configurations. An inductive approach is therefore warranted and can complement the more deductive research approach.
Cognitive scales	Single (homogenous preferences)	Multiple (heterogeneous preferences)	A number of indicators and results (ex. scales, standards, norms, assessments, sources) can be used for this meta-evaluation analysis. They are very distinct and numerous and need to be somehow "channelled" into new patterns and knowledge. A deductive approach can therefore use a fixed, or single, norm-based frame (such as that provided by the OECD/DAC evaluation criteria), and an inductive approach can be facilitated by an open frame of reference that enables multiple types and levels of "discovery".
Aggregation scales	Single (core aggregation scale)	Single or multiple (one or more aggregation scales)	There are a number of aggregation "models" or patterns that need to be taken into account. Sometimes the aggregation is very simple (ex. a single function of a ministry); sometimes not (ex. many departments focused on innovation). Sometimes it covers a single function, sometimes many. Sometimes there is good potential for contribution analysis, sometimes not. An induction approach can be applied but it may be difficult to "achieve understanding" due to the small number of events and their heterogeneity. The deductive approach, on the other hand, can help to quickly identify what level of aggregation is relevant for any particular criteria or component.
Predictive vs. Stochastic accuracy	High – Low (one likely future)	Low – High (many likely futures)	At this point in time it is too early to tell if the available data (including the nature of the data and the stability of the data across contexts) will support any mathematical (or other) modelling to predict future behaviour such as what is required for Bernoulli space functions required for probability prediction procedures. That being said, <i>one should assume that the small number of individual events, combined with the infinite variety of conditions that define the events (part of heterogeneity) will not allow for any form of accurate Stochastic-based accuracy.</i> The accuracy of any predictions made by the researchers will be based on statistical correlation analyses and statistical hypotheses confidence levels. <i>Again, given the nature and stability of the evaluation events analysed, it is possible that the level of confidence for any managerially-useful prediction will not be great.</i>

Attribute	Deductive	Inductive	This meta evaluation's specific applications of inductive and deductive approaches (examples)
Data intensity	Low (group or partial attributes)	High (individual or group attributes)	The unstable and variable definitions used for standards and norms will clearly mean that there will be a high level of interpretation involved in assessing evaluation reports. In that regard, using inductive approaches will mean that some form of "framework" be established to allow the researchers to have some form of 'idea' as to what to look for in terms of trends or directions that appear from the evaluation data. That will allow the researcher to identify issues and cases that will complement the research results of the deductive approach used in the meta-evaluation.

Source: Adapted by R. N. LeBlanc from a table originally prepared by Alexandiris, K.T. (2006) "Exploring Complex Dynamics in Multi Agent-Based Intelligent Systems", Pro Quest Publications

From the table above the Team concludes that the inductive approach is not only applicable in this particular meta-evaluation but will provide a solid source of complementary data and information relevant to the purpose of Phase 2.

ANNEX 5: PORTFOLIO MATRIX

Control no.	Title of the evaluation report	Score TOR	Score report	Type of report	TOR	ITT	OECD/DAC Sector	Year	Budget from Finland EUR	Country / area	Geographical scope	Budget evaluation EUR	Evaluation Commissioner	Evaluation Company	Modality of implementation (of project)	MFA Unit
9	Appr Innovation Partnership Programme (IPP) in Vietnam, Phase 2 2011-2018	80,5	66	APPR	YES	YES	Communications	2013	11.000.000 €	Vietnam	Country-wide	50.000 €	MFA	Finnish		ASA-10
10	APPR Rural Village Water, Phase III (RVRWP III) Nepal	42,5	38	APPR	YES	Contract	Water and sanitation	2015	15.000.000 €	Nepal	Sub-national	29.985 €	MFA	Finnish		ASA-40
7	APPR Scaling-up Participatory Sustainable Forest Management Project, SUFORD-SU	65	16	APPR	YES	NO	Forestry	2013	10.940.000 €	Laos	Country-wide	60.000 €	MFA	Finnish		ASA-10
11	APPR UNDP/ Evaluation of Aid for Trade Wider Europe project	77,5	32,5	APPR	YES	NO	Business support services	2013	9.032.000 €	Tajikistan, Kazakhstan, Uzbekistan, Kyrgyzstan, Armenia, Georgia, Belarus, Moldova, Ukraine	Regional/ multi-country	43.000 €	MFA	Finnish		ITÄ-20
8	APPR Water and Sanitation Programme Vietnam, Phase III	49,8	57,5	APPR	YES	Contract	Water and sanitation	2013	2.500.000 €	Vietnam	Country-wide	50.000 €	MFA	Finnish		ASA-10
2	Appraisal of ADPP Mozambique's project proposal: Farmers' Clubs for wealth creation among smallholder farmers in Mozambique	68	60,5	APPR	YES	NO	Agriculture	2013	7.985.434 €	Mozambique	Country-wide	n/a	MFA	Finnish		ALI-30
3	Appraisal of AWEPA Mozambique's project proposal: Strengthening the Legislative and Executive Oversight Function of Elected Organs in Mozambique 2015-2018	57	34,5	APPR	YES	YES	Government and civil society	2015	2.000.000 €	Mozambique	Country-wide	25.000 €	MFA	Individual local consultant (Africa)		ALI-30

Control no.	Title of the evaluation report	Score TOR	Score report	Type of report	TOR	ITT	OECD/DAC Sector	Year	Budget from Finland EUR	Country / area	Geo-graphical scope	Budget evaluation EUR	Evaluation Commissioner	Evaluation Company	Modality of implementation (of project)	MFA Unit
6	Appraisal of project extension "Strengthening National Geographic Services in Lao PDR (SNGS)"	66.5	63.5	APPR	YES	YES	Government and civil society	2014	1,000,000 €	Laos	Country-wide	50,000 €	MFA	Finnish		ASA-10
5	Appraisal of the Draft Project Document for the Second Phase of the Finnish – Southern Africa Partnership Programme to Strengthen NEPAD/SANBio Network (BioFISA II) 2013-2016	72	65	APPR	YES	NO	Communications	2013	6,000,000 €	Southern Africa	Regional/multi-country	40,000 €	MFA	Finnish-African mixed team		ALI-30
1	Appraisal of UNDP Dryland Development Centre's Project "Programme of Catalytic Support to Implement The Convention to Combat Desertification In West Asia And North Africa – Phase V"	67	29.5	APPR	YES	YES	Environment	2013	3,000,000 €	Northern Africa and Middle East	Regional/multi-country	16,000 €	MFA	in-house multi-lateral organisation		ALI-10
27	EVAL Finland's Support to Institutional Strengthening of IGAD 2011-2014	42	35	EVAL	YES	YES	Government and civil society	2014	1,430,000 €	IGAD	Regional/multi-country	20,000 €	MFA	Finnish	Bilateral, TA + direct funding	ALI-20
37	EVAL Independent Evaluation of the International Trade Centre	40	55	EVAL	YES	NO	Business support institutions	2014	1,400,000 €	Worldwide	World-wide	365,000 €	Joint	Non-Finnish company	Multilateral	TUO-10
19	EVAL MSI (Marie Stopes International), 2009-2014	60	83	EVAL	YES	NO	Reproductive health care	2015	2,700,000 €	Afghanistan	Country-wide	47,500 €	Marie Stopes International	Non-Finnish company	INGO	ASA-40
38	EVAL of Norad's support to UNIDO's Trade Capacity Building Programme 2005-2013		73	EVAL	NO	NO	Trade policy and regulation	2015	4,200,000 €	Worldwide	World-wide	n/a	NORAD	Non-Finnish company	Multilateral	TUO-10
33	EVAL Partnership for Market Readiness (PMR)	79	74	EVAL	YES	NO	Environment	2015	4,110,000 €	Worldwide	World-wide	n/a	World Bank	Non-Finnish company	Multilateral	KEO-60
17	EVAL Programme for Agriculture and Livelihoods for Western Communities (PALWECO)	77	61	EVAL	YES	YES	Agriculture	2014	27,000,000 €	Kenya	Country-wide	80,000 €	MFA	Finnish	Bilateral, Finnish TA	ALI-20
31	EVAL Regional biodiversity Programme for the Amazon Region (BioCAN)	75	74	EVAL	YES	YES	Environment	2015	6,275,000 €	Andes regional	Regional/multi-country	180,000 €	MFA	Finnish	Bilateral, Finnish TA + direct funding	ASA-30

Control no.	Title of the evaluation report	Score TOR	Score report	Type of report	TOR	ITT	OECD/DAC Sector	Year	Budget from Finland EUR	Country / area	Geo-graphical scope	Budget evaluation EUR	Evaluation Commission	Evaluation Company	Modality of implementation (of project)	MFA Unit
4	EVAL Small Scale Irrigation, SIP Zambia	57	66	EVAL	YES	YES	Agriculture	2015	10.000.000 €	Zambia	Sub-national	22.600 €	MFA	Non-Finnish company	Multilateral	ALI-30
29	EVAL The African Union Mediation Support Capacity Project 2012-2014, Phase II	71	63	EVAL	YES	NO	Conflict prevention, resolution, peace and security	2015	3.000.000 €	Regional AU	Regional/multi-country	n/a	Joint evaluation	Finnish-African mixed team	INGO	ALI-30
30	EVAL Environmental Management Lao PDR, PHASE I	75,5	64	EVAL	YES	NO	Environment	2015	9.500.000 €	Laos	Country-wide	90.000 €	MFA	Finnish	National implern.	ASA-10
25	Evaluation of the ACWL	46	51,5	EVAL	YES	NO	Trade policy and regulation	2014	900.000 €	Worldwide	World-wide	n/a	Switzerland	Non-Finnish company	Multilateral	TUO-10
21	Evaluation of the Afghanistan Sub-National Governance Programme (ASGP/UNDP), Phase II	84	78	EVAL	YES	NO	Government and civil society	2014	1.500.000 €	Afganistan	Sub-national	88.900 €	SIDA	Non-Finnish company	Multilateral	ASA-40
28	Evaluation of the Cooperation in Science, Technology and Innovation between Finland and Mozambique (STIFIMO) UHA2011-006469		30,5	EVAL	NO	YES	Communications	2015	12.714.161 €	Mozambique	Country-wide	70.000 €	MFA	Non-Finnish company	Bilateral, TA	ALI-30
35	Evaluation of the Enhanced Integrated Framework	65	54	EVAL	YES	NO	Trade policy and regulation	2014	14.000.000 €	Worldwide	World-wide	n/a	WTO secretariat	Non-Finnish company	Multilateral	TUO-10
36	Evaluation of the International Trade Centre, Report of the Office of Internal Oversight Services /E/AC.51/2015/		83	EVAL	NO	NO							UN ECOSOC	UN in-house evaluation	Multilateral	TUO-10
32	Evaluation on the Framework Agreement between the Government of Finland and the OSCE	47,5	45,5	EVAL	YES	NO	Unallocated/unspecified	2014	1.835.291 €	Armenia, Georgia, Azerbaijan, Tajikistan, Kazakhstan, Uzbekistan, Kyrgyzstan, Turkmenistan, Moldova, Ukraine, Belarus	Regional/multi-country	50.638 €	MFA	Finnish	Multilateral	ITÄ-20

Control no.	Title of the evaluation report	Score TOR	Score report	Type of report	TOR	ITT	OECD/DAC Sector	Year	Budget from Finland EUR	Country / area	Geo-graphical scope	Budget evaluation EUR	Evaluation Commissioner	Evaluation Company	Modality of implementation (of project)	MFA Unit
34	External Evaluation of UNCTAD's Development Account 7th Tranche Project: Strengthening capacities for policy-oriented analysis of key global development challenges at developing country universities	82	78	EVAL	YES	NO	Business support institutions	2014	4.650.000 €	Worldwide	World-wide	n/a	UN secretariat	Non-Finnish company	Multilateral	TUO-10
23	MTE Aid for Trade UNDP-Kosovo	75	38	MTE	YES	NO	Trade policy and regulation	2014	1.337.738 €	Kosovo	Country-wide	n/a	MFA	Individual Finnish consultant	Multilateral	EUR-40
15	MTE COWASH water sanit. Ethiopia 2011-2016	75	68	MTE	YES	YES	Water and sanitation	2015	22.000.000 €	Ethiopia	Country-wide	115.000 €	MFA	Finnish	Bilateral, Finnish TA + direct funding	ALI-20
16	MTE land admin Ethiopia, REILA	62	87,5	MTE	YES	YES	Agriculture	2015	13.900.000 €	Ethiopia	Country-wide	90.000 €	MFA	Finnish	Bilateral, Finnish TA	ALI-20
13	MTE Middle East and North Africa (MENA MDTF)	24	54,5	MTE	YES	NO	Government and civil society	2015	800.000 €	Northern Africa and Middle East	Regional/multi-country	n/a	World Bank, joint evaluation	Non-Finnish company	Multilateral	ALI-10
18	MTE OF ADB TA 7987(REG) Core Environment Program In The Greater Mekong Subregion, PHASE 2 2011-2015		60	MTE	NO	NO	Environment	2015	11.476.000 €	Mekong regional	Regional/multi-country	n/a	ADB (+1 consultant funded by Nordics)	Non-Finnish company	Multilateral	ASA-10
22	MTE of Nepal Multi-Stakeholder Forestry Programme	88	92	MTE	YES	NO	Forestry	2015	9.200.000 €	Nepal	Country-wide	n/a	SDC	Non-Finnish company	National implem.	ASA-40
39	MTE OF TECHNICAL ASSISTANCE FOR TEVT (SOFT SKILLS) DEVELOPMENT IN SCHOOL SECTOR REFORM PLAN IN NEPAL	61	66	MTE	YES	Contract	Education	2015	1.600.000 €	Nepal	Country-wide	74.000 €	MFA	Finnish	Bilateral, Finnish TA	TUO-10

Control no.	Title of the evaluation report	Score TOR	Score report	Type of report	TOR	ITT	OECD/DAC Sector	Year	Budget from Finland EUR	Country / area	Geo-graphical scope	Budget evaluation EUR	Evaluation Commission	Evaluation Company	Modality of implementation (of project)	MFA Unit
20	MTE of the Regional Programme for Afghanistan and Neighbouring Countries 2011 – 2015	78	80	MTE	YES	NO	Other social services	2015	1.000.000 €	Afghanistan	Regional/multi-country	n/a	UNODC	In-house Evaluation Unit	Multilateral	ASA-40
14	MTE Palestine Land Administration	50	60	MTE	YES	YES	Other social services	2014	5.200.000 €	Palestinian territories	Country-wide	29.500 €	Joint evaluation: World Bank & MFA	World Bank staff + non-Finnish consultant hired by MFA	Multilateral	ALI-10

APPR Appraisal

EVAL Evaluation

MTE Mid-term evaluation

ANNEX 6: QUALITY ASSESSMENT TOOL FOR EVALUATION ToR/ITT

Introduction to the ToR/ITT grid

STRUCTURE: The TOR/ITT assessment tool combines the report outline (Annex II) and the quality check list (Table 11) of MFA's Evaluation Manual.

RATING: Evaluators provide a rating for headline standards (dark orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on “expert judgment based on the inductive principle of research” and is not necessarily a mathematical sum of the ratings provided on the components.

The 5-point scale is as follow:

5 = Exceeds most key quality criteria and standards. Must exceed standards related to the answers to the EQs

4 = Meets all quality standards

3 = Has minor deficiencies, but not where EQ analyses are concerned

2 = Has minor deficiencies in any quality grid factor, including analysis of EQ

1 = Serious deficiencies

SCORING: At section level (dark orange) and heading level (light orange), the rating is translated into a score (extreme right column) based on the maximum score set in the immediate left column. The score at section level is the addition of scores at heading level.

WEIGHTING: Foremost importance is given to (i) the quality (coverage of evaluation criteria, adjustment to project specifics, clarity, and number) of evaluation questions (35 points), (ii) resources, particularly team composition and experts' profiles (25 points), and (iii) rationale, purpose and objectives of the mission (15 points). Those three factors are key for bidders to fit with the needs of the evaluation and the elaborate a methodology that will provide the evidence-based findings requires and adjust to the level of resources availed by MFA.

INSTRUCTIONS TO USERS: First, read carefully the TOR (and ITT) and scroll the annexes. Rate sub-headings and then headings by diverting from the arithmetic average to convey the quality assessment obtained during the analysis of the report. Diversion is hardly about more than 1 rate.

From there, provide scores with a diversion from the translation of ratings in scores that should not exceed 2-3 points for key criteria.

Short Title:

Year of report :

Type:

Meta-evaluation date: 2016

RATING: Evaluators provide a rating for headline standards (dark orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on “expert judgment based on the inductive principle of research” and is not necessarily a mathematical sum of the ratings provided on the components. The 5-scale is as follow:

5 = Exceeds most key quality criteria and standards. Must exceed standards related to the answers to the EQs

4 = Meets all quality standards

3 = Has minor deficiencies, but not where EQ analyses are concerned

2 = Has minor deficiencies in any quality grid factor, including analysis of EQ

1 = Serious deficiencies

SCORING: At heading (dark orange) level only, the rating is translated into a score (extreme right column) evaluated based on the maximum score set in the immediate left column.

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.	Meta-evaluator's notes and observations					5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
	1. There is sufficient background information to the evaluation/review provided					5					5			
1.1 The programme context (policy, country, regional, global, thematic context) is sufficiently describe														
a) Is the broader context of the programme described, including development objectives of partner countries?														
b) Is Finland's development policy and its linkages to the programme being evaluated described?														
1.2 There is a clear description of the programme to be evaluated														
a) Are programme objectives, implementing strategies, resources for implementation and intervention logic covered?														
b) Are the issues related to the promotion of human rights, gender equality, reduction of inequalities and promotion of climate change included in the background information provided?														
c) Are stakeholders involved and their roles described, including beneficiaries and institutions?														
d) Is there a description of the history of the programme, including how results and targeted outcomes (project purposes) have changed over time?														
1.3 The results of previous evaluations are presented														
This section should describe what is already known from previous evaluations and what this evaluation will add.														

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.		Meta-evaluator's notes and observations		5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
2. The rationale, purpose and objectives of the evaluation are clearly described											15	0
2.1 The rationale and purpose of the evaluation are stated clearly in the ToR.												
a) Do the ToR describe why the evaluation was undertaken at this particular time and for whom?												
b) Is the use of results of the evaluation adequately described? (e.g. learning and accountability purposes)												
c) Is it clear from the ToR what type of a study this is (review, mid-term, final, ex-post)?												
2.2 The specific objectives of the evaluation are clearly stated.												
Are the issues, analysis and recommendations the evaluation will focus on sufficiently described?												
3. There is an appropriate and sufficiently detailed description of the scope of the evaluation											3	0
3.1 The scope of the evaluation clearly describes the time span the evaluation covers, stakeholder groups involved, and geographical areas covered.												
a) Do the ToR clearly define what is included/excluded from the evaluation and the reasons why?												
b) Does the timeline offer an appropriate scope for achieving the evaluation objectives?												
c) Is the proposed range of stakeholders groups and sampling targets an appropriate scope for achieving the evaluation objectives?												
4. The evaluation objectives are translated into relevant and specific evaluation questions											35	0
4.1 The ToR apply the agreed OECD/DAC and MFA criteria for evaluation development assistance.												
a) For instance, are relevance, efficiency, effectiveness, impact and sustainability covered?												
b) If a particular criterion is not applied and/or any additional criteria added, is there an explanation in the ToR as to why?												
4.2 Relevance of the object of the evaluation is adequately addressed.												
Are there questions on the extent to which the objectives of the programme are consistent with beneficiaries' requirements, country priorities, global priorities and partners' and Finland's policies included?												

[illegible]

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.		Meta-evaluator's notes and observations		5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
5. The implementation of aid effectiveness commitments is described												
a) Ownership											5	0
b) Alignment												
c) Harmonisation												
d) Management for development results												
e) Mutual accountability												
6. The proposed methodology is appropriate and capable of addressing the evaluation questions												
6.1 General guidelines for the methodology are included in the ToR for data collection and analysis.											5	0
a) Is there general information on the methods to be used (e.g. qualitative or quantitative)?												
b) Is there information on the sources of data to be analysed?												
c) Is there information on how data analysis will be conducted, ensuring for example that data will be disaggregated by gender, age group or other relevant group?												
7. The evaluation process and management structure are adequately described in the ToR or ITT												
7.1 The ToR adequately describe the evaluation phases, their sequencing and approximate duration.												
a) Do the TOR describe phases such as the kick-off meeting, inception and desk study phase, inception meeting, interviews and field missions, reporting, presentation of the evaluation results in the field, reporting and presentation of the evaluation results?												
b) Are key milestones in the process described?												
7.2 The reports and outputs (results) to be submitted in each phase of the evaluation are clearly specified.												
Do the ToR specify types of deliverables required such as: inception report, presentation on the field findings, draft final report, final report, presentation of the evaluation findings?												
7.3 There is a specific request in the ITT or ToR to propose and implement a quality assurance system for the evaluation and its deliverables.												

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.		Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
8. The resources required for this evaluation are sufficiently described in the TOR or ITT											
8.1 The expertise required to conduct the mandate is sufficiently described by the ToR or ITT (overall composition of the evaluation team).										25	0
Is the desired expertise/knowledge of the team described in the ToR (or ITT)?											
Does the team composition require both international and national experts? Do the ToR request a team leader to be nominated?											
Does the evaluating team have to possess a mix of evaluative skills and thematic knowledge?											
Is the gender balance of the team considered?											
Are the following fields specified: programme evaluations and planning in the relevant sector, relevant sectors in developing countries, integrating cross-cutting objectives, HRBA etc.?											
Is the expertise required in terms of cross-cutting objectives/themes/issues described?											
8.2 The budget is specified and is adequate to execute the mandate											
8.3 The mandate to investigate is provided											
Authorization to speak to relevant individuals or organisations but not to seem to represent Finland											
8.4 Do the ToR give focus and direction to enable a good appraisal response?											
8.5 Is there enough information for evaluators to prepare a proposal?											
9. Annexes and structure of the ToR											
9.1 The ToR sections hold together in a logically consistent way that will allow for a coherent evaluation report										2	0
9.2 The style of the ToR is adequate (brief, to the point, logically structured and easy to understand).											
9.3 Do the ToR give focus and direction to enable a good evaluation response?											
9.4 Is there enough information for evaluators to prepare a proposal?											
9.5 Are appropriate/required annexes specified ?											
TOTALS										100	

ANNEX 7: QUALITY ASSESSMENT TOOL FOR EVALUATION REPORTS

Introduction to the Evaluation Report Assessment Grid

STRUCTURE: The evaluation report assessment tool combines, inter alia, the report outline (Annex IV) and the quality check list (Table 11) of MFA Evaluation Manual.

RATING: Evaluators provide a rating for headline standards (dark orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on “expert judgment based on the inductive principle of research” and is not necessarily a mathematical sum of the ratings provided on the components.

The 5-scale is as follow:

5 = Exceeds most key quality criteria and standards. Must exceed standards related to the answers to the EQs

4 = Meets all quality standards

3 = Has minor deficiencies, but not where EQ analyses are concerned

2 = Has minor deficiencies in any quality grid factor, including analysis of EQ

1 = Serious deficiencies

SCORING: At section level (dark orange) and heading level (light orange), the rating is translated into a score (extreme right column) based on the maximum score set in the immediate left column. The score at section level is the addition of scores at heading level.

WEIGHTING: Foremost importance is given to the reliability of the findings and the linkage between evidence-based findings and conclusions-recommendations. The reliability of findings relates (i) to the quality of the methodology (stakeholders mapping, sampling methods and evaluation framework e.g. evaluative questions or evaluation matrix), (ii) grounding findings on facts, figures, and documentary references. Evidence-based findings (5.), methodology (4.) and answer to evaluation questions (6.) are weighted 40 on the maximum score of 100.

Key judgement criteria for the 15 points attached to conclusions is the coverage of OECD criteria and demonstration a clear linkage with key findings. For recommendations (15 points), the main factor is the intrinsic quality of propositions: clear distinction between strategic and operational recommendations, congruence between those two categories, and consistency with project deadline and demonstrated ability to perform. In all, 75 points are attached to the quality of the content of the reports.

INSTRUCTIONS TO USERS: First, read carefully the report and scroll the annexes. The table of contents provides already a glance on the level of compliance with the Evaluation Manual and the level of understand of the evaluators of the role and function of OECD and MFA evaluation criteria. Underline key findings and main statements in conclusions and recommendations. Second, fill the column for notes, first for sub-headings (judgment criteria, in light orange), then synthesize key features at criteria level.

Care is to be taken to define precisely what is expected from different evaluation event: for example, MTE/MTR are not supposed to cover impact; in those cases, tick the NA (not applicable) box and later give all points to this sub-line).

Once all notes filled, rate sub-headings and then headings by diverting from the arithmetic average to convey the quality assessment obtained during the analysis of the report. Diversion is hardly about more than 1 rate.

From there, provide scores with a diversion from the translation of ratings in scores that should not exceed 2-3 points for key criteria.

EVALUATION REPORT										
QUALITY ASSESSMENT GRID										
Title:										
Year:										
Type:										
Meta-evaluation date: 2016										
<p>RATING: Evaluators provide a rating for headline standards (dark orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on "expert judgment based on the inductive principle of research" and is not necessarily a mathematical sum of the ratings provided on the components. The 5-scale is as follow:</p> <p>5 = Exceeds most key quality criteria and standards. Must exceed standards related to the answers to the EQs</p> <p>4 = Meets all quality standards</p> <p>3 = Has minor deficiencies, but not where EQ analyses are concerned</p> <p>2 = Has minor deficiencies in any quality grid factor, including analysis of EQ</p> <p>1 = Serious deficiencies</p> <p>SCORING: At section (dark orange) and heading (light orange), the rating is translated into a score (extreme right column) evaluated based on the maximum score set in the immediate left column. The score at section level is the addition of scores at heading level.</p>										
Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.	Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc	Not applicable	Max. Score	Score
PRELIMINARIES										
Table of Contents and Acronyms									7	0
Executive Summary that covers:									1	
a) A representative overview of the report									6	
b) Main findings, conclusions, recommendations, and their logical links in table format										
c) Overall lessons learned										
MAIN TEXT										
1. Introduction, including:									85	0
a) Evaluation's rationale, purpose and objectives									1	
b) Scope (temporal, geographical, any other scoping contained in ToR/ITT)										
c) Main evaluation questions										
2. Context, including:									4	
a) Description of the broader environment										
b) Influence of the context on the performance of the project/programme										
3. Description of programme or project being evaluated, including:									5	
a) Its objectives										
b) Implementation strategies										
c) Resources for implementation										

[illegible]

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.	Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
5. Evidence-based findings , including:									15	
a) An analysis of empirical data, facts providing a sound level of evidence to findings related to:										
b) Overall progress of the implementation (for expected outputs, outcomes and impacts)										
i) Relevance										
ii) Effectiveness										
iii) Impact										
iv) Sustainability										
v) Efficiency										
This section must also integrate:										
c) The temporal scope of the evaluation										
d) The target groups indicated in TOR										
e) The socio-geographical areas linked to the programme										
f) Achievements on Cross-Cutting Objectives and progress in HRB approach										
6. Answers to Evaluation Questions based on the assessment of findings									10	
7. Conclusions including:									15	
a) An assessment of overall performance of the programme based on findings and their correspondence to the evaluation criteria										
i) Relevance										
ii) Effectiveness										
iii) Impact										
iv) Sustainability										
v) Efficiency										
b) Each conclusion must present a means of knowing which EQ or indicator has shaped the conclusion										
c) Conclusions are prioritized and the basis for that ranking should be made clear										
d) Should enable the reader to understand, among other things, why the intervention or programme worked or not										

Base reference of the meta-evaluation	Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc	Not applicable	Max.	Score
This represents the standards against which reports will be evaluated.									15	
8. Recommendations , including:										
a) Proposed improvements, changes, actions that could remedy problems in performance or capitalise on strengths.										
A clear indication of:										
b) to whom is the recommendation directed (MFA, embassy, partner institutions, consultant providing support services, etc.)										
c) Who is responsible for implementing the recommendation										
d) When the recommendation should be implemented (immediate implementation, medium to long-term)										
9. Lessons learned , including:									5	
a) General conclusions that are likely to have the potential for wider application and use										
b) Lessons that were deduced from the evaluation concerning cause and effect relationships										
Annexes									5	0
Terms of Reference									0	
Detailed methodology, limitations of the study									3	
Lists of information sources e.g. people interviewed, documents reviewed, etc.									1	
Quality assurance statement, stakeholders' comments (pro and con)									1	
Non-content issues									3	0
Other evaluation concerns that are not included in an evaluation report but that are needed for quality purposes, including: -Is the written quality of the report such that it is presented in a clear manner that can be understood by its key audiences?									3	
								TOTAL SCORE	100	0

ANNEX 8: QUALITY ASSESSMENT TOOL FOR APPRAISAL TOR/ITT

APPRAISAL TERMS OF REFERENCE and ITT QUALITY ASSESSMENT GRID		RATING: Evaluators provide a rating for headline standards (light orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on "expert judgment based on the inductive principle of research" and is not necessarily a mathematical sum of the ratings provided on the components. The 5-scale is as follows: 5 = Exceeds most key quality criteria and standards. (Not only are they mentioned but in enough detail to enable the preparation of a comprehensive and high-quality tender) 4 = Meets all quality standards. (i.e. all items mentioned are present in the ToR) 3 = Has minor deficiencies in a few (less than four), but not where content issues or questions or feasibility analyses are concerned. (Some items may be missing but not those relating to feasibility and logic) 2 = Has minor deficiencies in any quality grid factor, (Has minor deficiencies in many sections, but no complete absence of direction in sections relating to feasibility and logic) 1 = Serious deficiencies SCORING: At heading (dark orange) level only, the rating is translated into a score (extreme right column) evaluated based on the maximum score set in the immediate left column.										
Short Title:												
Year:												
Type:												
Meta-evaluation date: 2016												
Base reference of the meta-evaluation												
This represents the standards against which reports will be evaluated.												
1. There is sufficient background information to the appraisal provided in the TOR or ITT												
1.1 There is clear description of the context		5	4	3	2	1	Not covered in doc.	Not applicable	Max.	Score		
a) Policy												
b) Country												
c) Region												
d) Global												
e) Thematic												
f) Sector												
g) Thematic and geographic priorities												
h) CC priorities												
i) Linkages to other partners												
j) Linkages to other interventions												

[illegible]

Base reference of the meta-evaluation		Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc.	Not applica- ble	Max.	Score
This represents the standards against which reports will be evaluated.											
4. The appraisal objectives are translated into relevant and specific appraisal issues										35	0
4.1 Issues addressed must include recc of MFA QA Group											
a) Max 12 AQ, presented by criteria (OECD or internal structure (see manual or items below in bold)											
b) If some criteria left out, explanations given											
c) Some other criteria may be added											
4.2 Coverage of meet needs, human rights, gender, inequalities, climate											
a) by relevance appraisal questions											
b) by feasibility appraisal questions											
4.3 Adequacy of background analysis											
incl. problems, existing strength and resources, stakeholder analysis. Must include all CC											
4.4 Analysis of programme logic											
a) Results chain, ToC, in terms of potential impact											
b) Effectiveness and efficiency of the programme and the proposed management design											
c) Appraisal of integration of human rights, gender, climate and inequality											
4.5 Sustainability											
a) Likely continuation of programme objectives when external support comes to an end											
b) Includes human rights, gender, climate and inequality											
4.6 Coherence											
Refers to issues beyond development Cooperation focusing on contradictions or mutual reinforcement with other policies to achieve the development Objective											
5. The implementation of aid effectiveness commitments is described										5	0
a) Ownership											
b) Alignment											
c) Harmonisation											
d) Management for development results											
e) Mutual accountability											

[illegible]

Base reference of the meta-evaluation		Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc.	Not applica- ble	Max.	Score
This represents the standards against which reports will be evaluated.											
8.2 The budget is specified and is adequate to execute the mandate											
8.3 The mandate to investigate is provided											
Authorization to speak to relevant individuals or organisations but not to seem to represent Finland											
10.1 Do the TOR give focus and direction to enable a good appraisal response?											
10.2 Is there enough information for evaluators to prepare a proposal?											
9. Annexes and structure of the TOR										1	0
9.1 The TOR sections hold together in a logically consistent way that will allow for a coherent appraisal report											
9.2 The style of the TOR is adequate (brief, to the point, logically structured and easy to understand).											
9.3 Do the TOR give focus and direction to enable a good appraisal response?											
9.4 Is there enough information for appraisors to prepare a proposal?											
9.5 Are appropriate/required annexes specified?											
a) Link to MFA Evaluation Manual											
b) Outline of the appraisal report											
c) Quality grid checklist											
										100	

ANNEX 9: QUALITY ASSESSMENT TOOL FOR APPRAISAL REPORTS

APPRAISAL REPORTS QUALITY ASSESSMENT GRID		RATING: Evaluators provide a rating for headline standards (in orange) and their numbered components by ticking the relevant column. Note that the overall judgement on the headline standard is based on "expert judgment based on the inductive principle of research" and is not necessarily a mathematical sum of the ratings provided on the components. The 5-scale is as follow: 5 = Exceeds most key quality criteria and standards. 4 = Meets all quality standards. 3 = Has minor deficiencies in a few (less than four), but not where EQ or outcome-related analyses are concerned. 2 = Has minor deficiencies in any quality grid factor, including analysis of EQ 1 = Serious deficiencies SCORING: At heading (dark orange) level only, the rating is translated into a score (extreme right column) evaluated based on the maximum score set in the immediate left column.									
		Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc.	Not applicable	Max. Score	Score
Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.											
Preliminaries										7	
Table of content and Acronyms										1	
Executive Summary										6	
The ES provides an overview of the report, highlighting the main findings, conclusions, recommendations and any overall lessons											
A summary table is included that presents main findings, conclusions and recommendations as well as their logical links											
Main Text										80	
1.Introduction Chapter contains:										1	
The appraisal's rationale, purpose and objectives,											
Scope and main appraisal questions (issues to examine)											
Methodology (incl. data collection and analysis)											
Indicators used and reasons why											
2.Context: This chapter contains:										4	
2.1 Broader context and its influence on the performance of the project or programme											
2.2 Introduction of the intervention being appraised											

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.		Meta-evalua- tor's notes and observations	5	4	3	2	1	Not covered in doc.	Not applicable	Max.	Score
2.3 Link between intervention and GoF strategies and country strategies											
2.4 Implementation strategies proposed											
2.5 Resources required for implementation											
3. Based on the existing version of the PD, description of the programme or intervention being appraised, including										5	
a) its objectives											
b) implementation strategies											
c) resources for implementation											
d) Introduction to the stakeholders and their role, including both beneficiaries and involved institutions											
e) a description of the intervention logic or Theory of Change that is proposed.											
f) The underlying assumptions that support or justify the Theory of Change or intervention logic											
4. Approach, methodology and limitations. The rating should be based on the following:										15	
a) A clear description of the overall approach, including risks and limitations											
b) A statement of the issues that were in the TOR/ITT, especially if the ones reported against are any different.											
c) Data collection and analysis methods											
d) Description of the sources of information or data.											
e) A critical assessment of the validity and reliability of the data and the analysis conducted upon it											
5. Evidence-based Findings:										15	
5.1 Empirical data, facts and evidence relevant to the indicators of the appraisal questions										3	
Overall progress in the preparation of the findings of the appraisal											
5.2 Findings by appraisal criteria or issue:										5	
i) Relevance											
ii) Feasibility Effectiveness											
iii) Feasibility Efficiency											
iv) Feasibility Impact											
v) Sustainability											

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.									
Meta-evaluator's notes and observations									
5	4	3	2	1	Not covered in doc.	Not applicable	Max.	Score	
5.3 Analysis of validity/feasibility of proposed intervention logic							4		
5.4 Findings related to HRBA and CCOs (climate, gender, etc)							3		
6. "Answers" or strategic analysis of issues based on findings							10		
7. Conclusions. The assessment of the performance of the project/programme based on the findings in relation to the set evaluation criteria, performance standards or policy issues							15		
Overall conclusions regarding the logic (LFA, ToC, IL)									
i) Relevance							3		
ii) Feasibility Effectiveness							3		
iii) Feasibility Efficiency							3		
iv) Feasibility Impact							3		
v) Sustainability							3		
8. Recommendations. Proposed improvements, changes, action to improve the project design or to capitalise on strengths.							15		
Recommendations are based on the findings and conclusions							10		
Clear indications are presented as to:							5		
to whom is the recommendation directed (MFA, partner institutions, consultant providing support services, etc.)									
who is responsible for implementing the recommendation									
9. Lessons learned							5		
The report identifies if there are any general conclusions that are likely to have the potential for wider application and use?									
C. Annexes. The following annexes, as a minimum, are included							5		
Terms of reference									
People interviewed									
Documents consulted									

Base reference of the meta-evaluation This represents the standards against which reports will be evaluated.		Meta-evaluator's notes and observations	5	4	3	2	1	Not covered in doc.	Not applicable	Max.	Score
D. NON-CONTENT QUALITY ISSUES NOT SPECIFICALLY MENTIONED IN EVAL MANUAL BUT REQUIRED FOR META_EVAL										3	
Does the appraisal describe, in detail, the intervention logic, the Theory of Change or the logical framework											
Does the appraisal contain an evaluability analysis and a detailed structure for M&E											
Does the appraisal contain a governance and management structure											
Does the appraisal contain a detailed risk assessment that identifies the risks, their probability of occurrence, and their impact?											
Does the appraisal contain a risk mitigation strategy and plan?											
Does the appraisal deal with institutional capability and capacity requirements that are required to enable target institutions to meet their expected outcomes?											
Is the intervention design suggested by the appraisal based on Results (RBM)?										100	
			Total points allocated by evaluator								0

ANNEX 10: QUALITY ASSESSMENT TOOL FOR SUMMARY OF FINNISH DEVELOPMENT COOPERATION

<div>Phase 2 Quality assessment grid</div> <div>with indicative approach categorization and relevant memos</div> <div>Title:</div> <div>Year:</div> <div>Type:</div> <div>Meta-evaluation date: 2016</div>				<div>RATING: Evaluators provide a rating for headline standards/DAC Eval. criteria (dark orange) and their numbered components (in orange and light orange)</div> <div>The 5-point scale is as follow:</div> <div>5 = component assessed positively (i.e. the project, as evaluated, met or exceeded its expected outcomes/results under the appropriate evaluation criteria, and therefore contributed to meeting Finnish development objectives)</div> <div>4 = Minor restrictions to a positive assessment of the component/criteria (i.e. no major setbacks in terms of meeting expected outcomes under the evaluation criteria being analysed)</div> <div>3 = No more than one serious restriction to a positive assessment of the component under the evaluation criteria. Intervention must have met sustainability and effectiveness criteria</div> <div>2 = Major restrictions to more than one positive assessment of the component under the evaluation criteria.</div> <div>1 = Component assessed negatively (i.e. did not met expectations or was not analysed sufficiently in the report to enable rating to take place)</div>			<div>NOTE: where the cell indicates "not applicable", meta-evaluators are not supposed to "Rate" the issue using the five-point rating system to the left. What is sought is an 'inductive' analysis where key comments that refer to the issue are posted in this grid</div>		<div>NOTE: All the points that refer to the priorities of the Finnish Government (2007 or 2012) will be assessed according to the ToR that accompanied/ generated the reports in which they are found.</div>				
Framework for Phase 2 rating assessment and inductive analysis				NA		Inductive reasoning memos		Rating given		Not in document		Methodological guidelines for reviewers	
RELEVANCE													
1. Degree of relevance of the intervention										Relevance: The extent to which the objectives of a development intervention are consistent with beneficiaries' requirements, country needs, global priorities and partners' and donors' policies.			

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
1.1 Consistency with the needs of the target group					<p>Could be ultimate target groups, beneficiaries or intermediary groups. Is target group clearly identified? Are there layers (national, organisation, sub-org, etc.) and is the relevance made clear for each layer? If there are more than one target group (public org and NGO for example), is relevance specified for each? Don't try to pre-define and force a rating based on the various types of target groups such as vulnerable and firms. Too many and we don't really know what they can be at this stage. So use the framework as a guide and, depending on the specifics of the report, speak to the needs of the targeted groups...</p> <p>We should also point out if the nature of the relevance is clear. What was planned in terms of relevance? Relevance must relate to the IL or ToC, or failing that the project plan. In particular, the meta-evaluator should seek out consistency with a) Vulnerable, rural/urban households, b) firms, c) public institutions or NSA and f) other (specify and categorize)</p>
1.2 Alignment with national development goals, policies and plans					<p>In order to be rated, the evaluation must be specific about how the intervention and its components are "aligned". In particular, the meta-evaluator should seek out alignment with a) national, b) sector planning c) sector /sub sector action plan, d) NSA and e) other (specify and categorize)</p>
1.4 What are the key factors that affected the relevance of the program in positive ways?			Not applicable		<p>What we are looking for are the key lessons for MFA concerning how relevance was managed. Was it part of the design? Was it evaluable? Was there indicators? Was there a special effort made to enable and encourage ownership? Was the design scoped wide enough to enable a significant outcome change?</p>
1.5 What are the key factors that affected the relevance of the program in negative ways?			Not applicable		<p>Refer only to factors that were not included in 3.1 to 3.3</p>
1.6 The report explicitly deals with and analyses the extent to which "relevance" is supported through CCO and HRBA					<p>This is a very important consideration and the OVERALL rating given to the headline standard is heavily dependent upon this type of analysis being present. If the analysis is not there then a low rating is to be given BECAUSE IT IS TO BE CONSIDERED AS A "SERIOUS" WEAKNESS.</p>

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
1.7 Link of objectives with Finland's aid priorities (refer to note in upper right-hand quadrant)					<p>In this relevance section, we are assessing intentions not results. So when we say "link of objectives" it means that that we assess whether the reports indicate the extent to which the interventions are going to generate results that will push forward the aid priorities of Finland.</p> <p>A democratic and accountable society that promotes human rights.: This relates to such issues as supporting partner countries' capacity for monitoring and implementing human rights, independent judiciary, freedom of expression, association and assembly, free and fair elections, an accountable government, freedom for civil society, access to decision-making in society, transparency of how public funds are spent, corruption, peace mediation activities, reforming the security sector, disarming combatants, etc.</p> <p>An inclusive green economy that promotes employment. This relates to generating human development with the sustainable use of natural resources, and creating equal possibilities for all for decent work. This may include such things as bringing microentrepreneurs into the formal economy through mobilizing small and medium-sized enterprises, integration of countries into international trade in a balanced way, preventing tax evasion, illicit capital flight, improved sharing of information, accounting standards, corporate social responsibility, and innovation.</p> <p>Sustainable management of natural resources and environmental protection. Refers to issues such as energy efficiency and renewable energy, sustainable urban development, food security, agricultural productivity, drinking water and sanitation, management of water resources, forests</p> <p>Human development. Refers to issues such as literacy, quality education (whether basic, vocational, or higher education), access to education, including people with special needs or from linguistic minorities, high quality research and innovation linked with producing skilled labour for entrepreneurship, and health issues such as maternal and child health, community-based preventive health work, and communicable and non-communicable diseases.</p>

EFFECTIVENESS				
2. Degree of achievement of stated objectives (or likelihood to do so) during the implementation period. Not only "during" but within planning horizon. Relevance needs to cover the future				Effectiveness: The extent to which the development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance. Take into account whether this is a mid-term evaluation and whether the progress towards objectives is congruent with the implementation period (actual implementation time vs. planned). Comments need to be grounded in programme objectives, i.e. to what extent were those realised?
	NA	Inductive reasoning memos	Rating given	Not in document
Framework for Phase 2 rating assessment and inductive analysis				Methodological guidelines for reviewers
2.1 Were the outputs achieved as planned in the implementation period for the following components?				This needs to be based on evidence of outputs. Need to be careful not to confuse inputs (ex. there was a TA in place) with outputs
				Base rating on what is reported in the evaluation report (preferred). When necessary, qualify with reviewer's judgment .
				Technical assistance
				Capital investment
				Trainings
2.2 Degree of achievement of the outcome level for				Policy dialogue
				Other, specify and categorize
				Base rating on what is reported in the evaluation report (preferred). When necessary, qualify with reviewer's. Needs to be based on evidence at outcome level
				Capacity development
				Policy development, reforms
2.3 Give an example of some outstanding outcome and the main related outputs/enabling factors				Social development, health, education
				Economic development
				Infrastructure development
				Other, specify and categorize
			Not applicable	Instead of noting specifics from the report, be more general in the types of successes you report. Try to focus on the design of the intervention. If the outcome selected was not within the control of the implementation team, please identify.

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
2.4 Identify one severe shortcoming that was identified in the report and the main related outputs/hindering factors			Not applicable		Instead of noting specifics from the report, be more general in the types of weaknesses you report. Reported shortcomings should be within the programme's control, e.g. was programme design one of the shortcomings? See if you can identify weaknesses in the design and not the project implementation.
2.5 Indicate main factors quoted by evaluators that enabled the outcomes to be achieved			Not applicable		Within or without programme control Existence of a social/institutional demand Flexibility in managing resources Coordination with other donors Other, specify and categorize
2.6 Indicate main factors quoted by evaluators that hindered the achievements of outcomes			Not applicable		Note the main implementing issues, or contextual issues outside or within programme control. Insufficient buy-in by implementing agencies or beneficiaries Insufficient resources to respond to needs Shortage of qualified expertise Other, specify and categorize
2.7 The report explicitly deals with and analyses the extent to which "EFFECTIVENESS" is supported through CCO and HRBA					This is a very important consideration and the OVERALL rating given to the headline standard is heavily dependent upon this type of analysis being present. If the analysis is not there then a low rating is to be given BECAUSE IT IS TO BE CONSIDERED AS A "SERIOUS" WEAKNESS.
IMPACT					
3. Contribution to the achievement of impact (intermediary) level results (even if not an "impact evaluation" per se)					Impacts: Positive and negative; primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended. Rating required for mid-term evaluations: N/A. If a section labelled "impacts" actually contains information on outputs or outcomes, these examples should not be included here.
3.1 Degree of achievement of main intended intermediary impacts					Not required for MTE/MTR. Look for "types" of impacts, but will need to see it is possible to establish some kind of typology once we see patterns from the first set of reviews. 3.1.1 to 3.1.5 are Finland aid objectives. A democratic and accountable society that promotes human rights see comments right An inclusive green economy that promotes employment see comments right Sustainable management of natural resources and environmental protection see comments right Human development see comments right Other, specify and categorize

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
3.2 Indication of the existence of unintended intermediate impacts			Yes/no		Not required for MTE/MTR. Look for "types" of impacts. Impacts must be evidence-based and have a solid means of measuring. Clearly avoid such terms as "improve" or "clarify" etc..
3.3 Indication of an impact on social inequality (poverty)?			Yes/no		
3.4 Extent to which the report explicitly deals with and analyses the extent to which "IMPACT" is supported through CCO and HRBA					This is a very important consideration and the OVERALL rating given to the headline standard is heavily dependent upon this type of analysis being present. If the analysis is not there then a low rating is to be given BECAUSE IT IS TO BE CONSIDERED AS A "SERIOUS" WEAKNESS.
EFFICIENCY					
4. Degree of performance of the intervention oversight and management					Efficiency: A measure of how economically resources/inputs (funds, expertise, time, etc.) are converted to results.
4.1 Extent of transformation efficiency					In order to be rated, the evaluation report must be specific on how the transformations were done and the extent to which they were efficient (compared to some other option). The report should also identify the extent to which these resources were sufficient to meet the objectives set for the intervention.
					Technical assistance
					Capital investment
					Trainings
					Policy dialogue
					Other, specify and categorize
4.2 Extent of time efficiency					Need to see if the report explained the consequences of poor timing if such was the case
					Technical assistance
					Capital investment
					Trainings
					Policy dialogue
					Other, specify and categorize

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
					To be rated, the report needs to be specific concerning these management mechanisms.
4.3 Degree of quality of the oversight, decision-making and management reactivity					
4.4 Indicate main enabling factors quoted by evaluators			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
					Flexible procedures
					Decentralized decision-making
					Responsive partners
					Quality of the appraisal
4.5 Indicate main hindering factors quoted by evaluators			Not applicable		Other, specify and categorize
					This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
					Cumbersome procedures
					Inadequate cost estimates
					Lack of ownership
4.6 Extent to which the report explicitly deals with and analyses the extent to which "EFFICIENCY" is supported through CCO and HRBA					Lack of enabling environment for CD
					Other, specify and categorize
					This is a very important consideration and the OVERALL rating given to the headline standard is heavily dependent upon this type of analysis being present.
					If the analysis is not there then a low rating is to be given BECAUSE IT IS TO BE CONSIDERED AS A "SERIOUS" WEAKNESS.

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
SUSTAINABILITY					
5. Degree or prospects for sustainability					Sustainability: The continuation of benefits from a development intervention after major development assistance has been completed.
5.1 Degree or prospects for social sustainability					Social sustainability refers to sensitivity to cultural issues, including the dynamics of relationships amongst stakeholders, and broad participation of beneficiaries through an inclusive approach. Rights to assess to information and freedom of opinion, and ensuring that a programme should do no harm to easily marginalised groups. Quoted factors enabling social sustainability can be inter alia community participation, community organisation, setting of a maintenance system, political backing... Specify the key factor if quoted in the report.
5.2 Degree or prospects for financial/economic sustainability					Financial sustainability refers to programme benefits being sustained after external financing ends, such as prospects for funding activities through incomes from improved services, beneficiary contribution, tax or other domestic revenues over longer term. Economic sustainability refers to the distribution of benefits and costs among programme stakeholders, as well as the flows of resources between them, analysed and depicted separately for each stakeholder, especially final beneficiaries (qualitative analysis is normally sufficient). Quoted factors enabling financial/economic sustainability can be inter alia the setting of a cost recovery system, guaranteed budget allocations, oversight and protection against leakages in place... Specify the key factor if quoted in the report.
5.3 Degree or prospects for environmental sustainability					Refers to issues such as ecological sustainability, sustainable use of natural resources, prevention of environmental pollution, biodiversity, desertification, soil impoverishment, sustainable use and control of chemicals and waste, social protection and adaptation for those who are most vulnerable to environmental disasters and climate change, and/or consideration of environmental regulations such as Environmental Impact Assessment and international conventions such as the United Nations Framework Convention on Climate Change. Quoted factors enabling environmental sustainability can be inter alia targeted plans, inclusion on environment in implementation contracts, progress reports, M&E systems, targeted resources in implementation plans... Specify the key factor if quoted in the report.
5.4 Findings related to technical sustainability					Refers to appropriate technology selections and building up capacity for sustainable operation and maintenance of technologies. Includes considerations such as whether the technology is socially acceptable, affordable, generates employment, and replicable.

Framework for Phase 2 rating assessment and inductive analysis		NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
5.5 Degree or prospects for organisational sustainability						Refers to whether the capacity of the implementing organisation(s) have the capability and capacity to continue to deliver their expected outcomes over the mid to long term. Considerations include external factors affecting the organisation (e.g. political climate in country); relationships with other organisations; capability at all levels (people and systems, resources and enabling environment). Sustainability also considers whether the established capability can deliver to all expected targeted populations.
5.6 Indicate main enabling factors quoted by evaluators				Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
5.7 Indicate main hindering factors quoted by evaluators				Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
5.8 The report explicitly deals with and analyses the extent to which "SUSTAINABILITY" is supported through CCO and HRBA						This is a very important consideration and the OVERALL rating given to the headline standard is heavily dependent upon this type of analysis being present. If the analysis is not there then a low rating is to be given BECAUSE IT IS TO BE CONSIDERED AS A "SERIOUS" WEAKNESS.
AID EFFECTIVENESS						
6.1 Degree of coordination and harmonisation with other donors, host country organisations, civil society, NSA etc.						In order to be rated, the evaluation must make an assessment or judgment on the effectiveness and efficiency gains provided through coordination and harmonisation. In particular, the meta-evaluator will examine coordination and harmonisation dealing with a) co-financing b) coordination, c) complementarity d) division of labour e) joint undertakings and f) other
6.2 Extent to which mutual accountability is assured						Use the Paris Declaration definition

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
6.3 Extent to which the partner country is in the lead (i.e. formulate and implement their own national development plans, according to their own national priorities, using, wherever possible, their own planning and implementation systems)					Self-evident, but the document must specifically refer to the issue
6.4 Extent to which the intervention actually focusses on the management for results: i.e. focus on the result of aid, the tangible difference it makes in poor people's lives. Develop better tools and systems to measure this impact					Self-evident, but the document must specifically refer to the issue
6.5 Extent to which the intervention contributes to building more effective and inclusive partnerships.					Self-evident, but the document must specifically refer to the issue
6.6 Extent to which the report deals with the sector targets and indicators for aid effectiveness					Self-evident, but the document must specifically refer to the issue

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
OVERALL MANAGEMENT AND CONSEQUENCES OF HRBA and CROSS-CUTTING OBJECTIVES					
Human Rights-Based Approach					
7. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the crosscutting objective of HRBA. (refer to note in upper right-hand quadrant)					Value-based development policy promotes the core human rights principles such as universality, self-determination, non-discrimination and equality. The human rights-based approach to development includes civil and political rights and freedoms as well as economic, social and cultural rights. Finland emphasises the rights of women, children, ethnic, linguistic and religious minorities and indigenous people, the rights of persons with disability, people living with HIV and AIDS, and the rights of sexual and gender minorities.
7.1 The project or programme is congruent with the MFA policy on HRBA.					To be rated the report must be explicit on how the intervention is HRBA -compliant
7.2 What are key reported successes related to HRBA?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
7.3 What are key reported shortcomings related to HRBA?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
7.4 Is there evidence in the report that achievements in terms of HRBA will be sustained after the project or programme is completed?			Yes/no		This is a sustainability issue but I agree that it should remain here. Otherwise it will require a new element in the Sustainability section , as will all other CCO

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
Gender Equality					
8. The evaluation report indicates that the project or programme supported by the MFA effectively contributes to the cross-cutting objective of gender equality (planned higher-level results in this area are realized).(refer to note in upper right-hand quadrant)					Gender equality includes strengthening the role of women, through economic development and the promotion of well-being. Also includes support to the participation of women in decision-making and rejects any form of discrimination giving rise to gender inequality.
8.1 The programme is congruent with the MFA policy on gender equality.					To be rated the report must be explicit on how the intervention is gender -compliant
8.2 What are key reported successes related to gender equality (including any level of results: impact, outcomes or outputs)?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
8.3 What are key reported shortcomings related to gender equality?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
8.4 Is there evidence in the report that higher-level results in terms of gender equality will be sustained after the programme is completed?			Yes/no		We are looking for an indication that there is evidence to support the report's conclusion on this issue

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
REDUCTION OF INEQUALITY					
9. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the cross-cutting objective of reduction of inequality. (planned higher-level results in this area are realized). (refer to note in upper right-hand quadrant)					Reduction of inequalities includes support to social policies that increase equal opportunities for social, economic, and political participation as well as access to basic services and a social protection floor. Good nutrition, health, education, decent work and basic social protection as well as the realisation of the basic labour rights have a key role.
9.1 The programme is congruent with the MFA policy on reduction of inequalities.					To be rated the report must be explicit on how the intervention is gender-compliant
9.2 What are key reported successes related to reduction of inequality?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
9.3 What are key reported shortcomings related to reduction of inequality?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
9.4 Is there evidence in the report that higher-level results in terms of reduction of inequalities will be sustained after the programme is completed?			Yes/no		We are looking for an indication that there is evidence to support the report's conclusion on this issue

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
CLIMATE SUSTAINABILITY					
10. The evaluation report indicates that the project or programme supported by the MFA effectively addresses the cross-cutting objective of climate sustainability. (planned higher-level results in this area are realized). (refer to note in upper right-hand quadrant)					Finland promotes low carbon development and the capacity of its partner countries to adapt to climate change, and furthers integration of these goals into partner countries' own development planning. Finland also supports long-term measures that reduce the vulnerability of people and communities to natural disasters.
10.1 The programme is congruent with the MFA policy on climate sustainability.					To be rated the report must be explicit on how the intervention is gender -compliant
10.2 What are key reported successes related to climate sustainability?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
10.3 What are key reported shortcomings related to climate sustainability?			Not applicable		This component is for inductive analysis only. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
10.4 Is there evidence in the report that achievements in terms of climate sustainability will be sustained after the programme is completed?			Yes/no		We are looking for an indication that there is evidence to support the report's conclusion on this issue

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
Other Cross-cutting Objectives or Emerging Themes					
11 The evaluation report indicates that the project or programme supported by the MFA effectively addresses another cross-cutting objectives or emerging themes. (refer to note in upper right-hand quadrant)					MFA is interested in whether there are cross-cutting objectives or emerging themes from projects OTHER than those listed in their manual. Please briefly describe it under Comments. Those may include: capacity building, innovation, etc. Leave blank if there is no emerging theme
11.1 What was/were the other cross-cutting objective(s) or emerging theme(s)?					Leave blank if there is no emerging themes. This component is for inductive analysis. Use text analysis to identify point of interest to the analysis of Finland's development cooperation
11.2 What are the reported successes related to the other cross-cutting objective(s) or emerging theme(s)?			Not applicable		
11.3 What are the reported shortcomings related to the other cross-cutting objective(s) or emerging theme(s)?			Not applicable		
11.4 Is there evidence in the report that achievements in terms of the other cross-cutting objective(s) or emerging theme(s) will be sustained after the programme is completed?			Yes/no		

Framework for Phase 2 rating assessment and inductive analysis	NA	Inductive reasoning memos	Rating given	Not in document	Methodological guidelines for reviewers
RISK MANAGEMENT					
12. Degree of implementation and success of the risk management strategy.					<p>What we are trying to bring forward here for senior MFA managers is whether or not the intervention's risks were known (identified) and if so, whether they were managed actively. That does not mean that there were no risks or that the implementers reduced the occurrence or impact of risks to zero; knowing that the intervention was given every change of success (in this case through risk management) is the focus. You may need to infer what risks are. Risk related to sustainability should be described in more detail, e.g. what is the underlying risk that may cause sustainability issues?</p> <p>Types of risks according to the bilateral manual include:</p> <ul style="list-style-type: none"> - Developmental risks - Financial of fiduciary risks - Reputational risks - Risks of action and avoiding actions particularly in fragile states <p>Based on research, we should also identify who tried to mitigate the risk, hoping that the host country took an active part in that.</p>
12.1 Existence of an articulated risks management strategy (i.e. a thorough and reasoned analysis with identified risks, their chance of occurrence, the impact of occurrence and the mitigation strategy).					
12.2 Findings demonstrating a gain in utilising the risks management strategy.					Again, it would be best if this were clearly reported against. If it is not, then see what major problems were encountered and then see if this was identified as a risk.
ADDITIONAL COMMENTS					
13. Include any additional comments or examples of best practices on how the programme contributed to Finland's development cooperation.					

ANNEX 11: KEY FINDINGS, CONCLUSIONS AND RECOMMENDATIONS (FINNISH)

Seuraavan taulukon on määrä toimia pohjana johdon vastaukselle tähän meta-evaluointiin.

Taulukon alussa esitetään tuloksia, jotka eivät suoraan liity varsinaisiin evaluointikysymyksiin, vaan esittelevät meta-evaluoinnin yleistä lähestymistapaa. Asiakirjan muut osat on jäsennetty siten, että se vastaa tehtävänmäärittämisessä esitettyjä evaluointikysymyksiä. Kuten raportissa on kuvattu, päätelmät perustuvat suureen määrään näyttöön perustuvia löydöksiä. Koska tässä on kyseessä strateginen evaluointi, suositukset koskevat lähes aina useampaa kuin yhtä päätelmää, kuten on laita myös raportin suositusosiossa. Tässä asiakirjassa laajennetaan jonkin verran näitä suosituksia ulkoasiainministeriön raporttien laadintaohjeiden mukaisesti.

Jäljempänä eriteltyjä tuloksia ei tule pitää evaluoinnin ainoina havaintoina ja päätelminä, etenkin käytettäessä niitä johdon vastausten valmisteluun. Koko raportti on otettava huomioon, ja toimintayksiköiden ja osastojen olisi poimittava raportista itselleen olennaiset havainnot ja valmisteltava vastauksensa niiden perusteella.

Osa 1: Havainnot, jotka eivät suoraan liity evaluoitinkysymyksiin

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Havaintoihin tai päätelmiin liittyvät suositukset (organisaatio, jolle suositus on pääasiassa osoitettu, on annettu sulkeissa)
Portfolioanalyysi, lähestymistapa ja metodologiset löydökset		
1.1) Tässä raportissa on käytetty monitasoisempaa lähestymistapaa ja metodologia kuin aikaisemmissa ulkoasiainministeriön meta-evaluoinneissa käytetty, ja sen avulla EVA-11 saa paremman käsityksen portfolioon yleisestä laadusta. Arviointityökaluja on muokattu huomattavasti, ja nyt ne mahdollistavat sisällön laatuun perustuvan kvantitatiivisen analyysin. Suomen kehitysyhteistyön laatua voidaan nyt arvioida sekä deduktiivisella että induktiivisella pohjalta, ja mikä tärkeintä, arviointi perustuu nyt siihen, missä määrin ulkoasiainministeriön politiikan standardeja ja vaatimuksia on noudatettu. Pitkittäisanalyysit eivät ole mahdollisia ennen seuraavaa meta-evaluointikierrosta.	1.1) Tässä meta-evaluoinnissa käytetty lähestymistapa tuottaa huomattavasti suuremman määrän näyttöön perustuvan analyysin kuin aiemmissa meta-evaluoinneissa käytetyt lähestymistavat, jotka olivat pääosin deskriptiivisiä.	1.1) Tätä lähestymistapaa tulisi käyttää paranneltuna seuraavilla meta-evaluointikierroksilla. (EVA-11 ja ulkoasiainministeriön ylemmät johtajat)
1.2) Odotettuun laatutasoon perustuvat arviointijärjestelmät (kuten tässä meta-evaluoinnissa käytetty) ovat erityisen hyödyllisiä johdon varmistustyökaluina, koska ne perustuvat raporttien läpinäkyviin ja helposti eteenpäin välitettäviin laatuksiteereihin (ne perustuvat tunnettuihin normeihin tai standardeihin).	1.2. Meta-evaluoinnissa käytetty lähestymistapa tuottaa näyttöön perustuvaa tietoa, joka tukee ulkoasiainministeriön ylempien virkailijoiden vastuuvuorollisuutta.	1.2) Tämän meta-evaluoinnin yleinen rakenne tulisi integroida ulkoasiainministeriön raportointirakenteen vastuuvuorollisuuden viitekehukseen. (Ulkoasiainministeriön ylempi johto)
1.3) Yleensä ottaen Suomen kehitysyhteistyössä käytettyjä konsepteja ja menettelytapoja ei noudateta johdonmukaisesti, eivätkä kaikki käyttäjät luultavasti ymmärrä niitä samalla tavalla. Standardit ovat usein tulkinnanvaraisia (evaluoinneissa käytetään usein tarkemmin määrittelemättömiä sanoja kuten "riittävästi" tai "parantunut"), ja normeja voidaan tulkita huomattavan vapaasti (esim. indikaattorien laatu tai loogisen viitekehksen rakenne). Monet Suomen politiikat ja konseptit ilmaistaan käsitteellisinä termeinä, mutta niitä ei määritellä täsmällisesti (esim. tulospohjainen, kestävä), mistä seuraa erilaisia tulkintoja ja raportointikäytäntöjä.	1.3) Ulkoasiainministeriön virkamiehet eivät käytä päivittäisessä työssään peruskäsitteitä samoista lähtökohdista. Johdonmukaisuuden puuttuminen vaikuttaa koko ulkoasiainministeriön mandaatin toimintaan ja aiheuttaa epätydellistä palautetta, epäjohtamuksesta suunnittelua ja huonoa laadunvalvontaa.	1.3) Ulkoasiainministeriön tulisi tutkia ongelman syitä ja syyt selvitettyään ryhtyä korjaaviin toimenpiteisiin, joilla varmistetaan, että käsitteet ymmärretään ja niitä käytetään yhdenmukaisesti. Muussa tapauksessa ministeriön kehityspoliittisen toteutus- ja valvovan tehtävän hoitaminen kärsii. (EVA-11 ja ulkoasiainministeriön ylemmät johtajat)

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Havaintoihin tai päätelmiin liittyvät suositukset (organisaatio, jolle suositus on pääasiassa osoitettu, on annettu sulkeissa)
<p>1.4) Monia Suomen kehitysyhteistyöhön liittyviä korkeamman tason tavoitteita ei käsitellä evaluointi- tai arviointiraporteissa. Myöskään interventioille (hankkeille) määritetyt seuranta- ja evaluointikriteerit eivät yleensä käsittele näihin kehityspoliittisiin tavoitteisiin liittyviä yksityiskohtia, joista olisi hyötyä ulkoasiainministeriön ylemmälle hallinnolle.</p>	<p>1.4) Evaluointiprosessi ja sen täytäntöönpanon laajuus eivät ole tarpeeksi kattavia, jotta saataisiin vakuuttavia näkemyksiä kehityspoliittikan toteutuksen koko kirjosta.</p>	<p>1.4) Vaadittu raporttien sisältö sekä evaluointien ja etukäteisarviointien laajuus on tarkistettava, jotta ne kattavat kaikki asianmukaiset kehityspoliittikan alueet. Ulkoasiainministeriössä kehyspolitiikasta vastaavien yksiköiden olisi määritettävä, minkä tyyppistä strategista ja operationaalista tietoa ne tarvitsevat.</p> <p>(EVA-11 ja politiikan kehityksestä vastaavat yksiköt ja osastot)</p>
<p>1.5) Ulkoasiainministeriön virkamiesten kesken on merkittäviä eroja siinä, miten he ohjaavat arviointien ja evaluointien hoitamista. Raporttien ja tehtävänkuvaustenlaatu vaihtelee huomattavasti, ja portfolioon sisältyy nyt huonolaatuisiaakin raportteja, mikä johtuu osittain viranomaisten laadunvalvonnan heikkoudesta.</p>	<p>1.5) Tässä meta-evaluoinnissa mainitut heikkommat suoritukset liittyvät suurelta osin suoraan ulkoasiainministeriön virkamiesten valmiuksiin ja kapasiteettiin ottaa käyttöön ja noudattaa ulkoasiainministeriön evaluointiprosesseja (etukäteisarvioinnit mukaan lukien).</p>	<p>1.5) Ulkoasiainministeriön tulee suorittaa arviointien valmiuksista ja kapasiteetista ja sitten kehitettävä ja rahoitettava strategia ja suunnitelma kyky- ja kapasiteettipuutteiden poistamiseksi.</p> <p>(Ulkoasiainministeriön ylempi johto)</p>
<p>1.6) Tässä meta-evaluoinnissa tarkastellun portfolioin hankkeiden osalta vain 12 (hieman yli kolmannes) evaluoiduista projekteista keskittyi Suomen kehitysyhteistyön virallisiin ja perinteisiin kahdenvälisiin kumppanimaihin (Etiopia, Kenia, Mosambik, Nepal, Tansania, Vietnam ja Sambia), vaikka vuoden 2004 kehityspoliittisessa ohjelmassa tehtiin sitova päätös keskittyä harvempiin maihin ja harvempiin sektoreihin näissä maissa.</p>	<p>1.6) Tehdyn politiikkaratkaisun ja kentällä vallitsevan todellisuuden välillä on ristiriita.</p>	<p>1.6) Ulkoasiainministeriön tulisi analysoida tämä tilanne ja päättää, mitä kehityspoliitiikkaa se haluaa noudatettavan.</p> <p>(Ulkoasiainministeriön ylempi johto)</p>
<p>1.7) Suomen kehitysyhteistyön tavoitteet, jotka vuoden 2007 kehityspoliittinen ohjelma määritteli painottaen vahvasti ympäristöön, maanviljelykseen, liiketoimintaan ja kaupankäyntiin liittyviä asioita, ovat vasta nyt vahvasti näkyvillä projektien ja evaluointiraporttien portfolioissa.</p> <p>Kun lisäksi otetaan huomioon vuoden 2007 kehityspoliittisen ohjelman laajempi tavoite, ekologisesti kestävä kehitys ja luonnonvarojen käyttö, Suomen avun keskittyminen ”laajemmalle” ympäristösektorille tulee vielä ilmeisemmäksi. Jos lasketaan yhteen varsinainen ympäristö (ilmasto mukaan lukien), maanviljely, vesi- ja jätehuolto sekä metsäala, luonnonvaroihin liittyvät hankkeet edustavat meta-evaluointiportfoliossa 40:tä prosenttia kaikista projekteista. Jos portfolio on edustava, se näyttää vahvistavan havainnon, jonka kehityspoliittinen toimikunta vuonna 2009 ilmaisi lausunnonsaan valtioneuvostolle: ”Suomen kehitysyhteistyötä ohjannut pyrkimys maakohtaiseen keskittymiseen näyttää olevan vaihtumassa uudessa [vuoden 2007] ohjelmassa sektorin- tai teemakohtaisuuteen” (Suomen kehityspoliittikan tila 2009, s. 20).</p>	<p>1.7) Vuoden 2007 kehityspoliittisen toimintaohjelman vaikutukset alkavat näkyä.</p>	<p>1.7) Toimenpiteitä ei tarvita.</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Havaintoihin tai päätelmiin liittyvät suositukset (organisaatio, jolle suositus on pääasiassa osoitettu, on annettu sulkeissa)
1.8) Yli puolet projekteista on pieniä (51 %; budjetti enintään 5 miljoonaa euroa). Kuusi prosenttia (kaksi projektia) sai käyttönsä yli 20 miljoonan euron budjetin. Budjetin kannalta kaksi suurinta projektia edustaa 20:tä prosenttia (n=35) kaikkien budjettimäärärahojen summasta (49 miljoonaa euroa, tai 28 %, kun etukäteisarvioinnit jätetään pois laskusta, n=26).	1.8) Suomen rahoituksen osalta portfolio koostuu suuresta määrästä pieniä projekteja ja pienestä määrästä hyvin suuria projekteja.	1.8) Koska useimmat avunantajat ja OECD pyrkivät siihen, että suuremista hankeinterventioista tulee vakiokäytäntö, ulkoasiainministeriön tulisi arvioida, tuottavatko pienemmän mittakaavan interventiot todella sellaisia vaikutuksia, joita Suomen hallitus kehitysyhteistyöpolitiikallaan hakee. (Ulkoasiainministeriön ylempi johto)
1.9) Vuoden 2012 OECD/DAC:in vertaisarviossa todettiin, että apu ei ollut keskittynyt tärkeimpiin kumppanimaihin (s. 47–48). Tekemämme portfolioanalyysi tukee tätä vuoden 2012 havaintoa.	1.9) Jos oletetaan, että meta-evaluointimme portfolio on edustava, toteamme, että tämä käytäntö (avun pirstaloituminen) on pysynyt ennallaan.	1.9) Ulkoasiainministeriön on tutkittava tarkemmin käytäntöään olla keskittämättä tukeaan tärkeimpiin kumppanimaihin. (Ulkoasiainministeriön ylempi johto)

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
ULKOASIAINMINISTERIÖN HAJAUTETUN EVALUOINTIPORTFOLION LAATU (EVALUOINTIRAPORTIT JA NIIHIN LIITTYVÄT TEHTÄVÄNKUVAUKSET) EK 1: Mikä on ulkoasiainministeriön hajautetun evaluointiportfolion laatu (evaluointiraportit ja niitä vastaavat tehtäväkuvaukset) perustuen vuosien 2014–2015 OECD/DAC:in evaluointistandardeihin sekä Evaluointikäsi kirjassa annettuihin ohjeisiin ja vaatimuksiin luokiteltuna maiden, sektorien, evaluointityyppien, ulkoasiainministeriön hallintoyksiköiden, toimeksiantavan tahon, konsultointiyritysten jne. mukaan? Onko ulkoasiainministeriön tilaamien evaluointien ja ulkoasiainministeriön kumppanitahojen tilaamien evaluointien laadulla eroa?		
<p>2.1) Evaluoinnin tehtäväkuvauksille annettu yleisarvio oli 64/100. Melko huonoiksi arvioitut osiot sisältävät monia ydinasioita, kuten evaluointikysymykset, edellytys arvioida avun tuloksellisuutta, metodologia ja konteksti (hankkeen ja evaluoinnin).</p> <p>Lähes kaikissa tehtäväkuvauksissa oli ohjeistuksessa vaaditut osiot, paitsi avun tuloksellisuuden arviointi, joka oli mukana vain kuudessa 22:sta tehtävännäärityksestä. Monien tehtäväkuvauksen osien (kategorioiden) laatu oli kuitenkin melko huono, etenkin kun otetaan huomioon, että tehtäväkuvaukset ohjaavat konsultteille ulkoistettua evaluointia ja että ulkoasiainministeriön tavoitteiden saavuttaminen riippuu näistä analyyseistä.</p> <p>Tehtäväkuvaukset saivat positiiviset arviot kohdista ”perustelut, tarkoitukset ja tavoitteet” ja ”resurssit” sekä evaluointiprosessin kuvauksesta. Ne saivat huonot arviot kriittisistä eli ydinasioista, kuten tutkittavista evaluointikysymyksistä.</p>	<p>2.1) Meta-evaluoinnissa todettiin, että tehtäväkuvaukset olivat heikkoja useilla tärkeillä alueilla, mukaan lukien ydinasiat, joissa tarvitaan erityistä interventioon liittyvää ohjausta, kuten evaluointikysymysten esittäminen, ohjeet avun tuloksellisuuteen sitoutumisen analysoinnista, metodologiaan liittyvät suositukset ja konteksti.</p> <p>Tehtäväkuvauksen laatu on paljon heikempi kuin se saisi olla, kun otetaan huomioon sen merkitys ulkoistamisessa, se että tehtäväkuvauksen laatu on täysin ulkoasiainministeriön virkamiesten käsissä sekä laadunvalvontatehtävä, joka ulkoasiainministeriön virkamiehillä on osana valtionhallinnon ja erityisesti ulkoasiainministeriön vastuuvollisuutta.</p>	<p>2.1) Ulkoasiainministeriön yleisesti ja EVA-11:n pitäisi kehittää voimavarojen kehitysstrategia ja rahoittaa tarkka kehityssuunnitelma, jotta voidaan huomattavasti parantaa evaluoinnin tehtäväkuvauksen laatua.</p> <p>Ulkoasiainministeriön johtotason tulisi pystyä suorittamaan alaisten tuotteiden laadunvarmistusta, ja EVA-11:n on määriteltävä keinot varmistaa, että johtajat pystyvät siihen.</p> <p>EVA-11:n on laadittava opetus-työkaluja ja verkkomateriaalia sekä esimerkkejä evaluointikysymysten keskeisestä tehtävästä ja siitä, miten niitä laaditaan.</p> <p>EVA-11:n olisi käytettävä tälle meta-evaluoinnille laadittuja arviointityökaluja ja muokattava ne laadunvalvonnan tarkistuluetteloksi ja muiksi työkaluiksi, joita ulkoasiainministeriön virkamiehet voivat käyttää.</p> <p>(EVA-11 ja kaikki toimintayksiköt ja jaostot)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
2.2) Ulkoasiainministeriön kehityspolitiikan toimeenpanosta vastaavien yksiköiden ja osastojen kirjoittamien tehtävänkuvauksien yleinen laatu on enemmän tai vähemmän samalla tasolla kuin kansainvälisten verrokeiden, jotka kirjoittivat ulkoasiainministeriön rahoittamien projektien tehtävänkuvaukset, sikäli kuin voimme olettaa, että meta-evaluoinnin portfolio on edustava.	2.2) Ei ole mitään syytä uskoa, että ulkoasiainministeriön tuottamien evaluointiasiakirjojen laatu on riittävä vain, koska kansainvälisille verrokeille annetut pisteet vastaavat ulkoasiainministeriön virkamiesten tuotteille annettuja pisteitä. Tämän ja muiden selvitysten tulosten perusteella voidaan päätellä, että kummankaan (sen enempää ulkoasiainministeriön kuin muidenkaan) laatu ei ole riittävä.	2.2) Pitkällä aikavälillä ulkoasiainministeriön tulisi pyrkiä aina tarkistamaan Suomen julkisista varoista rahoitettujen hankkeiden evaluoinneista vastaavien muiden tahojen laatimat evaluointiasiakirjat (eli suositamaan laadunvarmistusta). (Kaikki ulkoasiainministeriön toimintayksiköt)
2.3) Evaluointiraporttien yleisarvosana on lähes sama kuin tehtävänkuvauksille annettu. Useimmat osiot eivät täytä laatustandardeja, ja monia aiheita ei ole käsitelty ollenkaan.	2.3) Tässä portfolioissa tutkittujen raporttien laatu ei ole tyydyttävä, jos vastuuvälisyyttä, laadunvarmistusta, vaikutusellisuutta ja tuloksellisuutta käytetään pääasiallisina arviointikriteereinä.	2.3) Kuten muualla on todettu, EVA-11:n ja ulkoasiainministeriön ylempien johtajien tulisi kehittää kattavat laatustandardit, joita ne odottavat kaikkien ulkoasiainministeriön virkamiesten noudattavan, joko itse laatimalla tai valvomalla ulkoistettua työtä. (EVA-11 ja ulkoasiainministeriön ylempät johtajat)
2.4) Ulkoasiainministeriön alueellisten ja toimialakohtaisten yksiköiden tilaamat raportit saavat keskimäärin pisteiksi 56,95, kun taas muiden avunantajien tilaamat raportit saavat 69,09 pistettä, eli ero on lähes 13 pistettä.	2.4) Ulkoasiainministeriön tilaamat raportit ovat laadultaan huomattavasti (lähes neljänneksen) heikompiä kuin muiden avunantajien tilaamat.	2.4) Erityisiä suosituksia ei vaadita, koska aiheita on käsitelty jo muissa suosituksissa.
2.5) Suomen tilaamat evaluoinnit eroavat huomattavasti muiden avunantajien hallinnoimista evaluoinneista, ja ulkoasiainministeriön tuottamien tehtävänkuvauksien laadun ja hyväksytyjen raporttien välillä on eroja: ulkoasiainministeriön virkamiehet hyväksyivät hylkäämisen sijaan raportteja, jotka joissakin tapauksissa saivat jopa 20 pistettä vähemmän kuin tehtävänkuvaukset (jolloin ne ovat selkeästi huonolaatuisia).	2.5) Kehityspolitiikan toimeenpanosta vastaavilla yksiköillä on huomattavia puutteita evaluointien hallinnoinnissa.	2.5) EVA-11:n tulisi tutkia, mitkä tekijät ovat johtaneet siihen, että evaluointeja hallinnoivat virkamiehet hyväksyvät heikkolaatuisia tuotteita ja sitten laatia strategia laadunvarmistuksen parantamiseksi tällä alueella. (EVA-11 ja toimintayksiköt)

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>2.6) Toisin kuin edellisessä meta-evaluoinnissa, jossa ei todettu merkittävää korrelaatiota tehtävänkuvauksen laadun ja raporttien laadun välillä, nykyinen meta-evaluointi löysi tilastollisesti merkittävän, joskaan ei vahvan, korrelaation (0.58) näiden kahden välillä.</p> <p>Keskimmääiset tehtävänkuvauksille ja evaluointiraportteille annetut arvosanat vaihtelevat laajalti ulkoasiainministeriön yksiköiden välillä (38–83).</p>	<p>2.6) Evaluoinnissa on todennäköisesti järjestelmätason heikkouksia, koska sekä tehtävänkuvauksissa että niitä vastaavissa raporteissa on havaittavissa merkittävää laadun heikkoutta. Jotkin ulkoasiainministeriön yksiköt saivat muita paljon korkeammat pisteet laadun suhteen.</p>	<p>2.6) Ulkoasiainministeriön ylemmän johdon tulisi tutkia syitä siihen, miksi jotkin yksiköt tuottavat laadukkaampia tuotteita kuin toiset, sekä hyödyntää tästä saatuja oppeja ja todettuja parhaita käytäntöjä evaluointeihin ja arviointeihin liittyvän toiminnan parantamiseksi koko organisaation tasolla.</p> <p>Liittyen joihinkin meta-evaluointista tehtyihin johtopäätöksiin seuraavia tarkempia ehdotuksia tarjotaan harkittaviksi:</p> <p>a) Aloitusraporttien (inception reports) metodologisia vaatimuksia on kiristettävä huomattavasti (UM:lle tulisi esittää hyväksyttäväksi tarkka metodologia, johon sisältyvät tietolähteet, indikaattorit, tiedonkeräys- ja analyysityökalut, otantamenetelmät, haastatteluoppaat ja haastattelumuistiinpanomallit).</p> <p>b) On vaadittava tiukkaa todistusaineistoa kaikkien havaintojen tueksi.</p> <p>c) Evaluointeihin ja arviointeihin liittyviä odotuksia on määritettävä tarkemmin kolmen suomalaisen kriteerin – johdonmukaisuus, suomalainen lisäarvo ja avun tuloksellisuus – osalta.</p> <p>d) On laadittava erityinen ohjeasiakirja, jossa käsitellään erityisesti raporttien hyväksyttävää sisältöä ja tarjotaan raporteille normit ja standardit.</p> <p>(EVA-11 ja koko ulkoasiainministeriön hallintotiimi)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
EVALUOINNIN KATTAVUUS ULKOASIAINMINISTERIÖSSÄ		
EK 2: Mikä on ulkoasiainministeriön evaluoinnin kattavuus (evaluointisuunnitelmien ja toteutettujen evaluointien vertailu)?		
Meta-evaluointi ei löytänyt tähän kysymykseen vastaamiseen tarvittuja tietoja. Myös ulkoasiainministeriön EVA-11 on myöntänyt, että tietoa ei ollut saatavissa muodossa, joka olisi tehnyt evaluoinnin kattavuusanalyysin mahdolliseksi. Siitä syystä tämä EK jätettiin käsittelemättä.	--	--
ETUKÄTESARVIOINNIN TEHTÄVÄNKUVAUSTEN JA RAPORTTIEN LAATU		
EK 3: Mikä on arviointiraporttien ja niistä vastaavien tehtäväkuvauksen laatu?		
* Huomautuksena on todettava, että Suomen ulkoasiainministeriö käyttää etukätesarviointia hankesuunnitelman arviointiin ja analysoimiseen ennen kuin kyseinen asiakirja on hyväksytty hankedokumenttina (project document). Etukätesarvion tekijää ei siis vaadita laatimaan hankesuunnitelmasta uutta versiota, mutta hänen odotetaan esittävän ehdotuksia ja suosituksia, joiden avulla hankesuunnitelmasta voidaan muokata hankedokumentin muodolliset ja sisällölliset standardit ja normit täyttävä lopullinen asiakirja. Arvioinnin on kuitenkin perustuttava näyttöön, ja sen on tarjottava vastauksia (tai ”perusteltu arvio”) asiakkaan etukäteen määrittämiin kohtiin. Havaintojen on erityisesti oltava näyttöön perustuvia, ja niiden on toimitettava päätelmien ja suositusten kulmakivenä. Yleisenä havaintona voidaan todeta, että tässä meta-evaluoinnissa esiin noussut käytäntö ei vastaa Suomen valtionhallinnon ja ulkoasiainministeriön tavoitteita ja vaatimuksia monellakaan tapaa.		
2.7) Meta-evaluoinnissa todettiin, että etukätesarvioinnin tehtäväkuvauksien kon-sulleille antamat ohjeet eivät olleet riittävän täsmällisiä, kun otetaan huomioon, että ne teetetään ulkoisena hankintana. Todettiin myös, että tehtäväkuvaukset olivat heikkoja ilmaisemaan selkeästi määriteltynä arviointialheita. Niissä mainittiin usein suuri määrä tutkittavia asioita yhdistämättä niitä etukäteen määriteltyihin analyysikriteereihin (kuten ulkoasiainministeriön standardit edellyttäisivät). Tämä johtaa epätarkasti kohdistettuihin kysymyksiin ja epävarmuuteen niistä ulkoasiainministeriön prioriteeteista, joita toimeksi-saajien tulisi noudattaa. Kun kaikki arvioinnin tehtäväkuvauksen (ja/tai tarjouspyyntöjen) arvostelutalukon osiot otetaan huomioon, annettujen pisteiden keskiarvo on 64,78/100. Meta-evaluointitiimi antoi vain vähän pisteitä seuraavista kategorioista: a) Arvioinnin tavoitteiden muuntaminen (ehdotetun intervention erityisessä kontekstissa) arvioitaviksi relevanteiksi ja tarkkaan määritellyiksi aiheiksi (puolet 10 tehtäväkuvauksesta sai 20 pistettä tai vähemmän, kun suurin mahdollinen pistemäärä oli 35). b) Avun tuloksellisuus-tään sitoutumisen sisällyttäminen käsiteltäviin aihepiireihin. c) Edes vähimmäistason täyt-tävän ohjauksen ja opastuksen antaminen soveltuvimman laatijan ensi käden tiedot kontekstista ja logiasta ottaen huomioon tehtäväkuvauksen laatijan ensi käden tiedot kontekstista ja saatavilla olevista tietolähteistä, sekä d) Arviointiin suunnattujen resurssien riittävyys sekä toimenkuvauksen ja siihen tarvittavan asiantuntemuksen epäselvä tai olematon kohtaaminen.	2.7) Koska: a) näiden asiakirjojen laatijat ovat lähes aina ulkoasiain-ministeriön sisäisiä ja itse asiakir-joille tehdään laadunvalvontaa sisäisesti ulkoasiainministeriössä ja b) jotkin asiakirjoista on ehkä tarkistettu kehityskumppanien, vastaanottavien organisaatioiden tai täytäntöönpanosta vastaavien tahojen toimesta, keskimääräistä pistemäärää ei voida pitää hyväksyttävänä. Tämä päätelmä esittelee sitä silmällä pitäen, että tuotetuissa tehtäväkuvauksissa oli huomattavia heikkouksia erityi-sesti sellaisissa arvioinnin sei-koissa, joihin asiakirjan laadineiden virkamiesten piti kehittää hankesp-esifistä sisältöä.	2.7) Ulkoasiainministeriön pitäisi suunnitella ja toteuttaa kykyjen/ kapasiteetin kehitysaloitte (sekä siihen soveltuvat aikuispedagogi-set tukimekanismit ja instrumentit), joilla voitaisiin parantaa sen kykyä luoda ohjeita arvioinneista vastaaville toimeksisaajille. Sopivia menetelmiä voisivat olla men-torit, sisäiset/ulkoiset neuvojat ja laadunvalvonnan tuki, esimerkiksi-paukset jne. sekä internet-pohjain-en sisäinen ja ulkoinen suora tuki. (EVA-11 ja toimeenpanevat yksiköt)
Toisaalta tehtäväkuvaukset saivat hyvät pisteet a) ”perusteluissa, tarkoituksessa ja tavoitteissa” sekä b) kuvauksesta ”ulkoasiainministeriön sisäisestä” arviointiprosessista.		

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
2.8) Yksinkertaisen arvioinnin analyysin avulla tiimi totesi, että monissa kenttäsuuden sisältävissä arvioinneissa vietettiin alle 10 päivää kentällä samaan aikaan kun raporteissa puhuttiin suurista ongelmista, puutteista ja heikkouksista.	2.8) On oletettavaa, että resurssieja (aikaa ja budjettia) ei allokoita riittävästi arvioinneille, jotta ne voisivat tuottaa tarvittavaa todisteineistoa ulkoisten konsulttien tekemien päätelmien ja ehdotusten tueksi.	2.8) Ulkoasiainministeriön olisi myös tehtävä havaintoihin perustuva analyysi arviointien asianmukaisista resurssitasoista ja -tyypeistä niin, että ne voisivat tuottaa laadukasta ja riittävää tietoa hankealoitteista hankedokumentaation lopulliseen muotoon saamiseksi.
2.9) Arviointiraportit olivat yleisesti ottaen huonolaatuisia (keskimääräinen arvio 4,6/5), mikä saattaa heijastaa niiden taustalla ja ohjeena olleiden tehtäväkuvauksen laatua. Mikä tärkeintä, ne saivat huonot pisteet näyttöön perustuvien havaintojen esittämisessä ja tyydyttävien vastausten tarjoamisessa ongelmien, jotka oli esitelty tehtäväkuvauksissa; molemmat ovat kaikkien etukäteisarviointien ydinelementtejä. Arviointeja myös tehtiin, ennen kuin projektin suunnittelu oli saatu päätökseen (joskus jopa ilman loogista viitekehystä, tavoitehierarkiaa tai muutosteoraa ja siten, että senhetkessä hankesuunnitelma-dokumentissa oli monia puuttuvia, kriittisiä osia). Tästä syystä arviojakonsultit laativat epätyydellisiä raportteja, jotka evaluointikonsultit mainitsivat myöhemmin evaluointiraporteissa tärkeimmäksi intervention tavoitteiden saavuttamatta jäämisen syynä (ts. interventio oli huonosti suunniteltu).	2.9) Arvioinnit eivät täydet strategista rooliaan suunnittelu- ja ohjelmointisyklissä: ne eivät tarjoa ylemmälle johdolle varmistusta suunnitelman laadusta, eivätkä ne välttämättä tuota parasta laatua olevia asiakirjoja ylemmän johdon hyväksynnälle.	2.9) Ulkoasiainministeriön olisi kattavalla analyysillä selvitettävä, millaista tukea arvioinnit varsinaisesti antavat projektisuunnitteluun ja hankesykliin. Vertailukohtana voi toimia viittaus vuonna 2015 laadittuun konseptipaperiin ennakoarvioinneista.
2.10) Merkittävä määrä raportteja ei sisältänyt tärkeitä, ulkoasiainministeriön edellyttämiä osioita (tai jos sisälsivät, sisältö todettiin pinnalliseksi). Seuraavassa huomattavia havaintoja ja muutamia yhteenvetoja jakautumisesta: <ul style="list-style-type: none"> Vastaukset arviointikysymyksiin tai empiirisesti perustellut havainnot tai strateginen analyysi (3/10 puuttuu). Tässä on todettava, että 5/10 arvioinneista sai vain tuskin välttävän arvosanan 3, ja 2/10 arvioinneista sai vain 1 pisteen. Nämä havainnot ovat huomattavia, koska ne ovat arviointiraporttien ydinaluetta. Päätelmät (3/10 puuttuu). Laajemmat päätelmät opituista asioista (4/10 puuttuu). 	2.10) Käytetty laadunvalvonnan muoto ja organisaatio ei pääse käsiin ulkoistettujen raporttien tärkeisiin heikkouksiin. Monien vaadittavien raporttiosien laatuvaatimukset jäävät täyttymättä, ja monissa raporteissa ei ole lainkaan tiettyjä osioita (tai kriittisten asioiden analyyseja).	2.10) Ulkoasiainministeriön ylempään johdon tulisi määrittää selkeästi raporttien laadulle asetetamansa odotukset ja sitten valvoa laatustandardien noudattamista. EVA-11 voi tarjota vähintään koulutusta ja tukea suuntaviivoista, ja sen pitäisi valvoa laaturikkaita.
2.11) Tiimi totesi, että itse raportit kirjoitettiin usein suppeasta toimiala kohtaisesti sektorinäkökulmasta, eikä niissä käsitelty laajempia poliittisia ja kontekstuaalisia aiheita, jotka ovat todella tärkeitä hankedokumentin osia, kuten ihmisoikeusperustaisen lähestymistavan integrointi hankesuunnitelmaan, politiikkajohdonmukaisuuden ja avun tuloksellisuuden analysointi tai intervention vaikutuksellisuus osana maastrategiaa.	2.11) Strategiset (ulkoasiainministeriön) tarpeet eivät välttämättä täyty arviointiprosessissa. Palautetta ja tärkeitä poliittikan aiheista opittuja asioita (lessons learned) ei tuoteta prosessissa niin kuin pitäisi.	2.11) Katso suositus 2.10.

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
2.12) Tehtäväkuvauksen Arviointiaiheet-osion ja kahden ”aiheisiin” liittyvän raporttien osion – Havainnot- ja Vastaukset arviointikysymyksiin-osiot – pisteitä verrattiin. Nämä ovat arviointiin liittyvien asiakirjojen ydinasioita. Vastaukset-osio sai hyvin huonot pisteet kaikissa raporteissa, sen keskiarvo oli vain 32/100. Vain kerran se sai suuremmat pisteet kuin Arviointiaiheet-osio (joka oli myös huono, pisteiden keskiarvo 58,86/100). Vastaukset ylittivät ”Havainnot”-osion pistemäärän vain yhden raportin kohdalla.	2.12) Sekä arvioinnin tehtävänkuvauksen että arviointiraporttien kehityspoliittikan operationalisointia koskevat ydinasiat on kaiken kaikkiaan käsitelty huonosti.	2.12) Katso suositus 2.10.
SUOMEN KEHITYSYHTEISTYÖN LAATU		
EK 4: Mitä voidaan sanoa Suomen kehitysyhteistyön laadusta luotettavien hajautettujen evaluointiraporttien ja niihin liittyvien suunnitteluasiasiakirjojen perusteella kunkin OECD/DAC-kriteerin mukaan?		
2.13) Suomen kehitysyhteistyön todettiin olevan:		
A. Tarkoituksenmukaista sikäli, että se on sekä Suomen että edunsaajaorganisaatioiden politiikkojen ja strategioiden mukaista, vaikka interventiota ja sitä, miten se liittyy vastaantottajamaan prioriteetteihin, käsitellään hyvin ylmalkaisesti. Meta-evaluoinnissa todettiin, että Suomen kehitysyhteistyö oli tarkoituksenmukaista pääasiassa, koska analyysit tehdään aina kohtuullisen korkealla käsitteellisellä tasolla. Portfolion ”tarkoituksenmukaisuuden” analyysit harvoin tehdään ”paikallisen” tai ”alemmman tason” kontekstin vaikutusten tasolla.	A. Suomen kehitysyhteistyötä voidaan pitää tarkoituksenmukaisena, mutta tarkoituksenmukaisuuden etsinnässä havaitaan olevan paljon toimintavapautta ja tulkinnanvaraa. Tavoitteet ovat liian korkealla käsitteellisellä tasolla ilmaistuja, jotta varsinaisen toimeenpanon tavoitteiden saavuttaminen ja tätä kautta hallinnon vastuuvastuuvelvollisuus voisi toteutua.	A. Mitä tulee projektien ja ohjelmien evaluointien tarkoituksenmukaisuuden arviointiin, ulkoasiainministeriön pitäisi selvittää, minkälaisesta sisällöstä se haluaa saada tietoa ja missä laajuudessa (sekä kehittää ohjeistoja ja tulkintaesimerkkejä). (Ulkoasiainministeriön toiminta- ja politiikkayksiköt)

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>B. Vain kohtuullisen tuloksellista korkeamman tason tavoitteiden saavuttamisessa; osa vähäisestä pistemäärästä liittyy selvästi siihen, että monilla hankkeilla ei ole näiden tavoitteiden seuranta- ja evaluointia helpottavia tiedonkeruujärjestelmiä, jolloin raporteissa ei voitu raportoida niistä. Meta-evaluoinnissa todettiin, että ulkoasiainministeriön kehitysyhteistyöohjelmalla on todennäköisesti vaikeuksia raportoida ylöspäin (vastuuvollisuus) interventtioiden tuloksellisuudesta, etenkin lopputulosten tasolla. Todettiin myös, että interventiot eivät heijasta läpileikkaavien tavoitteiden (CCO) ja ihmisoikeusperustaisen lähestymistavan (HRBA) tukemaa tuloksellisuutta. Viimeksi mainittujen (HRBA/CCO) saama keskimääräinen arvio oli hyvin pieni, ja vain 12 asiakirjaa 18:sta käsittelee aihetta ollenkaan.</p> <p>Odotettujen alemman tason suoritustavoitteiden täyttymisessä onnistumisen taso on melko korkea useimpien interventtioiden useimmilla osa-alueilla. Monitahoisia haasteita tulee vastaan muutettaessa tuotoksia tuloksiksi, eikä evaluointiraporttien perusteella useimpia näistä haasteista selvästikään ollut osattu aavistaa suunnitteluvaiheessa.</p>	<p>B. Vaikka tieto alemman tason vaikutuksista on todennäköisesti saatavilla, ulkoasiainministeriön on luultavasti vaikea laatia havaintoihin perustuvia tiivelvöllisyyseraportteja strategisemmän tason vaikutuksista.</p> <p>Ulkoasiainministeriö saa vain vähän evaluointipohjaisia raportteja, jotka käsittelevät ihmisoikeusperustaista lähestymistapaa tai muita läpileikkaavia tavoitteita.</p> <p>Useimmat suunnitelluista tuotoksista saavutettiin, mutta raporteissa kehitetään vain heikko linkki tuotosten aikaasaamisen ja niiden tuloksiksi muuttamisen välille.</p> <p>Monia tuloksia ei tunnisteta tai määritellä selkeästi (jolloin tulospurustainen johtaminen on vaikeaa, ellei jopa mahdotonta).</p>	<p>B. Samaa suositusta kuin kohdassa 2.13 A sovelletaan, kun "tarkoituksenmukaisuuden" tilalle vaihdetaan "tuloksellisuus".</p> <p>Ulkoasiainministeriön tulisi määrittää paremmin raportointistandardinsa ihmisoikeusperustaisen lähestymistavan ja kaikkien läpileikkaavien tavoitteiden osalta.</p> <p>Näiden ja muiden päätelmien pohjalta ulkoasiainministeriön tulisi selvittää merkittävästi sekä tulosperustaiselta johtamiselta että tuloskeijuilta (mukaan lukien muutosteoriat) odottamaansa laatua (ja niiden käyttöä) ohjelmiansa laadinnassa ja suunnittelussa.</p> <p>(EVA-11 ja kaikki ulkoasiainministeriön ylemmät johtajat)</p>
<p>C. Tehokkuus sai vähän pisteitä (2,6/5) suureksi osaksi siksi, että raporteissa ei raportoitu siitä. Joissakin raporteissa arvioitiin budjettia ja maksatuksia, mutta ei varsinaista tehokkuutta. Ainoastaan mainittiin, että byrokratia ja monimutkaiset hankintamenetelmät hidastivat toteutusta huomattavasti. Yleisellä tasolla todettiin, että ulkoasiainministeriöllä ei ole vankkaa kykyä kaikkien sen ohjelmien tehokkuuden osa-alueiden raportointiin. Meta-evaluointi tulkitse tehokkuus-termiä laajasti tässä kontekstissa ja otti mukaan talouden tai kustannustehokkuuden lisäksi myös valitun strategiaketjun tehokkuuden panosten muuntamisessa vaikutuksiksi. Ulkoasiainministeriö voi todeta suoriutuvansa paljon paremmin vaikutuksiksi muuttamisen tehokkuudessa kuin reaktioajan ja kehitysmuutoksen tehokkuudessa, mikä on edelleen monien interventtioiden ongelma: muutokseen reagoimiseen kuluu liian pitkä aika, kun työskennellään avunantajaprosessien ja monia päätöksentekijöitä sisältävien rakenteiden kanssa.</p>	<p>C. Useimmissa raporteissa tehokkuutta tarkasteltiin vain taloudelliselta kannalta.</p> <p>Joskus raporteissa mainitaan, että ajoitus tai johonkin toimenpiteeseen tarvittu ajan pituus oli ongelma. Yhdessäkään raportissa ei mainittu, miten ongelmat ratkaistiin ja mikä vaikutus niillä oli.</p>	<p>C. Ulkoasiainministeriön tulisi määrittää selkeästi, mitä se odottaa kehitysyhteistyönsä tehokkuuden analysoinnilla. Tässä kontekstissa ulkoasiainministeriön tulisi tehdä selväksi, että tehokkuus ei rajoitu taloudellisiin tai yksinkertaisiin hallinnollisiin muuttujiin.</p> <p>(Ulkoasiainministeriön politiikkayksiköt ja ylempi johto)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>D. Kestävyvyyden saama arvosana oli melko alhainen (2,3/5) osittain siksi, että raporteista ei käynyt ilmi, miksi niiden mielestä interventiot olisivat kestäviä, tai niissä a) puhuttiin ”kestävyvyyden mahdollisuudesta” tai b) kestävyyttä pidettiin oletuksena, vaikka raportista oli käynyt ilmi, että tärkeimmät osa-alueet eivät täyttäneet tavoitteita, tai c) aiheita käsiteltiin vain pinnallisesti.</p> <p>Rahoituksen/taloudellisen kestävyvyyden saamat yleisarvosanat heijastavat raportteihin usein kirjoitettua mainintaa siitä, että (raportin kirjoitushetkellä) ei vielä ollut ryhdytty konkreettisiin toimenpiteisiin sen varmistamiseksi, että se ohjelma tai tavoite, johon interventio liittyy, saisi tarvitsemaansa jatkuvaa taloudellista tukea hallituksilta.</p> <p>Raporteissa annetuista tiedoista käy myös ilmi, että organisaatioilla on merkittäviä puutteita kapasiteetin ja voimavarojen sekä organisaatioiden vakauden suhteen.</p> <p>Raporteissa harvoin käsitellään sitä, miten hallitukset pystyvät täyttämään projektinhallintayksiköiden poistumisen tai projekti-infrastruktuurien purkamisen jälkeen syntyvän tyhjiön.</p>	<p>D. Kestävyyttä käsiteltiin huonosti, jos lainkaan. Silloin kun se mainittiin, todettiin, että vastaanottava hallitus (tai virasto) ei ollut huolehtinut edellytyksistä kestävyvyyden varmistamiseksi.</p>	<p>D. Ulkoasiainministeriön tulisi suorittaa tarkka evaluointi (tai ainakin perusteellinen havaintoihin perustuva tutkimus) sen määrittämiseksi, miten se voi paremmin parantaa interventioidensa kestävyttä. Osana tätä suositusta sen tulisi yhdistää kestävyys selkeästi riskinhallintakäytäntöihinsä.</p> <p>Ulkoasiainministeriön tulisi varmistaa, että se valvoo kumppanensa kestävyvyyteen liittyvien sitoumuksien noudattamista ja varmistaa, että tätä osa-aluetta valvotaan tarkasti heti interventioiden alusta alkaen. Vaatimusten noudattamatta jättämisestä aiheutuvat seuraamukset olisi tehtävä selkeiksi alusta alkaen, ja seuraamuksiin olisi ryhdyttävä väliarvioiden aikaan eikä vasta projektin lopussa, kuten tällä hetkellä.</p> <p>(Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>E: Vaikuttavuuden saama hyvin huono arvio heijastaa sitä seikkaa, että hyvin harvassa raportissa pystyttiin ilmaisemaan, mikä vaikutus olisi. Pieni pistemäärä johtuu suurelta osin vaikuttavuustietojen keräämiseen tarvittujen tietojärjestelmien puuttumisesta kautta linjan sekä selvästi ”ylevistä” vaikuttavuuden ilmaisuista.</p>	<p>E: Ulkoasiainministeriö ei luultavasti pysty tekemään havaintojen pohjalta johtopäätöksiä ohjelmansa vaikuttavuudesta.</p>	<p>E: Ulkoasiainministeriön pitäisi ilmoittaa selkeästi raporttien vaikuttavuuden (tai kaikkien korkean tason vaikutusten) käsittelylle asettamansa odotukset. Ellei se ole valmis valvomaan ja hallitsemaan vaikutuksia, sekin tulisi tehdä selkeäksi. Jos hyväksytään ohjelmadokumentteja, joilla odotetaan olevan vaikutuksia korkeammalle (hallinnon) tasolle, valvontaan ja hallintaan liittyvät valmiudet tulisi sisällyttää interventioon.</p> <p>(Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>
<p>F: Avun tuloksellisuudesta (aid effectiveness) meta-evaluointiimi totesi, että useimmissa raporteissa ei erityisesti mainittu avun tuloksellisuutta erillisenä käsitteenä. Tiimin oli sen vuoksi ”louhittava” raporteja jossakin määrin pystyäkseen löytämään tarkoituksenmukaisia tietoja näiden kuuden alueen analyysiin.</p> <p>Yhteenvetona sanoi näyttää siltä, että ulkoasiainministeriön on raportointi onnistuneen yhteensovittamisessa (alignment) ja kokeneen vaikeuksia mutta myös selviä onnistumisia harmonisaatiossa, mutta yhtä onnistuneena sitä ei pidetty keskinäisen vastuuvollisuuden hallitsemisessa. Mainittakoon, että ”keskinäisyyden” käsite edellyttää vähintään kaksi osapuolta, eikä evaluointiraporteissa koskaan käsitellä sitä, miten vastaanottajamaa yrittää noudattaa omia keskinäisen vastuuvollisuuden sitoumuksiaan. Raporteissa myös hyvin harvoin pohdittiin avun tuloksellisuutta erillisenä käsitteenä.</p> <p>Lopulta meta-evaluointiimi totesi, että evaluointiraporteissa puhuttiin harvoin keskitetyistä tai suunnitelluista pyrkimyksistä ”kehittää kumppanuuksia” tai mistään muutosta siihen liittyvästä käsitteestä. Poikkeuksena tähän havaintoon on tilanne, jolloin tämä tavoite ilmaistaan selkeästi osana intervention loogista viitekehystä tai komponentti-kohtaista toimintasuunnitelmaa. Joissakin tapauksissa mukana voi olla kymmenkunta organisaatiota tai useampi, mutta useimmiten raporteissa puhutaan ”koordinoinnista” vain intervention omien tavoitteiden näkökulmasta. Raporteissa ei evaluoida minkäänlaista kumppanuuden tuomaa lisäarvoa, eikä niissä käsitellä intervention ”twinning”-toiminnasta, erittäin pätevän teknisen (henkilö)avun tai konsultointiyhteyksien kanssa työskentelystä saamia tarkoituksenmukaisuusetuja. Itse asiassa hankkeissa käytettyä teknistä apua on hyvin harvoin evaluoitu ollenkaan evaluointiraporteissa, kun taas muut resurssien ja panosten muodot on kirjattu ylös ja kontekstualisoitu.</p>	<p>F: Useimmissa raporteissa avun tuloksellisuutta ei käsitellä erillisenä konseptina. Keskinäinen vastuuvollisuus ja kumppanuudet ovat kaksi osa-aluetta, joista ei ollut minkäänlaisia mainintoja toimitetuissa raporteissa.</p> <p>Jos ulkoasiainministeriö aikoo käyttää evaluointeja välineenä avun tuloksellisuuteen liittyvän tiedon keräämiseen, se ei ole vielä ottanut tätä aietta mukaan arvioinneille ja evaluoinneille asettamiin vaatimuksiinsa.</p>	<p>F: Ulkoasiainministeriön tulisi määrittää selkeästi odotuksensa, jotka liittyvät avun tuloksellisuuteen ja hallintaan liittyvistä tekijöistä raportointiin. Valvonta- ja seurantajärjestelmiä tulisi valvoa tarkasti sen varmistamiseksi, että nämä tekijät otetaan tarvittaessa mukaan.</p> <p>Ulkoasiainministeriön pitäisi järjestelmällisesti liittää keskinäisen tiivelvöllisuuden hallinta kehitysyhteistyöhönsä. Tämän suosituksen pitäisi koskea myös taustalla olevia Accran ja Busanin käytäntöjä. Kaikille asianmukaisille osapuolille olisi tarjottava ohjeistot, parhaat käytännöt, mallit ja parhaat tapaukset, jotka kuvaavat ja määrittelevät, miten ulkoasiainministeriö haluaa ottaa nämä sopimukset ja julkilausumat mukaan ohjelmiinsa ja projektisykliinsä.</p> <p>(Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>G. Mitä tulee ihmisoikeusperustaiseen lähestymistapaan tai läpileikkaaviin tavoitteisiin, vain kuusi raporttia käsittelee niitä merkityksellisellä tavalla. Monia toisarvoisia viittauksia on tehty, mutta vain vähän analyysia, mikä selittää huonon arvosanan (0,8/5).</p> <p>Raporteissa lähes aina todettiin, että interventio ei ollut vielä (väliraporteissa) tai lainkaan (loppuraporteissa) ottanut käyttöön menetelmää ihmisoikeuksiin perustuvan lähestymistavan suunnitteluun, valvontaan, toteutukseen, hallintaan jne. Tiimi arvioi raportit erityisesti siinä valossa, viitattiinko niissä yhtenäiseen ja suunniteltuun lähestymistapaan. Annettu arvosana antaa ulkoasiainministeriölle selvän käsityksen siitä, että sen ihmisoikeusperustaista lähestymistapaa ei noudateta käytännössä tai sitä ei ainakaan ole raportoitu otetuksi käyttöön.</p> <p>Vain pienessä määrässä raporteja käsiteltiin sukupuoleen liittyviä asioita, ja mainintojen laatu vaihteli. Vain kourallinen interventioita oli ottanut käyttöön sukupuoleen liittyvistä tiedosta kertomiseen tarvittavat valvonta- ja seurantajärjestelmät. Kahta projektia lukuun ottamatta missään muussa evaluointiraportissa ei esitetty raportin pääasiallisessa osiossa sukupuolen mukaan eriteltyjä tietoja (tai mitään vastaavaa muututtajaa), ja vain harvassa raportissa esitettiin päätelmiä ja suosituksia intervention sukupuoleen perustuvista kokemuksista.</p> <p>Yleisesti ottaen on selkeää, että evaluointiraporteissa ei epätasa-arvoa käsitellä erillisenä alueena. Termiä käytetään itse asiassa harvoin. Vain yksi projekti sai korkeat pisteet (5) tällä alueella, vaikka aihe on keskeinen Suomen kehityspoliittikan viitekehksessä. Havaittiin myös, että ilmausta ”epätasa-arvon vähentäminen” ei käytetty interventioiden tavoitteiden tai osa-alueiden kuvauksessa.</p> <p>Monissa raporteissa mainittiin ilmasto, mutta lähes kaikki jäivät pinnallisiksi viittauksiksi. Viittauksia vaikkapa kasvihuonekaasuihin tai hiilidioksidin talteenottoon (vain kaksi ilmastomuutokseen liittyvistä konseptista mainitaksemme) ei ollut yleisesti havaittavissa otoksessa. Muutamassa projektissa raportoitiin, että ilmastomuutoksella oli suurempi vaikutus projektiin kuin projektilla ilmastomuutokseen (esim. kastelujärjestelmät).</p>	<p>G. Voidaan perustellusti väittää ja esittää näyttöä siitä, että ulkoasiainministeriön ihmisoikeusperustaisen lähestymistavan käyttäntöjä ja prioriteetteja ei ole otettu onnistuneesti käyttöön portfolion projekteissa.</p> <p>Sukupuoleen liittyviä asioita ei ole kattavasti evaluoitu meta-evaluointiportfolion raporteissa.</p> <p>Epätasa-arvoon liittyviä asioita ei ole kattavasti evaluoitu meta-evaluointiportfolion raporteissa.</p> <p>Ilmastomuutosta erityisenä ympäristöllisen kestävyysalueena ei käsitellä kattavasti.</p>	<p>G. Ulkoasiainministeriön pitäisi tämentää odotuksiaan ihmisoikeusperustaisesta lähestymistavasta ja läpileikkaavista tavoitteista niiden täytäntöönpanon osalta. Sen johtajien pitäisi myös valvoa tarkemmin interventioita kaikissa suunnitellu- ja täytäntöönpanosyökljen vaiheissa, jotta voidaan varmistaa politiikan toteutuminen ja tiivielvolisuuden mahdollistaminen.</p> <p>Ulkoasiainministeriön pitäisi tarkentaa raportointivaatimuksiaan sukupuolen, ilmaston ja muiden läpileikkaavien tavoitteiden osalta.</p> <p>Ulkoasiainministeriön pitäisi perustaa mekanismeja, joilla voidaan valvoa sen ihmisoikeusperustaisen lähestymistavan ja läpileikkaavien tavoitteiden käyttäntöjen noudattamista.</p> <p>(Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>H. Yhdessäkään raportissa ei käsitelty aiheita, jotka liittyvät riskinhallinnan strategioiden osuuteen suunniteltujen tulosten toteutumisessa.</p> <p>Edellä mainituista tuloksista käy ilmi selkeä kuvio: Ulkoasiainministeriöllä voi olla riskinhallintakäytäntöjä, mutta projektit eivät toteuta niitä käytännössä. Joko niin, tai evaluointiraportit eivät järjestelmällisesti raportoi niistä. Havaittiin myös, että ainoa projekti, joka sai hyvät pisteet riskinhallinnasta (3,0), sai myös hyvät pisteet (3,5) tulosperustaisen johtamisen (RBM) käytöstä. Meta-evaluointiimme mielestä se ei todennäköisesti ole sattumaa.</p>	<p>H. Ulkoasiainministeriön riskinhallintakäytäntöjä ja niistä raportointia ei siis noudateta käytännössä.</p>	<p>H. Ulkoasiainministeriön pitäisi muotoilla paljon selkeämin riskinhallintakäytäntöihin liittyvät suoritusodotuksensa. Yhteys riskinhallinnan ja muiden käytäntöjen, kuten tulosperustaisen johtamisen, ihmisoikeus- ja tulosketjuloogikan välillä pitäisi tehdä selvemmäksi. Ulkoasiainministeriön tulisi tarjota jäsenmellymmät ja tarkemmat suunnat, tukityökalut ja mekanismit riskinhallinnan ja kehitysyhteistyön aloitteiden yhdistämiseen. Tähän tulisi sisällyä valvontaa, ohjausta, evaluointia ja muita hallintotoimia. (Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>
HALLINNON TARKASTUKSET JA EVALUOINNIT		
EK 5: Mitkä ovat syitä hallinnon tarkastuksen tilaamiselle evaluoinnin sijasta (jos mahdollista)?		
Missään analysoituista raporteista ei annettu mitään tutkimuslaatuista vastauksia tähän kysymykseen.	--	--
MERKITTÄVIIMMÄT SEIKAT		
EK 6: Mitkä ovat merkittävimmät hajautetuissa evaluointiraporteissa ilmenevät seikat? Mitä menestystarinoita, hyviä käytäntöjä ja haasteita on havaittavissa?		
Huomioita tämän EK:n löydöksistä		
<p>Kaikkia analysoituja asiakirjoja tarkasteltaessa on huomattavaa, että vaiheen 2 analyysissa käsiteltiin 32:ta erillistä seikkaa, ja otokseen kuului 18 projektia (mahdollisesti yhteensä 576 arvioitavaa seikkaa deduktiivista lähestymistapaa käyttäen). <i>Yhteensä 239 seikkaa (yli 41 %) jäi käsittelemättä raporteissa</i>; joitakin seikkoja ei käsitelty oikeastaan missään projekteissa (kuten riskinhallinta), kun taas joitakin käsiteltiin murto-osassa 18 projektista (esim. numero 6.4, "tulosperustainen hallinta", jossa vain 11 projektia käsittelee aiheetta). Tämä antaa ulkoasiainministeriön johtajille tärkeää tietoa, sillä siitä käy ilmi, missä laajuudessa raportit kokonaisuudessaan tarjoavat heille varmistuksessa käytettäviä tietoja.</p>		<p>Suosituksien yleisistä löydöksistä:</p> <p>Ulkoasiainministeriön laadunvalvonta- ja ohjaustoimia pitäisi tarkistaa, jotta voitaisiin ymmärtää paremmin, miksi olennaisista asioista kirjoitetaan puuttuu tärkeitä osioita. Ohjaustoimintaa pitäisi päivittää ja parantaa, mahdollisesti tarvittaessa ulkoisella tuella.</p> <p>(Ulkoasiainministeriön ylempi johto)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>2.14) Seuraavassa on pieni otos meta-evaluoinnin toisessa vaiheessa yksittäisissä raporteissa havaituista tärkeimmistä kohdista:</p> <p>A) Tarkoituksenmukaisuus:</p> <ul style="list-style-type: none"> Suomalaiset projektit ovat yleensä erittäin hyviä vastaamaan niiden ryhmien tarpeisiin, joiden kanssa varsinaisesti työskennellään, vaikka intervention suunnittelussa tarkoituksenmukaisuutta koskevat lausunnot ovatkin yleensä paljon korkeammalla strategiatasolla. Yleisesti ottaen raporteissa puhutaan yhdenmukaisuudesta Suomen kehityspoliittikan kanssa, mutta vain korkeimmalla abstraktitasolla, jolloin tietoa on vaikea käyttää politiikan kehittämiseen. <p>B) Tuloksellisuus:</p> <ul style="list-style-type: none"> Induktiivinen analyysi viittasi suhteellisen runsaaseen turhautumiseen minkä tahansa intervention kohtaamia haasteita kohtaan, mutta siinä havaittiin useita innovatiivisia toimenpiteitä, joita suunniteltiin ja otettiin käyttöön kontekstiin ja tekniikkaan liittyvien ongelmien ratkaisemiseksi. Monitahoisia haasteita tulee vastaan muutettaessa panostuksia tuloksiksi, eikä useimpia näistä haasteista selvästikään ollut osattu aavistaa suunnitteluvaiheessa. Vaikka Suomen yhteistyön ylipäättään ei osoitettu olevan tuloksellinen korkeamman tason tavoitteiden saavuttamisessa (katso sivu 4 edellä ja tämän raportin asianmukaiset sivut), Suomen interventiot saavuttavat yleensä suurimman osan (alemmman tason) tuloksista, jotka oli määritetty tuloksetjuanalyysissa (eli sellaiset, jotka syntyvät suoraan tuotoksista). Näitä alemman tason vaikutuksia ei jotenkin onnistuta muuntamaan korkeamman tason tuloksiksi (tai ainakaan niiden saavuttamisen tasoa ei seurata). Ilmiön syiden analysointi menee yli tämän meta-evaluoinnin toimeksiannon. Meta-evaluointianalyysissa käy kuitenkin ilmi, että interventiolla on useita vakavia haasteita, kuten ylimitoitettui tavoitteet ja hankintatöimistusten viipyminen yli sen ajankohdan, jolloin toimituksia olisi tarvittu (etenkin monenkeskisten toimijoiden kanssa hallinnon epäselvä määrittely johtui usein huonosta tulosten määrittelystä) (luettelo on osittainen ja vain suuntaa antava). Monissa teknistä avustusta sisältäneissä projekteissa todettiin, että TA ja sen vastapuolet tuottivat lopullisia luonnoksia ehdotetuista laeista, säädöksistä ja muista asiakirjoista, joita sitten ei koskaan viety eteenpäin ja esitely hyväksyttäväksi. Tästä voidaan laatia oletamus, että joko tekninen avustus työskenteli sellaisten tehtävien parissa, joita hyödynsääjäviro ei pitänyt tarkoituksenmukaisina ja TA:ta ei osattu käyttää tehokkaasti hyväksi, tai ehdotettuja ratkaisuja ei pidetty asianmukaisina tai halutun kaltaisina. 	<p>A) Interventoiden päämäärä/t ilmaistaan usein liian korkealla abstraktitasolla, jotta niitä voitaisiin käyttää valvonnan ja täytäntönnäpon ohjauksessa.</p> <p>B) Raporteista kävi usein ilmi, että huomattavasti innovatiivista ja luovutta käytettiin täytäntönnäpon yhteydessä sellaisten ongelmien ratkaisemiseksi, joita ei vielä ollut käsitelty suunnitteluvaiheessa. Järjestelmätason tekijät vaikuttavat silti yhä negatiivisesti interventioiden tuloksellisuuteen. Jotkin näistä ovat puhtaasti hallinnollisia ongelmia, jotka voidaan ratkaista (esim. ylimeritoitus). Monenkeskisten virastojen heikko suoriutuminen on usein esiin tuleva heikon tuloksellisuuden tekijä.</p> <p>On merkkejä siitä, että TA:n roolia ja odotuksia voitaisiin tutkia uudestaan, jotta voidaan varmistaa, että tämän strategian käyttöä tarvitaan ja halutaan (asiakkaan taholta).</p>	<p>2.14)Tämän EK:n luonne ei edellytä varsinaisesti suosituksia. Kuitenkin kaikkiin 2.14-osion kohtiin liittyvät suositukset käsiteltiin edellä aikaisemmassa evaluointikysymyksessä.</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
C) Tehokkuus: <ul style="list-style-type: none"> Interventiot eivät olleet tehokkaita ajankäytön suhteen, ja tärkeimmiksi ongelmiksi mainittiin pitkät viiveet hankinnassa ja päätöksenteossa. Suomen avusta huomioitiin sen kyky tarjota joustavuutta. Kansalliset hallitukset ja useimmat monenkeskiset virastot todettiin erityisesti liian jäykiksi. Interventioissa ei yleensä puututa varsinaiseen tehokkuuteen. Niissä keskitytään siihen, että "tehdään se, mitä suunniteltiin, siten kuin suunniteltiin" sekä budjetin ja maksatus-ten hallinnointi hyväksytyyn suunnitelman puitteissa. 	<p>C) Tehokkuus ei ole täytännönpanosta vastaavien johtajien pääasiallinen huolenaihe. Se tulee selvästi vasta "tuloksellisuuden" jälkeen.</p>	
D) Kestävyy: <ul style="list-style-type: none"> Suomen interventioissa käytetään yleisesti teknisiä ratkaisuja, jotka on sovitettu kohteena olevien avunsaajien tarpeisiin ja kykyihin. Avunsaajat ottavat ne helposti käyttöön ja "omakseen". Taloudellista kestävyyttä harvoin varmistetaan edes projektin lopussa. Vaikka kapasiteetin kehityskomponentti olisi osa interventiota, "organisatorinen kestävyys" hankkeen lopussa on liian vähäinen, jotta ao. taho voisi tuloksellisesti jatkaa hankkeen tavoitteisiin pyrkimistä. 	<p>D) Kestävyys on yhä merkittävä haaste Suomen kehitysyhteistyössä. Osa ongelmaa on riskinhallinta ja "omistus" (ownership).</p>	
E) Vaikutuksellisuus: <ul style="list-style-type: none"> On selvää, että Suomen kehitysyhteistyöllä ei ole käsitystä siitä, missä määrin sen interventiot osaltaan edistävät odotettua vaikutusta. Tarvittuja tietoja ei kerätä järjestelmällisesti, ja lausunnot vaikutuksesta tai jopa korkeamman tason lopputuloksista kirjoitetaan korkean tason käsitteellisillä termeillä, jotka eivät ole suoraan evaluoitavissa. 	<p>E) Ulkoasiainministeriön on luultavasti vaikea raportoida osallisuudesta haluttuun vaikutukseen, jota se pyrkii tuottamaan, mikä haittaa julkisen hallinnon vastuuvollisuuden toteutumista. Suuri osa ongelmaa on se, että vaikutukset ilmaistaan käsitteellisin ja ei-mitattavissa olevin termein, ja interventiot eivät usein perusta tiedonkeruumekanismeja, joilla voitaisiin tunnistaa vaikutuksia (tarkoitettuja tai ei).</p>	
F) Avun tuloksellisuus: <ul style="list-style-type: none"> Meta-evaluointiimi totesi, että useimmissa raporteissa ei erityisesti mainittu avun tuloksellisuutta erillisenä käsitteenä. Yhteensovittamista (alignment) esiintyy erityisen voimakkaasti etenkin korkeammilla tasoilla. Sitä ei koskaan käytetä käsitteenä, jolla kuvattaisiin yhteensovittamista alemman tason strategioiden tai yksityiskohtaisempien kansallisten suunnitelmien kanssa. Harmonisaatiosta harvoin raportoidaan sellaisenaan, vaikka raporteissa luettelaankin lyhyesti muita avunantajia, joihin interventio liittyy. Arvioiduissa raporteissa ei viitata keskinäiseen tiilivelvollisuuteen. 	<p>F) Yhteensovittamista ja harmonisaatiota käsitellään interventioiden ja ohjelman laadinnan puitteissa, mutta keskinäistä tiilivelvollisuutta ei mainita (liittyen Pariisin julistukseen). Raporteissa ei käsitellä Accran tai Busanin sitoumuksia.</p>	

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>G) Ihmisoikeusperustainen lähestymistapa (HRBA) ja läpileikkaavat tavoitteet (CCO)</p> <ul style="list-style-type: none"> Annettu arvio antaa ulkoasiainministeriölle selvän kuvan siitä, että sen ihmisoikeusperustaisista lähestymistapaa ei noudeta käytännössä tai sitä ei ainakaan ole raportoitu otetuksi käyttöön. Hyvin harvassa raportissa edes mainitaan ihmisoikeusperustaista lähestymistapaa. Meta-evaluointi totesi, että vaikka termi ”ihmisoikeusperustainen lähestymistapa” mainittiin lähes aina raporteissa, jotka oli kirjoitettu vuoden 2012 politiikan puitteissa, raporteissa ei koskaan evaluoitu tällaista lähestymistapaa. Sukupuolten tasa-arvoa käsitellään yksioikoisesti (joko sitä edistetään tai sitten ei). Suurella osalla raportteja todettiin, että osassa aktiviteeteista naiset olivat ”kohteina”, kuten koulutuksen osallistujina, mutta raporteissa myös huomioitiin, että naiset eivät olleet mukana päätöksenteossa tai suorita edunsaajia seurauksena tietoisesta päätöksestä hankkeessa. Vain muutamassa interventiossa oli mitään sukupuoleen liittyviä seurantajärjestelmiä. Evaluointiraporteissa ei käsitellä ”epätasa-arvoa” erillisenä alueena. Termiä käytetään itse asiassa harvoin. Monissa raporteissa mainittiin itse asiassa ilmasto, mutta lähes kaikki viittaukset jäivät pinnallisiksi. 	<p>G) Suomen ihmisoikeusperustaiseen lähestymistapaan ja läpileikkaaviin tavoitteisiin liittyviä käytäntöjä (esim. sukupuoli, epätasa-arvo, ilmasto) ei ole raportoitu hyvin ja niitä ei luultavasti ole otettu käyttöön niin hyvin kuin ulkoasiainministeriön politiikka edellyttäisi.</p>	
<p>ALKUSUUNNITTELUA KOSKEVAT, ENNAKKOARVIOINNEISTA SAATAVAT OPIT</p> <p>EK 7: Mitä voidaan oppia arviointiraporteista (ja niiden tehtäväkuvauksista) Suomen kehitysyhteistyön interventioiden alkusuunnittelussa?</p>		
<p>2.15) Yleisesti ottaen hankedokumenttien luonnokset eivät ole valmiita arviointeihin, koska suunnittelun olennaiset osat useimmiten puuttuvat, mukaan lukien kehityksinterventtion logiikka, tulosten viitekehys, yksityiskohtainen täytäntöönpanostrategia, väliaikatulosten ja lopputulosten määrittely sekä analyysi siitä, missä määrin on saatavilla informaatiota (tietokantoja) ja alkutilanneanalyysiä (baseline).</p> <p>Pienessä määrässä arviointeja todettiin, että siihen vaiheeseen mennessä intervention suunnittelua oli tehty vain vähän, ja koska toimeksisaajilla ei ollut valtuuksia muuttaa ohjelmadokumentin luonnosta, niiden suositukset olivat melko laajoja ja kaiken kattavia.</p> <p>Kiinnostavaa kyllä, joissakin evaluointiraporteissa mainittiin, että ”heidän” interventionensa kohtaamien ongelmien laajuus johtui huonosta suunnittelusta.</p>	<p>2.15) Monet sellaiset arviointi- ja evaluointiraporttien löydökset, jotka liittyvät intervention heikkouksiin tai vaikeuksiin noudattaa ulkoasiainministeriön käytäntöjä, standardoja ja normeja, ovat selvästi järjestelmätasoisia. Ongelmat todetaan usein ulkoasiainministeriön projektisyklin alkusuunnitteluvaiheiden aikana. Yksi raportti on saattanut antaa hyvin viisaan neuvon todetessaan: ”projektissa vastaan tulleita ongelmilta olisi vältetty paremalla alkuvaiheen analyysillä” (lainaus meta-evaluointitiimin muokkaama).</p>	<p>2.15) Kuten ennakkoevaluointia käsittelevässä konseptipaperissa jo todettiin, ulkoasiainministeriön pitäisi tarkastaa arviointien rooli ja toiminnot. Muutoksia pitäisi tehdä arviointeihin liittyviin prosesseihin, jotta voitaisiin lieventää prosessin heikkouksien vaikutuksia. Jotkin alustavat huolta aiheuttavat alueet on mainittu tässä raportissa. (Ulkoasiainministeriön politiikka- ja toimintayksiköt)</p>

Havainnot	Havaintojen pohjalta tehdyt johtopäätelmät	Tulosten tai päätelmien pohjalta annettavat suositukset
<p>2.16) Meta-evaluoinnissa todettiin myös seuraavaa:</p> <p>a) Monet ongelmat, jotka liittyvät arviointien toimintaan ulkoasiainministeriön projektisyklissä, käsiteltiin ja hoidettiin viimeaikaisessa EVA- 11:n tilaamassa, ennakkoevaluointia koskevassa konseptipaperissa. Monet tämän paperin päätelmät ovat yhä voimassa, ja ne on erityisesti määritetty ja kuvattu tällä meta-evaluoinnilla.</p> <p>b) Ulkoasiainministeriön kehityspolitiikan jäsentämisessä käytetyn tulosketjun logikka on harvoin valmiina arvioinnin aikaan, tai siihen mennessä tehdyssä työssä oli merkittäviä heikkouksia. Lyhyesti sanottuna siitä seuraa, että alustava intervention suunnitelma luotiin ilman loogista viitekehystä, ulkoasiainministeriön aiheesta antamista ohjeista huolimatta.</p> <p>c) Arvioinneista käy ilmi, että tulosperustaista johtamista ei käytetä projektisuunnittelussa, ulkoasiainministeriön ohjeiden vastaisesti. Evaluuatioreportit vahvistavat tämän havainnon myös.</p> <p>d) Ohjelmadokumenttien luonnokset eivät perustu ihmisoikeusperustaiseen lähestymistapaan, ja niissä käsitellään läpileikkaavia tavoitteita vain pinnallisesti. Tavoitteet ja indikaattorit ovat hyvin harvoin selvästi määritellyt. Tämä ei ole selvästikään ulkoasiainministeriön politiikan mukaista.</p> <p>e) Ohjelmadokumenttien luonnoksissa käsitellään harvoin tarkasti ja kattavasti tehokkuutta, kestävyyttä ja tuloksellisuutta. Sen sijaan niissä keskitytään tarkoituksenmukaisuuteen ja vaikutusellisuuteen. Tämä aiheuttaa lopulta ongelmia arviointiprosessissa ja rajoittaa "evaluointivuuden" politiikkaa. Ulkoasiainministeriön on vaikea raportoida OECD-kriteerien perusteista.</p> <p>f) Arvioinneissa todetaan johdonmukaisesti, että interventioiden hallintajärjestelmät ovat heikkoja, mukaan lukien valvontaan, seurantaan ja tarkkailuun liittyvät käytännöt. Tämä löydös on tärkeä, koska se saattaa osoittaa, että olivat intervention vahvuudet (tai heikkoudet) sitten mitä tahansa, ulkoasiainministeriöllä ei ole varhaisen varoituksen ja muutoksen hallintaan tai raportoinnin ja läpinäkyvyyden hallintaan tarvittavia tietoja.</p> <p>g) Avun tuloksellisuuden aiheetta ei ole käsitelty hyvin arvioinneissa (arvio vain 50 %), mikä osoittaa kenties sen, että arvioijat eivät saaneet ohjeita tehtäväkuvauksiin kohdistuvista odotuksista tai heille ei kerrottu ulkoasiainministeriön vaatimuksista, jotka olivat selkeitä muissa asiakirjoissa. Ulkoasiainministeriön on raportoitava kansallisesti ja kansainvälisesti avun tuloksellisuudesta, mutta sillä ei välttämättä ole aiheen tarkaan käsittelyyn tarvittavia tietoja.</p> <p>h) Yleisesti ottaen arviointiraportteja ei ole jäsennetty OECD/DAC-kriteerien tai ulkoasiainministeriön sektorikohtaisten ohjeiden ja suuntaviivojen mukaisesti. Näin politiikan/ohjeiden evaluoinnista tulee hyvin vaikeaa ilman tietoja.</p> <p>i) Arvioinnit (ja myöhemmin evaluoinnit) harvoin tarjoavat oppeja ulkoasiainministeriölle. Se on tärkeää kontekstissa, jossa ulkoasiainministeriö pitää itseään tietopohjaisena organisaationa.</p> <p>j) Arvioinneissa usein todetaan, että jonkinlainen riskinhallinta on lisättävä ohjelmadokumentteihin. Ulkoasiainministeriön ohjeistuksessa on mukana riskinhallinta.</p>	<p>2.16) Ulkoasiainministeriön tapa hallita arvioiteja projektisyklissään johtaa ennustettaviin ongelmiin ja heikkouksiin pidemmällä ketjussa. Vasemmanpuoleisessa sarakkeessa kerrotut löydökset ovat vain lyhyt kuvaus selkeimmistä ja usein toistuvista tilanteen ilmentymistä.</p> <p>Itse asiassa arvioinnit eivät aina onnistu hoitamaan ulkoasiainministeriön oppaissa ja ohjeissa kuvattuja tehtäviä.</p>	<p>2.16) Katso suositus 2.15.</p> <p>Arviointiin liittyvien asiakirjojen saamien huonojen yleisarvioiden perusteella ulkoasiainministeriön tulisi muuttaa arviointien roolia siten, että ne tehdään paljon myöhemmin projektisyklin varrella. Ohjelmadokumenttien luonnosten pitäisi olla lähes valmiissa vaiheessa ja niiden tulisi täyttää vähimmäisvaatimukset sisällön ja rakenteen osalta, ennen kuin niihin kohdistetaan arviointia, jota voidaan edellyttää ennakoarviolta. On kyseenalaista, onko tehokasta tai tuloksellista ulkoistaa merkittävä osa interventioiden rakenteesta yksilöille tai yrityksille, joille on annettu tehtävän hoitamiseen vain rajoitetut resurssit.</p> <p>(Ulkoasiainministeriön ylempi johto)</p>

ANNEX 12: KEY FINDINGS, CONCLUSIONS AND RECOMMENDATIONS (ENGLISH)

The following table was designed to serve as a basis for the Management Response to this Meta-evaluation.

The table begins with a few findings that are not directly related to evaluation questions per se but instead refer to the overall approach for the meta-evaluation. The rest of the document is structured in a way that matches the Evaluation Questions outlined in the ToR. As described in the report, conclusions are based on a large number of findings (evidence-based) and, since this is a strategic evaluation, the recommendations almost always cover more than one conclusion, as is the case in the recommendations section in the report. This document expands somewhat on those recommendations in keeping with the report preparation guideline of the MFA.

The findings outlined below must not be considered as the “only” important evaluation findings and conclusions, especially for the purpose of preparing management responses. The entire report needs to be considered and the operating units and departments should extract (from the report) the findings that are relevant to them and prepare their responses based on that.

Part 1: Findings not directly related to EQ

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions (the organisation to whom the recommendation is primarily addressed is in brackets)
Portfolio analysis, approach and methodological findings		
1.1) The approach and methodology used in this report is much more complex than that used in previous meta-evaluations within the MFA, and enables EVA-11 to much better understand the overall quality of the portfolio. Assessment tools have been significantly modified to now enable quantitative analysis based on the quality of content. The Quality of Finnish development cooperation is now able to be assessed on both a deductive and inductive basis, and, more importantly, is now based on the extent to which the standards and requirements of MFA policy have been complied with. Longitudinal analyses will not be possible until the next round of meta-evaluations takes place.	1.1) The approach used in this meta-evaluation provides much more evidence-based analysis than the approach used in previous meta-evaluations (which were primarily descriptive).	1.1) This approach should be used, with improvements, for the next rounds of meta-evaluations. (EVA-11 and senior MFA managers)
1.2) Assessment systems based on expected levels of excellence or performance (such as the one in this meta-evaluation) are particularly useful as management assurance tools because they are based on transparent and communicable levels of performance for deliverables (they are based on known norms or standards).	1.2. The approach used in meta-evaluation provides evidence-based insights to support the accountability framework of the senior MFA officers.	1.2) The overall design of this meta-evaluation should be integrated into the accountability framework reporting structure of the MFA (MFA senior management)
1.3) Overall, a good part of the concepts and policies used in Finnish development cooperation are not consistently applied and are not likely understood the same way by all users. Standards are often open to interpretation (ex. words such as "adequately" or "improved" are common and undefined), and norms are open to considerable interpretation (ex. the quality of the indicators or the structure of the Logical Framework). Many Finnish policies and concepts are stated in conceptual terms but not specifically defined (ex. Results-based, sustainable), resulting in quite different interpretations and reporting approaches.	1.3) MFA officials are not all on the same footing when it comes to the basic concepts that are used in their daily work. The lack of consistency affects the performance of the entire MFA mandate, and leads to incomplete feedback, inconsistent planning and design, and poor quality control.	1.3) MFA should examine why this problem exists and, once the causes are clear, take corrective action to develop consistent understanding and use of concepts. To do otherwise will undermine the policy function of the Ministry. (EVA-11 and MFA senior management)
1.4) Many higher-level Finnish policies relating to development cooperation are not treated within evaluation or appraisal reports. Nor do the M and E requirements set up for interventions generally capture the details concerning these policies that would be of use to senior MFA management.	1.4) The evaluation process and the scope of its implementation are not comprehensive enough to provide assurance insights on the entire policy spectrum.	1.4) Revise the required content and scope of evaluations and appraisals so that they cover all relevant policy areas. The units responsible for policy within MFA should identify what type of strategic and operational information they require. (EVA-11 and policy-development mandated units and departments)

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions (the organisation to whom the recommendation is primarily addressed is in brackets)
<p>1.5) There is a wide discrepancy between officials of the MFA in how they direct the conduct of appraisals and evaluations. The quality of the reports and the ToRs varies considerably and poor quality reports are now part of the portfolio, in part due to weaknesses in quality management by officials</p>	<p>1.5) Much of the causes for the poorer performances noted in this meta-evaluation relate directly to the capability and capacity of MFA officials to implement the evaluation processes of the MFA (including appraisals).</p>	<p>1.5) MFA needs to do a capability and a capacity assessment and then develop and fund a strategy and plan for eliminating the capability and capacity gaps. (MFA senior management)</p>
<p>1.6) In terms of the portfolio of initiatives examined in this meta-evaluation, only 12 (slightly over one third) of the evaluated projects took place in the official and traditional bilateral partner countries of Finnish development cooperation (Ethiopia, Kenya, Mozambique, Nepal, Tanzania, Vietnam and Zambia), despite the firm decision taken in the 2004 Development Policy Programme to concentrate on fewer countries, and fewer sectors in those countries.</p>	<p>1.6) There is a discrepancy between the policy decision taken and the reality on the ground.</p>	<p>1.6) MFA should analyse this situation and decide which policy it wants to follow. (MFA senior management)</p>
<p>1.7) The orientation of Finnish development cooperation introduced by the 2007 Development Policy Programme, with an important emphasis on environment, agriculture and business, and trade, only now is strongly visible in the portfolio of projects and evaluation reports. Moreover, when the larger objective of the 2007 Development Policy Programme, "ecologically sustainable development" and the use of natural resources are considered, the concentration of Finnish aid on the "larger" environment sector becomes even more salient. When clustering together environment proper (including climate), agriculture, water and sanitation and forestry, the meta-evaluation portfolio represents 40 percent of projects directly related to natural resources. If representative, this portfolio then seems to confirm the observation expressed in 2009 by the Development Policy Committee, an advisory board for the Government, that "Finland's development cooperation [...] appears to be shifting under the [2007] Development Policy Programme from country-specific to sector- or theme-specific cooperation" (The State of Finland's Development Policy 2009, p. 20).</p>	<p>1.7) The 2007 policy decisions appear to be taking hold.</p>	<p>1.7) No action required</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions (the organisation to whom the recommendation is primarily addressed is in brackets)
<p>1.8) Over one half of the projects are small (51%) with a budget of max. 5 MEUR, and six percent (two projects) benefitted from a budget of over 20 MEUR. The two largest projects budget-wise represent 20 percent (n=35) of the sum of all budget allocations (49 MEUR, or 28% when excluding appraisals, n=26).</p>	<p>1.8) The portfolio consists of a large number of small projects in terms of funding from Finland, and few very large projects.</p>	<p>1.8) Since most donors and the OECD are orienting themselves so that larger scale interventions become the norm, MFA should evaluate whether small-scale interventions really provide the impacts that the Finnish Government seeks in its development cooperation portfolio (MFA senior management)</p>
<p>1.9) The 2012 OECD-DAC peer review observed that aid was not concentrated in key partner countries (p. 47-48). The portfolio analysis we conducted supports this 2012 finding.</p>	<p>1.9) Assuming that the portfolio in our meta-evaluation is representative, we conclude that this practice (i.e. fragmentation) has not changed.</p>	<p>1.9) MFA should re-examine its practice of not focussing its support in key partner countries. (MFA senior management)</p>

Part 2: Findings related to EQ

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
QUALITY OF MFA's DECENTRALIZED EVALUATION PORTFOLIO (EVALUATION REPORTS AND THEIR CORRESPONDING TORs) EQ 1: What is the quality of MFA's decentralized evaluation portfolio (evaluation reports and their corresponding TORs) based on the OECD/DAC evaluation standards in 2014-2015 and the guidance given in the Evaluation Manual and the requirements classified by countries, sectors, budgets, evaluation types, managing units of MFA, commissioner, consultant companies etc.? Is there a difference between the quality of MFA commissioned evaluations and the quality of evaluations that are commissioned by MFA's partners?		
<p>2.1) The overall rating for evaluation TORs was 64 out of 100. Sections that were assessed as quite poor include many CORE sections such as the evaluation questions, the commitments to aid effectiveness, the required methodology and the context.</p> <p>Almost every TOR contained the sections it was supposed to have, save for the implementation of aid effectiveness which was included in only 6 of 22 TOR studied. However, the quality of the various parts (categories) of the TOR was quite poor, especially when it is remembered that these are key documents for the conduct of contracted-out analyses upon which the attainment of MFA objectives depends.</p> <p>The TORs scored positively for "rationale, purpose and objectives", "resources" and describing the evaluation process. They were weak on critical, or "Core" sections such as the statements of the EQ to investigate.</p>	<p>2.1) The meta-evaluation concluded that the TORs were weak in a number of important areas, including the "core" sections where specific direction on the intervention is required, such as the statements of EQ, instructions on aid effectiveness commitments, recommendations concerning methodology and context.</p> <p>The quality of TOR is much lower than it should be, considering its role in the contracting process; the complete control that MFA managers have over the quality of TORs; and the the quality assurance function that must be exercised by MFA officials as part of the accountability framework of the GoF generally and MFA specifically.</p>	<p>2.1) MFA generally and EVA-11 should develop a capability development strategy and fund a detailed development plan to significantly improve the quality of evaluation TORs.</p> <p>MFA managers must be better prepared to quality control the products of their subordinates, and EVA-11 must define how to ensure that managers are able to do this.</p> <p>EVA-11 must develop teaching tools and on-line references and examples on the central role of EQ and the way they are prepared.</p> <p>EVA-11 should use the assessment tools prepared for this meta-evaluation and transform them into quality control checklists and other tools that can be used by MFA officials.</p> <p>(EVA-11 and all operating units and divisions)</p>
<p>2.2) The overall quality of TOR written by the implementing units and departments of MFA is more or less at the level of their international peers who wrote the TOR for MFA-funded projects, as far as we can assume that the meta-evaluation portfolio is representative.</p>	<p>2.2) There is no reason to believe that the MFA quality is adequate just because the scores given to international peers are similar to those awarded to MFA officials.</p> <p>This finding, combined with others, leads to the conclusion that the quality of both (MFA and others) is not adequate.</p>	<p>2.2) In the long run, MFA should position itself to always double check (i.e. do a quality assurance verification) on evaluation documents prepared by other agencies that implement interventions funded in large part by GoF.</p> <p>(all MFA operating units)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>2.3) The overall rating for evaluation reports is almost the same as for the TORs. Most sections do not, on average, meet quality standards, and many issues are not dealt with at all.</p>	<p>2.3) The reports studied in this portfolio are not of a satisfactory quality if accountability, assurance, impact and effectiveness are the guiding performance objectives.</p>	<p>2.3) As noted elsewhere, EVA-11 and MFA senior managers should develop a comprehensive set of quality standards that they expect all MFA officials to meet, whether through their own drafting efforts or through the control of contracted-out work.</p> <p>(EVA-11 and MFA senior management)</p>
<p>2.4) Reports commissioned by the regional and thematic units of MFA score, on the average, 56.95, while the reports commissioned by some other agency score 69.09, almost thirteen points of difference.</p>	<p>2.4) MFA commissioned reports are of significantly lesser quality (by almost a quarter) than those commissioned by other agencies.</p>	<p>2.4) No specific recommendation required because the issues have already been dealt with in other recommendations</p>
<p>2.5) There are significant differences between Finnish commissioned evaluations and those managed by other donors, and between the quality of ToR produced by MFA and the reports approved: reports scoring over 20 points below the ToR score (therefore being clearly of poor quality) were approved by MFA officials instead of being rejected.</p>	<p>2.5) There are serious evaluation management deficiencies in implementing units' evaluations.</p>	<p>2.5) EVA-11 should undertake a study to determine the factors that led to the approval (by supervisors) of sub-standard products, and then draft a strategy to improve overall quality of the supervisory function in this domain.</p> <p>(EVA-11 and implementing units)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>2.6) Contrary to the previous meta-evaluation which did not find a significant correlation between the quality of ToR and the quality of reports, the current meta-evaluation found a statistically significant, although not robust, correlation (0.58) between those two.</p> <p>The average scores given to ToR and evaluation reports varies widely between MFA units (from 38 to 83).</p>	<p>2.6) There is likely to be systemic weaknesses in the execution of the evaluation function since both the ToRs and their corresponding reports show important quality weaknesses. Some MFA units scored much higher than others in terms of rated quality.</p>	<p>2.6) The management group of the MFA should study the reasons why some units generate higher quality products than others, and use the lessons learned and best practices identified to improve overall organisational performance dealing with evaluations and appraisals.</p> <p>With respect to some of the conclusions that arise from the meta-evaluation, the following specific suggestions are offered for consideration:</p> <ul style="list-style-type: none"> a) Significantly tighten methodology requirements for inception reports (the client should approve a detailed methodology that includes the data sources, indicators, tools for data collection and analysis, sampling methods, interview guides and interview notes). b) Insist that evidence be specifically provided to support all findings. c) Better define the expectations of evaluations and appraisals with respect to the three Finnish criteria coherence, Finnish value-added and aid effectiveness. d) Develop a separate guidance document that specifically addresses the acceptable content of reports, and provides norms and standards for them. <p>(EVA-11 and entire management team at MFA)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
EVALUATION COVERAGE WITHIN MFA		
EQ 2: What is MFA's evaluation coverage (comparison of evaluation plans and realized evaluations)?		
The meta-evaluation was not able to find the data required to answer this question. MFA EVA-11 has agreed that the information was not there in a form that would have enabled the team to help develop a "coverage" analysis; as a result, this EQ was dropped.	Not applicable	Not applicable
QUALITY OF APPRAISAL TOR AND REPORTS		
EQ 3: What is the quality of the appraisal reports and their corresponding ToRs?		
<p>* As a contextual note, it must be stressed that GoF MFA uses appraisals to examine and "critique" whatever documentation is proposed for use as a Programme Document (PD) before it is approved; the appraiser is not required to develop another version of the PD but must offer suggestions and recommendations that would allow the existing draft PD to meet the standards and norms that such documents must meet. Nevertheless, the appraisal must be based on evidence and must provide "answers" (or a "learned judgement") to a set of issues that are pre-defined by the client. The findings, specifically, must be evidence-based and must be the cornerstone for conclusions and recommendations. As an overall finding, the praxis examined in this meta-evaluation does not reflect the intent of the GoF MFA in many ways.</p>		
<p>2.7) The meta-evaluation found that appraisal ToRs were not specific in their direction to contractors. It also found that the ToR were weak in stating clearly defined issues. They regularly identified a large number of issues to study without (as requested by MFA standards) clustering them around pre-defined analysis criteria, leading ultimately to a loss of focus and uncertainty concerning the priorities of the MFA that the contractors should address. When all parts of the assessment grid for appraisal ToR/ITT are taken into account, the average number of points given is 64.78 out of a possible 100. Low ratings were given by the Meta-evaluation Team for the following categories: a) translating appraisal objectives (within the specific contexts of the proposed intervention) into relevant and specific issues to be appraised (half of the 10 ToR were scored at 20 points or less out of a possible 35); b) The integration of aid effectiveness commitments into the appraisal; c) Providing some minimum amount of direction and guidance on the most appropriate approach and methodology given the author's privileged knowledge of context and available information sources, and d) Insufficient levels of resources and inappropriate/unclear match of mandate and expertise.</p> <p>On the other hand, ToRs scored highly in a) "rationale, purpose and objectives", and b) the description of the "internal to MFA" appraisal process.</p>	<p>2.7) Given that: a) the authors of these documents are almost always internal to MFA and that the documents themselves are therefore subject to QA by MFA supervisors and b) some of the documents may have been subjected to review by development partners, recipient organizations or implementing agencies, the average score must be considered to be unacceptable. This conclusion is presented with the understanding that the ToRs produced had significant weaknesses in those categories or issues where the officials drafting the document needed to develop intervention-specific content.</p>	<p>2.7) MFA should design and execute a capability/capacity development initiative (along with appropriate pedagogical support mechanisms and instruments) to improve its ability to generate instructions to contractors hired to do appraisals. These could include mentors, internal/external advisors and QA support, examples, etc., and be part of the on-line support available both internally and externally. (EVA-11 and operating units)</p>
<p>2.8) Using a simple financial algorithm, the team found that many international researches has less than 10 days in the field to understand the context and then manage the appraisal, even though their reports spoke of major issues, gaps and weaknesses.</p>	<p>2.8) It is hypothesized that not enough field-based resources (time and budget) are allocated to these documents to generate the evidence required to support the conclusions and suggestions made by the contractors.</p>	<p>2.8) MFA should also undertake an evidence-based study of the appropriate levels and kinds of resources that are needed to execute appraisals so that they generate quality and sufficient insights into interventions before the Program document is approved. (EVA-11 and operating units)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>2.9) Appraisal reports generally were of poor quality (average rating of 46.5) which may (or may not) reflect, as it were, the ToR that generated and guided them. Importantly, they scored poorly on presenting evidence-based findings and on providing satisfactory answers to issues identified in the ToR (both are core elements of any appraisal). Appraisals were also conducted before the project was fully planned (sometimes without even a logic or results chain and where the existing programme document had many missing, but critical parts), so the appraisers generated incomplete reports that were later cited by others (in evaluation reports) as a key cause of failure to obtain results during the intervention (i.e. the intervention was poorly designed).</p> <p>2.10) A significant number of reports did not include important sections required by the MFA (or if they did the content was judged to have been superficial). The following are of note, with a few summary observations concerning the distribution:</p> <ul style="list-style-type: none"> • Answers or strategic analysis of issues based on findings (3 missing out of 10). It should be noted here that five out of 10 appraisals only scored a barely passable "3" and two of the 10 scored only a "1". These observations are noteworthy because these are the "Core" of an appraisal report. • Conclusions (3 missing out of 10). • Lessons learned (4 missing out of 10). <p>2.11) The team found that the reports themselves were often written in a "sector" perspective and were not preoccupied with the broader policy and contextual issues that are real and important parts of PDs, such as the integration of HRBA, the analysis of coherency and aid effectiveness or the focus on intervention impact as a factor of the Country Strategy.</p> <p>2.12) Scores given to the Issues section of the TORs and two sections of the Reports related to 'issues' – the Findings and Answers sections—were compared. These are the "core" sections of the appraisal related documents. The Answers section scored very poorly throughout the reports with an average of only 32%, and only one scoring higher than the Issues section (which is also poor at an average of 58.86%). Answers also only surpassed the "Findings" section once.</p>	<p>2.9) Appraisals are not fulfilling their strategic role in the planning and programming cycle: they are not providing assurance to senior management on the quality of design and they are not necessarily producing top-quality documents for senior management approval.</p> <p>2.10) The quality control function in place does not catch important weaknesses in the reports prepared under contract.</p> <p>Many of the report sections that are supposed to be generated do not meet quality standards, and many reports are missing sections (or analysis of critical issues) completely.</p> <p>2.11) Strategic needs (of MFA) are not necessarily being satisfied through the appraisal process. Feedback and lessons learned on key policy issues is not being generated through the process as it should.</p> <p>2.12) The core sections of both the appraisal TOR and the appraisal reports (i.e. those that deal with the operationalisation of the cooperation) are poorly addressed overall.</p>	<p>2.9) MFA should conduct a comprehensive analysis of the performance of support that appraisals actually provide within the project planning and programme document "cycles." Reference to the concept paper prepared in 2015 on ex-ante evaluations could serve as a baseline.</p> <p>(MFA senior management)</p> <p>2.10) MFA senior management should identify clearly its expectations in terms of the quality of reports and then enforce those quality standards. EVA-11 can provide at least training and guideline support, and should monitor for breaches of quality.</p> <p>(EVA-11 and MFA operating units)</p> <p>2.11) See recommendation 2.10</p> <p>2.12) See recommendation 2.10</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
QUALITY OF FINNISH DEVELOPMENT COOPERATION EQ 4: What can be said about the quality of Finnish development cooperation based on the reliable decentralized evaluation reports, and related planning documents, by each OECD/DAC criteria?		
2.13) Finnish development cooperation was found to be:		
<p>A. Relevant in that it is aligned to both Finnish and beneficiary organisations' policies and strategies, although the depth of treatment given to the intervention and how it relates to the priority of the recipient country is very cursory. The meta-evaluation found that Finnish development cooperation was relevant primarily because the analyses always take place at a relatively high level of conceptualization. The analyses of "relevance" in the portfolio rarely take place at the level of the effects in the "local" or "lower-level" contexts.</p>	<p>A. Finnish development cooperation can be considered relevant, but there is a great deal of latitude and interpretation assumed to exist in the search for relevance. Targets are too conceptual for real accountability to be exercised.</p>	<p>A. MFA should clarify (and develop guidelines and examples for the interpretation of...) the nature and scope of the content MFA wants to be informed of insofar as the assessment of relevance in project and programme evaluations is concerned. (MFA operating and policy units)</p>
<p>B. It was only moderately effective in meeting its higher-level objectives (part of the low score is clearly attributable to the fact that many interventions do not have information systems that monitor these objectives so the reports could not report on them). The meta-evaluation found that MFA's development cooperation programme is likely finding it difficult to report upward (accountability) on the extent to which interventions are effective, particularly at the outcomes level. It also found that the interventions do not reflect effectiveness as being supported by CCOs or HRBA. The average rating for the latter (HRBA/CCO) was very low, with only 12 of the 18 documents dealing with the topic at all.</p> <p>The level of success in meeting expected lower-level performance targets is fairly high in most components of most interventions. It is the transformation of outputs into outcomes that faces multifaceted challenges, most of which were apparently not foreseen in the design stage (according to the evaluation reports themselves).</p>	<p>B. While information on lower-end effects are likely to be available, MFA likely finds it difficult to generate evidence-based accountability reports on more strategic level effects.</p> <p>MFA receives little in the way of evaluation-based reports dealing with the effectiveness of HRBA or other CCO</p> <p>Most planned outputs were achieved, but reports develop only a weak link between the generation of outputs and their transformation into outcomes. Many outcomes are not clearly identified or qualified (making RBM difficult, if not impossible, to implement).</p>	<p>B. The same recommendation as 2.13 A applies when "relevance" is changed to "effectiveness"</p> <p>MFA should better define its reporting standards for HRBA and all CCOs.</p> <p>Using these conclusions and others, MFA should significantly clarify its expectations of quality (and use of) of both RBM and results chains (including Theory of Change) in its programming and planning. (EVA-11 and all MFA senior management)</p>
<p>C. Efficiency was awarded a low score (2.6 out of 5) largely because the reports did not report on it. Some reports measured budget and expenditures but not efficiency per se except to indicate that bureaucracy and complex procurement procedures slowed down execution considerably. As an overall finding, the MFA is not in a solid position to report on the all the elements of the efficiency of its programmes. The term efficiency was broadly interpreted in this context by the meta-evaluation team, and included not only financial or cost efficiency, but the efficiency of the chain of strategies chosen for transforming inputs into impacts. MFA can note that it registers much better with transformational efficiencies than it does with time of reaction and change, which continues to be a problem that many interventions have had: the time required to react to a change is too great when dealing with donor processes and multi-decision-maker structures.</p>	<p>C. Most reports only took a financial perspective when looking at efficiency.</p> <p>Reports sometime indicate that timing, or the length of time required to do something, was a problem. None of the reports identified how those problems were dealt with and the effect of those problems.</p>	<p>C. MFA should clearly define its expectations as to the analysis of efficiency within its development cooperation. In that context, MFA should make it clear that efficiency is not limited to financial or simple administrative variables. (MFA policy units and MFA senior management)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>D. The sustainability rating was rather low (2.3 out of 5) in part because the reports did not show why they believed that interventions would be sustainable, or they a) spoke of “potential sustainability” or b) assumed sustainability even if major components were reported as not going to meet objectives, or c) dealt with the topic in a superficial manner.</p> <p>The overall ratings for financial/economic sustainability reflect the often-written sections in the reports indicating that there are still (at the time of writing of the report) no concrete steps taken to ensure that the programme or objective to which the intervention has contributed will receive the ongoing financial support required from the governments. The reports also provide information showing that the organisations have important gaps in capacity and capability as well as in organisational stability. The issue of how the governments will be able to fill the shoes left when the PIUs or project infrastructures are disbanded is very rarely discussed in the reports.</p>	<p>D. Sustainability was poorly dealt with, if at all. When mentioned, it was reported that the recipient government (or agency) had not put in place the wherewithal to ensure sustainability.</p>	<p>D. MFA should undertake a detailed evaluation (or at least a rigorous evidence-based study) to identify how it can better improve the sustainability of its interventions. As part of this recommendation, it should clearly link sustainability to its policies on risk management.</p> <p>MFA should ensure that it monitors the execution of the sustainability-related commitments of its partners, and ensure that this aspect is monitored closely right from the beginning of the interventions. The consequences of non-execution should be clear right from the onset and the consequences should be invoked around the time of the mid-term reviews and not, as is the case, at project end.</p> <p>(MFA policy units and operating units)</p>
<p>E. The very low rating given to impact reflects the fact that very few reports were able to indicate what the impact would be. An across-the-board absence of information systems to gather required data on impact, coupled with what were very clearly “lofty” expressions of impact, together account for a large part of the low score.</p>	<p>E. MFA is likely unable to conclude, based on evidence, on the achievement of impact in its programmes overall.</p>	<p>E. MFA should be clear about its reporting expectations for impact (or, for that matter, all higher-level effects). If it is not prepared to monitor and manage impacts, it should make that clear. If it approves programme documents with higher-level impact expectations, the wherewithal to monitor that and manage it should be built into the intervention.</p> <p>(MFA policy units and operating units)</p>

F. With respect to **aid effectiveness**, the meta-evaluation Team found that most reports do not specifically address the issue of aid effectiveness as a separate concept. The Team has therefore had to “data mine” the reports to some extent to be able to identify relevant information to use in the analysis of these six areas.

As a cluster, it would appear that MFA is reported as having been successful at alignment, as having some difficulty but evident successes in harmonisation, but has not been seen as being successful in managing its mutual accountability commitments. It is recognised that the concept of “mutual” requires at least two parties; the evaluation reports never discuss how the recipient country tries to execute its own mutual accountability commitments). The reports also very rarely reflect on aid effectiveness as a concept in and of itself.

Finally, the meta-evaluation team found that the evaluation reports rarely spoke of concerted or planned efforts to “develop partnerships” or any other related concept. An exception to this observation is when such an objective is explicitly stated as part of the intervention’s log frame or component-based structure. In some cases, a dozen or more organisations may be involved, but the most the reports do is to speak of “coordination” for the purposes of the intervention’s own objectives. Reports do not evaluate any form of partnership value-added, nor do they discuss the benefits to the intervention (i.e. relevance) of “twinning”, working with highly-qualified TA or consulting firms. In fact, the TA used in the interventions is/are very rarely evaluated at all in evaluation reports, whereas other forms of resources and inputs are noted and contextualised.

Conclusions related to the findings

F. Most reports do not deal with aid effectiveness as a separate concept. Mutual accountability and partnerships are two domains where the reports provided were silent.

If the intent of MFA is to use evaluations as a means of gathering information on aid effectiveness, it has not yet included that intent in the requirements for appraisals and evaluations.

Recommendations related to the findings or conclusions

F. MFA should clearly define its expectations concerning the ongoing management and reporting of factors inherent in aid effectiveness. M and E systems should be closely monitored to include these factors, if required.

MFA should systematically include the management of mutual accountability in its development cooperation. This recommendation should also apply to the underlying practices of Accra and Busan. Guidelines, best cases, models and best cases that illustrate and define how MFA wants these agreements and declarations integrated into their programme and project cycle should be made available to all interested parties.

(MFA policy units and operating units)

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>G. In terms of HRBA or cross-cutting objectives, only six reports deal with HRBA or CCO in a meaningful way. There are many tangential references but little analysis, thus explaining the low rating given (0.8 out of 5).</p> <p>Reports almost always noted that the intervention was not yet (in the case of MTE) or had not (final reports) put in place the means to plan, monitor, execute, manage etc. an approach that was founded on human rights. The Team specifically rated the reports in the light of their reference to a concerted and planned approach; the rating given provides MFA with a clear indication that its HRBA policy is not being implemented or is not being reported upon as such.</p> <p>Only a small number of reports dealt with gender issues, with wide differences in quality. There were only a handful of interventions that had put into place the M and E systems required to provide information on gender. Except for two projects, no other evaluation reports presented, in the main part of the report, disaggregated data based on gender (or any other similar variable for that matter) and only a few reports presented conclusions and recommendations based on the intervention's experience with gender.</p> <p>Overall, it is clear that evaluation reports do not deal specifically with "inequality" as a specific domain. In fact, the term is rarely used. Only one project scored highly (5) in this area in spite of its centrality in Finland's development policy framework. It was also observed that the expression "reduction of inequality" was not used in the description of the interventions' objectives or components.</p> <p>It was noted that many reports mention climate but almost all were superficial references. References to concepts of greenhouse gasses, or carbon sequestration, for example (to name only two climate change related concepts) were not generally present in the sample. A few projects reported that climate change had more of an effect or impact on the project than the project had on climate change (ex. irrigation schemes).</p>	<p>G. There is reason (evidence) to believe that the MFA's policies and priorities on HRBA have not been successfully implemented within the portfolio projects.</p> <p>Gender is not comprehensively evaluated within those reports in the meta-evaluation portfolio.</p> <p>Inequality is not comprehensively evaluated within those reports in the meta-evaluation portfolio.</p> <p>Climate change, as a specific manifestation to environmental sustainability, is not dealt with comprehensively.</p>	<p>G. MFA should be much more specific on its HRBA and CCO expectations with respect to their implementation. Its managers should also more closely monitor interventions at all points in the planning and implementation cycles to ensure that policy is implemented and accountability is enabled.</p> <p>MFA should clearly elaborate its reporting requirements on gender, climate and other CCOs.</p> <p>MFA should set up mechanisms to enforce its HRBA and CCO policies. (MFA policy units and operating units)</p>
<p>H. Not a single report dealt with the issues of the contribution of risk management strategies to the realisation of planned results.</p> <p>The results noted above paint a clear picture: MFA may have policies on risk management but the projects are not implementing them. Either that, or the evaluation reports are systematically not reporting on them. It was also observed that the only project to score well (3.0) on risk management also scored well (3.5) on the use of RBM. That, in the opinion of the meta-evaluation team, is not likely to be a coincidence.</p>	<p>H. The MFA'S policies on risk management and the reporting thereof are not being implemented.</p>	<p>H. MFA should much more clearly elaborate its performance expectations concerning its policies on risk management. The link between risk management and other policies such as RBM, HBRA the use of results chain logic should be made much clearer. MFA should provide much more elaborated and specific guidelines, support tools and mechanisms on how to tie risk and development cooperation initiatives together; this should include monitoring, supervision, evaluation and other management functions.</p> <p>(MFA policy units and operating units)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
MANAGEMENT REVIEWS AND EVALUATIONS		
EQ 5: What are the reasons to commission a management review instead of an evaluation (if possible)?		
None of the reports analysed provided any research-quality insights into this question.	Not applicable	Not applicable
MAJOR ISSUES		
EQ 6: What are the major issues emerging from the decentralized evaluation reports? What are success stories, good practices and challenges?		
Note on findings for this EQ		
When considering the entire set of documents analyzed, it is noteworthy that 32 separate issues were considered in the Phase 2 analysis; and 18 projects were part of the sample (a possible total of 576 issues that needed to be rated using the deductive approach). A total of 239 issues (or over 41%) were not dealt with in the reports; some issues were not dealt with at all by any project (i.e. risk) while others were dealt with by a fraction of the 18 projects (ex. number 6.4, "management for results" where only 11 projects dealt with the issue). This provides some insight for MFA executives as it indicates the extent to which the reports as a whole provide them assurance information.		Recommendation on the overall findings: MFA's quality control and supervision functions should be reviewed to better understand why key documents have important sections missing. The supervisory function should be updated and improved, possibly with external support if required. (MFA senior management)
2.14) The following are but a small sample of the key points identified in the second phase of the Meta-evaluation as identified in the individual reports: A) Relevance: <ul style="list-style-type: none"> Finnish projects tend to be very good in specifically addressing the needs of the groups actually worked with, even though the statements of relevance in the intervention design tend to be at a much higher strategic level. Generally, reports tend to speak of alignment with Finnish policy, but at the highest levels of abstraction only, making the information difficult to use for policy development. 	A) The intent and <i>raison d'être</i> of interventions are often stated at too high a level of abstraction to be able to be used as guideposts for monitoring and implementation.	2.14) The nature of this EQ does not call for recommendations per se. However, recommendations covering all the sections under 2.14 were covered in previous EQ above.

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>B) Effectiveness:</p> <ul style="list-style-type: none"> The inductive analysis indicated a relatively high degree of frustration with the challenges facing any intervention, but noted many innovative measures that were designed and implemented to resolve context and technical problems. It is the transformation of outputs into outcomes that faces multifaceted challenges, most of which were apparently not foreseen in the design stage. While Finnish cooperation overall was not shown to be effective at meeting higher-end objectives (see page 4 above and the relevant sections of this report), Finnish interventions tend to achieve a majority of the (lower-lever) effects that were identified in the results chain analysis (i.e. those that are directly generated by the outputs). These lower-end effects somehow do not get transformed into higher-end effects (or at least they are not followed); the analysis that would shed light on the reasons for this are way beyond the scope of this meta-evaluation. The meta-evaluation analysis does show that interventions have many serious challenges however, including over-scoping, mismatching of procurement deliverables and the time they were needed (especially with multilateral agencies, a lack of management focus that often resulted from poor result definition (an illustrative and partial list only). Many projects that involved Technical Assistance noted that the TA and their counterparts produced final drafts of proposed laws, regulations and other documents that were never brought forward for adoption. The hypothesis that can be drawn here is that either the TA were working on tasks that were not seen as relevant to the "client" and they (i.e. the TA) were not efficiently used, or that the solutions proposed were not seen as appropriate or wanted. 	<p>B) Reports tended to show that a considerable amount of innovation and creativity are present during implementation to resolve issues that were not dealt with during the design phase. But systemic factors still negatively impact on the effectiveness of the interventions. Some of these are purely management issues that can be resolved (ex. over-scoping). The weak performance of multilateral agencies is an often mentioned factor of low levels of effectiveness.</p> <p>There are indications that the role and expectations of TA could be re-examined in order to ensure that the use of that strategy is required and wanted ("pulled" by the client).</p>	
<p>C) Efficiency:</p> <ul style="list-style-type: none"> Interventions were not time efficient, with long delays for procurement and decision-making noted as key problems. Finland aid was noted for its ability to provide flexibility. National governments and most multilateral agencies were specifically identified as being overly rigorous. Interventions do not generally manage efficiency per se. They are much more concerned about "doing what was planned the way it was planned", and managing the budget and the expenditures within an approved disbursement plan. 	<p>C) Efficiency is not a key concern for implementation managers. It clearly takes second place to "effectiveness".</p>	

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
D) Sustainability: <ul style="list-style-type: none"> • Finnish interventions generally use technical solutions that are adapted to the needs and capabilities of the target beneficiaries. Beneficiaries easily adopt and “own” them • Financial sustainability is rarely assured, even at project end. • Even if a capacity development component is part of the intervention, the “organisational sustainability” required to continue towards outcome achievement is very low. 	<p>D) Sustainability is still a major challenge in Finnish Development Cooperation. Part of the issue is risk management and ownership.</p>	
E) Impact: <ul style="list-style-type: none"> • It is clear that Finnish development cooperation does not have a handle on the extent to which its interventions contribute to expected impact. The information required is not gathered systematically and the statements of impact or even of higher-level outcomes are written in high-level conceptual terms and are not readily “evaluable”. 	<p>E) MFA likely finds it difficult to report on (accountability framework) the contribution it has made to the desired impact it was seeking to generate. A large part of the issue is that impacts are stated in conceptual and non-measurable terms, and the interventions do not often set up data-gathering mechanisms to identify effects, intentional or not.</p>	
F) Aid Effectiveness: <ul style="list-style-type: none"> • The meta-evaluation Team found that most reports do not specifically address the issue of aid effectiveness as a separate concept. • Alignment is particularly strong, especially at higher levels. It is never used as a concept to indicate alignment with sub-strategies or detailed national plans • Harmonisation is very rarely reported against as such, although reports briefly list other donors with which the intervention interfaces • There are no references to mutual accountability in the reports assessed. 	<p>F) Alignment and harmonisation are dealt with within interventions and programming, but mutual accountability is not mentioned. (re: Paris declaration). Reports do not deal with Accra or Busan commitments.</p>	
G) Human Rights Based Approach (HRBA) and Cross-cutting Objectives (CCO) <ul style="list-style-type: none"> • The rating given provides MFA with a clear indication that its HRBA policy is not being implemented or is not being reported upon as such. Very few reports even mention HRBA. • The Meta-evaluation Team found that while the term “HRBA” was almost always mentioned in the reports that were written under the 2012 policy umbrella, the reports never evaluated such an “approach”. • Gender equality is treated either as a “do-or-do-not” issue. A large proportion of reports noted that some activities involved women as “targets”, such as including women in training course, but also noted that they were not involved in decision-making or were not the direct beneficiaries as the result of an overt decision. Only a handful of interventions had monitoring systems concerned with gender at all. • Evaluation reports do not deal specifically with “inequality” as a specific domain. In fact, the term is rarely used. • Many reports did in fact mention climate but almost all were superficial references. 	<p>G) Finland’s policies on HRBA and CCO (ex. Gender, Inequality, Climate) are not well reported on and are likely not as well implemented as MFA policy would require.</p>	

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
LESSONS TO BE LEARNED FROM APPRAISALS CONCERNING INITIAL DESIGN EQ 7: What can be learned from appraisal reports (and their ToRs) on the quality of the initial design of Finnish development cooperation interventions?		
<p>2.15) Overall, the draft Programme Documents are not ready for appraisals because key parts of the design are most often missing, including the development intervention logic, the results framework, the detailed implementation strategy, the statement of intermediate results and outcomes and the analysis of the extent to which information database and baselines are available.</p> <p>A small number of appraisals noted how little had been done in terms of intervention design up to that point and, since contractors were not mandated to change the draft PD, their recommendations were rather broad and all-inclusive.</p> <p>Interestingly, some evaluation reports identified the extent to which the problems “their” interventions faced were the result of poor design.</p>	<p>2.15) Many of the findings from appraisal and evaluation reports that deal with intervention weaknesses or difficulties in meeting MFA’s policies, standards and norms are clearly systemic. Problems are often identified during the initial planning stages of the MFA’S project cycle. One report may have provided very wise advice when it noted that:</p> <p>“the problems encountered in the project would not have occurred if a more solid front-end analysis would have taken place” (this quotation has been paraphrased by the Meta-evaluation team).</p>	<p>2.15) As already noted in the concept paper on ex-ante evaluation, MFA should review the role and functions of appraisals. Adjustments should be made to the processes that surround appraisals in order to mitigate against the impact of the weaknesses in the process. Some initial areas of concern to examine are spelled out in this report:</p> <p>(MFA policy units and operating units)</p>

Statement of findings	Conclusions related to the findings	Recommendations related to the findings or conclusions
<p>2.16) The meta-evaluation also found that:</p> <ul style="list-style-type: none"> a) Many of the issues that deal with the function of appraisals within the MFA project cycle were examined and dealt with in a recent concept paper on ex-ante evaluation commissioned by EVA-11. Most of the conclusions of that paper still apply and have been specifically defined and described through this meta-evaluation. b) The results-chain logic on which MFA policy is structured is rarely prepared at the time of the appraisal, or whatever was done up to that time had important weaknesses. In short, that means that a preliminary intervention design was generated without a logical framework in spite of MFA guidance on that topic. c) Appraisals (confirmed by evaluations) indicate that RBM is not applied in project design, contrary to MFA instructions. d) The draft PDs are not based on HRBA and only deal superficially with CCOs. Targets and indicators are very rarely available. This is clearly not in line with MFA policy. e) The draft PDs rarely explicitly and comprehensively deal with efficiency, sustainability or effectiveness, but they do focus on relevance and impact. This, eventually, will cause problems with the approval process and will constrain the policy on "evaluability". MFA will find it hard to report on the basis of the OECD criteria. f) Appraisals consistently identify that the management systems for interventions are weak, including those for monitoring, supervision, and oversight. This finding is important because it may indicate that no matter what the success (or weakness) of an intervention may be, the MFA will not have the data for early-warning and change management, or for reporting and transparency management. g) The topic of aid effectiveness is not well treated in appraisals (given a rating of only 50%), indicating perhaps that the appraisers were either not instructed on expectations in that regard in the ToR or were not made aware of the MFA's requirements made explicit in other documents. MFA must report nationally and internationally on aid effectiveness, but does not necessarily have the information it needs to deal with the issue in detail. h) Overall, the appraisal reports are not structured along the lines of the OECD/DAC criteria or the MFA policy domains. Evaluating policy/guidance then becomes very difficult without information. i) Appraisals (and later evaluations) rarely provide MFA with lessons learned. This is important in the context where MFA sees itself as a knowledge-based organisation j) Appraisals often indicate that some form of risk management needs to be included in programme documents. MFA guidance includes the management of risk. 	<p>2.16) The way MFA manages appraisals in its project cycle leads to predictable problems and weaknesses down-stream. The findings in the left-hand column are only a brief description of the more obvious and often-occurring manifestations of that situation.</p> <p>In fact, appraisals do not always accomplish their role as described in MFA manuals and guidelines.</p>	<p>2.16) See recommendation 2.15</p> <p>Based on the conclusion dealing with the poor overall ratings given to appraisal-related documents, MFA should change the role of appraisals so that they take place considerably later on in the project cycle. Draft PDs should be in a near-complete state and meet minimum content and design standards before being subjected to the critique that can only be rendered through an ex-ante appraisal. It is questionable whether it is efficient or effective to contract out a major part of the design of interventions to individuals or firms who only have limited resources to do so.</p> <p>(MFA senior management)</p>

**META-EVALUATION OF PROJECT AND
PROGRAMME EVALUATIONS IN 2014-2015**



**MINISTRY FOR FOREIGN
AFFAIRS OF FINLAND**