# META-EVALUIERUNG ZUR QUALITÄT VON (PROJEKT-) EVALUIERUNGEN IN DER DEUTSCHEN ENTWICKLUNGS-ZUSAMMENARBEIT

*Onlineanhang zum Evaluierungsbericht*

*2022*

*Dr. Kerstin Guffler*
*Laura Kunert*
*Marian Wittenberg*
*Dr. Nico Herforth*

DEval
DEUTSCHES EVALUIERUNGSINSTITUT DER ENTWICKLUNGS-ZUSAMMENARBEIT

# INHALT


1. Ergänzende Informationen zur Einleitung ... 1
2. Ergänzende Informationen zur theoretischen und empirischen Herleitung ... 2
   - 2.1 Analyseraster ... 2
     - 2.1.1 OECD-DAC- und DeGEval-Standards ... 2
     - 2.1.2 Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ... 17
   - 2.2 Definitionen und Wirkungszusammenhänge der Faktoren ... 21
3. Ergänzende Informationen zum methodischen Vorgehen ... 26
   - 3.1 Stichprobenziehung ... 26
   - 3.2 Datenmanagement ... 28
   - 3.3 Kodierung und Inter- beziehungsweise Intra-Kodierenden-Übereinstimmung ... 29
   - 3.4 Datengrundlage und -analyse ... 37
   - 3.5 Bewertung ... 39
4. Ergänzende Ergebnisse ... 40
   - 4.1 Anwendung der Qualitätsstandards ... 40
     - 4.1.1 Anwendung der OECD-DAC- und der DeGEval-Standards ... 40
     - 4.1.2 Good Practices zu ausgewählten OECD-DAC- und DeGEval-Qualitätskriterien ... 51
     - 4.1.3 Elemente von Qualitätssicherungsprozessen ... 58
     - 4.1.4 Darstellung der Gründe für eine Nichtanwendung der OECD-DAC- und DeGEval-Standards in der Onlinebefragung ... 59
     - 4.1.5 Darstellung der Verschriftlichung von OECD-DAC- und DeGEval-Standards in Organisationsdokumenten ... 61
     - 4.1.6 Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in GIZ- und KfW-Evaluierungen ... 61
     - 4.1.7 Anwendung von Text Mining für ausgewählte Datenerhebungsmethoden ... 63
     - 4.1.8 Nebeneinanderstellung der Erkenntnisse aus dem DEval-Umsetzungsmonitoring der Systemprüfung und der vorliegenden Meta-Evaluierung ... 64
   - 4.2 Erklärung der Anwendung der Qualitätskriterien ... 68
5. Literatur ... 78
6. Kodierleitaden ... 80

# Abbildungen


Abbildung 1 Mehrstufiges Verfahren zur Auswahl der (Projekt-)Evaluierungen ..... 27

Abbildung 2 Good Practice „Darstellung der Wirkungszusammenhänge“ ..... 53

Abbildung 3 Good-Practice-Beispiel „Einbindung der Stakeholder*innen“ ..... 54

Abbildung 4 Good-Practice-Beispiel „Adressat*innen der Empfehlungen“ ..... 56

Abbildung 5 Good-Practice-Beispiel „Darstellung organisationale Unabhängigkeit des Gutachtenden“ ..... 57

Abbildung 6 Verschriftlichung der Qualitätsstandards in Organisationsdokumenten ..... 61

Abbildung 7 Ergebnisse des Strukturgleichungsmodells zur Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ..... 62

Abbildung 8 Ergebnisse der zum Teil über Text Mining ermittelten Datenerhebungsmethoden ..... 64


# Tabellen


Tabelle 1 Übersicht über bereits durchgeführte Meta-Evaluierungen ..... 1

Tabelle 2 Beispiel der Herleitung der Überschneidungen des Qualitätskriteriums „Beschreibung des Evaluierungsgegenstands“ ..... 2

Tabelle 3 Überblick über die Herleitung der 37 Qualitätskriterien aus OECD-DAC- und DeGEval-Standards ..... 3

Tabelle 4 Übersicht über die Zuordnung zwischen OECD-DAC- und DeGEval-Standards ..... 15

Tabelle 5 Nebeneinanderstellung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierungen ..... 17

Tabelle 6 Definition, Wirkungszusammenhänge und Hypothesen der Faktoren ..... 21

Tabelle 7 Informationen zu den Sprachen der Evaluierungsdokumente ..... 29

Tabelle 8 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2a ..... 33

Tabelle 9 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2b ..... 35

Tabelle 10 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2c ..... 36

Tabelle 11 Übersicht über die abhängigen Variablen (Qualitätskriterien/Indices) ..... 38

Tabelle 12 Übersicht über die Lang- und Kurzformen der Namen der Qualitätskriterien und Faktoren ..... 41

Tabelle 13 Übersicht über die Ergebnisse der OECD-DAC- und DeGEval-Qualitätsstandards/-kriterien ..... 47

Tabelle 14 Good Practice „Berücksichtigung Kapazitätsentwicklung“ (GIZ) ..... 55

Tabelle 15 Bestehende Qualitätssicherungsprozesse bei den teilnehmenden Organisationen ..... 58

Tabelle 16 Gründe für eine Nichtanwendung der Qualitätsstandards ..... 59

Tabelle 17 Informationen zum DEval-Umsetzungsmonitoring und der vorliegenden Meta-Evaluierung ..... 65

Tabelle 18 Nebeneinanderstellung der Ergebnisse ausgewählter Qualitätskriterien ..... 66

Tabelle 19 Korrelationsmatrix zwischen Faktoren und Qualitätskriterien (1/3) ..... 68
Tabelle 20 Korrelationsmatrix zwischen Faktoren und Qualitätskriterien (2/3) ..... 69
Tabelle 21 Korrelationen für die Indices im Standardcluster „Berichtslegung und Methoden“ (3/3)..... 70
Tabelle 22 Ergebnisse der Regressionsanalysen (1/3) ..... 71
Tabelle 23 Ergebnisse der Regressionsanalysen (2/3) ..... 73
Tabelle 24 Ergebnisse der Regressionsanalysen (3/3) ..... 75
Tabelle 25 Odds Ratios für die signifikanten Effekte ..... 76

# ABKÜRZUNGEN UND AKRONYME

| | |
|---|---|
| ADA | Austrian Development Agency |
| BGR | Bundesanstalt für Geowissenschaften und Rohstoffe |
| BMZ | Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung |
| CARE | CARE International e. V. |
| DAC | Development Assistance Committee (Entwicklungsausschuss) |
| DeGEval | Gesellschaft für Evaluation e. V. |
| DRK | Deutsches Rotes Kreuz |
| DVV | Deutscher Volkshochschul-Verband International |
| EWDE | Evangelisches Werk für Diakonie und Entwicklung e. V. |
| EZ | Entwicklungszusammenarbeit |
| GIZ | Deutsche Gesellschaft für Internationale Zusammenarbeit |
| hbs | Heinrich-Böll-Stiftung |
| iLCM | interactive Leipzig Corpus Miner |
| InWent | Internationale Weiterbildung und Entwicklung |
| IR | Inception Report |
| KAS | Konrad-Adenauer-Stiftung |
| KfW | KfW Entwicklungsbank |
| OECD | Organisation for Economic Co-operation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) |
| PTB | Physikalisch-Technische Bundesanstalt |
| QK | Qualitätskriterien |
| ToR | Terms of References (Leistungsbeschreibung) |
| UNEG | United Nations Evaluation Group |
| UNFPA | United Nations Population Fund |
| USAID | United States Agency For International Development |
| vgl. | vergleiche |

# 1. ERGÄNZENDE INFORMATIONEN ZUR EINLEITUNG

Tabelle 1 gibt einen Überblick über bisher durchgeführte Meta-Evaluierungen und zeigt, welche Organisationen untersucht wurden sowie welche Qualitätsstandards als Grundlage für die Untersuchung dienten.

**Tabelle 1 Übersicht über bereits durchgeführte Meta-Evaluierungen**

| Nr. | Autor*innen (Jahr) | untersuchte Organisation/en | genutzte Qualitätsstandards |
|---|---|---|---|
| 1 | Caspari (2010) | InWEnt | OECD-DAC-, DeGEval- und UNEG-Standards |
| 2 | Caspari (2011) | GIZ | OECD-DAC- und DeGEval-Standards |
| 3 | Friedrich-Ebert-Stiftung (2015) | Friedrich-Ebert-Stiftung | OECD-DAC-Kriterien, eigene Qualitätsstandards |
| 4 | Freimann et al. (2016) | GIZ | DeGEval-Standards |
| 5 | Freimann et al. (2017) | GIZ | DeGEval-Standards |
| 6 | Krämer et al. (2019) | GIZ | DeGEval-Standards |
| 7 | Hageboeck et al. (2013) | USAID | eigene Qualitätsstandards |
| 8 | HTSPE LTD. (2011) | European Union | eigene Qualitätsstandards |
| 9 | Koy et al. (2016) | MISEREOR | DeGEval-Standards |
| 10 | Mauthofer und Silvestrini (2018) | World Vision Germany | Bond Evidence Principles |
| 11 | Noltze et al. (2018) | GIZ, KfW | OECD-DAC-Kriterien, eigene Qualitätsstandards |
| 12 | Queiroz de Souza (2017) | Welthungerhilfe | DeGEval-Standards |
| 13 | Silvestrini et al. (2018) | Ministry for Foreign Affairs Finland | OECD-DAC-Kriterien, eigene Qualitätsstandards |
| 14 | Silvestrini und Bäthge (2019) | ADA | eigene Qualitätsstandards |
| 15 | UNFPA (2020) | UNFPA | UNEG-Standards |
| 16 | Väth et al. (2022) | Ministry for Foreign Affairs Finland | OECD-DAC-Standards |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: UNEG = United Nations Evaluation Group, USAID = United States Agency for International Development, ADA = Austrian Development Agency, InWent = Internationale Weiterbildung und Entwicklung, UNFPA = United Nations Population Fund*

# 2. ERGÄNZENDE INFORMATIONEN ZUR THEORETISCHEN UND EMPIRISCHEN HERLEITUNG

## 2.1 Analyseraster

### 2.1.1 OECD-DAC- und DeGEval-Standards

Tabelle 2 stellt dar, wie die Überschneidung zwischen Teilaspekten der OECD-DAC- und der DeGEval-Standards identifiziert wurde. Dabei zeigen die gelb markierten Textstellen die sich überschneidenden Teilaspekte zwischen den beiden Qualitätsstandards auf. Die blau markierten Textstellen sind Teilaspekte, die im Rahmen der Herleitung der Überschneidung für weitere Qualitätsstandards herangezogen wurden. Teilaspekte, die nicht farbig unterlegt sind, wurden nicht untersucht.

**Tabelle 2 Beispiel der Herleitung der Überschneidungen des Qualitätskriteriums „Beschreibung des Evaluierungsgegenstands“**

| OECD-DAC-Standards 2.3 – Gegenstand und Reichweite der Evaluierung | DeGEval-Standard G1 – Beschreibung des Evaluationsgegenstandes |
|---|---|
| **Die zu evaluierende Entwicklungsmaßnahme (der Gegenstand der Evaluierung) wird klar definiert,** was auch **eine Beschreibung der Wirkungslogik bzw. -theorie beinhaltet**, die ihr zu Grunde liegt. | Sowohl **das Konzept des Evaluationsgegenstands** als auch seine **Umsetzung sollen genau und umfassend beschrieben und dokumentiert werden**. |

*Quelle: DEval, eigene Darstellung basierend auf OECD DAC (2010) und DeGEval (2016)*
*Anmerkung: gelb = Beschreibung des Evaluierungsgegenstands; dunkelblau = Darstellung der Wirkungszusammenhänge*

Tabelle 3 stellt die Herleitung aller Qualitätskriterien der OECD-DAC- und der DeGEval-Standards vor (inkl. der Überschneidungen).

**Tabelle 3 Überblick über die Herleitung der 37 Qualitätskriterien aus OECD-DAC- und DeGEval-Standards**

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 1 | Beschreibung des Evaluierungs-gegenstands | OECD DAC 2.3: Die zu evaluierende Entwicklungsmaßnahme (der Gegenstand der Evaluierung) wird klar definiert[.]<br><br>DeGEval G1: [...] Konzept des Evaluationsgegenstands [...] genau und umfassend beschrieben und dokumentiert [...] | Das Qualitätskriterium ist erfüllt, wenn 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteur*innen (politische Partner*innen und/oder Träger) der EZ-Maßnahme genannt sind. | vgl. Noltze et al. (2018); Qualitäts-kriterium 01 |
| 2 | (Einbindung des Kontexts) Beschrei-bung des Kontexts der Entwicklungs-maßnahme | OECD DAC 3.7: Der Evaluierungsbericht beschreibt den Kontext der Entwick-lungsmaßnahme [...] In der Evaluierung wird der Einfluss der Kontextfaktoren auf die Ergebnisse der Entwicklungsmaßnahme [...] bewertet.<br><br>DeGEval G2: Der Kontext des Evaluationsgegenstandes soll ausreichend umfas-send und detailliert analysiert sowie bei der Interpretation von Ergebnissen be-rücksichtigt werden. | Das Qualitätskriterium ist erfüllt, wenn der Kontext der Entwicklungsmaßnahme beschrieben wird (zum Beispiel politische, ökonomische, soziale, [ag-rar-]ökologische, [agrar-]klimatische, gesetzliche, technologische, historische, kulturelle Kontextfak-toren). | eigenes Quali-tätskriterium |
| 3 | (Einbindung des Kontexts) Berücksichtigung des Kontexts bei Ergebnissen | OECD DAC 3.7: Der Evaluierungsbericht beschreibt den Kontext der Entwick-lungsmaßnahme [...] In der Evaluierung wird der Einfluss der Kontextfaktoren auf die Ergebnisse der Entwicklungsmaßnahme [...] bewertet.<br><br>DeGEval G2: Der Kontext des Evaluationsgegenstandes soll ausreichend umfas-send und detailliert analysiert sowie bei der Interpretation von Ergebnissen be-rücksichtigt werden. | Das Qualitätskriterium ist erfüllt, wenn der Einfluss des Kontexts auf die Evaluierungsergebnisse be-rücksichtigt wird. | eigenes Quali-tätskriterium |
| 4 | Darstellung der Wirkungs-zusammenhänge | OECD DAC 2.3: Die zu evaluierende Entwicklungsmaßnahme (der Gegenstand der Evaluierung) wird klar definiert, was auch eine Beschreibung der Wirkungslo-gik bzw. -theorie beinhaltet, die ihr zu Grunde liegt. [...]<br><br>OECD DAC 3.8: Im Evaluierungsbericht wird die Logik bzw. Theorie, die der Ge-staltung der Entwicklungsmaßnahme zu Grunde liegt, beschrieben[.] | Das Qualitätskriterium ist erfüllt, wenn bei der Be-schreibung der intendierten Wirkungszusammen-hänge der Entwicklungsmaßnahme zwischen den Wirkungsstufen Input-Output-Outcome-Impact un-terschieden wird. | vgl. Noltze et al. (2018); Qualitäts-kriterium 03 |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | DeGEval G1: [...] Umsetzung [...] des Evaluationsgegenstands [...] genau und umfassend beschrieben und dokumentiert [...] | | |
| 5 | Beschreibung des Erkenntnisinteresses | OECD DAC 2.1: [...] Zweck [...] der Evaluierung [wird] klar dargelegt[.]<br><br>OECD DAC 2.2: Die spezifischen Ziele der Evaluierung geben Aufschluss darüber, was mit der Evaluierung erreicht werden soll[.]<br><br>OECD DAC 2.7: Die Evaluierungsziele werden durch zweckdienliche und spezifische Evaluierungsfragen konkretisiert. [...]<br><br>OECD DAC 3.12: [...] Die ursprünglichen Evaluierungsfragen [...] werden im Bericht dokumentiert[.]<br><br>DeGEval G3: Zwecke, Fragestellungen [...] der Evaluation [...] sollen so genau dokumentiert und beschrieben werden, dass sie nachvollzogen und beurteilt werden können[.] | Das Qualitätskriterium ist erfüllt, wenn 1. Zweck, 2. Ziele und 3. Evaluierungsfrage(n) der Evaluierung genannt werden. | vgl. Noltze et al. (2018); Qualitätskriterium 02 |
| 6 | Nachvollziehbarkeit der Informationsquellen | OECD DAC 3.9: Im Evaluierungsbericht werden die verwendeten Informationsquellen (Dokumente, befragte Personen, öffentliche Statistiken, Fachliteratur usw.) hinreichend genau beschrieben[.]<br><br>DeGEval G4: Die im Rahmen einer Evaluation genutzten Informationsquellen sollen hinreichend genau dokumentiert werden[.]<br><br>DeGEval N6: Evaluationsberichte sollen [...] wesentliche Informationen zur Verfügung stellen und [...] verständlich und nachvollziehbar sein. | Das Qualitätskriterium ist erfüllt, wenn die in der Evaluierung genutzten Informationsquellen ausreichend detailliert im Bericht dargestellt werden. | vgl. Noltze et al. (2018) und Lücking et al. (2015) |
| 7 | Darstellung der Angemessenheit des methodischen Vorgehens | OECD DAC 2.9: [...] Die Ausführungen zu den Methoden beinhalten die [...] zur Erfassung [...] der Daten eingesetzten Techniken. Mit den ausgewählten Methoden werden die Evaluierungsfragen auf der Grundlage glaubwürdiger Belege beantwortet. [...] | Das Qualitätskriterium ist erfüllt, wenn a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden. | vgl. Noltze et al. (2018); Qualitätskriterium 06 |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | OECD DAC 3.10: [...] Die zur Erhebung und Analyse der Daten eingesetzten Techniken werden genau aufgeführt. Ihre Auswahl wird begründet[,] etwaige Einschränkungen und Unzulänglichkeiten werden erklärt.<br><br>OECD DAC 3.13: Im Evaluierungsbericht werden etwaige Behinderungen und Einschränkungen [...] bei den angewandten Methoden [...] dargestellt[.]<br><br>DeGEval G3: [...] Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen so genau dokumentiert und beschrieben werden, dass sie nachvollzogen und beurteilt werden können.<br><br>DeGEval G5: Erhebungsverfahren und Datenquellen sollen so gewählt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. [...]<br><br>DeGEval G7: Qualitative und quantitative Informationen einer Evaluation sollen nach fachlichen Maßstäben angemessen und systematisch analysiert werden, damit die Fragestellungen der Evaluation beantwortet werden können.<br><br>DeGEval N4: Auswahl und Umfang der erfassten Informationen sollen die adäquate Beantwortung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen[.] | | |
| 8 | Qualitätssicherung mit Inception Report | OECD DAC 1.7: Während des gesamten Evaluierungsprozesses wird eine Qualitätskontrolle durchgeführt. [...]<br><br>DeGEval G6: Die in einer Evaluation gesammelten, aufbereiteten, analysierten [...] Informationen sollen systematisch auf Fehler geprüft werden. | Das Qualitätskriterium ist erfüllt, wenn ein Inception Report erstellt wird. | Eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 9 | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen | OECD DAC 3.11: [...] Die Feststellungen folgen logisch aus der Analyse der Daten, so dass eine klare Beweislinie zu erkennen ist, auf die sich die Schlussfolgerungen stützen. Die Schlussfolgerungen werden durch Feststellungen und Analysen gestützt. [...]<br><br>DeGEval G8: [...] Schlussfolgerungen sollen ausdrücklich auf Grundlage der erhobenen und analysierten Daten begründet werden[.] | Das Qualitätskriterium ist erfüllt, wenn die Datenanalyse, die Ergebnisse und die Schlussfolgerungen logisch miteinander in Bezug gesetzt werden beziehungsweise aufeinander aufbauen. | vgl. Noltze et al. (2018); Qualitätskriterium 14 |
| 10 | Informationsgehalt der Leistungsbeschreibung | OECD DAC 2.12 : Die Planungs- und Konzeptionsphase endet mit der Ausarbeitung einer […] schriftlichen Unterlage, zumeist Leistungsbeschreibung („Terms of Reference[s]“ – TOR) genannt[.]<br><br>DeGEval F1: Die Rechte und Pflichten der an einer Evaluation beteiligten Parteien (was, wie, von wem, wann getan werden soll und darf) sollen schriftlich festgehalten werden. | Das Qualitätskriterium ist erfüllt, wenn die Leistungsbeschreibung der Evaluierung 1. Zweck, 2. Nutzende, 3. Ziele, 4. Methoden, 5. Mittel, 6. den zeitlichen Rahmen, 7. Veröffentlichungsrechte und 8. mitwirkende Personen der Evaluierung beinhalten. | eigenes Qualitätskriterium |
| 11 | (Zugänglichkeit) Veröffentlichung des Evaluierungsberichts | OECD DAC 4.1: [...] Es wird eine systematische Verbreitung [...] des Berichts gewährleistet, um allen Entwicklungspartnern einen problemlosen Zugang zu ermöglichen [...]<br><br>OECD DAC 4.3: Die Evaluierungsergebnisse werden in einem leicht zugänglichen Format präsentiert [...]<br><br>DeGEval G9: [...] Evaluationen [sollen] in geeigneter Form [...] soweit wie möglich zugänglich gemacht werden. | (Onlinebefragung) Werden die Evaluierungen Ihrer Organisation auf einer der Öffentlichkeit zugänglichen Webseite/Datenbank veröffentlicht? | eigenes Qualitätskriterium |
| 12 | (Zugänglichkeit) Veröffentlichung der Zusammenfassung | OECD DAC 4.1: [...] Es wird eine systematische Verbreitung [...] des Berichts gewährleistet, um allen Entwicklungspartnern einen problemlosen Zugang zu ermöglichen [...]<br><br>OECD DAC 4.3: Die Evaluierungsergebnisse werden in einem leicht zugänglichen Format präsentiert [...] | (Onlinebefragung) Werden die Evaluierungen Ihrer Organisation auf einer der Öffentlichkeit zugänglichen Webseite/Datenbank veröffentlicht? | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | DeGEval G9: [...] Evaluationen [sollen] in geeigneter Form [...] soweit wie möglich zugänglich gemacht werden. | | |
| 13 | (Zugänglichkeit) Zugänglichkeit für Stakeholder*innen | OECD DAC 4.1: [...] Es wird eine systematische Verbreitung [...] des Berichts gewährleistet, um allen Entwicklungspartnern einen problemlosen Zugang zu ermöglichen[.]]<br><br>OECD DAC 4.3: Die Evaluierungsergebnisse werden in einem leicht zugänglichen Format präsentiert [...]<br><br>DeGEval G9: [...] Evaluationen [sollen] in geeigneter Form [...] soweit wie möglich zugänglich gemacht werden.<br><br>DeGEval F5: Evaluationsergebnisse- und berichte sollen [...] soweit wie möglich zugänglich gemacht werden. | Das Qualitätskriterium ist erfüllt, wenn der finale Evaluierungsbericht oder eine schriftliche Zusammenfassung des finalen Berichts (auch in Form einer Ergebnispräsentation) den Stakeholder*innen der Evaluierung zugänglich gemacht werden soll. | eigenes Qualitätskriterium |
| 14 | Kompetenz der Gutachtenden | OECD DAC 3.1: [...] Die Mitglieder des Evaluierungsteams kombinieren Qualifikationen im Bereich der Evaluierung mit fachlichen Kenntnissen zur jeweiligen Thematik. [...]<br><br>DeGEval N3: Wer Evaluationen durchführt, soll fachlich und methodisch kompetent sein, damit für die Evaluation und ihre Ergebnisse ein Höchstmaß an Glaubwürdigkeit und Akzeptanz erreicht wird. | (Onlinebefragung) Sind in Ihren Evaluierungen alle drei aufgeführten Kompetenzen bei den Gutachter*innen vertreten? (1. Evaluierungsexpertise, 2. Fach-/Sektor-expertise und 3. Regional-/Länderexpertise.) | eigenes Qualitätskriterium |
| 15 | Rechtzeitigkeit der Erkenntnisse | OECD DAC 4.1: [...] Die Evaluierung wird rechtzeitig abgeliefert, um eine optimale Nutzung ihrer Ergebnisse zu gewährleisten. [...]<br><br>DeGEval N7: Evaluationsvorhaben sollen so rechtzeitig [...] abgeschlossen werden, dass ihre Ergebnisse in anstehende Entscheidungs-, Verbesserungs- oder sonstige Nutzungsprozesse einfließen können. | (Onlinebefragung) Werden die Evaluierungen Ihrer Organisation rechtzeitig zum vereinbarten Zeitpunkt (oder früher) abgeschlossen? | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 16 | (Nützlichkeit der Empfehlungen) Adressat*innen der Empfehlungen | OECD DAC 4.1: [...] Schlussfolgerungen, Empfehlungen und übergeordnete Erkenntnisse sind klar, zweckdienlich, zielgerichtet und handlungsorientiert, damit die Evaluierung für die beabsichtigten Lern- und Rechenschaftszwecke eingesetzt werden kann.<br><br>DeGEval N8: [...] Berichterstattung einer Evaluation sollen die Beteiligten und Betroffenen dazu ermuntern, die Ergebnisse [...] zu nutzen. | Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen der Evaluierung (alternativ: ihre Schlussfolgerungen) adressat*innenenorientiert sind. | eigenes Qualitätskriterium |
| 17 | (Nützlichkeit der Empfehlungen) Handlungsorientierung der Empfehlungen | OECD DAC 4.1: [...] Schlussfolgerungen, Empfehlungen und übergeordnete Erkenntnisse sind klar, zweckdienlich, zielgerichtet und handlungsorientiert, damit die Evaluierung für die beabsichtigten Lern- und Rechenschaftszwecke eingesetzt werden kann.<br><br>DeGEval N8: [...] Berichterstattung einer Evaluation sollen die Beteiligten und Betroffenen dazu ermuntern, die Ergebnisse [...] zu nutzen. | Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen (alternativ: Schlussfolgerungen) der Evaluierung handlungsorientiert sind. | eigenes Qualitätskriterium |
| 18 | Evaluierungseffizienz | OECD DAC 3.4: [...] Die Evaluierung wird effizient [...] vorgenommen.<br><br>DeGEval D1: Evaluationsverfahren [...] sollen so gewählt werden, dass [...] der Aufwand [...] in einem adäquaten Verhältnis zum intendierten Nutzen der Evaluation gehalten wird[.]<br><br>DeGEval D3: Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen. | (Onlinebefragung) Wird in Ihrer Organisation sichergestellt, dass die Effizienz (Aufwand-Nutzen-Verhältnis) einer Evaluierung reflektiert wird? | eigenes Qualitätskriterium |
| 19 | Transparenz von Meinungsverschiedenheiten | OECD DAC 3.14: Die Mitglieder des Evaluierungsteams haben die Möglichkeit, sich von bestimmten Urteilen oder Empfehlungen zu distanzieren, mit denen sie nicht einverstanden sind. Auf ungelöste Meinungsverschiedenheiten zwischen den Teammitgliedern wird [...] hingewiesen.<br><br>OECD DAC 3.15: Relevante betroffene Akteure haben die Möglichkeit, Kommentare zum Entwurf des Berichts abzugeben. Die endgültige Fassung des Berichts trägt diesen Kommentaren Rechnung und weist auf etwa bestehende größere | (Onlinebefragung) Werden Gutachter*innen darauf hingewiesen, Meinungsverschiedenheiten, wenn es welche gibt, im Evaluierungsbericht transparent zu machen? | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | Meinungsverschiedenheiten hin.<br><br>DeGEval N5: Werthaltungen der Beteiligten und Betroffenen, die sich in deren Perspektiven und Annahmen manifestieren und einen Einfluss haben auf die Evaluation, sollten transparent dokumentiert werden, um Evaluationsergebnisse besser einordnen zu können. | | |
| 20 | (Einbindung der Stakeholder*innen) Einbindung der internen und externen Stakeholder*innen | OECD DAC 1.4: Um die Eigenverantwortung für Entwicklung zu erhöhen und die gegenseitige Rechenschaftspflicht für Ergebnisse zu fördern, wird in einem frühen Stadium des Prozesses systematisch die Möglichkeit der Verfolgung eines partnerschaftlichen Ansatzes untersucht. Das Konzept der Partnerschaft beinhaltet einen inklusiven Prozess, an dem verschiedene Akteure, wie Regierung, Parlament, Zivilgesellschaft, Zielgruppen und internationale Partner, beteiligt sind.<br><br>OECD DAC 2.5: Die betroffenen Akteure werden frühzeitig in den Evaluierungsprozess einbezogen und erhalten Gelegenheit, an der Konzeption der Evaluierung mitzuwirken, u.a. durch Identifizierung zu behandelnder Themen und zu beantwortender Evaluierungsfragen.<br><br>OECD DAC 3.3: Das gesamte Spektrum der betroffenen Akteure, sowohl auf Seiten der Geber als auch der Partnerländer, wird während des Evaluierungsprozesses konsultiert und erhält Gelegenheit, Beiträge einzubringen.<br><br>OECD DAC 4.1: Die Konzeption, die Durchführung ebenso wie der Bericht zur Evaluierung tragen den Anforderungen der vorgesehenen Nutzer Rechnung.<br><br>DeGEval N1: Die am Evaluationsgegenstand oder [...] Evaluation Beteiligten [...] sollen vorab identifiziert werden, damit deren Interessen und Informationsbedürfnisse geklärt und [...] bei der Anlage der Evaluation berücksichtigt werden können.<br><br>DeGEval N2: Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen | Das Qualitätskriterium ist erfüllt, wenn Stakeholder*innen der Evaluierung in den Evaluierungsprozess (1. Konzeption/Planung, 2. Durchführung, 3. Berichtslegung) einbezogen werden. | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | [...] können.<br><br>DeGEval D2: Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen [...] der Evaluation erreicht werden kann. | | |
| 21 | (Einbindung der Stakeholder*innen) Einbindung der Partner | OECD DAC 1.4: Um die Eigenverantwortung für Entwicklung zu erhöhen und die gegenseitige Rechenschaftspflicht für Ergebnisse zu fördern, wird in einem frühen Stadium des Prozesses systematisch die Möglichkeit der Verfolgung eines partnerschaftlichen Ansatzes untersucht. Das Konzept der Partnerschaft beinhaltet einen inklusiven Prozess, an dem verschiedene Akteure, wie Regierung, Parlament, Zivilgesellschaft, Zielgruppen und internationale Partner, beteiligt sind.<br><br>OECD DAC 2.5: Die betroffenen Akteure werden frühzeitig in den Evaluierungsprozess einbezogen und erhalten Gelegenheit, an der Konzeption der Evaluierung mitzuwirken, u.a. durch Identifizierung zu behandelnder Themen und zu beantwortender Evaluierungsfragen.<br><br>OECD DAC 3.3: Das gesamte Spektrum der betroffenen Akteure, sowohl auf Seiten der Geber als auch der Partnerländer, wird während des Evaluierungsprozesses konsultiert und erhält Gelegenheit, Beiträge einzubringen.<br><br>OECD DAC 4.1: Die Konzeption, die Durchführung ebenso wie der Bericht zur Evaluierung tragen den Anforderungen der vorgesehenen Nutzer Rechnung.<br><br>DeGEval N1: Die am Evaluationsgegenstand oder [...] Evaluation Beteiligten [...] sollen vorab identifiziert werden, damit deren Interessen und Informationsbedürfnisse geklärt und [...] bei der Anlage der Evaluation berücksichtigt werden können.<br><br>DeGEval N2: Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen [...] können. | (Onlinebefragung) Binden Sie externe Stakeholder*innen in die Konzeption, Durchführung und Berichtslegung der Evaluierungen Ihrer Organisation ein? | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| | | DeGEval D2: Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen [...] der Evaluation erreicht werden kann. | | |
| 22 | Evaluierungsethik | OECD DAC 1.3: Die Evaluierung folgt den einschlägigen [...] ethischen Leitlinien[.]<br><br>OECD DAC 3.3: [...] Die Rechte und das Wohl der Evaluierungsteilnehmer werden geschützt. Auf Wunsch oder im Bedarfsfall werden die Anonymität und das Recht auf Vertraulichkeit der einzelnen Informationsgeber gesichert.<br><br>DeGEval F2: Evaluationen sollen so [...] durchgeführt werden, dass Rechte, Sicherheit [...] der in eine Evaluation einbezogenen Personen geschützt sind. | (Onlinebefragung) Werden Vorgaben zu Sicherheit und Rechten der Evaluierungsteilnehmenden gemacht? | eigenes Qualitätskriterium |
| 23 | (Unabhängigkeit) Darstellung organisationale Unabhängigkeit der Gutachtenden | OECD DAC 1.2: Der Evaluierungsprozess ist [...] von Programmmanagement und Politikgestaltung unabhängig[.]<br><br>OECD DAC 3.2: Die Evaluatorinnen und Evaluatoren sind von der Entwicklungsmaßnahme, d.h. insbesondere von den Politikverantwortlichen, den durchführenden Stellen, dem Management sowie den Zielgruppen der Entwicklungsmaßnahme unabhängig. [...] Das Evaluierungsteam kann frei und ohne Einmischung von außen arbeiten. [...]<br><br>OECD DAC 3.10: Während der Evaluierung möglicherweise aufgetretene Hemmnisse werden erwähnt, wobei dargelegt wird, wie sie die Evaluierung beeinflusst haben, auch was deren Unabhängigkeit und Unparteilichkeit betrifft.<br><br>DeGEval F4: Der gesamte Evaluationsprozess soll die unparteiische Position der Evaluierenden erkennen lassen. | Das Qualitätskriterium ist erfüllt, wenn die Gutachtenden der Evaluierung eine ausreichende Distanz zur EZ-Maßnahme haben, das heißt, wenn sie weder 1) politisch 2) noch operativ 3) noch beratend an der EZ-Maßnahme beteiligt sind und auch 4) nicht zur Zielgruppe der EZ-Maßnahme gehören. | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 24 | (Unabhängigkeit) Darstellung Unvoreingenommenheit der Gutachtenden | OECD DAC 1.2: Der Evaluierungsprozess ist [...] von Programmmanagement und Politikgestaltung unabhängig[.]<br><br>OECD DAC 3.2: Die Evaluatorinnen und Evaluatoren sind von der Entwicklungsmaßnahme, d.h. insbesondere von den Politikverantwortlichen, den durchführenden Stellen, dem Management sowie den Zielgruppen der Entwicklungsmaßnahme[,] unabhängig. [...] Das Evaluierungsteam kann frei und ohne Einmischung von außen arbeiten. [...]<br><br>OECD DAC 3.10: Während der Evaluierung möglicherweise aufgetretene Hemmnisse werden erwähnt, wobei dargelegt wird, wie sie die Evaluierung beeinflusst haben, auch was deren Unabhängigkeit und Unparteilichkeit betrifft.<br><br>DeGEval F4: Der gesamte Evaluationsprozess […] [soll] die unparteiische Position der Evaluierenden erkennen lassen. | (Onlinebefragung) Wird für Evaluierungen Ihrer Organisation von allen Gutachter*innen eine „Erklärung zur Unabhängigkeit“ unterschrieben, die Aspekte zur persönlichen Unvoreingenommenheit beinhaltet? | eigenes Qualitätskriterium |
| 25 | Informationsgehalt der Zusammenfassung | OECD DAC 3.6: Ein schriftlicher Bericht enthält eine Zusammenfassung. Diese Zusammenfassung liefert einen Überblick über den Bericht und stellt dessen Hauptergebnisse, Schlussfolgerungen, Empfehlungen und sonstige allgemeine Erkenntnisse heraus. | Das Qualitätskriterium ist erfüllt, wenn eine Zusammenfassung vorliegt und diese 1. Ergebnisse und entweder 2. Schlussfolgerungen oder 3. Empfehlungen der Evaluierung nennt. | eigenes Qualitätskriterium |
| 26 | Berücksichtigung Kapazitätsentwicklung | OECD DAC 1.6: Die positiven Effekte, die der Evaluierungsprozess auf die Evaluierungskapazitäten der Entwicklungspartner haben kann, werden maximiert. Eine Evaluierung kann beispielsweise die Kapazitätsentwicklung unterstützen, indem sie zur Verbesserung der Evaluierungskenntnisse und -qualifikationen beiträgt, das Evaluierungsmanagement stärkt [...] und ein Klima der Rechenschaftspflicht und des Lernens fördert. | Das Qualitätskriterium ist erfüllt, wenn im Evaluierungsprozess Aktivitäten zur Entwicklung von Evaluierungskapazitäten der Akteur*innen im Partnerland durchgeführt werden (1. Verbesserung individueller Evaluierungskenntnisse, 2. Stärkung des Evaluierungsmanagements, 3. Förderung der Nutzung von Evaluierungsergebnissen, 4. Unterstützung eines Klimas/Umfelds von Rechenschaftspflicht und Lernen in den Partnerländern) | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 27 | Berücksichtigung von Gemeinschaftsevaluierungen | OECD DAC 2.6: Um zu Harmonisierung und Partnerausrichtung sowie zu einer effizienten Arbeitsteilung beizutragen, untersuchen die zuständigen Stellen der Geber und die Partnerländer systematisch die Möglichkeit einer gemeinsamen Evaluierung, die von mehr als einer Entwicklungsorganisation und/oder mehr als einem Partnerland kooperativ durchgeführt wird. Gemeinschaftsevaluierungen befassen sich sowohl mit Fragen, die für alle Partner von Interesse sind, als auch mit solchen, die für einzelne Partner wichtig sind. | (Onlinebefragung) Prüfen Sie bei den Evaluierungen Ihrer Organisation, ob diese als Gemeinschaftsevaluierung mit Gebern und Partnerregierungen durchgeführt werden können? | eigenes Qualitätskriterium |
| 28 | Berücksichtigung partnerschaftliche Ansätze | OECD DAC 1.4: Um die Eigenverantwortung für Entwicklung zu erhöhen [...], wird in einem frühen Stadium des Prozesses systematisch die Möglichkeit der Verfolgung eines partnerschaftlichen Ansatzes untersucht. [...]<br><br>OECD DAC 1.5: Zur Verbesserung der Koordinierung der Entwicklungsevaluierung und Stärkung der Systeme der Partnerländer werden im Evaluierungsprozess nationale und lokale Evaluierungspläne, -aktivitäten und -regeln berücksichtigt. | (Onlinebefragung) Berücksichtigen Evaluierungen Ihrer Organisation nationale Evaluierungsaktivitäten und Evaluierungspläne/ -politiken der Partnerländer, in denen die Evaluierung durchgeführt wird? | eigenes Qualitätskriterium |
| 29 | (Zusammensetzung der Gutachtenden) Genderbalance im Team | OECD DAC 3.1: […] Bei der Zusammenstellung des Teams wird auf einen ausgewogenen Anteil von Männern und Frauen geachtet und unter seinen Mitgliedern befinden sich Fachleute aus den betroffenen Partnerländern bzw. -regionen. | (Onlinebefragung) Wird in Evaluierungen, die von mindestens zwei Gutachter*innen durchgeführt wird, auf einen ausgewogenen Anteil von Männern und Frauen im Gutachter*innen-Team geachtet? | eigenes Qualitätskriterium |
| 30 | (Zusammensetzung der Gutachtenden) Einbezug von Gutachtenden aus Partnerland | OECD DAC 3.1: […] Bei der Zusammenstellung des Teams wird auf einen ausgewogenen Anteil von Männern und Frauen geachtet und unter seinen Mitgliedern befinden sich Fachleute aus den betroffenen Partnerländern bzw. -regionen. | Das Qualitätskriterium ist erfüllt, wenn sich im Gutachtenden-Team Fachleute aus den betroffenen Partnerländern befinden. | eigenes Qualitätskriterium |
| 31 | Ausreichende Ressourcen vorhanden | OECD DAC 2.10: Die für die Evaluierung bereitgestellten Ressourcen sind in Bezug auf Finanzierung, Personalausstattung […] ausreichend, um zu gewährleisten, dass die Ziele der Evaluierung effektiv erreicht werden. | (Onlinebefragung) Sind die für die Evaluierungen zur Verfügung stehenden Ressourcen ausreichend, um die Ziele der Evaluierungen zu erreichen? | eigenes Qualitätskriterium |

| Nr. | Qualitätskriterium | OECD-DAC- und/oder DeGEval-Standards | Qualitätskriterium / Frage in Onlinebefragung | Quelle |
|---|---|---|---|---|
| 32 | Vorhandensein einer Management-Response | OECD DAC 4.2: Auf die Empfehlungen wird systematisch reagiert, wobei die jeweils angesprochenen Personen/Stellen entsprechende Maßnahmen ergreifen. Dies beinhaltet eine formelle Stellungnahme („Management Response")[.] | (Onlinebefragung) Wird bei den Evaluierungen Ihrer Organisation eine „Management Response" verfasst? | eigenes Qualitätskriterium |
| 33 | Anwendung OECD-DAC-Kriterium – Relevanz | OECD DAC 2.8: Bei der Evaluierung werden die vereinbarten Kriterien des DAC für die Evaluierung der Entwicklungszusammenarbeit angewandt. [...] | Das Qualitätskriterium ist erfüllt, wenn das Kriterium angemessen bearbeitet wurde. | eigenes Qualitätskriterium |
| 34 | Anwendung OECD-DAC-Kriterium – Effektivität | OECD DAC 2.8: Bei der Evaluierung werden die vereinbarten Kriterien des DAC für die Evaluierung der Entwicklungszusammenarbeit angewandt. [...] | Das Qualitätskriterium ist erfüllt, wenn das Kriterium angemessen bearbeitet wurde. | eigenes Qualitätskriterium |
| 35 | Anwendung OECD-DAC-Kriterium – Effizienz | OECD DAC 2.8: Bei der Evaluierung werden die vereinbarten Kriterien des DAC für die Evaluierung der Entwicklungszusammenarbeit angewandt. [...] | Das Qualitätskriterium ist erfüllt, wenn das Kriterium angemessen bearbeitet wurde. | eigenes Qualitätskriterium |
| 36 | Anwendung OECD-DAC-Kriterium – Wirkungen | OECD DAC 2.8: Bei der Evaluierung werden die vereinbarten Kriterien des DAC für die Evaluierung der Entwicklungszusammenarbeit angewandt. [...] | Das Qualitätskriterium ist erfüllt, wenn das Kriterium angemessen bearbeitet wurde. | eigenes Qualitätskriterium |
| 37 | Anwendung OECD-DAC-Kriterium – Nachhaltigkeit | OECD DAC 2.8: Bei der Evaluierung werden die vereinbarten Kriterien des DAC für die Evaluierung der Entwicklungszusammenarbeit angewandt. [...] | Das Qualitätskriterium ist erfüllt, wenn das Kriterium im Bericht angewandt wurde. | eigenes Qualitätskriterium |

*Quelle: DEval, eigene Darstellung basierend auf OECD DAC (2010), DeGEval (2016) und Noltze et al. (2018)*
*Anmerkung: In Klammern gesetzte Namen in der Spalte „Qualitätskriterien" benennen Qualitätsstandards, die aus zwei oder mehreren Qualitätskriterien bestehen. In der Spalte „Qualitätskriterium/ Frage in Onlinebefragung" steht entweder die Beschreibung des Qualitätskriteriums, das über die Evaluierungsdokumente untersucht wurde oder die Frage aus der Onlinebefragung für die Verantwortlichen der Evaluierungseinheiten/-stellen (die Onlinebefragung wurde für die Qualitätskriterien genutzt, für die nicht ausreichend Informationen in den Evaluierungsdokumenten ermittelt werden konnten). Weitere Details finden sich im Kodierleitfaden in Kapitel 6. TOR = Terms of References.*

Tabelle 4 zeigt, inwieweit Teilaspekte aus den OECD-DAC- und den DeGEval-Standards dem jeweils anderen Standarddokument zugeordnet wurde.

**Tabelle 4 Übersicht über die Zuordnung zwischen OECD-DAC- und DeGEval-Standards**

| OECD-DAC-Standard (Teilaspekt) | in DeGEval-Standard (als Teilaspekt) erwähnt | DeGEval-Standard (Teilaspekt) | in OECD-DAC-Standard (als Teilaspekt) erwähnt |
|---|---|---|---|
| 1.1 | Einleitung – wurde nicht beachtet | G1 | 2.3, 3.8 |
| 1.2 | F4 | G2 | 3.7 |
| 1.3 | F2 | G3 | 2.1, 2.2, 2.7, 2.9, 3.10, 3.12, 3.13 |
| 1.4 | N1, N2, D2 | G4 | 3.9 |
| 1.5 | Teilaspekt wurde bei OECD DAC only aufgenommen. | G5 | 2.9, 3.10, 3.13 |
| 1.6 | Teilaspekt wurde bei OECD DAC only aufgenommen. | G6 | 1.7 |
| 1.7 | G6 | G7 | 2.9, 3.10, 3.13 |
| 2.1 | G3 | G8 | 3.11 |
| 2.2 | G3 | G9 | 4.1, 4.3 |
| 2.3 | G1 | N1 | 1.4, 2.5, 3.3, 4.1 |
| 2.4 | Dieser Standard wurde inhaltlich in verschiedenen anderen Standards und deren Teilaspekten beachtet (OECD DAC: 3.4; DeGEval: D1, D3). | N2 | 1.4, 2.5, 3.3, 4.1 |
| 2.5 | N1, N2, D2 | N3 | 3.1 |
| 2.6 | Teilaspekt wurde bei OECD DAC only aufgenommen. | N4 | 2.9, 3.10, 3.13 |
| 2.7 | G3 | N5 | 3.14, 3.15 |
| 2.8 | Teilaspekt wurde bei OECD-DAC-Kriterien aufgenommen. | N6 | 3.9 |
| 2.9 | G3, G5, G7, N4 | N7 | 4.1 |
| 2.10 | Teilaspekt wurde bei OECD DAC only aufgenommen. | N8 | 4.1 |
| 2.11 | Dieser Standard wurde inhaltlich in verschiedenen anderen Standards und deren Teilaspekten beachtet (OECD DAC: 1.2, 1.4, 2.5, 2.9, 3.1, 3.2, 3.3, 3.10, 3.13, 3.14, 3.15; DeGEval: G3, G5, G7, N1, N2, D2, N4, N5, F4). | D1 | 3.4 |
| 2.12 | F1 | D2 | 1.4, 2.5, 3.3, 4.1 |

| OECD-DAC-Standard (Teilaspekt) | in DeGEval-Standard (als Teilaspekt) erwähnt | DeGEval-Standard (Teilaspekt) | in OECD-DAC-Standard (als Teilaspekt) erwähnt |
|---|---|---|---|
| 3.1 | N3, ein weiterer Teilaspekt wurde bei OECD DAC only aufgenommen. | D3 | 3.4 |
| 3.2 | F4 | F1 | 2.12 |
| 3.3 | N1, N2, D2, F2, | F2 | 1.3, 3.3 |
| 3.4 | D1, D3 | F3 | Dieser Standard wurde inhaltlich in verschiedenen anderen Standards und deren Teilaspekten beachtet (OECD DAC: 2.9, 3.10, 3.13; DeGEval: G3, G5, G7, N4). |
| 3.5 | Dieser Standard wurde inhaltlich in verschiedenen anderen Standards und deren Teilaspekten beachtet (OECD DAC: 2.3, 3.6, 3.7, 4.1 u. a.; DeGEval: G1, G2). Darüber hinaus stellt er eher eine Einleitung dar. | F4 | 1.2, 3.2, 3.10 |
| 3.6 | Teilaspekt wurde bei OECD DAC only aufgenommen. | F5 | 4.1, 4.3 |
| 3.7 | G2 | / | / |
| 3.8 | G1 | / | / |
| 3.9 | G4, N6 | / | / |
| 3.10 | G3, G5, G7, N4, F4 | / | / |
| 3.11 | G8 | / | / |
| 3.12 | G3 | / | / |
| 3.13 | G3, G5, G7, N4 | / | / |
| 3.14 | N5 | / | / |
| 3.15 | N5 | / | / |
| 4.1 | G9, N1, N2, N7, N8, D2, F5 | / | / |
| 4.2 | Teilaspekt wurde bei OECD DAC only aufgenommen. | / | / |
| 4.3 | G9, F5 | / | / |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: Sofern OECD-DAC- oder DeGEval-Standards keine Überschneidungen aufweisen und nicht alleinstehend aufgenommen wurden, bedeutet dies, dass der Inhalt dieser Standards aus Sicht des Evaluierenden-Teams durch Qualitätskriterien mehrerer anderer Standards abgedeckt wird.*

### 2.1.2 Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit

Tabelle 5 stellt die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierung nebeneinander dar, sodass die Zuordnung aller und Transformation einiger Qualitätskriterien nachvollzogen werden kann. Ein Vergleich zwischen den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierung ist nicht immer uneingeschränkt möglich. Sofern in der letzten Spalte „Transformiert“ steht, stellt dies dar, inwieweit die in der vorliegenden Meta-Evaluierung herangezogenen Qualitätskriterien umgerechnet wurden, um den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit zu entsprechen. Die Vergleichbarkeit der Ergebnisse der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierung ist besonders für die Qualitätskriterien „Methodisches Vorgehen beschrieben (5)“ und „Befragte Gesprächspartner identifiziert (7)“ aus der Meta-Evaluierung Nachhaltigkeit eingeschränkt, da diese mit demselben Qualitätskriterium „Nachvollziehbarkeit der Informationsquellen (6)“ aus der vorliegenden Meta-Evaluierung verglichen wurden. Dasselbe gilt für die zwei Qualitätskriterien „Schlussfolgerungen durch Daten überwiegend referenziert (14)“ und „Schlussfolgerungen aus Daten plausibel begründet (15)“, die mit dem Qualitätskriterium „Kohärenz von Daten-Ergebnissen-Schlussfolgerungen (9)“ verglichen wurden.

**Tabelle 5 Nebeneinanderstellung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierungen**

| Bereich | Name Qualitätskriterium Meta_Q (Nummer) | Name Qualitätskriterium Meta_N (Nummer) | Qualitätskriterium Meta_Q: Das Qualitätskriterium ist erfüllt, wenn … | Qualitätskriterium der Meta_N: Das Kriterium ist erfüllt, wenn … | Umgang mit Qualitätskriterium der Meta_N in der Meta_Q |
|---|---|---|---|---|---|
| Hintergrund | Beschreibung des Evaluierungs-gegenstands (1) | Gegenstand (Vorhaben) beschrieben (1) | … 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteur*innen (politische Partner*innen und/oder Träger) der EZ-Maßnahme genannt sind. | … 1) Ziele, 2) Zielgruppe, 3) Kontext sowie 4) Relevante Akteure (politischer Partner und/oder Träger) der EZ-Maßnahme dargestellt sind und somit eine Eingrenzung des Gegenstandes vorgenommen wurde. | Transformiert: 1 = wenn „Beschreibung des Kontexts der Entwicklungsmaßnahme“ und „Beschreibung des Evaluierungsgegenstands“ mit 4 bewertet wurde; 0 = wenn „Beschreibung des Kontexts der Entwicklungsmaßnahme“ und „Beschreibung des Evaluierungsgegenstandes“ mit 1, 2 oder 3 bewertet wurde |
| | Beschreibung des Erkenntnisinteresses (2) | Erkenntnisinteresse formuliert/operationalisiert (2) | … 1. Zweck, 2. Ziele und 3. Evaluierungsfrage(n) der Evaluierung genannt werden. | … das Erkenntnisinteresse und/oder Evaluierungsfragen spezifiziert bzw. konkretisiert wurden. | Transformiert: 1 = wenn „Beschreibung des Erkenntnisinteresses“ mit 3 und 4 bewertet wurde; 0 = wenn mit 1 und 2 bewertet wurde |
| Darstellung der Wirkungszusammenhänge | Darstellung der Wirkungszusammenhänge (4) | Wirkungslogik/Wirkungskette dargestellt (3) | … bei der Beschreibung der intendierten Wirkungszusammenhänge der EZ-Maß- | … bei der Darstellung der intendierten Wirkungen der EZ-Maßnahme zwischen verschiedenen Wirkungsebenen | Transformiert: 1 = wenn „Darstellung der Wirkungszusammenhänge“ mit 2, 3, oder 4 bewertet wurde; 0 = wenn mit 1 bewertet wurde |

| Bereich | Name Qualitätskriterium Meta_Q (Nummer) | Name Qualitätskriterium Meta_N (Nummer) | Qualitätskriterium Meta_Q: Das Qualitätskriterium ist erfüllt, wenn … | Qualitätskriterium der Meta_N: Das Kriterium ist erfüllt, wenn … | Umgang mit Qualitätskriterium der Meta_N in der Meta_Q |
|---|---|---|---|---|---|
| | | | nahme zwischen den Wirkungsstufen Input-Output-Outcome-Impact unterschieden wird. | (Input-Output-Outcome-Impact) unterschieden wird und diese logisch aufeinander aufbauen (und/oder gegebenenfalls Wirkungshypothesen formuliert werden). | |
| | | Wirkungslogik überwiegend durch Indikatoren operationalisiert (4) | siehe Meta_N | … der Zielerreichungsgrad der Mehrheit der Programmziele messbar gemacht/anhand von Indikatoren abgeschätzt wird. | Qualitätskriterium der Meta_N wurde übernommen. |
| Methodisches Vorgehen | Nachvollziehbarkeit der Informationsquellen (6) | Methodisches Vorgehen beschrieben (5) | … die in der Evaluierung genutzten Informationsquellen ausreichend detailliert im Evaluierungsbericht dargestellt werden. | … die in der Evaluierung zur Anwendung kommenden Arbeitsschritte zur Datenerhebung und Auswertung beschrieben und operationalisiert sind. | Transformiert: 1 = wenn „Nachvollziehbarkeit der Informationsquellen“ mit 3 und 4 bewertet wurde; 0 = wenn mit 1 und 2 bewertet wurde |
| | | Befragte Gesprächspartner identifiziert (7) | | … die zur Datenerhebung konsultierten/befragten Gesprächspartner identifiziert wurden. | |
| | | Auswahlverfahren der Gesprächspartner*innen beschrieben (8) | siehe Meta_N | … die Auswahl der Gesprächspartner beschrieben wurde bzw. die Auswahlkriterien dargestellt wurden. | Qualitätskriterium der Meta_N wurde übernommen. |
| | Darstellung der Angemessenheit des methodischen Vorgehens (7) | Stärken und/oder Limitationen des methodischen Vorgehens identifiziert (6) | … a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden. | … begründet wird, warum die angewandten Methoden dem Gegenstand der Evaluierung angemessen sind. Vorteile und Limitationen des methodischen Vorgehens werden diskutiert. | Transformiert: 1 = wenn „Darstellung der Angemessenheit des methodischen Vorgehens“ mit 2, 3, oder 4 bewertet wurde; 0 = wenn mit 1 bewertet wurde |

| Bereich | Name Qualitätskriterium Meta_Q (Nummer) | Name Qualitätskriterium Meta_N (Nummer) | Qualitätskriterium Meta_Q: Das Qualitätskriterium ist erfüllt, wenn … | Qualitätskriterium der Meta_N: Das Kriterium ist erfüllt, wenn … | Umgang mit Qualitätskriterium der Meta_N in der Meta_Q |
|---|---|---|---|---|---|
| Evaluierungs-design | Darstellung der Angemessenheit des methodischen Vorgehens (7) | Vorher-nachher-Vergleich (9) | siehe Meta_N | … die Wirkungen des EZ-Vorhabens anhand eines Vergleichs der Mehrzahl aller Indikatoren vor Vorhabenbeginn und nach Vorhabenende ermittelt werden. | Qualitätskriterium der Meta_N wurde übernommen. |
| | | Kontroll-/Vergleichsgruppe einbezogen (10) | siehe Meta_N | … die Wirkungen des EZ-Vorhabens anhand eines Vergleichs zwischen Kontroll- (außerhalb des Einflussbereichs der EZ-Maßnahme) und Interventions-Gruppe (innerhalb des Einflussbereichs der EZ-Maßnahme) ermittelt werden. | Qualitätskriterium der Meta_N wurde übernommen. |
| | | Kausalität über Plausibilitäten hergeleitet (11) | siehe Meta_N | … die Wirkungen des EZ-Vorhabens auf der Grundlage eines systematischen Verfahrens anhand von Plausibilitäten (insbesondere theoriebasierter Ansätze, z.B. durch Kontributionsanalysen) ermittelt werden. | Qualitätskriterium der Meta_N wurde übernommen. |
| Robustheit der Ergebnisse | Darstellung der Angemessenheit des methodischen Vorgehens (7) | Daten-Triangulation (12) | Daten-Triangulation wurde inhaltlich über Methoden-Triangulation abgedeckt. | / | / |
| | | Methoden-Triangulation angewandt (13) | siehe Meta_N | … die Auswertung der Daten derselben Quelle durch verschiedene Methoden erfolgt (> 1 Methode). | Qualitätskriterium der Meta_N wurde übernommen. |
| Auswertung und Schlussfolgerungen | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen (9) | Schlussfolgerungen durch Daten überwiegend referenziert (14) | … die Datenanalyse, die Ergebnisse und die Schlussfolgerungen logisch miteinander in Bezug gesetzt werden beziehungsweise aufeinander aufbauen. | … der überwiegende Anteil der Ergebnisse und Schlussfolgerungen mit der Datengrundlage/-analyse in der Mehrheit der Schlussfolgerungen in Bezug gesetzt werden. | Transformiert: 1 = wenn „Kohärenz von Daten-Ergebnissen-Schlussfolgerungen“ mit 3 und 4 bewertet wurde; 0 = wenn mit 1 und 2 bewertet wurde |

| Bereich | Name Qualitätskriterium Meta_Q (Nummer) | Name Qualitätskriterium Meta_N (Nummer) | Qualitätskriterium Meta_Q: Das Qualitätskriterium ist erfüllt, wenn … | Qualitätskriterium der Meta_N: Das Kriterium ist erfüllt, wenn … | Umgang mit Qualitätskriterium der Meta_N in der Meta_Q |
|---|---|---|---|---|---|
| | | | | Blick auf die Wirkung auf der Grundlage der verwendeten Daten nachvollziehbar sind. | |
| | | Datengrundlage ausreichend hinsichtlich Schlussfolgerungen (16) | siehe Meta_N | …. die Datengrundlage und die methodische Vorgehensweise qualitativ und quantitativ ausreichend sind, um die ausgesprochenen Schlussfolgerungen (im Sinne von erreichten Wirkungen) zu ziehen. | Qualitätskriterium der Meta_N wurde übernommen. |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: Meta_N = Meta-Evaluierung Nachhaltigkeit; Meta_Q = vorliegende Meta-Evaluierung*

## 2.2 Definitionen und Wirkungszusammenhänge der Faktoren

In Tabelle 6 werden für die untersuchten Faktoren deren Definitionen sowie Wirkungszusammenhänge mit der Anwendung der Qualitätskriterien und den daraus abgeleiteten Hypothesen dargestellt.[1]

**Tabelle 6 Definition, Wirkungszusammenhänge und Hypothesen der Faktoren**

| Faktor/Proxy | Definition | Wirkungszusammenhang/-hänge | Hypothesen |
|---|---|---|---|
| Konflikt | • Bewaffneter Konflikt, der eine Regierung und/oder ein Gebiet betrifft, über das der Einsatz von Waffengewalt zwischen den Streitkräften von zwei Parteien und der mindestens 25 kampfbedingte Todesopfer pro Jahr zur Folge hat (UCDP, o. D.). | • Konflikte erschweren die Durchführung von Evaluierungen – insbesondere die Datenerhebung und Zusammenarbeit mit Organisationen vor Ort – aufgrund von Zugangsbeschränkungen, einem erhöhten Sicherheitsrisiko und einer eingeschränkten Infrastruktur.<br>• Bei Konflikten werden öfter lokale Gutachtende aus dem Partnerland einbezogen. | • Wenn ein „Konflikt“ besteht, beeinflusst dies das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC negativ.<br>• Wenn ein „Konflikt“ besteht, beeinflusst dies die QKs „Einbindung der internen und externen Stakeholder*innen“ und „Berücksichtigung Kapazitätsentwicklung“ negativ und das QK „Einbezug von Gutachtenden aus Partnerland“ positiv. |
| Sozialkapital | • Sozialkapital[2] ist das Vertrauen (innerhalb einer Gemeinde oder Gesellschaft) in verschiedene Gruppen: Familie, Nachbarschaft; Personen, die man persönlich kennt; Personen, die man das erste Mal trifft; Personen einer anderen Religion; und Personen einer anderen Nationalität (WVS, 2022). | • Wenn das Sozialkapital höher ist, findet eine effektivere Datenerhebung statt sowie eine bessere Einbindung der Organisationen/Partner vor Ort. | • Ein höheres „Sozialkapital“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC sowie den QK „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Pandemie (Jahr 2020) | • Stark ausbreitende (meist weltweite) Krankheiten, zum Beispiel von Viren, gegen die große Teile der Bevölkerung noch keinen Immunschutz besitzt (RKI, 2009). | • Pandemien erschweren die Durchführung von Evaluierungen – insbesondere die Datenerhebung und Zusammenarbeit mit Organisationen vor Ort – aufgrund von Zugangsbeschränkungen und einer eingeschränkten Infrastruktur. | • Wenn eine „Pandemie“ besteht, beeinflusst dies das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC negativ.<br>• Wenn eine „Pandemie“ besteht, beeinflusst dies die QKs „Einbindung der internen und externen |

[1] Obwohl die Wirkungszusammenhänge und Hypothesen (zum Teil) kausal beziehungsweise gerichtet formuliert wurden, wurden die Ergebnisse der Regressionsanalysen korrelativ interpretiert, da mit dem untersuchten Design keine Messung von rigorosen Wirkungszusammenhängen möglich war.

[2] Es gibt eine Vielzahl an unterschiedlichen inhaltlichen Dimensionen von „Sozialem Kapital“. Aufgrund der Datenverfügbarkeit wurde im Rahmen dieser Meta-Evaluierung auf Informationen und somit auch die Definition des World Values Surveys zurückgegriffen.

| Faktor/Proxy | Definition | Wirkungszusammenhang/-hänge | Hypothesen |
|---|---|---|---|
| | | • Bei Pandemien werden öfter lokale Gutachtende aus dem Partnerland einbezogen. | Stakeholder*innen“ und „Berücksichtigung Kapazitätsentwicklung“ negativ und das QK „Einbezug von Gutachtenden aus Partnerland“ positiv. |
| Anzahl der internen und externen Gutachtenden | • Anzahl der externen und internen Gutachtenden[a] | • Wenn mehr Gutachtende die Evaluierung durchführen, dann liegen unterschiedliche Kompetenzen beim Evaluierungsteam vor, zum Beispiel bei Evaluierungs- und Methodenkenntnissen, und desto besser ist die Berichtslegung und Methoden und die Kapazitätsentwicklung der Stakeholder*innen in die Evaluierung.<br>• Je mehr Gutachtende die Evaluierung durchführen, desto wahrscheinlicher ist es, dass ein*e nationale*r (lokale*r) Gutachtende Teil des Evaluierungsteams ist. | • Eine höhere „Anzahl der internen und externen Gutachtenden“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Eine höhere „Anzahl der internen und externen Gutachtenden“ beeinflusst die QKs „Einbezug von Gutachtenden aus Partnerland und „Berücksichtigung Kapazitätsentwicklung“ positiv. |
| Tagessatz externe Gutachtende | • Durchschnittlicher Tagessatz aller externen Gutachtenden[a] | • Je höher der Tagessatz der externen Gutachtenden ist, desto höhere Qualifikationen liegen vor und desto besser sind die Berichtslegung und Methoden.<br>• Je höher der Tagessatz der externen Gutachtenden ist, desto mehr Kompetenzen bestehen, um die Partner*innen und weitere Stakeholder*innen einzubeziehen und die Kapazitäten im Bereich der Evaluierung voranzutreiben. | • Ein „höherer Tagessatz für externe Gutachtende“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Ein „höherer Tagessatz für externe Gutachtende“ beeinflusst die QKs „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Informationsgehalt der Leistungsbeschreibung | • Die Leistungsbeschreibung beinhaltet grundlegende Informationen zum Evaluierungsgegenstand und gibt die Rahmendaten der Evaluierung vor (zum Beispiel Zeitraum der Evaluierung, einzuhaltende Standards, geforderte Kompetenzen). Mithilfe der Leistungsbeschreibung werden Gutachtende gesucht, die die Evaluierung durchführen. | • Je informativer die Leistungsbeschreibung ist, desto besser ist die Orientierung für Gutachtende hinsichtlich der Berichtslegung und der Methoden. Da sich Leistungsbeschreibung an die Gutachtenden richtet, kann diese jene QSs beeinflussen, deren Anwendung bei den Gutachtenden liegt (alle QSs des SCs „Berichtslegung und Methoden“ und QS „Nützlichkeit der Empfehlungen“). | • Eine „informative Leistungsbeschreibung“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Eine „informative Leistungsbeschreibung“ beeinflusst den QS „Nützlichkeit der Empfehlungen“ positiv. |

| Faktor/Proxy | Definition | Wirkungszusammenhang/-hänge | Hypothesen |
|---|---|---|---|
| Qualitätssicherung mit Inception Report | • Ein Inception Report wird vor der Datenerhebung beziehungsweise zu Beginn einer Evaluierung – meist nach der Sichtung der vorhandenen Dokumente – erstellt. Er beinhaltet beispielsweise Vorkenntnisse zum Evaluierungsgegenstand, Evaluierungsdesign, Evaluierungsfragen sowie Angaben zur Datenerhebung/methodischen Herangehensweise. | • Wenn ein IR verfasst wird, dann findet eine stärkere und frühzeitige Auseinandersetzung mit dem Evaluierungsgegenstand, den Partner*innen sowie mit der Berichtslegung und Methoden statt. | • Wird ein „Inception Report“ verfasst, beeinflusst das den SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Wird ein „Inception Report“ verfasst, werden die QKs „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv beeinflusst.<br>• Wird ein „Inception Report“ verfasst, wird der QS „Nützlichkeit der Empfehlungen“ positiv beeinflusst. |
| Einbindung der internen und externen Stakeholder*innen | • Einbindung der Evaluierungsmitwirkenden zum Beispiel Evaluierungsverantwortliche (aus der Evaluierungseinheit) oder Partner*innen in unterschiedliche Phasen der Evaluierung | • Je stärker beziehungsweise öfter die Evaluierungsmitwirkenden in die Evaluierung eingebunden sind, desto mehr Qualitätskontrollen finden statt, desto mehr unterschiedliche Perspektiven werden berücksichtigt und desto besser wird die Berichtslegung und Methoden. | • Eine stärkere „Einbindung der internen und externen Stakeholder*innen“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SCs positiv. |
| GAs-Tage/Kosten Maßnahme | • Anzahl der Gutachtenden-Tage durch die Kosten der EZ-Maßnahme | • Je mehr Tage für die Gutachtenden pro Euro der EZ-Maßnahme vorliegen, desto mehr Geld steht für die Durchführung der Evaluierung zur Verfügung und desto besser ist die Berichtslegung und Methoden. | • Eine höheres „Verhältnis der Gutachtenden-Tagen zu den Kosten der EZ-Maßnahme“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv. |
| Remote-Datenerhebung | • Für den Einflussfaktor „Remote-Datenerhebung“ wurden die folgenden drei Kategorien definiert:<br>○ Remote: Die Daten – zum Beispiel Interviews oder Fokusgruppendiskussionen – werden durch die Gutachtenden ausschließlich „aus der Ferne“ beziehungsweise „virtuell“ erhoben (unabhängig davon, ob sich diese im Partnerland befinden). *Face to face*-Erhebungen finden nicht statt.<br>○ Semi-remote: Ein Teil der Daten wird „aus der Ferne“ und ein Teil der Daten wird vor Ort im Partnerland *face to face* erhoben. | • Wenn Daten remote erhoben werden, kann das zu Einschränkungen in Bezug auf den Informationsgehalt und das methodische Vorgehen und somit zu einer geringeren Angemessenheit des methodischen Vorgehens der Evaluierung führen.<br>• Wenn Daten remote erhoben werden, ist öfter ein*e nationale*r/lokale*r Gutachtende eingesetzt, da diese gut vernetzt sein sowie den Projekt- beziehungsweise Länderkontext gut kennen müssen. Darüber hinaus wird die Evaluierungskapazität der*s lokalen Gutachtende gefördert. | • Werden Daten remote erhoben, wird das QK „Darstellung der Angemessenheit des methodischen Vorgehens“ negativ beeinflusst.<br>• Werden Daten remote erhoben, werden die QKs „Einbezug von Gutachtenden aus Partnerland“ und „Berücksichtigung Kapazitätsentwicklung“ positiv beeinflusst. |

| Faktor/Proxy | Definition | Wirkungszusammenhang/-hänge | Hypothesen |
|---|---|---|---|
| | ○ Kein remote: Die Daten werden ausschließlich vor Ort und somit *face to face* erhoben. | | |
| Jahr der Evaluierung | • Jahr, in dem die Evaluierung durchgeführt wurde | • Die Evaluierungssysteme der Organisationen entwickeln sich weiter und verbessern sich, wodurch sich die Qualität von Evaluierungen – vor allem im Bereich „Berichtslegung und Methoden“ – mit den Jahren verbessert. | • Mit den „Jahren“ wird das SC „Berichtslegung und Methoden“ beziehungsweise werden die einzelnen QKs des SC positiv beeinflusst. |
| Größe der Evaluierungseinheiten/-stellen[b] absolut | • Vollzeitäquivalente der Evaluierungseinheiten/-stellen je Jahr | • Je mehr Vollzeitäquivalente der Evaluierungseinheiten/-stellen für die (Mit-)Verantwortung von Evaluierungen zur Verfügung stehen, desto mehr Ressourcen stehen zur Verfügung, um die gesamte Evaluierung stärker und enger mit zu begleiten, beispielsweise die Berichtslegung und die Methoden sowie Aspekte der Nutzbarkeit und Partizipation und Fairness. | • Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen“ beeinflussen das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen“ beeinflussen die QKs „Nützlichkeit der Empfehlungen“, „Einbezug von Gutachtenden aus Partnerland“, „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Größe der Evaluierungseinheiten/-stellen[b] im Verhältnis zur Anzahl der Evaluierungen | • Vollzeitäquivalente der Evaluierungseinheiten/-stellen im Verhältnis zu Gesamtzahl aller (mit-)verantworteten Evaluierungen je Jahr | • Je mehr Vollzeitäquivalente der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen für die (Mit-)Verantwortung zur Verfügung stehen, desto mehr Ressourcen stehen je Evaluierung zur Verfügung, um die gesamte Evaluierung stärker und enger mit zu begleiten, beispielsweise die Berichtslegung und die Methoden sowie Aspekte der Nutzbarkeit und Partizipation und Fairness. | • Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen“ beeinflussen das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen“ beeinflussen die QKs „Nützlichkeit der Empfehlungen“, „Einbezug von Gutachtenden aus Partnerland“, „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Größe der Evaluierungseinheiten/- | • Vollzeitäquivalente der Evaluierungseinheiten/-stellen im Verhältnis zu den Vollzeitäquivalenten der Organisation je Jahr | • Je mehr Vollzeitäquivalente der Evaluierungseinheiten/-stellen im Verhältnis zu den Vollzeitäquivalenten der Organisation zur Verfügung stehen, desto mehr Bedeutung wird der (Mit-)Verantwortung von Evaluierungen beigemessen, und es | • Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen im Verhältnis zu den Vollzeitäquivalenten der Organisation“ beeinflussen das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv. |

| Faktor/Proxy | Definition | Wirkungszusammenhang/-hänge | Hypothesen |
|---|---|---|---|
| stellen[b] im Verhältnis zur Größe der Organisation | | ist davon auszugehen, dass dann eine stärkere Auseinandersetzung mit der Qualität von Evaluierungen erfolgt. Dies beeinflusst die Berichtslegung und die Methoden sowie Aspekte der Nutzbarkeit und Partizipation und Fairness. | • Mehr „Vollzeitäquivalente in der Evaluierungseinheiten/-stellen im Verhältnis zu den Vollzeitäquivalenten der Organisation“ beeinflussen die QKs „Nützlichkeit der Empfehlungen“, „Einbezug von Gutachtenden aus Partnerland“, „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Evaluierungstätigkeit | • Durchschnittlicher Prozentsatz der Arbeitstätigkeit der Mitarbeitenden der Evaluierungseinheiten/-stellen**, der für die Evaluierungstätigkeit aufgewendet wird | • Je mehr Evaluierungstätigkeiten Mitarbeitende in den/r Evaluierungseinheiten/-stellen haben, desto mehr Evaluierungskompetenz besitzen diese und können somit die Evaluierungen qualitativ besser mitbegleiten. Dies beeinflusst die Berichtslegung und die Methoden sowie Aspekte der Nutzbarkeit und Partizipation und Fairness. | • Mehr „Evaluierungstätigkeit der Mitarbeitenden“ beeinflusst das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv.<br>• Mehr „Evaluierungstätigkeit der Mitarbeitenden“ beeinflusst die QKs „Nützlichkeit der Empfehlungen“, „Einbezug von Gutachtenden aus Partnerland“, „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der internen und externen Stakeholder*innen“ positiv. |
| Organisation | • Die elf teilnehmenden Organisationen, die im Rahmen der Meta-Evaluierung untersucht werden | • Unterschiedliche Aspekte der Organisationen (zum Beispiel die strukturierten Planungsprozesse und die Lernkultur von Evaluierungen) beeinflussen die untersuchten QKs sowohl positiv als auch negativ. | • Die „Organisationen“ beeinflussen das SC „Berichtslegung und Methoden“ beziehungsweise die einzelnen QKs des SC positiv beziehungsweise negativ.<br>• Die „Organisationen“ beeinflussen die QKs „Nützlichkeit der Empfehlungen“, „Einbezug von Gutachtenden aus Partnerland“, „Berücksichtigung Kapazitätsentwicklung“ und „Einbindung der Stakeholder*innen“ positiv beziehungsweise negativ. |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: SC = Standardcluster; QS = Qualitätsstandard; QK = Qualitätskriterium; IR = Inception Report. [a] Definition: Als externe Gutachtende gelten Personen, die mit der Durchführung der Evaluierung (in der Regel im Rahmen einer externen Ausschreibung) beauftragt wurden. Externe Gutachtende sind strukturell unabhängig von der auftraggebenden Organisation oder der Partnerorganisation und können sowohl lokal/national als auch international sein. Interne Gutachtende umfassen Gutachtende von (un-)beteiligten Stellen innerhalb der Organisation oder Partnerorganisation, die die Evaluierung durchführen. [b] Da einige Organisationen keine Evaluierungseinheit haben, fallen hierunter die Mitarbeitenden, die in der jeweiligen Organisation Evaluierungen (mit-)verantworten.*

# 3. ERGÄNZENDE INFORMATIONEN ZUM METHODISCHEN VORGEHEN

## 3.1 Stichprobenziehung

Im Rahmen der Meta-Evaluierung wurde ein zweistufiges Verfahren zur Identifikation der relevanten (Projekt-)Evaluierungen gewählt (Abbildung 1). Zuerst wurden die staatlichen und nichtstaatlichen Organisationen (Primäreinheiten) ausgewählt. Dabei wurden die staatlichen Organisationen vollständig aufgenommen und sieben nichtstaatliche Organisationen entlang der nachfolgenden Kriterien ausgewählt:

*1. Kriterium – Höhe der BMZ-Zuwendungen:* Zwei Organisationen mit den **höchsten** und **niedrigsten absoluten BMZ-Zuwendungen wurden ausgewählt** (Annahme: Höhe der BMZ-Zuwendungen spiegelt sich in der Höhe der Kosten für die Evaluierung wider; Evaluierungen mit unterschiedlichen Volumen erfüllen unterschiedliche Qualitätsstandards; Evaluierungen mit hohem Volumen erfüllen eher mehr Qualitätsstandards und/oder einzelne Qualitätsstandards umfassender)**.**

*2. Kriterium – Relative Evaluierungstätigkeit:* Eine Organisation mit **relativ hoher** und **niedriger Evaluierungstätigkeit wurde ausgewählt** (niedrige Anzahl an BMZ-Zuwendungen pro Evaluierung; Annahme: Organisationen mit relativer hoher Evaluierungstätigkeit sind erfahrener in ihrer Evaluierungstätigkeit und erfüllen mehr Qualitätsstandards, Organisationen mit niedriger Evaluierungstätigkeit erfüllen weniger Qualitätsstandards).

*3. Kriterium – Haushaltstitel und Förderrichtlinien:* Mindestens **eine Organisation je Haushaltstitel wurde ausgewählt** (Annahme: Organisationen, die aus unterschiedlichen **Haushaltstiteln** BMZ-Zuwendungen erhalten, haben unterschiedliche BMZ-Vorgaben hinsichtlich ihrer Evaluierungstätigkeit und erfüllen gegebenenfalls unterschiedliche Qualitätsstandards).[3]

*4. Kriterium – Evaluierungshäufigkeit:* mindestens **ca. zwei Evaluierungen pro Jahr** (Annahme: Organisationen ohne regelmäßige Evaluierungstätigkeit haben weniger Anlass, sich mit Qualitätsstandards auseinandergesetzt zu haben). [4]

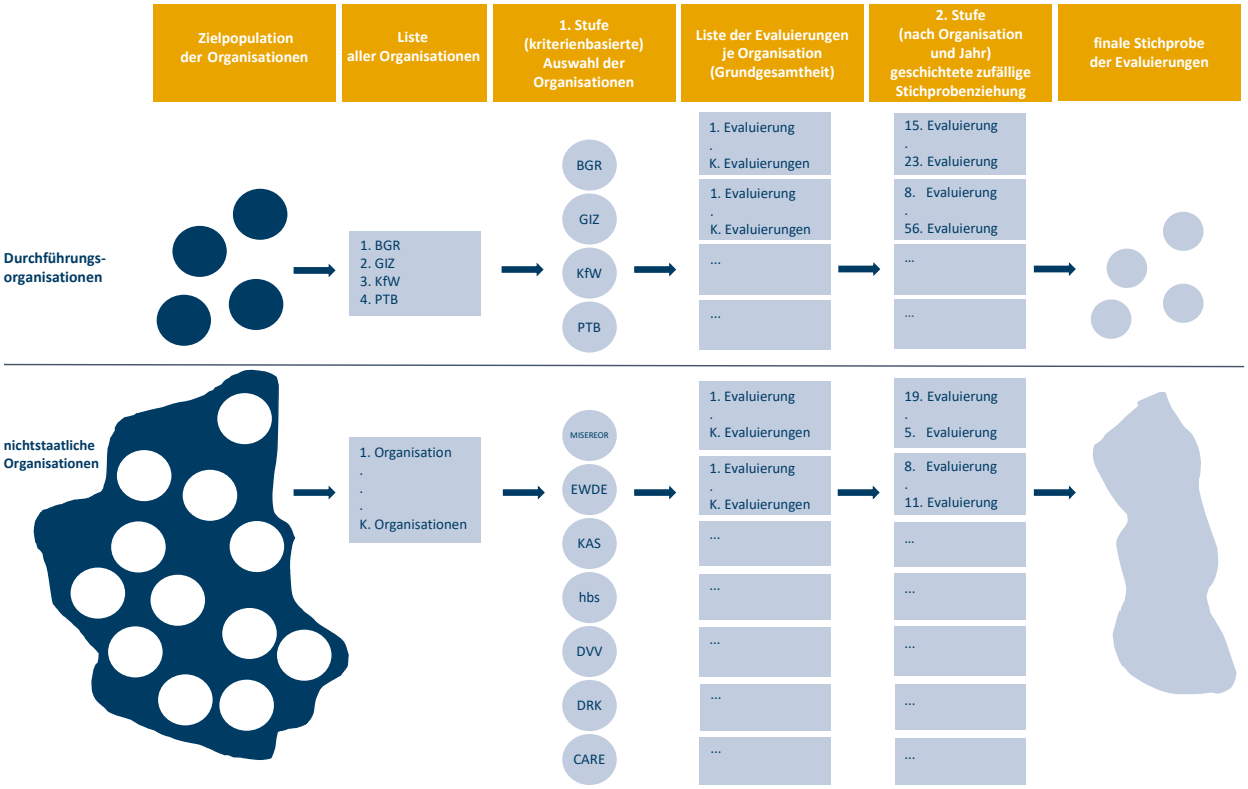
Im Anschluss wurden die relevanten (Projekt-)Evaluierungen (Sekundäreinheiten) der teilnehmenden Organisationen entlang der im Evaluierungsbericht beschriebenen Auswahlparameter identifiziert. Bei Organisationen mit weniger als zehn Evaluierungen wurde keine Zufallsstichprobe gezogen, sondern eine Vollerhebung (Untersuchung aller Evaluierungen) vorgenommen. Bei Organisationen mit mehr als zehn Evaluierungen wurden je Organisation so viele Evaluierungen identifiziert, dass bei einem Konfidenzniveau von 95 Prozent eine Fehlerspanne von 10 Prozent nicht überstiegen wurde. Diese statistischen Messwerte erklären die Aussagekraft der Stichprobe. Die Fehlerspanne steht dabei für eine mögliche Abweichung der gefundenen

[3] Konkret wurden bei den nichtstaatlichen Organisationen die ausgewählt, die im Kapitel „Zivilgesellschaftliches, kommunales und wirtschaftliches Engagement“ des Einzelplans 23 in den vier BMZ-Haushaltstiteln „Förderung entwicklungswichtiger Vorhaben der Sozialstruktur“, „Förderung entwicklungswichtiger Vorhaben der politischen Stiftungen“, „Förderung entwicklungswichtiger Vorhaben der Kirchen“ und „Förderung entwicklungswichtiger Vorhaben privater deutscher Träger“ (Bundestag, 2020) verortet werden konnten. Nicht aufgenommen wurden nichtstaatliche Organisationen aus den Haushaltstiteln „Institutionelle Förderung von Engagement Global“ und „Förderung der entwicklungspolitischen Bildung“, da beide einen Fokus auf Aktivitäten im Inland haben; den Haushaltstiteln „Entwicklungspartnerschaft mit der Wirtschaft“ und „Förderung des kommunalen Engagements“, da beide im Rahmen anderer DEval-Evaluierungen abgedeckt werden; den Haushaltstiteln „Ziviler Friedensdienst“ und „Entwicklungspolitischer Austausch und Freiwilligendienst“, da beide einen Fokus auf Entsendungen legten und nicht auf „klassische“ Entwicklungsmaßnahmen; und dem Haushaltstitel „Gesellschaftliche Krisenreaktions- und Stabilisierungshilfe“, da diesem 0 Prozent der BMZ-Mittel hinterlegt war.

[4] Ausnahmen bildeten hierbei GIZ und CARE, bei denen ausschließlich Evaluierungen für die Jahre 2018 bis 2020 vorlagen, da in den Jahren 2016 und 2017 entweder keine oder nur Evaluierungen eines älteren Evaluierungstyps durchgeführt wurden.

Ergebnisse von +/− 10 Prozent vom realen Wert. Das Konfidenzlevel besagt, wie sicher die gefundenen Ergebnisse sind. Zusammen mit der Fehlerspanne kann man also Aussagen treffen, dass zum Beispiel bei einem im Sample gefundenen Prozentanteil von 40 Prozent zu einem bestimmten Qualitätskriterium mit 95 Prozent Wahrscheinlichkeit der Wert in der Grundgesamtheit zwischen 30 Prozent und 50 Prozent liegt. Die Stichprobe lässt somit valide Aussagen über die Grundgesamtheit zu. Gewisse Abweichungen sind durch die relativ weite, aber geläufig angewandte Fehlerspanne möglich. Die Gesamtzahl von 576 Evaluierungen stellt die Grundgesamtheit[5] der zentralen (Projekt-)Evaluierungen der teilnehmenden Organisationen dar.

**Abbildung 1 Mehrstufiges Verfahren zur Auswahl der (Projekt-)Evaluierungen**




*Quelle: DEval, eigene Darstellung angelehnt an Döring und Bortz (2016, S. 316)*

**Durch eine Reihe von Maßnahmen – Festlegung der Fehlerspanne und des Konfidenzintervalls bei der Stichprobenziehung, Gewichtung der Daten bei der Auswertung, Pilotierung des Analyserasters in der Inter-Kodierenden-Phase und Gewährleistung der Inter-Kodiererenden-Übereinstimmung – konnte das Risiko eines totalen Stichprobenfehlers[6] als eher gering eingestuft werden.** Der totale Stichprobenfehler ergab sich aus mehreren möglichen Fehlern, den sogenannten Stichprobenfehlern und Nichtstichprobenfehlern. Letztere ließen sich in die Fehlerquellen Spezifikationsfehler, Beobachtungsfehler und Nichtbeobachtungsfehler unterteilen (Faulbaum, 2019). Die Herausforderungen und Maßnahmen zur Reduzierung dieser Fehlerquellen werden im Folgenden näher erläutert.

[5] Da davon ausgegangen wurde, dass die Auswahlgesamtheit (Auswahlpopulation), das heißt alle vordefinierten Evaluierungen, in die Gesamtliste der Evaluierungen integriert war (gegebenenfalls mit einzelnen nicht systematischen Ausnahmen), wurde im Rahmen der Meta-Evaluierung der Begriff „Grundgesamtheit“ (Zielpopulation der Evaluierungen) mit dem Begriff „Auswahlgesamtheit“ (Auswahlpopulation der Evaluierungen) gleichgesetzt (Döring und Bortz, 2016, S. 294).

[6] Hierbei war es die Absicht, die „Abweichung der Parameterschätzung (Schätzung von Populationsparametern wie Mittelwert, Varianz etc.) auf der Basis des (bereinigten) Datensatzes […] auf verschiedene Fehlerquellen zurückzuführen, die auf unterschiedlichen Stufen des Erhebungsprozesses in unterschiedlichem Ausmaß zur Gesamtabweichung der Parameterschätzung vom wahren Populationswert beitragen“ (Faulbaum, 2019, S. 505).

Der *Stichprobenfehler* bewirkt eine Abweichung der Datenstruktur der Stichprobe vom wahren Wert der Grundgesamtheit. Da im Rahmen der Meta-Evaluierung für alle Organisationen eine geschichtete Zufallsauswahl auf Basis einer Fehlerspanne von 10 Prozent und einem Konfidenzintervall von 95 Prozent aufgenommen wurde (Ausnahme CARE: Vollerhebung)[7], wurde ein Stichprobenfehler möglichst niedrig gehalten und der quantitative Schätzer bleibt möglichst erwartungstreu. Aufgrund der zugrunde gelegten Konfidenzintervalle und Fehlerspannen wurden allerdings Evaluierungen von Organisationen mit wenig Evaluierungen eher in die Stichprobe aufgenommen als Evaluierungen von Organisationen mit vielen Evaluierungen. Diese Verzerrung wurde über eine Gewichtung in der Auswertung entsprechend berücksichtigt.

*Spezifikationsfehler* bestehen, wenn das, was gemessen wurde, nicht dem entspricht, was gemessen werden sollte. Diese Fehlerquelle trat auf, wenn die Qualitätskriterien des Analyserasters nicht genau das maßen, was darunter verstanden wurde (zum Beispiel bestimmte Aspekte eines Qualitätsstandards). In der Meta-Evaluierung wurde diese Fehlerquelle reduziert, indem zum einen eine Pilotierung des Analyserasters in der Inter-Kodierenden-Phase umgesetzt wurde (das heißt die Inhalte der Qualitätskriterien wurden in den zu untersuchenden Dokumenten identifiziert) und zum anderen wurde durch die Durchführung einer Onlinebefragung gewährleistet, dass Qualitätskriterien, die nur eine eingeschränkte Ausprägung in den Dokumenten aufwiesen, durch eine ergänzende Methode aufgenommen und dadurch ebenfalls untersucht wurden.

*Beobachtungsfehler* sind Messfehler, Verarbeitungsfehler, menschliche Fehler bei der Dateneingabe oder technische Fehler. Da in der Meta-Evaluierung die Dokumente von mehreren Kodierenden analysiert wurden, hätten Messfehler unter anderem durch unterschiedliche Vorgehensweisen und individuelle punktuelle Fehler im Kodierungsprozess entstehen können. Der Beobachtungsfehler wurde daher durch einen möglichst präzisen Kodierungsleitfaden und die Sicherstellung einer zufriedenstellenden Inter-Kodierenden-Übereinstimmung reduziert.

*Nichtbeobachtungsfehler* entstehen, wenn die Auswahlgesamtheit nicht mit der Grundgesamtheit übereinstimmt. Ein Nichtbeobachtungsfehler im Sinne der *unit nonresponse* wurde in der Meta-Evaluierung als gering eingestuft, da auch unveröffentlichte Evaluierungen durch die Organisationen zur Verfügung gestellt und damit in die Grundgesamtheit aufgenommen wurden. Dennoch konnte nicht ausgeschlossen werden, dass einzelne Organisationen unbewusst oder aufgrund niedriger Dokumentationsgüte einzelne (Projekt-)Evaluierungen nicht zur Verfügung stellten. Da zum einen die Wahrscheinlichkeit dieser Vorkommnisse als niedrig und zum anderen auch als unsystematisch eingestuft wurde, war die Fehlerquelle als gering einzuschätzen.

## 3.2 Datenmanagement

**Im Rahmen des Datenmanagements wurden die Schritte Datenerhalt und Datenaufbereitung umgesetzt.** Die Evaluierungsdokumente wurden von den meisten Organisationen über die Plattform NextCloud zur Verfügung gestellt. Anschließend wurden die Dokumente auf ihre Vollständigkeit überprüft und aufbereitet. Konkret wurden die bis zu vier Dokumente einer Evaluierung in einer PDF-Datei zusammengefasst und eine Identifikationsnummer für jede Evaluierung vergeben (das heißt Evaluierungsbericht, [Berichts-]Anhänge, Leistungsbeschreibung, Inception Report). Insgesamt lagen ca. 1.000 Dokumente für 296 Evaluierungen vor). Parallel wurde eine Mastertabelle angelegt, die eine Übersicht über alle Evaluierungen, die jeweils vorliegenden Dokumenttypen sowie die Sprachen dieser Dokumente enthält. Tabelle 7 gibt einen Überblick über die Sprachen der Evaluierungen je Organisation.

[7] Im Falle der Vollerhebung der CARE-Evaluierungen ist eine Verzerrung der Werte zwischen Stichprobe und Grundgesamtheit auszuschließen.

**Tabelle 7 Informationen zu den Sprachen der Evaluierungsdokumente**

| Organisation | # Evaluierungen | # DE | # DE/EN | # EN | # FR | # SP |
|---|---|---|---|---|---|---|
| BGR | 18 | 8 | 9 | 0 | 1 | 0 |
| CARE | 6 | 0 | 0 | 4 | 0 | 2 |
| DRK | 17 | 0 | 1 | 15 | 1 | 0 |
| DVV | 17 | 0 | 2 | 10 | 4 | 1 |
| EWDE | 13 | 0 | 1 | 10 | 0 | 2 |
| GIZ | 38 | 1 | 18 | 16 | 2 | 1 |
| hbs | 19 | 13[a] | 5 | 1 | 0 | 0 |
| KAS | 17 | 15 | 2 | 0 | 0 | 0 |
| KfW | 68 | 50[b] | 16 | 2 | 0 | 0 |
| MISEREOR | 55 | 0 | 0 | 38 | 12 | 5 |
| PTB | 28 | 3 | 17 | 2 | 2 | 4 |
| **Gesamt** | **296** | **90** | **71** | **98** | **22** | **15** |

*Quelle: DEval, eigene Darstellung*
*Anmerkung: # = Anzahl; DE = Deutsch, EN = Englisch, FR = Französisch, SP = Spanisch. [a] Ein Anhang lag ausschließlich in Portugiesisch vor und wurde nicht kodiert. [b] Die Evaluierungsdokumente von zwei Evaluierungen waren zum Teil auf Deutsch und zum Teil auf Französisch.*

## 3.3 Kodierung und Inter- beziehungsweise Intra-Kodierenden-Übereinstimmung

**Im Anschluss an die Datenaufbereitung wurden die Prozessschritte Planung der Kodierung (das heißt Erwerb Softwarekenntnisse und Kenntnisse der Kodierregeln und des Kodierleitfadens), Kodierung der Qualitätskriterien sowie die Berechnung der Inter- beziehungsweise Intra-Kodierenden-Übereinstimmung durchgeführt.** Bevor mit der Kodierung begonnen wurde, wurde eine adäquate Nutzung der Software MAXQDA anhand einer Schulung durch einen externen Dienstleister gewährleistet.[8] Den Kodierenden wurden im Anschluss ein MAXQDA-Projekt zur Verfügung gestellt, in dem sowohl die Evaluierungsdokumente als auch ein Kodebaum (mit den zu kodierenden Qualitätskriterien, ihrer Beschreibung und Bewertungsstufen) hinterlegt waren. Anschließend wurden die Kodierenden vier Qualitätskriterien-Gruppen zugeordnet, in denen jeweils ausgewählte Qualitätskriterien geschult wurden. Um die Evaluierungen einheitlich entlang der Qualitätskriterien zu kodieren, wurden ein Dokument mit allgemeinen Kodierregeln und ein Kodierleitfaden verschriftlicht und vergemeinschaftet (zum Beispiel Umgang mit unsicheren Kodes, Hinweise zur Selektion

[8] Einem*r Kodierenden, der*die nicht an der Schulung teilnehmen konnte, wurden die Dokumentation der Schulung bereitgestellt, sodass er*sie die relevanten Informationen im Selbststudium und mit Rückfragen erarbeiten konnte.

und Klassifikation[9]). Im Anschluss an die Schulungen wurden 30 Evaluierungen (10 Prozent) der Evaluierungen der elf teilnehmenden Organisationen jeweils von allen Kodierenden für eine ausgewählte Qualitätskriterien-Gruppe kodiert.[10] Um eine einheitliche Bewertung der Qualitätskriterien zwischen den jeweiligen Kodierenden[11] zu gewährleisten, wurde eine Inter-Kodierenden-Übereinstimmung oder ein diskursiver Austausch zur Selektion und Klassifikation und/oder eine Intra-Kodierenden-Übereinstimmung durchgeführt. Im Folgenden wird das Vorgehen für die Evaluierungsfragen 2a, 2b, 2c und 3 dargestellt.

### Inter-Kodierenden-Übereinstimmung für die Untersuchung, inwieweit international geltende Qualitätsstandards angewandt wurden (Evaluierungsfrage 2a)

**Die durchschnittliche Inter-Kodiererenden-Übereinstimmung für die 22 über die Evaluierungsdokumente erfassten Qualitätskriterien lag bei einem guten Wert von 0,77 – der niedrigste Wert lag bei 0,63 und der höchste bei 0,97.** In den vier Gruppen wurden die Qualitätskriterien in Kodierschulungen auf Deutsch vorgestellt, unter den Kodierenden diskutiert und anhand ausgewählter Evaluierungen besprochen. Anschließend wurden die Qualitätskriterien überarbeitet und den Kodierenden im Nachgang erneut zur Verfügung gestellt. Aus den 296 Evaluierungen wurden zur Berechnung der Inter-Kodiererenden-Übereinstimmung 30 Evaluierungen (10 Prozent) geschichtet nach Organisation und randomisiert gezogen (Deutsch = 4; Englisch = 13; Deutsch/Englisch = 13), von allen Kodierenden einer Gruppe kodiert[12] und im Anschluss für alle Kodierenden der Koeffizient Krippendorffs Alpha[13] mit der Statistik-Software Stata berechnet.[14] Der Koeffizient von Krippendorffs Alpha kann zwischen 0,00 (fehlende Reliabilität) und 1,00 (perfekte Reliabilität) liegen (Werte von $1{,}00 \geq \alpha \geq 0{,}80$ sind als gut und Werte von $0{,}80 > \alpha \geq 0{,}67$ als akzeptabel anzusehen; Krippendorff, 2012, S. 241). Die durchschnittliche Inter-Kodierenden-Übereinstimmung für die Qualitätskriterien, die mit Krippendorffs Alpha berechnet wurden, lag bei 0,73 (niedrigster Wert: 0,63; höchster Wert: 0,89).[15] Die Höhe des Koeffizienten wurde unter anderem von der inhaltlichen Komplexität der Qualitätskriterien (je komplexer desto niedriger der Koeffizient), dem Skalenniveau (wenn ordinal anstatt binär dann ein niedrigerer Koeffizient) und der Anzahl der Kodierenden (je mehr Kodierende desto niedriger der Koeffizient) beeinflusst. Da diese Merkmale in der Meta-Evaluierung bei den meisten Qualitätskriterien allein oder in Kombination vorlagen, lagen die Werte häufig im akzeptablen Bereich. Bei insgesamt zehn Qualitätskriterien lag der Inter-

[9] Während bei der Selektionskodierung festgelegt wird, dass das Qualitätskriterium – vollständig beziehungsweise zu einem gewissen Teil – in einer Textstelle beschrieben wird, wird bei der Klassifikationskodierung festgelegt, welcher Wert der Textstelle (oder gegebenenfalls mehrere Textstellen) entlang der vorab festgelegten Bewertungsstufen/Skalenniveaus für das Qualitätskriterium zugeordnet wird. Da sich in der ersten Kodierphase gezeigt hat, dass Informationen zu Qualitätskriterien innerhalb einer Evaluierung wiederholt aufgeführt wurden und dieselbe Bewertung (Klassifikation) entlang unterschiedlicher Textstellen vorgenommen werden konnte, wurde die Übereinstimmung der Klassifikation und nicht zusätzlich die Selektion der Textstellen berechnet.

[10] In der Planung der Kodierungen wurden unterschiedliche Aspekte beachtet: späte finale Festlegung einiger Qualitätskriterien, unterschiedliche Verfügbarkeiten und Sprachkenntnisse der Kodierenden sowie unterschiedliche Schwierigkeitsgrade der Qualitätskriterien.

[11] Von diesen acht Personen unterstützte eine externe Person die Kodierung.

[12] Da 71 (24,0 Prozent) Evaluierungen zum Teil auf Deutsch und zum Teil auf Englisch, weitere 90 (30,4 Prozent) nur auf Deutsch und 98 (33,1 Prozent) nur auf Englisch vorlagen, wurde keine separate Berechnung der Inter-Kodierenden-Übereinstimmung für beide Sprachen vorgenommen. In allen Kodierschulungen wurden die Qualitätskriterien anhand von ausgewählten Evaluierungen beider Sprachen besprochen (alle Kodierenden sprachen Deutsch und Englisch). Die 22 französischsprachigen Evaluierungen wurden von einer und die 15 spanischsprachigen Evaluierungen von zwei Personen kodiert. Da sowohl für Französisch als auch für Spanisch weniger als 10 Prozent der Gesamtevaluierungen vorhanden waren, wurde für diese beiden Sprachen beziehungsweise für diese Evaluierungen keine Inter-Kodiererenden-Übereinstimmung berechnet. Unsicherheiten bei diesen Evaluierungen wurden zwischen den Kodierenden diskursiv gelöst beziehungsweise Fragen in der Kodierschulung thematisiert.

[13] Krippendorffs Alpha kann für eine beliebige Anzahl an Kodierenden berechnet und bei unterschiedlichen Skalenniveaus sowie fehlende Werten angewandt werden. Darüber hinaus berücksichtigt Krippendorffs Alpha die zufällige Übereinstimmung von Kodierungen und minimiert diesen Effekt (Nili et al., 2017).

[14] Qualitätskriterien, die in mehr als 24 Evaluierungen und bei mehr als neun Organisationen durchgehend nicht kodiert werden konnten, wurden mithilfe der Onlinebefragung untersucht (Kapitel 3.4).

[15] Bei der Berechnung der Inter-Kodierenden-Übereinstimmung mit Krippendorffs Alpha ist bei sechs Qualitätskriterien das sogenannte Alpha-Paradox aufgetreten (Gwet, 2015). Dies tritt auf, wenn entweder kaum oder keine Varianz bei den vergebenen Bewertungen vorliegt und (fast) ausschließlich eine Bewertungskategorie vergeben wurde. Für die entsprechenden Qualitätskriterien wurde daraufhin der Gwets AC berechnet (Gwet, 2015). Der Wert des Gwets AC lag bei diesen Qualitätskriterien durchschnittlich bei 0,89 (niedrigster Wert: 0,73; höchster Wert: 0,97).

Kodierenden-Übereinstimmungskoeffizient zunächst nicht über einem Cut-off-Wert von 0,67, sodass diese nachgeschult und nachkodiert wurden, bis ein Koeffizient von 0,67 oder höher erreicht wurde.[16] Tabelle 8 zeigt die wichtigsten Informationen zu den Qualitätskriterien und den Inter-Kodierenden-Übereinstimmungswerten. Neben einer Darstellung der erreichten Koeffizienten-Werte ist es auch sinnvoll, die Werte mit der Inter-Kodierenden-Übereinstimmung vergleichbarer Studien in Bezug zu setzen. In der organisationsübergreifenden Meta-Evaluierung von Noltze et al. (2018) wurde ein durchschnittlicher Koeffizient (Kappa-Wert nach Cohen) von 0,63 (= moderate Übereinstimmung) erreicht.

### Diskursiver Austausch und Intra-Kodierenden-Übereinstimmung für die Untersuchung, inwieweit organisationsspezifische Qualitätsstandards angewandt wurden (Evaluierungsfrage 2b)

**Organisationsspezifische Qualitätskriterien wurden im Rahmen eines diskursiven Austauschs hinsichtlich Selektion und Klassifikation besprochen. Zusätzlich wurde die Intra-Kodierenden-Übereinstimmung berechnet, sofern eine Kodierende die Evaluierungsdokumente überwiegend oder ausschließlich kodierte. Diese lag durchschnittlich bei 0,91 lag.** Insgesamt wurden für vier teilnehmende Organisationen elf Qualitätskriterien zur Untersuchung der Anwendung der organisationsspezifischen Qualitätsstandards kodiert (Tabelle 9). Für acht der elf Qualitätskriterien beziehungswiese drei der vier Organisationen haben zwei Kodierschulungen stattgefunden, im Rahmen derer mindestens 10 Prozent der Evaluierungen je Organisation hinsichtlich der Selektion und Klassifikation besprochen und Qualitätskriterien gegebenenfalls leicht angepasst wurden.[17] Die Qualitätskriterien einer Organisation wurden ausschließlich von einer Person kodiert, somit hat keine Schulung stattgefunden. Da alle organisationsspezifischen Qualitätskriterien – bis auf wenige französisch- und spanischsprachige Evaluierungen – ausschließlich von einer Person kodiert wurden, wurde die Selektion und Klassifikation bei mindestens 10 Prozent der Evaluierungen je Organisation nach einigen Wochen wiederholt. Im Anschluss wurde für jedes Qualitätskriterium die Intra-Kodierenden-Übereinstimmung berechnet, die durchschnittlich bei 0,91 lag mit dem niedrigsten Wert bei 0,68 und dem höchsten Wert bei 1,00.[18]

### Inter- und Intra-Kodierenden-Übereinstimmung für die Untersuchung, inwieweit die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit von 2018 in (Projekt-)Evaluierungen der KfW und GIZ in der vorliegenden Meta-Evaluierung angewandt wurden (Evaluierungsfrage 2c)

**Die durchschnittliche Inter-Kodierenden-Übereinstimmung für die sieben Qualitätskriterien lag bei einem Wert von 1,00.** Die Qualitätskriterien wurden in Kodierschulungen auf Deutsch vorgestellt, unter den beiden Kodierenden diskutiert und anhand ausgewählter Evaluierungen besprochen. Anschließend wurden die Qualitätskriterien überarbeitet und den Kodierenden im Nachgang erneut zur Verfügung gestellt. Aus den 106 Evaluierungen von KfW und GIZ (Deutsch = 51; Englisch = 18; Deutsch/Englisch = 34; Französisch = 2; Spanisch = 1) wurden zur Berechnung der Inter-Kodierenden-Übereinstimmung elf Evaluierungen (10 Prozent) geschichtet nach Organisation und randomisiert gezogen, von beiden Kodierenden kodiert und im Anschluss wurde der Koeffizient Krippendorffs Alpha mit der Statistik-Software Stata berechnet (Tabelle 10).

**Für die sieben Qualitätskriterien wurde zusätzlich eine Intra-Kodierenden-Übereinstimmung berechnet.** Da die sieben Qualitätskriterien für alle GIZ- und KfW-Evaluierungsberichte nach der Berechnung der Inter-

[16] Bei den Qualitätskriterien „Beschreibung des Erkenntnisinteresses“ und „Handlungsorientierung der Empfehlungen“ wurde ein Koeffizient von 0,64 beziehungsweise 0,63 erreicht. Dieser konnte im Rahmen von Nachschulungen nicht weiter verbessert werden, da die Kodierenden nicht mehr verfügbar waren. Da die Werte knapp unter dem genannten Cut-off-Wert liegen, wurden diese Qualitätskriterien dennoch analysiert.

[17] Die insgesamt 87 Evaluierungen wurden von drei Kodierenden kodiert. Zwei der Kodierenden haben jeweils (wenn vorhanden) die einzelnen spanisch- und französischsprachigen Evaluierungen kodiert, die dritte Person die deutsch- und englischsprachigen Evaluierungen. Die spanisch- und französischsprachigen Evaluierungen wurden immer im Rahmen der Kodierschulungen besprochen.

[18] Für die Berechnung der Intra-Kodierenden-Übereinstimmung wurde ebenfalls Krippendorffs Alpha herangezogen, der Koeffizient musste auch hier je Qualitätskriterien mindestens einen akzeptablen Wert erreichen (0,67). Da bei zwei Qualitätskriterien das Alpha-Paradox aufgetreten ist, wurde der Koeffizient Gwets AC berechnet. Alle berechneten Werte lagen über dem Cut-off-Wert, somit mussten keine Nachschulungen vorgenommen werden.

Kodierenden-Übereinstimmung von einer Person kodiert wurden, wurde für diese zusätzlich eine Intra-Kodierenden-Übereinstimmung für alle Qualitätskriterien berechnet. Die Intra-Kodierenden-Übereinstimmung wurde für eine weitere – nach Organisation geschichtete und randomisierte – gezogene Stichprobe (10 Prozent) der 106 Evaluierungen zwischen zwei Zeitpunkten für dieselbe kodierende Person berechnet. Die Intra-Kodierenden-Übereinstimmung lag für alle sieben Qualitätskriterien bei 1,0. Die hohen Werte der Inter-Kodierenden-Übereinstimmung für die Qualitätskriterien der Evaluierungsfrage 2c kamen aufgrund eines intensiven Austauschs zwischen dem Gutachtenden-Team der vorliegenden Meta-Evaluierung und einem Teammitglied der Meta-Evaluierung Nachhaltigkeit von 2018 zustande.

### Kodierung für die Untersuchung, inwieweit verschiedene Faktoren die Anwendung der Qualitätsstandards beeinflussten (Evaluierungsfrage 3)

**Informationen zu einigen Faktoren konnten über die Evaluierungsdokumente ermittelt werden. Da es sich dabei überwiegend um formale Qualitätskriterien handelte, wurde keine Inter- oder Intra-Kodierenden-Übereinstimmung berechnet", sondern Kodierungen stichprobenhaft überprüft. Für die multivariaten Regressionsanalysen wurden neben einer Datenanfrage bei den Organisationen und Informationen aus Sekundärdatenbanken einige Qualitätskriterien (zum Beispiel „Volumen der EZ-Maßnahme“ oder „Remote-Datenerhebung“) zusätzlich in den Evaluierungen kodiert.** Für die Einflussfaktoren „Informationsgehalt der Leistungsbeschreibung“ und „Einbindung der internen und externen Stakeholder*innen“ war die Inter-Kodierenden-Übereinstimmung über die Kodierung für Evaluierungsfrage 2a sichergestellt.

**Tabelle 8 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2a**

| Nr. QS | Nr. QK/Frage | Name Qualitätskriterium | Standardcluster | OB | IK | Skalenniveau | Gruppe (# Kodierende) | Nachschulung |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Beschreibung des Evaluierungsgegenstands | B & M | Nein | 0,81[a] | ordinal | Gruppe 1 (6) | Nein |
| 2 | / | Einbindung des Kontexts | B & M | / | / | / | / | / |
| 2a | 2 | Beschreibung des Kontexts der Entwicklungsmaßnahme | B & M | Nein | 0,81 | ordinal | Gruppe 4 (3) | Ja |
| 2b | 3 | Berücksichtigung des Kontexts bei Ergebnissen | B & M | Nein | 0,81 | ordinal | Gruppe 4 (3) | Ja |
| 3 | 4 | Darstellung der Wirkungszusammenhänge | B & M | Nein | 0,69 | ordinal | Gruppe 4 (3) | Ja |
| 4 | 5 | Beschreibung des Erkenntnisinteresses | B & M | Nein | 0,64 | ordinal | Gruppe 1 (6) | Nein |
| 5 | 6 | Nachvollziehbarkeit der Informationsquellen | B & M | Nein | 0,71 | ordinal | Gruppe 3 (3) | Nein |
| 6 | 7 | Darstellung der Angemessenheit des methodischen Vorgehens | B & M | Nein | 0,75 | ordinal | Gruppe 4 (3) | Ja |
| 7 | 8 | Qualitätssicherung mit Inception Report | B & M | Nein | / | binär | Gruppe 2 (5) | / |
| 8 | 9 | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen | B & M | Nein | 0,70 | ordinal | Gruppe 4 (3) | Ja |
| 9 | 10 | Informationsgehalt der Leistungsbeschreibung | B & M | Nein | 0,67 | ordinal | Gruppe 1 (6) | Ja |
| 10 | / | Zugänglichkeit | N | / | / | / | / | / |
| 10a | 11 | Veröffentlichung des Evaluierungsberichts | N | Ja | / | / | / | / |
| 10b | 12 | Veröffentlichung der Zusammenfassung | N | Ja | / | / | / | / |
| 10c | 13 | Zugänglichkeit für Stakeholder*innen | N | Nein | 0,89 | binär | Gruppe 4 (3) | / |
| 11 | 14 | Kompetenz der Gutachtenden | N | Ja | / | ordinal | Gruppe 1 (6) | / |
| 12 | 15 | Rechtzeitigkeit der Erkenntnisse | N | Ja | / | binär | Gruppe 2 (5) | / |
| 13 | / | Nützlichkeit der Empfehlungen | N | / | / | / | / | / |
| 13a | 16 | Adressat*innen der Empfehlungen | N | Nein | 0,68 | ordinal | Gruppe 3 (3) | Ja |
| 13b | 17 | Handlungsorientierung der Empfehlungen | N | Nein | 0,63 | ordinal | Gruppe 3 (3) | Ja |
| 14 | 18 | Evaluierungseffizienz | N | Ja | / | binär | Gruppe 4 (3) | / |
| 15 | 19 | Transparenz von Meinungsverschiedenheiten | P, U & F | Ja | / | binär | Gruppe 3 (3) | / |
| 16 | / | Einbindung der Stakeholder*innen | P, U & F | / | / | / | / | / |

| Nr. QS | Nr. QK/Frage | Name Qualitätskriterium | Standardcluster | OB | IK | Skalenniveau | Gruppe (# Kodierende) | Nachschulung |
|---|---|---|---|---|---|---|---|---|
| 16a | 20 | Einbindung der internen und externen Stakeholder*innen | P, U & F | Nein | 0,72 | ordinal | Gruppe 4 (3) | Ja |
| 16b | 21 | Einbindung der Partner*innen | P, U & F | Ja | / | / | / | / |
| 17 | 22 | Evaluierungsethik | P, U & F | Ja | / | binär | Gruppe 3 (3) | / |
| 18 | / | Unabhängigkeit der Gutachtenden | P, U & F | / | / | / | / | / |
| 18a | 23 | Darstellung organisationale Unabhängigkeit der Gutachtenden | P, U & F | Nein | 0,72 | binär | Gruppe 3 (3) | Nein |
| 18b | 24 | Darstellung Unvoreingenommenheit der Gutachtenden | P, U & F | Ja | / | binär | Gruppe 3 (3) | / |
| 19 | 25 | Informationsgehalt der Zusammenfassung[b] | B & M | Nein | 0,73 [a] | binär | Gruppe 2 (5) | Nein |
| 20 | 26 | Berücksichtigung Kapazitätsentwicklung[b] | P, U & F | Nein | 0,71 | ordinal | Gruppe 3 (3) | Ja |
| 21 | 27 | Berücksichtigung von Gemeinschaftsevaluierungen[b] | P, U & F | Ja | / | binär | Gruppe 3 (3) | / |
| 22 | 28 | Berücksichtigung partnerschaftlicher Ansätze[b] | P, U & F | Ja | / | binär | Gruppe 3 (3) | / |
| 23 | / | Zusammensetzung der Gutachtenden[b] | P, U & F | / | / | / | / | / |
| 23a | 29 | Genderbalance im Team[b] | P, U & F | Ja | / | binär | Gruppe 2 (5) | / |
| 23b | 30 | Einbezug von Gutachtenden aus Partnerland [b] | P, U & F | Nein | 0,82 | binär | Gruppe 4 (3) | Nein |
| 24 | 31 | Ausreichende Ressourcen vorhanden[b] | N | Ja | / | binär | Gruppe 4 (3) | / |
| 25 | 32 | Vorhandensein einer Management-Response[b] | N | Ja | / | binär | Gruppe 1 (6) | / |
| 26 | / | Anwendung der OECD-DAC-Kriterien[c] | eigener Bereich | / | / | / | / | / |
| 26a | 33 | Anwendung OECD-DAC-Kriterium – Relevanz[c] | eigener Bereich | Nein | 0,92[a] | binär | Gruppe 1 (6) | Nein |
| 26b | 34 | Anwendung OECD-DAC-Kriterium – Effektivität[c] | eigener Bereich | Nein | 0,93[a] | binär | Gruppe 1 (6) | Nein |
| 26c | 35 | Anwendung OECD-DAC-Kriterium – Effizienz[c] | eigener Bereich | Nein | 0,71 | binär | Gruppe 2 (5) | Nein |
| 26d | 36 | Anwendung OECD-DAC-Kriterium – Wirkungen[c] | eigener Bereich | Nein | 0,95[a] | binär | Gruppe 2 (5) | Nein |
| 26e | 37 | Anwendung OECD-DAC-Kriterium – Nachhaltigkeit[c] | eigener Bereich | Nein | 0,97[a] | binär | Gruppe 2 (5) | Nein |
| | | | | Durchschnitt | 0,77 | / | / | / |

*Quelle: DEval, eigene Darstellung*

*Anmerkung; QS = Qualitätsstandard; QK = Qualitätskriterium; OB = Onlinebefragung; IK = Inter-Kodierenden-Übereinstimmung; # = Anzahl; B & M = Berichtslegung und Methoden; P, U & F = Partizipation, Unabhängigkeit und Fairness; N = Nutzbarkeit; [a] Gwets AC, [b] OECD-DAC-only-Bereich; [c] OECD-DAC-Kriterien-Bereich*

**Tabelle 9 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2b**

| Name Qualitätskriterium | Beschreibung | SC | Intra-K | SN | # K |
|---|---|---|---|---|---|
| Mandat und Kapazitäten des*der Partners*in | Das Qualitätskriterium ist erfüllt, wenn in der Leistungsbeschreibung das Mandat und die Kapazitäten des*der Partners*in als wichtige Kontextinformation für die Durchführenden der Evaluierung beschrieben wird. | P, U & F | 0,90 | ordinal | 2 |
| Bezug zum Mandat des*der Partners*in bei den Evaluierungsergebnissen | Das Qualitätskriterium ist erfüllt, wenn bei den Evaluierungsergebnissen Bezug auf das Mandat des*der Partners*in genommen wird. | P, U & F | 1,00 | binär | 2 |
| Berücksichtigung des*der Partners*in bei den Empfehlungen | Das Qualitätskriterium ist erfüllt, wenn Empfehlungen gezielt an den*der Partner*in und an die Zusammenarbeit zwischen XXX und dem*der Partner*in gerichtet werden. | P, U & F | 1,00 | ordinal | 2 |
| Gender und Inklusion 1 (ToR) | Das Qualitätskriterium ist erfüllt, wenn in der Leistungsbeschreibung die Querschnittsthemen Gender und Inklusion angemessen angewandt werden. | P, U & F | 1,00 | ordinal | 2 |
| Gender und Inklusion 2 (Evaluierungsbericht) | Das Qualitätskriterium ist erfüllt, wenn die in der Leistungsbeschreibung thematisierten Kategorien zu den Querschnittsthemen Gender und Inklusion im Evaluierungsbericht aufgegriffen werden. | P, U & F | 1,00 | ordinal | 2 |
| Genderkompetenz und interkulturelle Kompetenz | Das Qualitätskriterium ist erfüllt, wenn in der Leistungsbeschreibung als Expertise für die Gutachtenden Genderkompetenz und interkulturelle Kompetenz genannt werden. | P, U & F | 1,00 | ordinal | 1 |
| Rahmenbedingungen der geschlechterpolitischen Realität | Das Qualitätskriterium ist erfüllt, wenn die politischen Rahmenbedingen mit Bezug zur Geschlechterdemokratie beschrieben werden. | P, U & F | 0,75 | ordinal | 1 |
| Wirkungen des Programms mit Bezug zur Geschlechterdemokratie | Das Qualitätskriterium ist erfüllt, wenn die Wirkungen des Programms mit Bezug zur Querschnittsaufgabe Geschlechterdemokratie beschrieben werden. | P, U & F | 1,00 | ordinal | 1 |
| Kontributionsanalyse Outcome-Ebene (GIZ) | Das Qualitätskriterium ist erfüllt, wenn es eine Kontributionsanalyse, also eine Analyse des Beitrags des Projekts zu Outcome anhand ausgewählter Wirkungshypothesen gibt. | B & M | 0,68 [a] | ordinal | 3 |
| Kontributionsanalyse Impact-Ebene (GIZ) | Das Qualitätskriterium ist erfüllt, wenn es eine Kontributionsanalyse, also eine Analyse des Beitrags des Projekts zu Impact an-hand ausgewählter Wirkungshypothesen gibt. | B & M | 0,68 [a] | ordinal | 3 |
| Effizienzanalyse (Follow-the-Money-Approach) (GIZ) | Das Qualitätskriterium (Follow-the-Money-Approach) ist erfüllt, wenn es eine Analyse der Kosten je Output gibt. | B & M | 1,00 | ordinal | 3 |
| | | Durchschnitt | 0,91 | / | / |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: B & M = Berichtslegung und Methoden; P, U & F = Partizipation, Unabhängigkeit und Fairness; ToR = Terms of References; SN = Skalenniveau; SC = Standardcluster; Intra-K = Intrakodierenden-Übereinstimmung; #K = Anzahl Kodierend;e [a] Gwets AC. Da kein Intra-Kodierenden-Wert unter 0,67 vorlag, wurden keine Nachschulungen vorgenommen. Die Zuordnung der Qualitätskriterien zu nichtstaatlichen Organisationen wird ebenso wie im Evaluierungsbericht mit Ausnahme der GIZ-Kriterien anonymisiert dargestellt.*

**Tabelle 10 Informationen zu den Qualitätskriterien für Evaluierungsfrage 2c**

| Nr. | Name QK (Nr. QK in Meta-Evaluierung Nachhaltigkeit) | Standardcluster | Inter-Kodierenden-Übereinstimmung | Intra-Kodierenden-Übereinstimmung | Skalenniveau | Nachschulung |
|---|---|---|---|---|---|---|
| 1 | Wirkungslogik überwiegend durch Indikatoren operationalisiert (4) | B & M | 1,00 | 1,00 | binär | Nein |
| 2 | Auswahlverfahren der Gesprächspartner beschrieben (8) | B & M | 1,00 | 1,00 | binär | Nein |
| 3 | Vorher-Nachher-Vergleich (9) | B & M | 1,00 | 1,00 | binär | Nein |
| 4 | Kontroll-/Vergleichsgruppe einbezogen (10) | B & M | 1,00 | 1,00 | binär | Nein |
| 5 | Kausalität über Plausibilitäten hergeleitet (11) | B & M | 1,00 | 1,00 | binär | Nein |
| 6 | Methoden-Triangulation angewandt (13) | B & M | 1,00 | 1,00 | binär | Nein |
| 7 | Datengrundlage ausreichend hinsichtlich Schlussfolgerungen (16) | B & M | 1,00 | 1,00 | binär | Nein |
| | | Durchschnitt | 1,00 | 1,00 | / | / |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: QK = Qualitätskriterium; B & M = Berichtslegung und Methoden. Die hohen Werte der Inter-Kodierenden-Übereinstimmung für die Qualitätskriterien der Evaluierungsfrage 2c kamen aufgrund eines intensiven Austauschs zwischen dem Gutachtenden-Team der vorliegenden Meta-Evaluierung und einem Teammitglied der Meta-Evaluierung Nachhaltigkeit von 2018 zustande.*

## 3.4 Datengrundlage und -analyse

### Datengrundlage

**Die ausgewählten 296 Evaluierungen schlossen unterschiedliche Jahre, Länder, Sektoren und Evaluierungstypen ein und wurden in vier verschiedenen Sprachen verfasst**. Die Länder mit der höchsten Anzahl an Evaluierungen waren Südafrika mit 14, Uganda mit elf, Philippinen mit acht und die Demokratische Republik Kongo, Indien und Tansania mit je sieben Evaluierungen. Die durchschnittliche Laufzeit der evaluierten EZ-Maßnahmen betrug 40 Monate.

**Die Onlinebefragung wurde durchgeführt, wenn für Qualitätskriterien keine ausreichenden Informationen in den Evaluierungsdokumenten verfügbar waren.** Sofern für ein Qualitätskriterium in der Inter-Kodierenden-Phase mehr als 24 Mal bei mindestens neun Organisationen durchgehend keine Textstellen kodiert werden konnten, galt das Qualitätskriterium als nicht ausreichend dargestellt. Da die niedrigste Anzahl an Evaluierungen für eine Organisation bei sechs lag, wurde durch den Schwellenwert größer 24 ausgeschlossen, dass durch eine Organisation ein Qualitätskriterium nicht in die Onlinebefragung aufgenommen wurde. Von den insgesamt 37 Qualitätskriterien in Evaluierungsfrage 2a wurden so 15 in die Onlinebefragung aufgenommen.[19] Qualitätskriterien, die per Definition in bestimmten Evaluierungsdokumenten dargestellt werden mussten (zum Beispiel die „Beschreibung des Evaluierungsgegenstands“), wurden nicht in die Onlinebefragung aufgenommen, wenn keine Informationen kodiert werden konnten. Das Qualitätskriterium galt somit als „nicht erfüllt“. In der Onlinebefragung schätzten die Verantwortlichen der Evaluierungseinheiten/-stellen ein, in wie vielen Evaluierungen die 15 Qualitätskriterien jeweils umgesetzt wurden, gaben Begründungen an, warum ein Qualitätskriterium gegebenenfalls nicht angewandt wurde und wiesen auf Organisationsdokumente hin, in denen die (Nicht-)Anwendung der Qualitätsstandards verschriftlicht wurde.

### Datenanalyse

**Die explorative Faktorenanalyse ist eine statistische Methode, mit der untersucht wird, ob bestimmte Qualitätskriterien empirisch zu latenten Faktoren (hier Standardclustern) zusammengefasst werden können** (vgl. Backhaus et al., 2015; Brown, 2006). Für die Faktorenanalyse wurde die Software Stata 15 genutzt. Bei den im Anschluss an die explorative Faktorenanalyse berechneten Faktorwerten gingen die Informationen der Qualitätskriterien gewichtet ein (abhängig von der Stärke ihres Zusammenhangs mit dem Standardcluster). Informationen der Qualitätskriterien, bei denen kein empirischer Zusammenhang mit dem Faktor (Standardcluster) bestand, wurden nicht in die Faktorwerte mit aufgenommen. Vor dem Hintergrund, dass die Standardwerke für Evaluierungsqualität des OECD DAC und der DeGEval weitere Teilaspekte abbildeten, die in der Analyse der Meta-Evaluierung nicht berücksichtigt wurden, wurden reflexive und nicht formative Faktoren der Berechnung zugrunde gelegt.

**In den Regressionsanalysen wurden verschiedene lineare und nicht lineare Modelle geschätzt, Standardfehler geclustert und gewichtet.** Insgesamt wurden 16 abhängige Variablen (14 Qualitätskriterien und zwei

[19] Von den 15 Qualitätskriterien wurden die Ergebnisse von 14 Qualitätskriterien hinsichtlich der Rückmeldung der Verantwortlichen aus den Evaluierungseinheiten/-stellen dargestellt. Das Qualitätskriterium „Qualitätssicherung mit Inception Report“ wurde abschließend nicht auf Basis der Rückmeldung aus der Onlinebefragung, sondern des Vorhandenseins eines Inception Reports auf Evaluierungsebene untersucht und bewertet.

Indices, 16 Modelle) geschätzt und je Modell bis zu 26 unabhängige Variablen (erklärende Faktoren) aufgenommen.[20] Basierend auf dem Skalenniveau der abhängigen Variablen (metrisch, binär und ordinal) wurden vier Regressionsmodelle der kleinsten Quadrate, drei logistische und neun geordnete logistische Regressionsmodelle gerechnet. In allen Regressionsmodellen wurden zudem Standardfehler auf der Organisationsebene geclustert[21], entsprechend der Stichprobenziehung (der Designgewichte) gewichtet[22] und der wirtschaftliche Entwicklungsstand des Landes der EZ-Maßnahme als Kontrollvariable aufgenommen[23]. Für die Analyse wurde die Software Stata 15 genutzt. Die folgende Tabelle 11 gibt einen Überblick über die Modelle und Informationen zu den abhängigen Variablen:

**Tabelle 11 Übersicht über die abhängigen Variablen (Qualitätskriterien/Indices)**

| Modell Nr. | abhängige Variable | Standardcluster | Skalenniveau |
|---|---|---|---|
| 1 | Beschreibung des Evaluierungsgegenstands | B & M | ordinal[a] |
| 2 | Einbindung des Kontexts | B & M | metrisch |
| 3 | Darstellung der Wirkungszusammenhänge | B & M | ordinal |
| 4 | Beschreibung des Erkenntnisinteresses | B & M | ordinal |
| 5 | Nachvollziehbarkeit der Informationsquellen | B & M | ordinal |
| 6 | Darstellung der Angemessenheit des methodischen Vorgehens | B & M | ordinal |
| 7 | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen | B & M | ordinal |
| 8 | Informationsgehalt der Leistungsbeschreibung | B & M | ordinal |
| 9 | Informationsgehalt der Zusammenfassung | B & M | binär[b] |
| 10 | Qualitätssicherung mit Inception Report | B & M | binär |

[20] Mithilfe der Software G*Power wurde sichergestellt, dass für die durchgeführten geordneten logistischen Regressionsmodelle und Regressionsmodelle der kleinsten Quadrate eine ausreichende Power von 0,80 oder höher vorlag. Darüber hinaus wurde für einen der imputierten Datensätze (siehe nächster Absatz) beziehungsweise für einen Teil der Regressionsmodelle (Nr. 1-8, 11, 12; alle als Regressionen der kleinsten Quadrate gerechnet) ein „Multiple Hypothesis Testing“ durchgeführt. Das „Multiple Hypothesis Testing“ wird angewandt, wenn eine größere Anzahl einzelner Hypothesen getestet wird und somit die Wahrscheinlichkeit höher ist, signifikante Effekte zu finden (McKenzie, 2020). Es ergab, dass die angepassten p-Werte für die zuvor signifikanten Effekte in den Regressionsmodellen (Tabellen 22, 23, 24) nicht mehr signifikant waren. Die Ergebnisse in Kapitel 4.3 des Berichts wurden vorsichtig interpretiert und entsprechend als Grundlage für Umsetzungshinweise und nicht für Empfehlungen herangezogen.

[21] Die Standardfehler wurden geclustert, da die Datenstruktur auf zwei Ebenen basierte: Die erste Ebene war die der Evaluierungen und die zweite die der Organisationen. Das Clustern der Standardfehler berücksichtigte, dass Evaluierungen über die Organisationen hinweg unabhängig voneinander waren, jedoch abhängig innerhalb der Organisation. Ein Mehrebenenmodell konnte nicht berechnet werden, da die Anzahl der zweiten Ebene (Organisation) zu klein war (Maas und Hox, 2005).

[22] Eine Gewichtung der Regressionsmodelle wurde vorgenommen, da die Stichprobe für die im Rahmen der Meta-Evaluierung untersuchten Evaluierungen nicht zufällig über alle Organisationen hinweg, sondern je Organisationen zufällig gezogen wurde. Über die Gewichtung wurde sichergestellt, dass von der untersuchten Stichprobe auf die Grundgesamtheit geschlossen werden kann. Der Gewichtungsfaktor wurde wie folgt berechnet: der Anteil an Evaluierungen der Organisation „a“ an der Grundgesamtheit der Evaluierungen aller Organisationen wurde geteilt durch den Anteil an Evaluierungen der Organisation „a“ an der Stichprobe der Evaluierungen aller Organisationen.

[23] Hierfür wurde auf Daten zum Bruttoinlandsprodukt des Jahres der Evaluierung im jeweiligen Partnerland zugrückgegriffen (Weltbank, 2022). In alle Regressionsmodelle, in denen der Faktor „Tagessatz für externe Gutachtende“ aufgenommen wurde, wurde zusätzlich eine Kontrollvariable aufgenommen, die beinhaltet, an welchen Evaluierungen ausschließlich interne Gutachtende beteiligt waren.

| Modell Nr. | abhängige Variable | Standardcluster | Skalenniveau |
|---|---|---|---|
| 11 | Summenindex Standardcluster Berichtslegung und Methoden | / | metrisch |
| 12 | Faktorindex Standardcluster Berichtslegung und Methoden | / | metrisch |
| 13 | Nützlichkeit der Empfehlungen | N | metrisch |
| 14 | Berücksichtigung Kapazitätsentwicklung | P, U & F | ordinal |
| 15 | Einbezug Gutachtenden aus Partnerland | P, U & F | binär |
| 16 | Einbindung der internen und externen Stakeholder*innen | P, U & F | ordinal |

*Quelle: DEval, eigene Darstellung*
*Anmerkung: [a] Bei einem ordinalen Skalenniveau liegen vier Abstufungen vor: 1 (= „nicht erfüllt"), 2 (= „eher nicht erfüllt), 3 (= „eher erfüllt") und 4 (= „vollständig erfüllt"), [b] bei einem binären Skalenniveau die Abstufungen 1 und 4. Die Qualitätskriterien aus dem Standardcluster „Berichtslegung und Methoden" wurden jeweils separat als abhängige Variablen in den Regressionsanalysen untersucht. Darüber hinaus wurde für sie ein Summen- und ein Faktorindex berechnet, die ebenfalls jeweils als abhängige Variablen untersucht wurden. Für die Qualitätskriterien aus den Standardclustern „Nutzbarkeit" und „Partizipation, Unabhängigkeit und Fairness" wurden keine Indices berechnet, da für viele der Qualitätskriterien keine Informationen auf Ebene der einzelnen Evaluierung vorhanden waren und Informationen über die Onlinebefragung erhoben wurden.*

**Da bei einigen erklärenden Faktoren fehlende Werte vorlagen, wurde zunächst untersucht, ob dies zufällige (*random*) fehlende Informationen darstellten und diese anschließend mithilfe einer multiplen Imputation ersetzt**. Bei zehn erklärenden Faktoren (beziehungsweise Kontrollvariablen) lagen fehlende Werte vor[24]. Um so viele Evaluierungen wie möglich in die Regressionenanalysen aufnehmen zu können, wurde zunächst mithilfe des „Little Test" untersucht, ob die fehlenden Werte vollständig zufällig fehlten (*missing completely at random*, MCAR) (Li, 2013) und falls ja anschließend die fehlenden Werte mithilfe einer „Mulitvariate Imputation by Chained Equations" Algorithmus (MICE) ersetzt.[25]

## 3.5 Bewertung

Über die Website der DeGEval wurde überprüft, ob die teilnehmenden Organisationen im Untersuchungszeitraum DeGEval-Mitglieder waren. Seit 2009 verpflichtet sich eine Organisation mit der Unterschrift der Beitrittserklärung zur Orientierung an den DeGEval-Standards. Zur aktuellen, revidierten Fassung der DeGEval-Standards verpflichteten sich allerdings nur die Organisationen, die seit deren Verabschiedung am 21. September 2016 die (aktuelle) Beitrittserklärung unterzeichneten (DeGEval, 2021). Für teilnehmende Organisationen, die vor 2016 (beziehungsweise vor 2009) der DeGEval beigetreten sind, wurde eine Verpflichtung zu den DeGEval-Standards über die Organisationsdokumente und/oder in Gesprächen mit den Verantwortlichen ermittelt.

[24] Für den Faktor „Konflikt" fehlten Werte bei 26 Evaluierungen; Sozialkapital: 157; Tagessatz externe Gutachtende: 31; Remote-Datenerhebung: 7; Gutachtenden-Tage im Verhältnis zu den Kosten der EZ-Maßnahme: 30; Evaluierungseinheiten/-stellen absolut: 12; Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen: 25; Evaluierungseinheiten/-stellen im Verhältnis zur Größe der Organisation: 18; Evaluierungstätigkeit: 19 und wirtschaftlicher Entwicklungsstand des Partnerlandes der EZ-Maßnahme: 33.

[25] Der Little Test wurde für jede Organisation mit einer ausreichend großen Anzahl an Evaluierungen separat durchgeführt. Da dies nicht alle Organisationen eingeschlossen hat, wurde der Test weiterhin über alle Evaluierungen aller Organisationen hinweg berechnet.

# 4. ERGÄNZENDE ERGEBNISSE

## 4.1 Anwendung der Qualitätsstandards

### 4.1.1 Anwendung der OECD-DAC- und der DeGEval-Standards

Tabelle 12 zeigt die Lang- und Kurzformen der Namen der untersuchten Qualitätskriterien und Faktoren. Tabelle 13 zeigt eine Übersicht über die Ergebnisse dieser Qualitätsstandards/-kriterien.

**Tabelle 12 Übersicht über die Lang- und Kurzformen der Namen der Qualitätskriterien und Faktoren**

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| **Evaluierungsfrage 2a: Anwendung OECD-DAC- und DeGEval-Qualitätskriterien/-standards** | | | | |
| 1 | 1 | Beschreibung des Evaluierungsgegenstands | Beschreibung Evaluierungsgegenstand | Evaluierungsgegenstand |
| 2 | / | Einbindung des Kontexts | Einbindung Kontext | Kontext |
| 2a | 2 | Beschreibung des Kontexts der Entwicklungsmaßnahme | Beschreibung Kontext EZ-Maßnahme | Kontext EZ-Maßnahme |
| 2b | 3 | Berücksichtigung des Kontexts bei Ergebnissen | Berücksichtigung Kontext Ergebnisse | Kontext Ergebnisse |
| 3 | 4 | Darstellung der Wirkungszusammenhänge | Darstellung Wirkungszusammenhänge | Wirkungszusammenhänge |
| 4 | 5 | Beschreibung des Erkenntnisinteresses | Beschreibung Erkenntnisinteresse | Erkenntnisinteresse |
| 5 | 6 | Nachvollziehbarkeit der Informationsquellen | Nachvollziehbarkeit Informationsquellen | Informationsquellen |
| 6 | 7 | Darstellung der Angemessenheit des methodischen Vorgehens | Darstellung methodisches Vorgehen | Darstellung Methodik |
| 7 | 8 | Qualitätssicherung mit Inception Report | Qualitätssicherung IR | Vorhandensein IR |
| 8 | 9 | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen | Kohärenz Schlussfolgerungen | Kohärenz |
| 9 | 10 | Informationsgehalt der Leistungsbeschreibung | Informationsgehalt Leistungsbeschreibung | Leistungsbeschreibung |
| 10 | / | Zugänglichkeit | Zugänglichkeit | / |
| 10a | 11 | Veröffentlichung des Evaluierungsberichts | Veröffentlichung Evaluierungsbericht | Veröffentlichung Bericht |
| 10b | 12 | Veröffentlichung der Zusammenfassung | Veröffentlichung Zusammenfassung | / |
| 10c | 13 | Zugänglichkeit für Stakeholder*innen | Zugänglichkeit Stakeholder*innen | / |
| 11 | 14 | Kompetenz der Gutachtenden | Kompetenz Gutachtende | Kompetenz |

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| 12 | 15 | Rechtzeitigkeit der Erkenntnisse | Rechtzeitigkeit Erkenntnisse | Rechtzeitigkeit |
| 13 | / | Nützlichkeit der Empfehlungen | Nützlichkeit Empfehlungen | Empfehlungen |
| 13a | 16 | Adressat*innen der Empfehlungen | Adressat*innen Empfehlungen | / |
| 13b | 17 | Handlungsorientierung der Empfehlungen | Handlungsorientierung Empfehlungen | Handlung Empfehlungen |
| 14 | 18 | Evaluierungseffizienz | / | / |
| 15 | 19 | Transparenz von Meinungsverschiedenheiten | Transparenz Meinungsverschiedenheiten | Meinungsverschiedenheiten |
| 16 | / | Einbindung der Stakeholder*innen | Einbindung Stakeholder*innen | / |
| 16a | 20 | Einbindung der internen und externen Stakeholder*innen | Einbindung Stakeholder*innen (intern/extern) | / |
| 16b | 21 | Einbindung der Partner*innen | Einbindung Partner*innen | / |
| 17 | 22 | Evaluierungsethik | Ethik | / |
| 18 | / | Unabhängigkeit der Gutachtenden | Unabhängigkeit GA | Unabhängigkeit |
| 18a | 23 | Darstellung organisationale Unabhängigkeit der Gutachten-den | Darstellung organisationale Unabhängigkeit GA | Darstellung Unabhängigkeit Organisa-tion |
| 18b | 24 | Darstellung Unvoreingenommenheit der Gutachtenden | Darstellung Unvoreingenommenheit GA | Darstellung Unvoreingenommenheit |
| 19 | 25 | Informationsgehalt der Zusammenfassung | Informationsgehalt Zusammenfassung | Zusammenfassung |
| 20 | 26 | Berücksichtigung Kapazitätsentwicklung | Kapazitätsentwicklung | / |
| 21 | 27 | Berücksichtigung von Gemeinschaftsevaluierungen | Berücksichtigung Gemeinschaftsevaluierungen | Gemeinschaftsevaluierung |
| 22 | 28 | Berücksichtigung partnerschaftliche Ansätze | Partnerschaftliche Ansätze | / |
| 23 | / | Zusammensetzung der Gutachtenden | Zusammensetzung Gutachtende | Zusammensetzung GA |

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| 23a | 29 | Genderbalance im Team | Genderbalance Team | Genderbalance |
| 23b | 30 | Einbezug von Gutachtenden aus Partnerland | Einbezug Gutachtende Partnerland | GA Partnerland |
| 24 | 31 | Ausreichende Ressourcen vorhanden | Ausreichende Ressourcen | Ressourcen |
| 25 | 32 | Vorhandensein einer Management Response | Vorhandensein Management-Response | Management-Response |
| 26 | / | Anwendung der OECD-DAC-Kriterien | Anwendung OECD-DAC-Kriterien | OECD-DAC-Kriterien |
| 26a | 33 | Anwendung OECD-DAC-Kriterium – Relevanz | Relevanz-Kriterium | / |
| 26b | 34 | Anwendung OECD-DAC-Kriterium – Effektivität | Effektivitäts-Kriterium | / |
| 26c | 35 | Anwendung OECD-DAC-Kriterium – Effizienz | Effizienz-Kriterium | / |
| 26d | 36 | Anwendung OECD-DAC-Kriterium – Wirkungen | Wirkungen-Kriterium | / |
| 26e | 37 | Anwendung OECD-DAC-Kriterium – Nachhaltigkeit | Nachhaltigkeits-Kriterium | / |
| **Evaluierungsfrage 2b: Anwendung organisationsspezifischer Qualitätskriterien** | | | | |
| / | / | Mandat und Kapazitäten des*der Partners*in | / | / |
| / | / | Bezug zum Mandat des*der Partners*in bei den Evaluierungsergebnissen | / | / |
| / | / | Berücksichtigung des*der Partners*in bei den Empfehlungen | / | / |
| / | / | Gender und Inklusion 1 (ToRs) | / | / |
| / | / | Gender und Inklusion 1 (Evaluierungsbericht) | / | / |
| / | / | Genderkompetenz und interkulturelle Kompetenz | / | / |
| / | / | Rahmenbedingungen der geschlechterpolitischen Realität | / | / |

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| / | / | Wirkungen des Programms mit Bezug zur Geschlechterdemokratie | / | / |
| / | / | Kontributionsanalyse Outcome-Ebene | / | / |
| / | / | Kontributionsanalyse Impact-Ebene | / | / |
| / | / | Effizienzanalyse (Follow-the-Money-Approach) | / | / |
| **Evaluierungsfrage 2c: Anwendung Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit** | | | | |
| / | 1 | Gegenstand (Vorhaben) beschrieben | Gegenstand beschrieben | / |
| / | 2 | Erkenntnisinteresse formuliert/ operationalisiert | Erkenntnisinteresse formuliert | / |
| / | 3 | Wirkungslogik/ Wirkungskette dargestellt | Wirkungslogik dargestellt | / |
| / | 4 | Wirkungslogik überwiegend durch Indikatoren operationalisiert | Indikatoren verwendet | / |
| / | 5 | Methodisches Vorgehen beschrieben | / | / |
| / | 7 | Befragte Gesprächspartner identifiziert | Befragte Stakeholder*innen identifiziert | / |
| / | 8 | Auswahlverfahren der Gesprächspartner beschrieben | Auswahlverfahren beschrieben | / |
| / | 6 | Stärken und/oder Limitationen des methodischen Vorgehens identifiziert | Stärken/Limitationen diskutiert | / |
| / | 9 | Vorher-Nachher-Vergleich | / | / |
| / | 10 | Kontroll-/ Vergleichsgruppe einbezogen | Kontroll-/Vergleichsgruppen | / |
| / | 11 | Kausalität über Plausibilitäten hergeleitet | Kausalität über Plausibilitäten | / |
| / | 12 | Daten-Triangulation | Daten-Triangulation | / |
| / | 13 | Methoden-Triangulation angewandt | Methoden-Triangulation | / |

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| / | 14 | Schlussfolgerungen durch Daten überwiegend referenziert | Schlussfolgerungen referenziert | / |
| / | 15 | Schlussfolgerungen aus Daten überwiegend plausibel begründet | Schlussfolgerungen plausibel | / |
| / | 16 | Datengrundlage ausreichend hinsichtlich Schlussfolgerungen | Datengrundlage ausreichend | / |
| **Evaluierungsfrage 3: Zusammenhänge Faktoren und Anwendung der Qualitätskriterien** | | | | |
| / | / | **Faktor** | **Proxy** | **Kurzname** |
| / | / | Fragilität | Konflikt | / |
| / | / | Kultureller Kontext | Sozialkapital-Index | Sozialkapital |
| / | / | Pandemie | Jahr 2020 | / |
| / | / | Kompetenz Gutachtende | Anzahl interne und externe Gutachtende | Anzahl GA |
| | | | Tagessatz externe Gutachtende | Tagessatz GA |
| / | / | Qualitätssicherungsprozesse | Informationsgehalt LB | Leistungsbeschreibung |
| | | | Qualitätssicherung mit Inception Report | Vorhandensein IR |
| | | | Einbindung der internen und externen Stakeholder*innen | Einbindung Stakeholder*innen |
| / | / | Remote-Datenerhebung | / | Remote |
| / | / | Kosten der Evaluierung | Gutachtenden-Tage im Verhältnis zu Kosten der EZ-Maßnahme | GA-Tage/Kosten Maßnahme |
| / | / | Jahr der Evaluierung | / | Jahr |
| / | / | Strukturierter Planungsprozess | / | / |
| / | / | Evaluierungs- und Lernkultur | Organisation | / |

| Nr. QS | Nr. QK | Langform Name Qualitätskriterium | Kurzform Name Qualitätskriterium | Kurzform Name Grafik |
|---|---|---|---|---|
| / | / | Evaluierungseinheit vorhanden | / | / |
| / | / | Größe der Evaluierungseinheiten/-stellen | Anzahl Vollzeitäquivalente | Größe Evaluierungseinheit |
| / | / | Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen | Anzahl der Vollzeitäquivalente in der Evaluierungseinheit im Verhältnis zur Evaluierungsanzahl | Verhältnis Evaluierungseinheit Evaluierungen |
| / | / | Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Größe der Organisation | Anzahl der Vollzeitäquivalente in der Evaluierungseinheit im Verhältnis zu Vollzeitäquivalenten der Organisation | Verhältnis Evaluierungseinheit Organisation |
| / | / | Evaluierungstätigkeit | / | / |

*Quelle: DEval, eigene Darstellung*

**Tabelle 13 Übersicht über die Ergebnisse der OECD-DAC- und DeGEval-Qualitätsstandards/-kriterien**

| Nr. QS | Nr. QK | Name QK | SC | OB | MW G1/G2 | Median G1/G2 | Min/Max G1 | Differenz (> 50 %) G1 | G2 > G1/ G1 > G2 (> 20 %) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Beschreibung des Evaluierungsgegenstands | B & M | Nein | 95,9/98,9 | 100,0/100,0 | 88,2/100,0 | Nein | Nein |
| 2 | / | Einbindung des Kontexts | B & M | / | 70,4/46,7 | / | 51,0/79,4 | Nein | G1 > G2 |
| 2a | 2 | Beschreibung des Kontexts der Entwicklungsmaßnahme | B & M | Nein | 83,4/50,3 | 100,0/50,0 | 45,1/96,6 | Ja | G1 > G2 |
| 2b | 3 | Berücksichtigung des Kontexts bei Ergebnissen | B & M | Nein | 57,4/43,1 | 66,7/33,3 | 40,7/77,2 | Nein | Nein |
| 3 | 4 | Darstellung der Wirkungszusammenhänge | B & M | Nein | 43,5/31,2 | 33,3/33,3 | 23,5/86,8 | Ja | Nein |
| 4 | 5 | Beschreibung des Erkenntnisinteresses | B & M | Nein | 74,8/81,5 | 100,0/83,3 | 42,6/99,1 | Ja | Nein |
| 5 | 6 | Nachvollziehbarkeit der Informationsquellen | B & M | Nein | 83,8/80,7 | 100,0/91,7 | 77,0/98,2 | Nein | Nein |
| 6 | 7 | Darstellung der Angemessenheit des methodischen Vorgehens | B & M | Nein | 34,6/35,3 | 33,3/33,3 | 1,9/65,8 | Ja | Nein |
| 7 | 8 | Qualitätssicherung mit Inception Report | B & M | Nein | 63,7/20,4 | 100,0/0,0 | 0,0/100,0 | Ja | G1 > G2 |
| 8 | 9 | Kohärenz von Daten-Ergebnissen-Schlussfolgerungen | B & M | Nein | 61,2/39,3 | 66,7/25,0 | 35,3/95,6 | Ja | G1 > G2 |
| 9 | 10 | Informationsgehalt der Leistungsbeschreibung | B & M | Nein | 53,0/69,6 | 66,7/83,3 | 3,9/82,4 | Ja | Nein |
| 10 | / | Zugänglichkeit | N | / | 72,1/68,2 | / | 36,8/100,0 | Ja | Nein |
| 10a | 11 | Veröffentlichung des Evaluierungsberichts | N | Ja | 44,4/75,0 | 0,0/75,0 | 0,0/100,0 | Ja | G2 > G1 |
| 10b | 12 | Veröffentlichung der Zusammenfassung | N | Ja | 75,0/50,0 | 100,0/50,0 | 0,0/100,0 | Ja | G1 > G2 |

| Nr. QS | Nr. QK | Name QK | SC | OB | MW G1/G2 | Median G1/G2 | Min/Max G1 | Differenz (> 50 %) G1 | G2 > G1/ G1 > G2 (> 20 %) |
|---|---|---|---|---|---|---|---|---|---|
| 10c | 13 | Zugänglichkeit für Stakeholder*innen | N | Nein | 84,4/73,8 | 100,0/75,0 | 23,5/100,0 | Ja | Nein |
| 11 | 14 | Kompetenz der Gutachtenden | N | Ja | 83,3/75,0 | 75,0/75,0 | 75,0/100,0 | Nein | Nein |
| 12 | 15 | Rechtzeitigkeit der Erkenntnisse | N | Ja | 72,2/75,0 | 75,0/75,0 | 50,0/75,0 | Nein | Nein |
| 13 | 16 | Nützlichkeit der Empfehlungen | N | / | 58,7/57,1 | / | 35,8/82,7 | Nein | Nein |
| 13a | 16 | Adressat*innen der Empfehlungen | N | Nein | 50,6/47,4 | 66,7/25,0 | 17,6/98,8 | Ja | Nein |
| 13b | 17 | Handlungsorientierung der Empfehlungen | N | Nein | 66,9/66,9 | 66,7/83,3 | 53,9/87,2 | Nein | Nein |
| 14 | 18 | Evaluierungseffizienz | N | Ja | 90,6/50,0 | 100,0/50,0 | 50,0/100,0 | Nein | G1 > G2 |
| 15 | 19 | Transparenz von Meinungsverschiedenheiten | P, U & F | Ja | 67,9/75,0 | 75,0/75,0 | 0,0/100,0 | Ja | Nein |
| 16 | / | Einbindung der Stakeholder*innen | P, U & F | / | 58,2/77,3 | / | 11,8/98,2 | Ja | Nein |
| 16a | 20 | Einbindung der internen und externen Stakeholder*innen | P, U & F | Nein | 58,2/77,3 | 66,7/58,3 | 11,8/98,2 | Ja | Nein |
| 16b | 21 | Einbindung der Partner*innen | P, U & F | Ja | 72,2/100,0 | 100,0/100,0 | 0,0/100,0 | Ja | G2 > G1 |
| 17 | 22 | Evaluierungsethik | P, U & F | Ja | 95,8/87,5 | 100,0/87,5 | 75,0/100,0 | Nein | Nein |
| 18 | / | Unabhängigkeit der Gutachtenden | P, U & F | / | 46,7/- | / | 14,3/91,2 | Ja | / |
| 18a | 23 | Darstellung organisationale Unabhängigkeit der Gutachtenden | P, U & F | Nein | 53,5/17,2 | 100,0/0,0 | 21,4/86,8 | Ja | G1 > G2 |

| Nr. QS | Nr. QK | Name QK | SC | OB | MW G1/G2 | Median G1/G2 | Min/Max G1 | Differenz (> 50 %) G1 | G2 > G1/ G1 > G2 (> 20 %) |
|---|---|---|---|---|---|---|---|---|---|
| 18b | 24 | Darstellung Unvoreingenommenheit der Gutachtenden | P, U & F | Ja | 33,3/- | 0,0/- | 0,0/100,0 | Ja | / |
| 19 | 25 | Informationsgehalt der Zusammenfassung[a] | B & M | Nein | 71,1/88,4 | 100,0/100,0 | 44,1/100,0 | Ja | Nein |
| 20 | 26 | Berücksichtigung Kapazitätsentwicklung[a] | P, U & F | Nein | 22,5/62,3 | 33,3/66,7 | 3,9/75,4 | Ja | G2 > G1 |
| 21 | 27 | Berücksichtigung von Gemeinschaftsevaluierungen[a] | P, U & F | Ja | 41,7/45,0 | 25,0/50,0 | 0,0/100,0 | Ja | Nein |
| 22 | 28 | Berücksichtigung partnerschaftlicher Ansätze[a] | P, U & F | Ja | 33,3/50,0 | 37,5/50,0 | 0,0/75,0 | Ja | Nein |
| 23 | / | Zusammensetzung der Gutachtenden[a] | P, U & F | / | 34,9/61,1 | / | 20,4/60,5 | Nein | / |
| 23a | 29 | Genderbalance im Team[a] | P, U & F | Ja | 70,0/56,3 | 75,0/50,0 | 50,0 / 100,0 | Nein | Nein |
| 23b | 30 | Einbezug von Gutachtenden aus Partnerland[a] | P, U & F | Nein | 17,4/63,5 | 0,0/100,0 | 3,6/65,8 | Ja | G2 > G1 |
| 24 | 31 | Ausreichende Ressourcen vorhanden[a] | N | Ja | 75,0/65,0 | 75,0/75,0 | 50,0/100,0 | Nein | Nein |
| 25 | 32 | Vorhandensein einer Management Response[a] | N | Ja | 79,2/45,0 | 87,5/50,0 | 50,0/100,0 | Nein | G1 > G2 |
| 26 | / | Anwendung OECD-DAC-Kriterien[b] | eigener Bereich | / | 95,3 | / | 78,9/100,0 | Nein | / |
| 26a | 33 | Anwendung OECD-DAC-Kriterium – Relevanz[b] | eigener Bereich | Nein | 95,9 | 100,0 | 82,4/100,0 | Nein | / |
| 26b | 34 | Anwendung OECD-DAC-Kriterium – Effektivität[b] | eigener Bereich | Nein | 96,4 | 100,0 | 83,3/100,0 | Nein | / |

| Nr. QS | Nr. QK | Name QK | SC | OB | MW G1/G2 | Median G1/G2 | Min/Max G1 | Differenz (> 50 %) G1 | G2 > G1/ G1 > G2 (> 20 %) |
|---|---|---|---|---|---|---|---|---|---|
| 26c | 35 | Anwendung OECD-DAC-Kriterium – Effizienz[b] | eigener Bereich | Nein | 90,1 | 100,0 | 42,1/100,0 | Ja | / |
| 26d | 36 | Anwendung OECD-DAC-Kriterium – Wirkungen[c] | eigener Bereich | Nein | 97,6 | 100,0 | 88,2/100,0 | Nein | / |
| 26e | 37 | Anwendung OECD-DAC-Kriterium – Nachhaltigkeit[b] | eigener Bereich | Nein | 96,4 | 100,0 | 78,9/100,0 | Nein | / |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: [a] OECD-DAC-only-Bereich, [b] OECD-DAC-Kriterien-Bereich; MW = Mittelwert; B & M = Berichtslegung und Methoden; P, U & F = Partizipation, Unabhängigkeit und Fairness; N = Nutzbarkeit; QS = Qualitätsstandard; QK = Qualitätskriterium; SC = Standardcluster; OB = Online-Befragung; Min/Max = Minimal-/Maximalwert; G = Gruppe*

### 4.1.2 Good Practices zu ausgewählten OECD-DAC- und DeGEval-Qualitätskriterien

#### *Good Practice „Beschreibung des Evaluierungsgegenstands“ (EWDE)*

*Das Qualitätskriterium ist erfüllt, wenn 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteur*innen (politische Partner und/oder Träger) der Entwicklungsmaßnahme genannt sind.*

Ziele: „A range of subjects are covered through these initiatives, including peace building and trauma healing, inter-religious dialogue, strengthening community structures, health infrastructure, adult literacy, agricultural extension in the context of food security as well as income-generating activities.“

Zielgruppen: „[E]ach addresses more than 35 dispersed communities, mostly poor, rural and disadvantaged people.“

Relevante Akteur*innen: „Partners include the Church of Christ in Nation (COCIN), the Evangelical Church Winning All (ECWA), the Church of the Brethren in Nigeria (Ekklesiyar Yan’uwa a Nigeria – EYN) and the Justice Peace and Reconciliation Movement (JPRM).“

#### *Good Practice „Beschreibung des Erkenntnisinteresses“ (MISEREOR)*

*Das Qualitätskriterium ist erfüllt, wenn 1. Zweck, 2. Ziel(e) und 3. Evaluierungsfrage(n) der Evaluierung genannt werden.*

Zweck: „The objectives of the evaluation, in line with the MISEREOR Evaluation Guidelines, have been twofold: a) Learning and quality development b) Performance assessment and accountability.“

Ziel: „The evaluation objectives follow the OECD/DAC dimensions for a project evaluation: evaluating the design and relevance of the project, effectiveness, efficiency, outcomes and impact and sustainability. Also, to be evaluated were the appropriateness of the internal systems of the SDMIC.“

Evaluierungsfrage: „To what extent do the interventions of the project comply with the needs and interests of the target group?“

#### *Good Practice „Beschreibung des Kontexts der Entwicklungsmaßnahme“ (KfW)*

*Das Qualitätskriterium ist erfüllt, wenn der Kontext der EZ-Maßnahme beschrieben wird (zum Beispiel politische, ökonomische, soziale, [agrar-]ökologische, [agrar-]klimatische, gesetzliche, technologische, historische, kulturelle Kontextfaktoren).*

„Mit rund 57 Mio. Einwohner und im Bevölkerungswachstum von 3,1 % im Jahr 2017 gehört Tansania zu den größten und gleichzeitig zu den am stärksten wachsenden Ländern Sub-sahara-Afrikas [sic]. Seit 1990 hat sich die Zahl der Einwohner Tansanias verdoppelt. Knapp 45 % der Einwohner sind jünger als 15 Jahre. Zwei Drittel der Tansanier leben noch immer in ländlichen Gebieten, wenn gleich [sic] die Bevölkerung in Städten mit 5,3 % überdurchschnittlich stark wächst. […] Auch im Gesundheitssektor hat es in den letzten Jahren Fortschritte gegeben. So stieg die Lebenserwartung zwischen 1990 und 2016 von 50 auf 6 Jahre an. Die HIV-Prävalenz und die Malaria-Inzidenz sind landesweit gesunken. […] Qualitätskriterien wie die Mütter- oder die Neugeborenensterblichkeit haben sich hingegen kaum verbessert. Sozial benachteiligte Bevölkerungsgruppen haben nach wie vor […] schlechter Zugang zu medizinischen Diensten. Die Qualität der Gesundheitsdienstleistungen ist weiterhin unzureichend. Insbesondere in abgelegenen ländlichen Regionen mangelt es in den Gesundheitseinrichtungen an notwendigen Medikamenten und qualifiziertem medizinischen Personal.“

### *Good Practice „Berücksichtigung des Kontexts bei den Ergebnissen“ (PTB)*

*Das Qualitätskriterium ist erfüllt, wenn der Einfluss des Kontexts auf die Evaluierungsergebnisse berücksichtigt wird.*

„Several factors have challenged project implementation, some of them do still persist: an unstable economic and political situation that caused uncertainties, financial cutbacks, staff and management fluctuation, rotation of political offices, rivalries between institutions; controversial debates (e.g. on standards and technical regulations, on European versus US American OSH standards, on accreditation and QM procedures) due to different cultural backgrounds and concepts that restrained a common understanding amongst the project partners and slowed down implementation.“

### *Good Practice „Kohärenz Daten-Ergebnisse-Schlussfolgerung“ (politische Stiftung)*

*Das Qualitätskriterium ist erfüllt, wenn die Datenanalyse, die Ergebnisse und die Schlussfolgerungen logisch miteinander in Bezug gesetzt werden beziehungsweise aufeinander aufbauen.*

Datenanalyse: „Die Untersuchung der Leistungserbringung des Länderprojekts basiert auf der Auswertung der Jahresplanungen des Länderprojekts sowie der Interviews mit Mitarbeitenden des Länderprojekts.“

Ergebnis: „Während die Wirkung dieser Trainings auf Basis der Datenlage heute nicht mehr messbar ist, ist es auf Basis der geführten Interviews plausibel, dass diese zu einer Stärkung der Fähigkeiten […] beigetragen haben.“

Schlussfolgerung: „Die Analyse der erbrachten Leistungen und kurz- und mittelfristigen Wirkungen veranschaulicht jedoch, dass eine systematische Umsteuerung ratsam erscheint, um […] wirksamer zu gestalten. Das Länderprojekt verfolgt einige vielversprechende Ansätze, um […] zu vertiefen. Aktivitäten […] können aber noch zielgerichteter verfolgt sowie systematischer mit anderen Arbeitsfeldern des Länderprojekts verzahnt werden.“

### *Good Practice „Darstellung der Wirkungszusammenhänge“ (GIZ)*

*Das Qualitätskriterium ist erfüllt, wenn bei der Beschreibung der intendierten Wirkungszusammenhänge der EZ-Maßnahme zwischen den Wirkungsstufen Input-Output-Outcome-Impact unterschieden wird (Abbildung 2).*

„For each of the 3 thematic Work packages, two interrelated hypotheses were formulated on two different levels: one hypothesis linked the activities to the outputs and outcomes (results hypothesis) and the other linked up the outcomes and impacts (impact hypothesis).“

„Results hypothesis 1.1 (activity > output > outcome/for analysis of OECD-DAC criterion ‘effectiveness’)

a) Activities:

- Raise awareness of participation potentials
- Generate willingness to participate among stakeholders
- Assess risks/demands of vulnerable groups (slum-specific vulnerability mapping)
- Define vulnerable groups’ priorities for community-based adaptation (Participatory Community Development Process, PCDP)

b) Output A:

- Decision-makers in 4 partner cities are better informed about the specific needs and challenges of the poor urban population, especially those of women and children, with respect to the adverse impacts of climate change

c) Outcome:

- Selected partner cities are implementing measures, on their own initiative and with the active involvement of the poor population, to increase adaptive capacities in slum settlements (module objective)

Impact hypothesis 1.2 (outcome > impact/for analysis of OECD-DAC criterion ‘impact’)

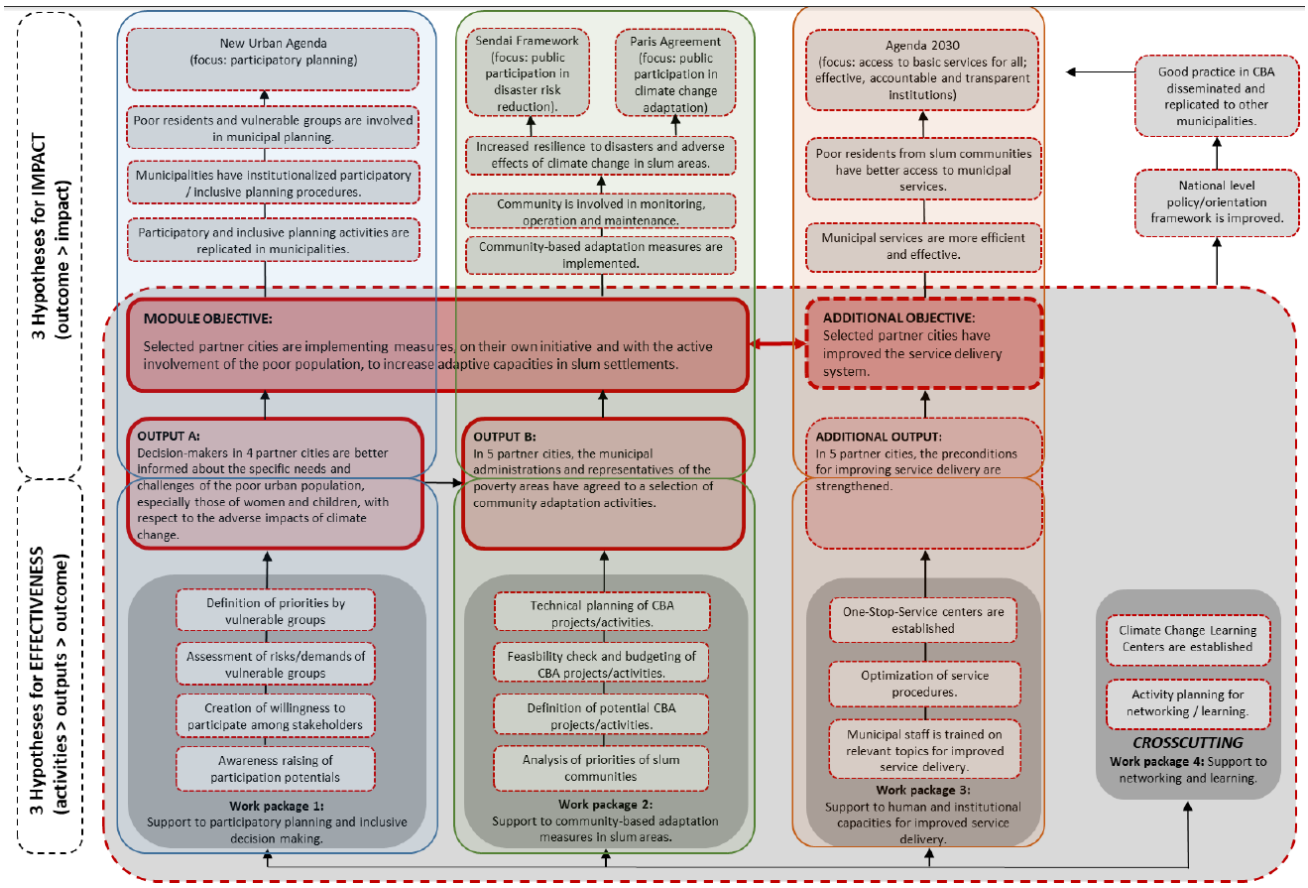
c) Outcome:

- Selected partner cities are implementing measures, on their own initiative and with the active involvement of the poor population, to increase adaptive capacities in slum settlements (module objective)

d) Impact:

- Participatory and inclusive planning activities are replicated in municipalities
- Municipalities have institutionalised participatory/inclusive planning procedures
- Poor residents and vulnerable groups are regularly involved in municipal planning
- Contributions to the objectives of the New Urban Agenda (focus: participatory planning)“

**Abbildung 2 Good Practice „Darstellung der Wirkungszusammenhänge“**




*Quelle: GIZ Evaluierung*

### *Good Practice „Darstellung der Angemessenheit des methodischen Vorgehens“ (DVV)*

*Das Qualitätskriterium ist erfüllt, wenn a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden.*

„When defining the samples of students and teachers to be interviewed, the evaluation team had to consider time constraints, security issues and logistical challenges.

[…]

Taking into consideration that the first field mission which was reserved for the collection of primary data lasted only two weeks, the team in consultation with the management of ANAFAE decided to implement interviews in ALCs and CLCs in Kabul and in Mazar-e Sharif, where security and transportation challenges were manageable, plus one ALC outside the two urban centres, namely ALC Aybak in Samangan Province.

The evaluation team acknowledges, that this selection of ANAFAE Centres may have caused a bias of primary data since rural Centres are underrepresented in the survey.

[…]

While the methodological limitations addressed above were unavoidable, the evaluation team applied the triangulation approach to assure the quality of primary data. Triangulation, in a nutshell, is a research strategy in empirical social sciences. Different methods are applied to balance the strengths and weaknesses of the different approaches. Weaknesses of the semi-structured interviews with the Managers of ANAFAE Centres, Master Trainers, Head of Departments, and of the Focal Group Discussions with the teachers in the visited centres were:

- Quality of interpretation
- Distractions during interviews (e.g. mobile phone calls, people coming and going)
- Time limitations“

The triangulation approach aims to increase the validity of data and to avoid systematic mistakes. The evaluation team has applied three triangulation strategies: triangulation of information through submission of selected interview notes to interviewees (especially ANAFAE managers, head teachers and Centre managers) for quality control purposes; triangulation of investigators through cross-checking of collected data within the evaluation team and triangulation of methods through the application of different data collection methods, including focal groups discussions, structured and semi-structured interviews, on-site inspections and photographic documentation.“

### *Good Practice „Einbindung der internen und externen Stakeholder*innen“ (GIZ)*

*Das Qualitätskriterium ist erfüllt, wenn Stakeholder*innen der Evaluierung in den Evaluierungsprozessen (1. Konzeption/Planung, 2. Durchführung, 3. Berichtslegung) einbezogen werden (Abbildung 3).*

**Abbildung 3 Good-Practice-Beispiel „Einbindung der Stakeholder*innen“**

| Work step | When | Responsible | Collaborating | To be informed |
|---|---|---|---|---|
| **Preliminary clarifications** including agreement on timing of evaluation | Dec 2017 | Evaluation Unit | AV, partner(s) | |
| **Quality check 2 of IR** | Feedback to contractor: 23.03.2018 | Evaluation Unit | AV, partner(s) (for material accuracy) | |
| **Final presentation, debriefing/ final meeting, local** | 20.04.2018 | International and local evaluator | AV/project team, country director, partner(s), embassy | |

*Quelle: GIZ-Evaluierung*

### *Good Practice „Berücksichtigung Kapazitätsentwicklung“ (GIZ)*

*Das Qualitätskriterium ist erfüllt, wenn im Evaluierungsprozess Aktivitäten zur Entwicklung von Evaluierungskapazitäten der Akteur*innen im Partnerland durchgeführt werden (1. Verbesserung individueller Evaluierungskenntnisse, 2. Stärkung des Evaluierungsmanagements, 3. Förderung der Nutzung von Evaluierungsergebnissen, 4. Unterstützung eines Klimas/Umfelds von Rechenschaftspflicht und Lernen in den Partnerländern, Tabelle 14).*

**Tabelle 14 Good Practice „Berücksichtigung Kapazitätsentwicklung“ (GIZ)**

| | Work step | When | Responsible | Collaborating |
|---|---|---|---|---|
| | Verbesserung individueller Evaluierungskenntnisse: Mindestens ein*e lokale*r/nationale*r Gutachter*in wird in Entscheidungsprozesse der Evaluierung einbezogen. | | | |
| 1 | **Draft inception report (IR)** in accordance with GIZ specifications and template, report language: **English** | **Submission of IR 04.01.2018** | International evaluator | Local evaluator |
| | Stärkung des Evaluierungsmanagements: Mindestens eine Partnerorganisation wird in den Planungs- und Konzeptionsprozess der Evaluierung einbezogen. | | | |
| 2 | **Quality check 2 of IR** | Feedback to contractor: 08.02.2019 | Evaluation Unit | AV, partner(s) (for material accuracy) |
| | Förderung der Nutzung von Evaluierungsergebnissen: Vorstellungen der Ergebnisse der Evaluierung bei den Partnerorganisationen. | | | |
| 3 | **Final presentation, debriefing/ final meeting, local** | 15.03.2019 | International and local evaluator | AV/project team, country director, partner(s), embassy |

*Quelle: GIZ-Evaluierung*

### *Good Practice „Handlungsorientierung der Empfehlungen“ (DRK)*

*Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen (alternativ: Schlussfolgerungen) der Evaluierung handlungsorientiert sind.*

„1. Recommendations for the design of the project

- Ensure a follow-up between phase 2 and phase 3

(See above under Relevance with regards to the priorities of the donor) It is especially recommended to keep the current seven involved chapters, at least with activities aiming at either complementing or consolidating the acquired knowledge and capacities during the present phase. In an ideal case the two chapters involved in phase 1 should be reintegrated into the programme with phase 3. To the point of view of the consultant the argument of excluding a chapter of the project if there is no back-up from the chapter’s leadership cannot be valid. It is actually contradicting with the priority of mainstreaming DRR in the chapters. If PRC NHQ decides to mainstream DRR to all chapters, this decision will have eventually to be followed by all chapters’ leadership.

[…]

2. Recommendations for the content

- Plan necessary adjustments (as compared to phase 2) to the changed institutional environment (new Act creating the department of disaster resilience)

As mentioned above it is recommended to keep the overall goal. When planning for phase 3 it will have to be remembered that the original German wording of the overall goal refers to the Pilipino DRRM law, meaning the DRRM Act of 2010. If this Act is complemented or even replaced by the new drafted Act on the creation of a department of disaster resilience, it will imply changes to the institutional and operational framework for the sector of DRRM. As a consequence the content of phase 3 will have to be planned accordingly to the new context in or to still 'contribute to the national consolidation and operationalization process' of the legislative framework.

- Ensure comprehensive operationalization of the meso-level

(See above under Relevance with regards to the priorities of the donor) As mentioned above the meso-level does not only mean the institutional strengthening of the 'partner organizations' (in this case PRC) but also of their networks. […]

- Strengthen advocacy and policy level

(See above under Directions for the concept note) It is strongly recommended to include more activities, measures of political dialogue, lobby and advocacy nature. The first step could be to strengthen the advocacy capacities of relevant units/ staff of PRC."

### *Good Practice „Adressat*innen der Empfehlungen" (MISEREOR)*

*Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen der Evaluierung (alternativ: ihre Schlussfolgerungen) adressat*innenorientiert sind (Abbildung 4).*

**Abbildung 4 Good-Practice-Beispiel „Adressat*innen der Empfehlungen"**

*15. Summary of recommendations*

| No. | Recommendation | Responsibility | Perspective | | |
|---|---|---|---|---|---|
| | | | Short-term | Mid-term | Long-term |
| | APPROACH | | 2019 | 2022 | 2025 |
| 1. | Have a clear, cohesive approach aimed at achieving outcomes and [redacted] | OCR (together with tenant representatives) | | | |
| 2. | Build on the ethnic cohesion [redacted] | OCR (together with tenant representatives) | | | |
| 3. | Explore opportunities to build social cohesion through interaction [redacted] | OCR (together with tenant representatives) | | | |
| 4. | Develop a structured training scheme using a participatory and [redacted] | OCR (together with tenant representatives and other NGOs) | | | |
| 5. | Define criteria for the clear utilisation of survey results and [redacted] | OCR (together with tenant representatives) | | | |
| 6. | Explore all options to increase availability of legal advice, [redacted] | OCR (together with tenant representatives, other NGOs and Legal Resource Centres) | | | |

*Quelle: MISEREOR-Evaluierung*

### *Good Practice „Informationsgehalt der Zusammenfassung" (PTB)*

*Das Qualitätskriterium ist erfüllt, wenn eine Zusammenfassung vorliegt und diese 1. Ergebnisse und entweder 2. Schlussfolgerungen oder 3. Empfehlungen der Evaluierung nennt.*

„Executive summary

[…]

Findings

1. Relevance

The project interventions are […]

2. Effectiveness

According to the data and impressions obtained during the evaluation, the indicators have been achieved in a satisfactory way. […]

[…]

Recommendations

1. Project management

Revise the Objective/Outcome/Output/Indicator logic with SMART indicators: Define indicators according to the SMART principle, use precise wording in formulation of the objectives and goals. […]

2. Risk Assessment

Use of Government/Safety Net existing data to reduce data collection effort.“

### *Good Practice „Darstellung organisationale Unabhängigkeit des Gutachtenden“ (KfW)*

*Das Qualitätskriterium ist erfüllt, wenn die Gutachtenden der Evaluierung eine ausreichende Distanz zur EZ-Maßnahme haben, das heißt, wenn sie sind weder 1) politisch 2) operativ 3) und beratend an der EZ-Maßnahme beteiligt sind und 4) nicht zur Zielgruppe der EZ-Maßnahme gehören (Abbildung 5).*

**Abbildung 5 Good-Practice-Beispiel „Darstellung organisationale Unabhängigkeit des Gutachtenden“**

**Anlage 2 – Erklärung zur Unabhängigkeit**

Hiermit versichern die an der Verfassung des Ex-post-Evaluierungsberichts beteiligten Personen, weder in der Planung noch in der Durchführung des evaluierten Vorhabens eine aktive Rolle eingenommen zu haben, sei es seitens der KfW Entwicklungsbank, eines anderen beteiligten Gebers, beim Projektträger, einem am Projekt beteiligten Consultingunternehmen, oder durch eine anderweitige Tätigkeit im Zusammenhang mit dem Vorhaben.

**An der Evaluierung Beteiligte** **Ort, Datum, Unterschrift**

*Quelle: KfW-Evaluierung*

### *Good Practice „Einbezug von Gutachtenden aus Partnerland“ (CARE)*

*Das Qualitätskriterium ist erfüllt, wenn sich im Gutachtenden-Team Fachleute aus den betroffenen Partnerländern befinden.*

„**Phase 1: Preparation (x until x of x of 2019)**

[…]

Phase 2: Conducting the Evaluation (field visit from x until x of x 2019)

The field visit, conducted by xxx from Kenya (local expert) and xxx (German expert), took place from the x to the x of x 2019.“

### 4.1.3 Elemente von Qualitätssicherungsprozessen

Durch die Onlinebefragung zeigte sich, dass neben dem Erstellen eines Inception Report zahlreiche weitere Aspekte eines Qualitätssicherungsprozesses von den Organisationen umgesetzt werden (Tabelle 15).

**Tabelle 15 Bestehende Qualitätssicherungsprozesse bei den teilnehmenden Organisationen**

| Qualitätssicherungsprozess | # durchführende Organisationen |
|---|---|
| ToR-Vorlagen | 10 |
| Kommentierung der ToR durch Gutachtende | 5 |
| Richtlinien oder Checklisten zur Durchführung von Evaluierungen | 10 |
| Anfertigung eines Inception Reports (oder Evaluierungskonzepts), der von der Organisation (zentrale Evaluierungsverantwortliche/Auftraggeber) kommentiert wird (mündlich oder schriftlich) | 9 |
| Zeitplan für die Durchführung der Evaluierung, der klare Verantwortlichkeiten und Fristen für alle wichtigen Zwischenschritte enthält | 10 |
| Annotierte Gliederung des Evaluierungsberichts | 8 |
| Handreichungen zu diversen Qualitätsaspekten einer Evaluierung (zum Beispiel zur Durchführung von Remote-Befragungen oder Kontributionsanalysen) | 6 |
| Fachliche Vorbesprechung mit Gutachtenden (über einen Inception Report hinaus) | 8 |
| Methodische Vorbesprechung mit Gutachtenden (über einen Inception Report hinaus) | 8 |
| Regelmäßiger Austausch zwischen dem Gutachtenden-Team und den zentralen Evaluierungsverantwortlichen während der gesamten Evaluierung (zum Beispiel Beratung während der Erhebungsphase) | 6 |
| Peer-Review des Evaluierungsberichts durch an der Evaluierung unbeteiligte Personen | 2 |
| Kommentierung des Evaluierungsberichts (oder des Entwurfs) durch die zentralen Evaluierungsverantwortlichen (schriftlich oder mündlich) | 10 |
| Kommentierung des Evaluierungsberichts (oder des Entwurfs) durch die Partnerorganisation(en) (schriftlich oder mündlich) | 5 |
| Kommentierung des Evaluierungsberichts (oder des Entwurfs) durch das operative Management der EZ-Maßnahme seitens ihrer Organisation (schriftlich oder mündlich) | 7 |
| Weitere nicht aufgelistete Komponente(n) | 4 |

*Quelle: DEval, eigene Darstellung*
*Anmerkung: ToR = Terms of References*

### 4.1.4 Darstellung der Gründe für eine Nichtanwendung der OECD-DAC- und DeGEval-Standards in der Onlinebefragung

Im Folgenden werden einige von den teilnehmenden Organisationen zurückgemeldete Gründe für die Nichtanwendung ausgewählter Qualitätskriterien aufgeführt (Tabelle 16). Die genannten Gründe auf Organisationsebene waren bei einigen Qualitätsstandards zwischen den Organisationen ähnlich, bei anderen unterschiedlich. Gründe für die Nichtanwendung lagen auf Organisations- und/oder Evaluierungsebene sowie in der Messung der Qualitätsstandards.

**Tabelle 16 Gründe für eine Nichtanwendung der Qualitätsstandards**

| QK | SC | mehrheitliche Begründung | Einzelbegründungen |
|---|---|---|---|
| Qualitätssicherung IR[a] | B & M | • fehlende Ressourcen<br>• Evaluierung zu klein für Qualitätssicherung<br>• Verantwortlichkeit liegt außerhalb der Evaluierungseinheit<br>• andere Formen als im QK gefordert bestehen (zum Beispiel Handreichungen) | • Sonstiges (zum Beispiel unterschiedliches Qualitätsverständnis) |
| Veröffentlichung Evaluierungsbericht/ Zusammenfassung | N | • Vertraulichkeit (zum Beispiel Schutz der Partnerorganisation/eigenen Arbeit vor Ort/Rechte Dritter, fehlende Anonymisierungsmöglichkeiten, Betriebs- und Geschäftsgeheimnisse, Beeinträchtigung von nat./internat. Verhandlungen)<br>• andere Formen als im QK gefordert bestehen (zum Beispiel Austausch mit Partnerorganisation, auf Nachfrage, gegebenenfalls im Jahresbericht, geteilt mit Arbeitsgruppe) | • Sonstiges (zum Beispiel Qualität des Berichts) |
| Kompetenz Gutachtende | N | • Verfügbarkeit | • Sonstiges (zum Beispiel keine eigenen Kenntnisse über Kompetenzen, hoher Aufwand für kleinere Organisationen, andere Organisationsstruktur) |
| Rechtzeitigkeit Erkenntnisse | N | • interne Gründe (zum Beispiel Abstimmungsprozesse, Partnerverfügbarkeit, geänderte Projektabläufe, Krankheiten, fehlende Ressourcen, geänderte Priorisierung, Evaluierungstyp)<br>• externe Gründe (zum Beispiel Verfügbarkeit Auftragnehmer, Sicherheit im Partnerland, Corona-Pandemie) | / |
| Evaluierungseffizienz | N | / | • Sonstiges (Nützlichkeit hat gegebenenfalls Vorrang, fehlende Ressourcen, fehlender Prozess, fehlende Qualität der Evaluierung, Prozesse auf Effizienz angelegt) |
| Ausreichende Ressourcen vorhanden | N | / | • Sonstiges (zum Beispiel nicht institutionalisierte Prozesse, Kontext, Angebot liegt über Budget) |

| QK | SC | mehrheitliche Begründung | Einzelbegründungen |
|---|---|---|---|
| Vorhandensein Management-Response | N | • fehlender Prozess<br>• andere Formen als im QK gefordert bestehen (zum Beispiel Ergebnisse fließen in Konzeption der Folgephase/Beendigung des Projekts; Austausch zur Nutzung und Nützlichkeit im Abschlussgespräch) | • Sonstiges (zum Beispiel fehlende Ressourcen, Abhängigkeit von Kontexten) |
| Transparenz Meinungsverschiedenheiten | P, U & F | • Konsens wird gesucht beziehungsweise Meinungsunterschiede werden geklärt | • Sonstiges (zum Beispiel Evaluierungsteam entscheidet) |
| Einbindung Partner*innen | P, U & F | / | • Sonstiges (zum Beispiel nicht sinnvoll in einer Evaluierung, hoher Koordinierungsaufwand) |
| Ethik | P, U & F | / | • Sonstiges (zum Beispiel Partnerorganisationen haben keinen Passus in Verträgen, Angaben eher generell) |
| Darstellung Unvoreingenommenheit GAs | P, U & F | • andere Formen als im QK gefordert bestehen (zum Beispiel ähnliche Inhalte, Ausschließen von Interessenskonflikten, in Gesprächen thematisiert, indirekt über Vorgabe der DeGEval-Standards, in Auswahlprozess bzgl. der Verbindung zur Organisation nicht zur Evaluierung) | / |
| Berücksichtigung Gemeinschaftsevaluierungen | P, U & F | • nicht möglich/angemessen (zum Beispiel keine anderen Evaluierungen in dem Bereich, Evaluierung ist auf Projektebene, nur kleinere Projekte werden evaluiert, selten Teil eines EZ-Programms, Kooperationsverhältnisse) | / |
| Partnerschaftliche Ansätze | P, U & F | • nicht möglich/angemessen (zum Beispiel spezieller Bereich beziehungsweise vorgegebener Zeitrahmen) | • Sonstiges (zum Beispiel nicht systematisch, indirekt über lokale Gutachtende, fehlendes Wissen) |
| Genderbalance Team | P, U & F | • geringere Kompetenzen beziehungsweise Verfügbarkeit | • Sonstiges (zum Beispiel kein Qualitätsstandard, manchmal keine Gutachtenden, Wirtschaftlichkeit ist prioritär) |

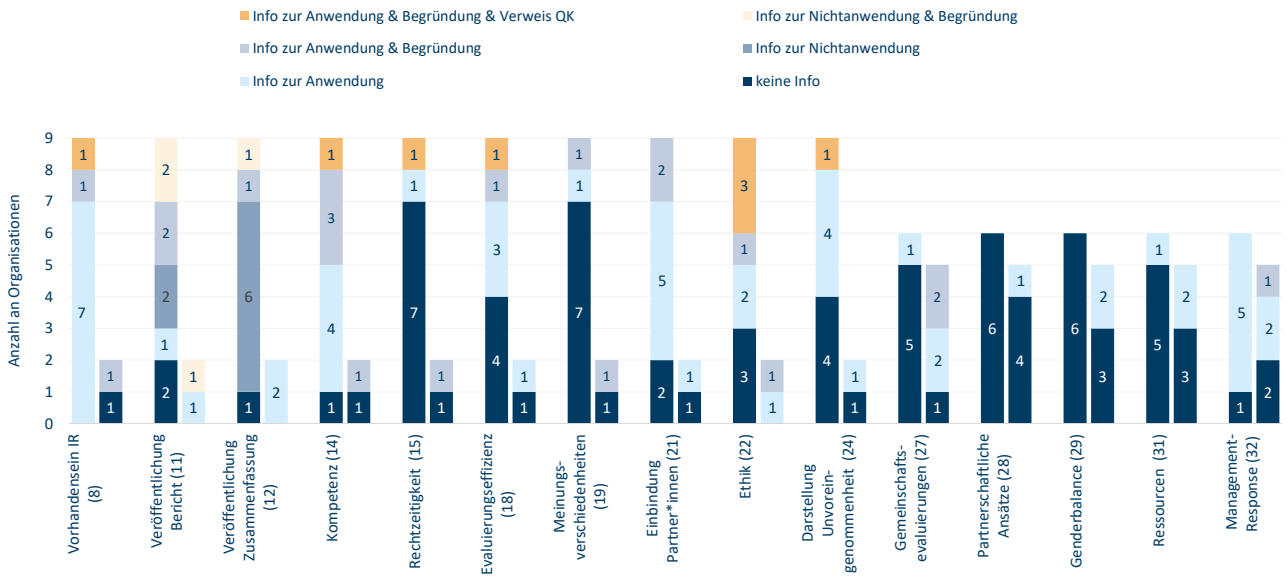
*Quelle: DEval, eigene Darstellung*

*Anmerkung: QK = Qualitätskriterium; SC = Standardcluster; B & M = Berichtslegung und Methoden; N = Nutzbarkeit; P, U & F = Partizipation, Unabhängigkeit und Fairness. [a] Das QK wurde sowohl auf Organisations- (Onlineanhang) als auch auf Evaluierungsebene erfasst.*

### 4.1.5 Darstellung der Verschriftlichung von OECD-DAC- und DeGEval-Standards in Organisationsdokumenten

Abbildung 6 stellt die Verschriftlichung der OECD-DAC- und DeGEval-Qualitätsstandards aus der Onlinebefragung in Organisationsdokumenten für Gruppe 1 und 2 dar.

**Abbildung 6 Verschriftlichung der Qualitätsstandards in Organisationsdokumenten**



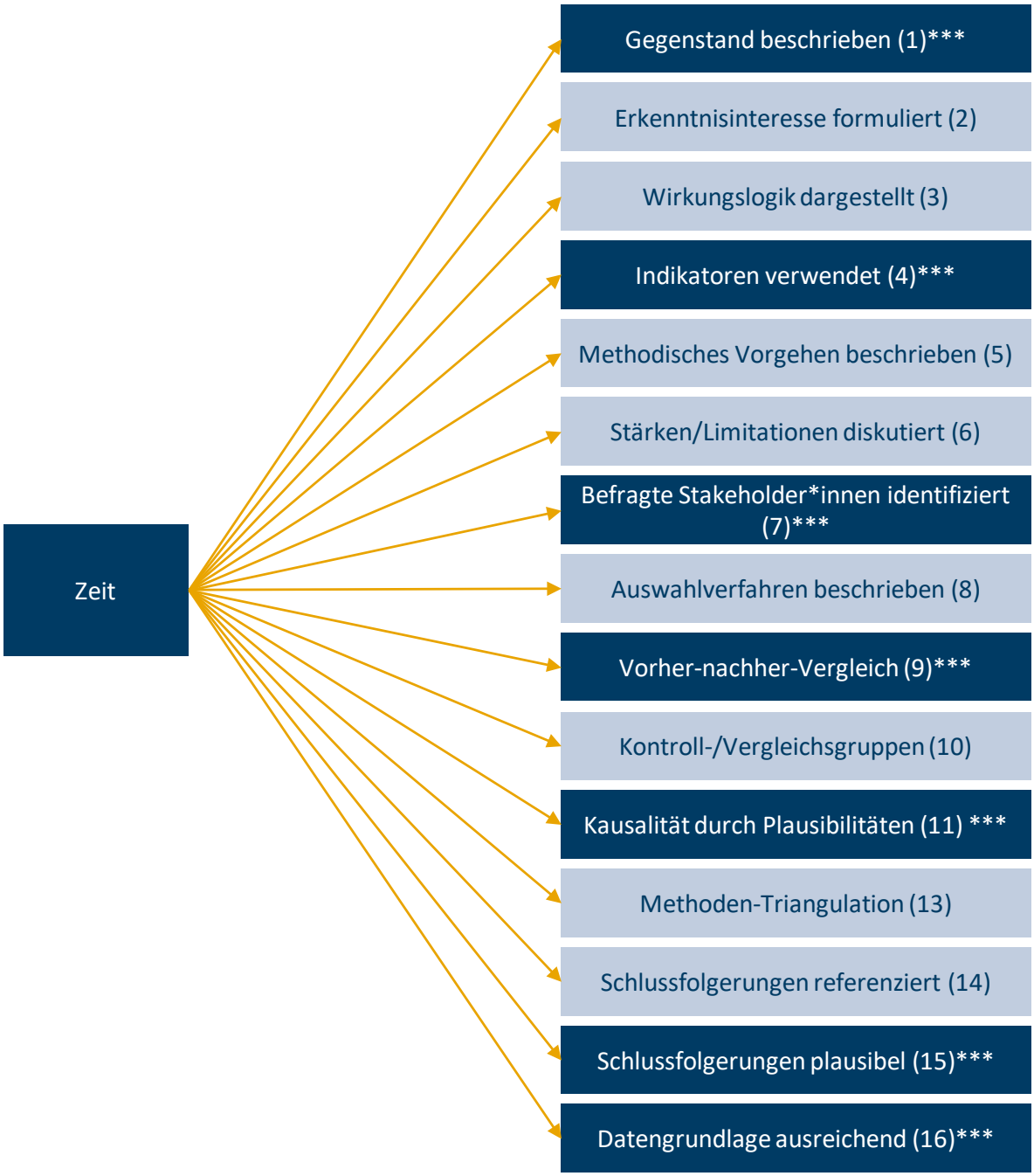

*Quelle: DEval, eigene Darstellung*

*Anmerkungen: IR = Inception Report, QK = Qualitätskriterium. „Info zur Nichtanwendung & Begründung & Verweis QK" wurde ebenfalls untersucht, konnte aber für kein Qualitätskriterium ermittelt werden.*

### 4.1.6 Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in GIZ- und KfW-Evaluierungen

Abbildung 7 zeigt die Ergebnisse des Strukturgleichungsmodells für die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den GIZ- und KfW-Evaluierungen.

**Abbildung 7 Ergebnisse des Strukturgleichungsmodells zur Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit**




*Quelle: DEval, eigene Darstellung*

*Anmerkung: Die hellblauen Qualitätskriterien sind nicht statistisch signifikant/bedeutsam. Im Strukturgleichungsmodell wurden darüber hinaus Korrelationen zwischen allen Qualitätskriterien zugelassen (hier aus Gründen der Lesbarkeit nicht abgebildet). Sternchen geben die statistische Signifikanz an: * $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$.*

### 4.1.7 Anwendung von Text Mining für ausgewählte Datenerhebungsmethoden

Über die Erkenntnisse in Evaluierungsfrage 2c hinaus wurde die Anwendung von den zwei Datenerhebungsmethoden „Leitfadeninterviews“ und „Fokusgruppendiskussionen“ der Meta-Evaluierung Nachhaltigkeit mithilfe von Text Mining, einem computergestützten Prinzip des maschinellen Lernens (Dumm, 2018), in Evaluierungen der vorliegenden Meta-Evaluierung ermittelt.[26]

**Die Text-Mining-Analyse wurde mit dem interactive Leipzig Corpus Miner (iLCM)[27] für die Anwendung der beiden Datenerhebungsmethoden „Leitfadeninterviews“ und „Fokusgruppendiskussionen“ in den Evaluierungen durchgeführt.** In einem ersten Schritt wurde entlang der bereits kodierten GIZ- und KfW-Evaluierungen aus der Meta-Evaluierung Nachhaltigkeit ein sogenannter „trainierter Klassifikator“ erstellt. Die qualitative Güte des Klassifikators wurde mit dem F-Score bestimmt. Dieser wurde über alle Evaluierungen, die mittels maschinellen Lernens klassifiziert wurden, und je Datenerhebungsmethode berechnet. Der F-Score lag für beide Datenerhebungsmethoden im akzeptablen Bereich: Leitfadeninterviews: 0,85; Fokusgruppendiskussionen: 0,80.[28] Im Anschluss erfolgte die automatisierte Klassifikation der beiden Datenerhebungsmethoden in den Methodenkapiteln der deutsch- und englischsprachigen[29] Evaluierungen der vorliegenden Meta-Evaluierung (Leitfadeninterviews: 259 Evaluierungen; Fokusgruppendiskussionen: 126 Evaluierungen).[30] Abschließend wurden die Ergebnisse des maschinellen Lernens für alle Evaluierungen qualitativ überprüft.

**Leitfadeninterviews wurden in 98,3 Prozent der Evaluierungen durchgeführt, Fokusgruppendiskussionen in 46,6 Prozent.** Im Vergleich kamen in den GIZ- und KfW-Evaluierungen der Meta-Evaluierung Nachhaltigkeit bei 83 Prozent Leitfadeninterviews und bei 26 Prozent Fokusgruppendiskussionen vor. Abbildung 7 zeigt, dass Leifadeninterviews in 291 Evaluierungen angewandt wurden, Fokusgruppendiskussionen in 138 und somit in etwas weniger als der Hälfte der insgesamt 296 Evaluierungen.

[26] Die zusätzliche Analyse wurde durchgeführt, um praktische Erfahrungen mit dem iLCM-Tool zu sammeln.

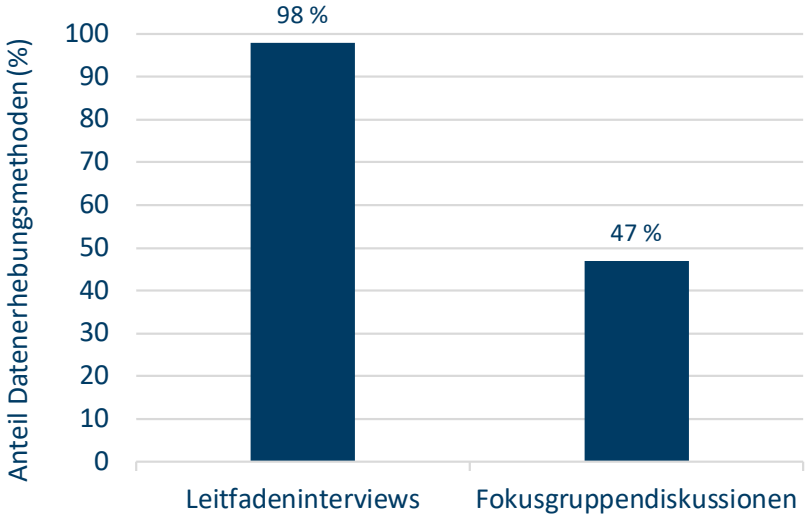
[27] Der iLCM ist eine auf der Programmiersprache R basierte Software-Infrastruktur der Universität Leipzig (Niekler, 2018).

[28] Der F-Score kann zwischen 0 und 1 liegen und gibt an, wie präzise – auf der Basis der Trainingsdaten – die über Text Mining zu ermittelten Informationen bestimmt werden können. Je höher der Wert des F-Score, desto geringer ist der Anteil an fehlerhaften maschinell ermittelten Kodierungen. Ein Wert von mindestens 0,8 entspricht einer hohen qualitativen Güte des Klassifikators (Allwright, 2022; Grüne, 2020).

[29] Spanisch- und französischsprachige Evaluierungen konnten nicht für die Text-Mining-Analyse berücksichtigt werden, da nicht ausreichend Evaluierungen für das maschinelle Lernen vorlagen.

[30] Für die Datenerhebungsmethode „Leitfadeninterviews“ konnten für alle Evaluierungen – bis auf die 37 spanisch- und französischsprachigen – Informationen über Text Mining erhoben werden (N = 259). Aufgrund des hohen Ressourcenaufwands für die Erstellung des Klassifikators (der für Deutsch und Englisch separat erstellt werden musste), konnten für die Datenerhebungsmethode „Fokusgruppendiskussionen“ lediglich Informationen für 126 deutschsprachige Evaluierungen über Text Mining erhoben werden. Evaluierungen, für die keine Informationen über Text Mining erhoben werden konnten (37 Evaluierungen bei „Leitfadeninterviews“ und 170 bei „Fokusgruppendiskussionen“), wurde eine händische Kodierung vorgenommen. Dabei wurden die Evaluierungen von einer Person kodiert und von einer weiteren überprüft.

**Abbildung 8 Ergebnisse der zum Teil über Text Mining ermittelten Datenerhebungsmethoden**




*Quelle: DEval, eigene Darstellung*
*Anmerkung: N = 296 Evaluierungen*

### 4.1.8 Nebeneinanderstellung der Erkenntnisse aus dem DEval-Umsetzungsmonitoring der Systemprüfung und der vorliegenden Meta-Evaluierung

Insgesamt zeigte sich, dass die Erkenntnisse aus dem DEval-Umsetzungsmonitoring der Systemprüfung und der Meta-Evaluierung keine Längsschnittanalyse beziehungsweise Trendanalyse zulassen, da einige relevante Faktoren über die beiden Erhebungen hinweg nicht aufeinander abgestimmt werden konnten (Tabelle 17). In Tabelle 18 werden die Qualitätskriterien, die sich ähneln, in ihrer Bewertung nebeneinandergestellt.

**Die vorliegende Meta-Evaluierung schließt an frühere Systemprüfungen des deutschen EZ-Systems und das Umsetzungsmonitoring der Systemprüfung an.** Da sich die vorliegende Meta-Evaluierung in ihrem Qualitätsverständnis ebenfalls an den OECD-DAC- und den DeGEval-Standards orientierte, ergaben sich stellenweise Anknüpfungsmöglichkeiten an die Untersuchungen der Systemprüfung von 2009 beziehungsweise das Umsetzungsmonitoring der Systemprüfung aus dem Jahr 2015, die in der Nebeneinanderstellung berücksichtigt wurden.

**Inhaltliche Anknüpfungspunkte zwischen der Meta-Evaluierung und der Systemprüfung von 2009 sowie dem Umsetzungsmonitoring der Systemprüfung fanden sich vor allem in den Standardclustern „Berichtslegung und Methoden“ sowie „Nutzbarkeit“.** Aufgrund der Anknüpfungspunkte konnten zu einigen Qualitätskriterien Erkenntnisse der vorliegenden Meta-Evaluierung neben Erkenntnissen aus den vorangegangenen Untersuchungen gestellt werden. Anknüpfungspunkte fanden sich unter anderem in den Qualitätskriterien/-standards „Beschreibung des Erkenntnisinteresses“, „Darstellung Angemessenheit des methodischen Vorgehens“, „Zugänglichkeit“ sowie „Nützlichkeit der Empfehlungen“. Beispielsweise lautet das Qualitätskriterium „Beschreibung des Erkenntnisinteresses“: „Das Qualitätskriterium ist erfüllt, wenn 1. Zweck, 2. Ziele und 3. Evaluierungsfrage(n) der Evaluierung genannt werden.“ Dieses Qualitätskriterium wies eine Ähnlichkeit – aber keine Deckungsgleichheit – mit dem Qualitätskriterium des DEval-Umsetzungsmonitorings der Systemprüfung auf: „Das Kriterium ist erfüllt, wenn die Evaluierungsfragen in Terms of References dokumentiert sind“ (Lücking et al., 2015).

**Tabelle 17 Informationen zum DEval-Umsetzungsmonitoring und der vorliegenden Meta-Evaluierung**

| | Meta-Evaluierung | DEval-Umsetzungsmonitoring der Systemprüfung |
|---|---|---|
| Untersuchungszeitraum | 10/2016 – 12/2020 | 2012 – 2013 |
| Organisationen | 11 deutsche EZ-Organisationen (kriterienbasierte Auswahl) | 32 deutsche EZ-Organisationen, die als Mittelempfänger 2012 und 2013 jeweils > 1 Mio. EUR vom BMZ erhalten haben, inklusive der EZ-Organisationen der Systemprüfung |
| Evaluierungsgegenstand | (Projekt-)Evaluierungen, die … in Partnerländern umgesetzt, … von BMZ-Mitteln (mit-)finanzierte Entwicklungsprojekte/Evaluierungen betreffen, … von Evaluierungsverantwortlichen in Deutschland (mit-)verantwortet wurden | strategische und/oder projektübergreifende Evaluierungen |
| Anzahl Evaluierungen | Grundgesamtheit: 576 Evaluierungen<br>Stichprobe: 296 Evaluierungen (geschichtet nach Organisation und Jahr) | Grundgesamtheit: 213 Evaluierungen<br>Stichprobe: 45 Evaluierungen (geschichtet nach Akteursgruppe und Evaluierungstyp) |
| Dokumente | verschiedene Evaluierungsdokumente | Evaluierungsberichte |

*Quelle: DEval, eigene Darstellung*

**Tabelle 18 Nebeneinanderstellung der Ergebnisse ausgewählter Qualitätskriterien**

| Name QK in Meta-Evaluierung (Nummer) | QK in Meta-Evaluierung<br>Das Qualitätskriterium ist erfüllt, wenn … | % der Fälle (N = 296) | QK in DEval-Umsetzungsmonitoring der Systemprüfung<br>Das Kriterium ist erfüllt, wenn … | % der Fälle (N = 45) |
|---|---|---|---|---|
| Beschreibung Evaluierungsgegenstand (1) | … 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteur*innen (politische Partner*innen und/oder Träger) der Entwicklungsmaßnahme genannt sind. | 99 % | … wesentliche Aspekte (wie Akteursstruktur und Implementierungskontext) des Evaluierungsgegenstands erläutert werden. (4) | 87 % |
| Beschreibung Kontext EZ-Maßnahme (2) | … der Kontext der Entwicklungsmaßnahme beschrieben wird (zum Beispiel politische, ökonomische, soziale, [agrar-]ökologische, [agrar-]klimatische, gesetzliche, technologische, historische, kulturelle Kontextfaktoren). | 83 % | … wesentliche Aspekte (wie Akteursstruktur und Implementierungskontext) des Evaluierungsgegenstands erläutert werden. (4) | 87 % |
| Darstellung Wirkungszusammenhänge (4) | … bei der Beschreibung der intendierten Wirkungszusammenhänge der Entwicklungsmaßnahme zwischen den Wirkungsstufen Input-Output-Outcome-Impact unterschieden wird. | 33 % | … die Wirkungskette der untersuchten Maßnahme dargestellt und nach Inputs, Outputs, Outcomes und Impacts unterschieden wird. (5)<br>… für die Evaluierung (Wirkungs-)Hypothesen formuliert wurden. Diese Hypothesen müssen begründete Vermutungen über Sachverhalte oder Zusammenhänge darstellen. (8) | 32 %<br>20 % |
| Beschreibung Erkenntnisinteresse (5) | … 1. Zweck, 2. Ziele und 3. Evaluierungsfrage(n) der Evaluierung genannt werden. | 82 % | … im Bericht die Adressaten und Zielgruppen der Evaluierung sowie deren Zweck und Zielsetzung eindeutig benannt werden. (23)<br>… die Evaluierungsfragen in Terms of References dokumentiert sind. (24) | 93 %<br>65 % |
| Nachvollziehbarkeit Informationsquellen (6) | … wenn die in der Evaluierung genutzten Informationsquellen ausreichend detailliert im Evaluierungsbericht dargestellt werden. | 92 % | … dem Lesenden bei der Analyse durchgehend verständlich ist, welche die methodischen Quellen der Informationen sind (zum Beispiel Dokumente, Interviews, schriftliche Fragebögen). (26) | 53 % |
| Darstellung methodisches Vorgehen (7) | … a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden. | 26 % | … die Auswahl des angewandten methodischen Vorgehens (bzgl. Erhebungsdesign, Methodenmix etc.) argumentiert wird. (2) | 16 % |

| Name QK in Meta-Evaluierung (Nummer) | QK in Meta-Evaluierung<br>Das Qualitätskriterium ist erfüllt, wenn … | % der Fälle (N = 296) | QK in DEval-Umsetzungsmonitoring der Systemprüfung<br>Das Kriterium ist erfüllt, wenn … | % der Fälle (N = 45) |
|---|---|---|---|---|
| Kohärenz-Schlussfolgerungen (9) | … die Datenanalyse, die Ergebnisse und die Schlussfolgerungen logisch miteinander in Bezug gesetzt werden beziehungsweise aufeinander aufbauen. | 61 % | … eine klare Trennung zwischen Beschreibung, Analyse, Bewertung und gegebenenfalls Handlungsempfehlungen vorgenommen wurde. (25) | 67 % |
| Nützlichkeit Empfehlungen (16, 17) | … die Empfehlungen der Evaluierung (alternativ: ihre Schlussfolgerungen) adressat*innenorientiert sind | 54 % | … für die Handlungsempfehlungen (größtenteils) ein Ziel beziehungsweise Zweck definiert worden ist. (31) | 67 % |
| | | | … die Ziele der Handlungsempfehlungen (größtenteils) konkret messbare Veränderungen benennen. (32) | 67 % |
| | … die Empfehlungen (alternativ: Schlussfolgerungen) der Evaluierung handlungsorientiert sind. | 76 % | … die Realisierbarkeit der Handlungsempfehlungen (größtenteils) methodisch oder inhaltlich abgesichert wurde (methodisch beispielsweise, indem Empfehlungen in Zusammenarbeit mit relevanten Akteuren entwickelt wurden; inhaltlich beispielsweise, wenn auf die konkrete Umsetzung der Empfehlung eingegangen wird). (33) | 67 % |
| | | | … projektübergreifende Empfehlungen zur entwicklungspolitischen Steuerung gegeben werden. (35) | 47 % |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: Qualitätskriterien des DEval-Umsetzungsmonitorings sind aus Lücking et al. (2015); Transformation der Bewertungsstufen 1 bis 4 zu 0 und 1, wobei die Stufen 1 und 2 zu 0 umgewandelt werden und die Stufen 3 und 4 zu 1. QK = Qualitätskriterium.*

## 4.2 Erklärung der Anwendung der Qualitätskriterien

Die Tabellen 19, 20 und 21 geben einen Überblick über die Korrelationen (Zusammenhänge) zwischen den Faktoren und den Qualitätskriterien. Die Tabellen 22, 23 und 24 zeigen die Ergebnisse der Regressionskoeffizienten. Tabelle 25 stellt die Odds Ratios für die signifikanten Effekte der logistischen und geordneten logistischen Regressionen dar.

### Tabelle 19 Korrelationsmatrix zwischen Faktoren und Qualitätskriterien (1/3)

| | Evaluierungs-gegenstand | Kontext | Wirkungs-zusammenhang | Kohärenz | Informtions-quellen | Darstellung Methodik | Ergebnisse-Schlussfolg. | Konflikt | Sozialkapital | Jahr 2020 | Anzahl GAs | Tagessatz externe GAs | Remote-Datenerhebung | GAs-Tage/ Kosten Maßnahme | Jahr der Evaluierung | Evaluierungs-einheit | Evaluierungs-einheit/Eval. | Evaluierungs-einheit/Org. | Evaluierungs-tätigkeit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluierungsgegenstand | 1 | | | | | | | | | | | | | | | | | | |
| Kontext | 0,05 | 1 | | | | | | | | | | | | | | | | | |
| Wirkungszusammenhänge | 0,12* | −0,06 | 1 | | | | | | | | | | | | | | | | |
| Erkenntnisinteresse | 0,15* | −0,07 | 0,19*** | 1 | | | | | | | | | | | | | | | |
| Informationsquellen | 0,11† | 0,02 | 0,22*** | 0,23*** | 1 | | | | | | | | | | | | | | |
| Darstellung Methodik | 0,08 | 0,01 | 0,25*** | 0,31*** | 0,18*** | 1 | | | | | | | | | | | | | |
| Kohärenz | 0,10 | 0,15* | 0,26*** | 0,11† | 0,25*** | 0,32*** | 1 | | | | | | | | | | | | |
| Konflikt | 0,03 | −0,03 | −0,04 | 0,00 | 0,04 | 0,05 | −0,00 | 1 | | | | | | | | | | | |
| Sozialkapital | −0,24*** | −0,06 | −0,01 | −0,05 | −0,10† | −0,13** | −0,13 | −0,28*** | 1 | | | | | | | | | | |
| Jahr 2020 | 0,11† | −0,07 | 0,23*** | 0,15** | 0,07† | 0,09 | 0,20*** | −0,07 | −0,03 | 1 | | | | | | | | | |
| Anzahl GAs | 0,03 | −0,06 | 0,12* | 0,22*** | 0,09† | 0,35*** | 0,22*** | 0,10† | −0,09† | 0,13* | 1 | | | | | | | | |
| Tagessatz externe GAs | 0,01 | 0,05 | 0,21*** | 0,13† | 0,08† | 0,17** | 0,08 | 0,09 | −0,21*** | 0,11† | 0,12† | 1 | | | | | | | |
| Remote-Datenerhebung | −0,02 | −0,07 | −0,05 | 0,03 | 0,02 | 0,02 | −0,06 | 0,00 | −0,04 | 0,28*** | 0,01 | 0,18** | 1 | | | | | | |
| GAs-Tage/Kosten Maßnahme | 0,01 | 0,11† | 0,01 | −0,47*** | −0,17*** | −0,24*** | −0,15* | 0,03 | −0,28*** | 0,08 | −0,25*** | −0,05 | 0,02 | 1 | | | | | |
| Jahr der Evaluierung | 0,16** | −0,10† | 0,31*** | 0,23*** | 0,05 | 0,18*** | 0,19*** | 0,03 | 0,03 | 0,69*** | 0,14* | 0,13*** | 0,20*** | 0,03 | 1 | | | | |
| Evaluierungseinheiten | 0,17** | 0,20*** | 0,45*** | −0,05 | 0,15** | 0,20 | 0,25 | 0,01 | −0,10 | 0,32*** | 0,07 | 0,08 | −0,13 | 0,39*** | 0,30*** | 1 | | | |
| Evaluierungseinheiten/ Evaluierungen | 0,08 | 0,12* | 0,38*** | 0,09 | 0,16* | 0,26* | 0,14† | 0,04 | −0,09 | 0,23*** | −0,17** | 0,25*** | −0,06 | 0,19*** | 0,27*** | 0,61*** | 1 | | |
| Evaluierungseinheiten/ Organisation | −0,01 | −0,14* | −0,14* | −0,07 | −0,06 | −0,28*** | −0,10† | 0,03 | 0,01 | −0,18** | −0,13* | 0,12† | 0,03 | −0,05 | −0,09 | −0,35*** | −0,34** | 1 | |
| Evaluierungstätigkeit | −0,05 | 0,06 | 0,14* | −0,08 | 0,04 | 0,01 | 0,10* | −0,00† | −0,15* | 0,06 | −0,012 | −0,07 | −0,05 | 0,45*** | 0,02 | 0,53*** | 0,15*** | −0,30 | 1 |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs = Gutachtende. Sternchen geben die statistische Signifikanz an: † p < 0,1; * p < 0,05; ** p < 0,01; *** p < 0,001.*

## Tabelle 20 Korrelationsmatrix zwischen Faktoren und Qualitätskriterien (2/3)

| | Zusammenfassung | Vorhandensein IR | Einbindung Stakeholder*innen | Leistungsbeschreibung | Empfehlungen | Kapazitäts-entwicklung | GAs Partnerland | Konflikt | Sozialkapital | Jahr 2020 | Anzahl GAs | Tagessatz GAs | Remote-Datenerhebung | GA-Tage/ Kosten EZ-Maßnahme | Jahr der Evaluation | Evaluierungseinheit | Evaluierungs-einheit/Eval. | Evaluierungs-einheit/Org. | Evaluierungstätigkeit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zusammenfassung | 1 | | | | | | | | | | | | | | | | | | |
| Vorhandensein IR | −0,18** | 1 | | | | | | | | | | | | | | | | | |
| Einbindung Stakeholder*innen | 0,27*** | −0,15* | 1 | | | | | | | | | | | | | | | | |
| Leistungsbeschreibung | 0,29*** | −0,28*** | 0,61*** | 1 | | | | | | | | | | | | | | | |
| Empfehlungen | 0,27*** | −0,04 | 0,43*** | 0,35*** | 1 | | | | | | | | | | | | | | |
| Kapazitätsentwicklung | 0,21*** | −0,23*** | 0,62*** | 0,48*** | 0,38*** | 1 | | | | | | | | | | | | | |
| GAs Partnerland | 0,17** | −0,26*** | 0,39*** | 0,28*** | 0,24*** | 0,67*** | 1 | | | | | | | | | | | | |
| Konflikt | −0,07 | −0.03 | 0,02 | 0,02 | −0,07 | 0,01 | 0,08 | 1 | | | | | | | | | | | |
| Sozialkapital | 0,05** | −0,16** | 0,14 | 0,16* | 0,22** | 0,26*** | 0,05 | −0,28*** | 1 | | | | | | | | | | |
| Jahr 2020 | −0,03 | 0,23*** | 0,13* | −0,03 | 0,05 | 0,02 | 0,03 | −0,07 | −0,03 | 1 | | | | | | | | | |
| Anzahl GAs | 0,17** | 0,10† | 0,14* | 0,11 | 0,16** | 0,27*** | 0,22*** | 0,10† | −0,10† | 0,13* | 1 | | | | | | | | |
| Tagessatz GAs | 0,10 | 0,02 | 0,14* | 0,11 | 0,09 | 0,02 | 0,02 | 0,09 | −0,21*** | 0,12† | 0,12† | 1 | | | | | | | |
| Remote-Datenerhebung | −0,01 | 0,06 | −0,04 | 0,05 | −0,01 | −0,20*** | −0,10† | 0,00 | −0,04 | 0,28*** | 0,01 | 0,17** | 1 | | | | | | |
| GA-Tage/ Kosten Maßnahme | −0,39*** | 0,38*** | −0,48*** | −0,53*** | −0,42*** | −0,53*** | −0,43*** | 0,03 | −0,28*** | 0,08 | −0,25*** | −0,05 | 0,02 | 1 | | | | | |
| Jahr der Evaluierung | 0,01 | 0,24*** | 0,19*** | 0,08 | 0,01 | 0,13* | 0,09 | 0,03 | −0,03 | 0,69*** | 0,14* | 0,13* | 0,20*** | 0,03 | 1 | | | | |
| Evaluierungseinheiten | −0,24*** | 0,42*** | −0,09 | −0,36*** | −0,09 | 0,12* | 0,13* | 0,01 | −0,10† | 0,32*** | 0,07 | 0,08 | −0,13 | 0,39* | 0,30*** | 1 | | | |
| Evaluierungseinheiten/Evaluie-rungen | −0,02 | 0,28 | −0,03 | −0,02 | −0,01 | −0,09 | −0,02 | 0,04 | −0,09 | 0,33*** | −0,17* | 0,25*** | −0,06 | 0,19*** | 0,27*** | 0,61*** | 1 | | |
| Evaluierungen/Organisation | 0,03 | −0,44*** | 0,07 | 0,00 | −0,10 | −0,08 | −0,06 | 0,03 | 0,01 | −0,18** | −0,13† | 0,12 | 0,03 | −0,05 | −0,09 | −0,35*** | −0,34** | 1 | |
| Evaluierungsstätigkeit | −0,24*** | 0,17*** | −0,12* | −0,14* | −0,21*** | −0,06 | −0,01 | −0,00 | −0,15* | 0,06 | −0,12 | −0,07 | −0,05 | 0,45*** | 0,02 | 0,53*** | 0,15*** | −0,30*** | 1 |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs = Gutachtende; IR = Inception Report. Sternchen geben die statistische Signifikanz an: † p < 0,1; * p < 0,05; ** p < 0,01; *** p < 0,001.*

**Tabelle 21 Korrelationen für die Indices im Standardcluster „Berichtslegung und Methoden“ (3/3)**

| | Summenindex Berichtslegung und Methoden | Faktorindex Berichtslegung und Methoden |
|---|---|---|
| Summenindex Berichtslegung und Methoden | 1 | |
| Faktorindex Berichtslegung und Methoden | 0,88** | 1 |
| Konflikt | −0,02 | 0,01 |
| Sozialkapital | −0,12† | −0,05 |
| Jahr 2020 (Pandemie) | 0,23*** | 0,19*** |
| Anzahl GAs | 0,32*** | 0,31*** |
| Tagessatz GAs | 0,21*** | 0,20** |
| Qualitätssicherung IR | 0,34*** | 0,00 |
| Einbindung Stakeholder*innen | 0,52*** | 0,67*** |
| Informationsgehalt der Leistungsbeschreibung | 0,45*** | 0,71*** |
| Remote-Datenerhebung | 0,00 | 0,02 |
| Tage GAs/Kosten Maßnahme | −0,32*** | −0,52*** |
| Jahr der Evaluierung | 0,32*** | 0,30*** |
| Evaluierungseinheiten absolut | 0,25*** | 0,04 |
| Evaluierungseinheiten/Anzahl Evaluierungen | 0,33 | 0,21 |
| Evaluierungseinheiten/Organisation | −0,31*** | −0,17** |
| Evaluierungsstätigkeit | 0,00 | −0,06 |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs= Gutachtende; IR = Inception Report; Sternchen geben die statistische Signifikanz an: † p < 0,1; * p < 0,05; ** p < 0,01; *** p < 0,001.*

**Tabelle 22 Ergebnisse der Regressionsanalysen (1/3)**

| | Evaluierungsgegenstand | Kontext | Wirkungszusammenhänge | Erkenntnisinteresse | Informationsquellen | Darstellung Methodik |
|---|---|---|---|---|---|---|
| **Regressionsmodell Nr. und Typ** | **1 (ologit)** | **2 (ols)** | **3 (ologit)** | **4 (ologit)** | **5 (ologit)** | **6 (ologit)** |
| Konflikt | −1,15 (2,38) | −0,03 (0,13) | −0,59 (0,35)† | −0,38 (0,34) | 0,03 (0,21) | 0,01 (0,41) |
| Sozialkapital | −8,76 (10,67) | −0,09 (0,29) | 0,06 (0,98) | −2,15 (1,65) | −1,70 (1,24) | −2,16 (0,92)* |
| Jahr 2020 (Pandemie) | 0,49 (2,32) | −0,01 (0,15) | −0,13 (0,45) | 0,07 (0,54) | 0,25 (0,43) | −0,88 (0,68) |
| Anzahl GAs | −0,55 (0,65) | 0,02 (0,04) | 0,21 (0,18) | 0,07 (0,20) | −0,21 (0,21) | 0,28 (0,14)* |
| Tagessatz externe GAs | −0,00 (0,00) | −0,00 (0,00) | 0,00 (0,00) | 0,00 (0,00) | −0,00 (0,00) | 0,00 (0,00) |
| Leistungsbeschreibung | 0,85 (0,74) | −0,08 (0,05) | 0,12 (0,12) | 0,29 (0,21) | 0,24 (0,18) | −0,06 (0,21) |
| Vorhandensein IR | −1,54 (1,15) | −0,16 (0,08) | −0,32 (0,46) | 0,23 (0,36) | 0,60 (0,32)† | 0,59 (0,45) |
| Einbindung Stakeholder*innen | 0,78 (0,72) | 0,17 (0,04)** | 0,32 (0,17)† | −0,30 (0,28) | 0,53 (0,12)*** | 0,10 (0,18) |
| Remote-Datenerhebung | / | / | / | / | / | 0,29 (0,18) |
| GAs-Tage/Kosten Maßnahme | −0,46 (0,69) | 0,03 (0,07) | 0,04 (0,15) | 0,24 (0,16) | −0,22 (0,19) | −0,19 (0,16) |
| Jahr der Evaluierung | 0,40 (0,40) | 0,05 (0,02)* | 0,29 (0,08)*** | 0,58 (0,22)** | −0,23 (0,11)* | 0,33 (0,13)* |
| Evaluierungseinheiten absolut | 0,30 (0,82) | −0,02 (0,06) | 0,08 (0,15) | −0,14 (0,24) | 0,11 (0,15) | −0,42 (0,30) |
| Evaluierungseinheiten/ Evaluierungen | −3,24 (5,08) | −0,76 (0,80) | −2,67 (2,42) | −3,20 (3,89) | −0,17 (2,63) | 1,40 (1,66) |
| Evaluierungseinheiten/ Organisationen | −0,01 (2,25) | 0,14 (0,26) | 0,34 (0,98) | −0,37 (1,20) | 0,33 (1,07) | −0,49 (0,89) |
| Evaluierungstätigkeit | −0,23 (0,15) | 0,01 (0,01) | −0,04 (0,03) | 0,00 (0,03) | −0,04 (0,03) | −0,04 (0,02)* |
| wirtschaftliche Entwicklung | 1,81 (3,67) | 2,11 (1,08) | 8,18 (3,10) | −2,51 (2,70) | −2,53 (1,54) | −1,17 (2,47) |
| Organisationen[a] | + | + | + | +† | + | +* |
| **Anzahl Beobachtungen** | **268** | **268** | **268** | **268** | **268** | **268** |

| | Evaluierungsgegenstand | Kontext | Wirkungszusammenhänge | Erkenntnisinteresse | Informationsquellen | Darstellung Methodik |
|---|---|---|---|---|---|---|
| **Pseudo-$R^2$/$R^2$ adjusted[b]** | **0,40** | **0,22** | **0,18** | **0,40** | **0,14** | **0,19** |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs = Gutachtende; IR = Inception Report. Dargestellt sind Koeffizienten logistischer (geordneter) Regressionen sowie Regressionen der kleinsten Quadrate mit den dazugehörigen Standardfehlern in Klammern. [a] Die Organisationen wurden einzeln in die Regressionen aufgenommen, aus Gründen der Anonymisierung werden nicht die Ergebnisse der einzelnen Organisationen abgebildet. [b] Bei einer linearen Regression wird das $R^2$ adjusted angegeben, bei einer nicht linearen das Pseudo-$R^2$. / = kein Zusammenhang getestet, da keine Hypothese vorhanden; Sternchen geben die statistische Signifikanz an: † $p < 0{,}1$; * $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$. ologit = geordnetes logistisches Regressionsmodell; logit = logistisches Modell; ols = Regressionsmodell der kleinsten Quadrate (*ordinary least squares*)*

**Tabelle 23 Ergebnisse der Regressionsanalysen (2/3)**

| | Ergebnisse-Schlussfolgerungen | Leistungsbeschreibung | Zusammenfassung | Vorhandensein Inception Report | Summenindex Berichtslegung und Methoden | Faktorindex Berichtslegung und Methoden |
|---|---|---|---|---|---|---|
| Regressionsmodell Nr. und Typ | 7 (ologit) | 8 (ologit) | 9 (logit) | 10 (logit) | 11 (ols) | 12 (ols) |
| Konflikt | −0,12 (0,29) | 0,04 (0,41) | −0,42 (0,53) | 0,34 (0,68) | −0,05 (0,05) | −0,10 (0,08) |
| Sozialkapital | −1,78 (1,87) | 0,10 (1,31) | 0,06 (1,98) | −0,88 (3,35) | −0,33 (0,13)† | −0,60 (0,24)† |
| Jahr 2020 (Pandemie) | 0,44 (0,54) | −0,16 (0,93) | 0,16 (0,09) | 1,78 (2,08) | 0,02 (0,07) | 0,01 (0,11) |
| Anzahl GAs | 0,18 (0,27) | 0,05 (0,36) | 0,92 (0,17)*** | 0,05 (0,44) | 0,06 (0,02)† | 0,06 (0,04) |
| Tagessatz externe GAs | −0,00 (0,00) | −0,00 (0,00) | 0,00 (0,00) | −0,00 (0,00) | −0,00 (0,00) | −0,00 (0,00) |
| Leistungsbeschreibung | 0,38 (0,15)* | / | −0,57 (0,27)* | −0,88 (0,93) | / | / |
| Vorhandensein IR | 0,32 (0,34) | −0,98 (1,02) | −0,09 (0,50) | / | / | 0,09 (0,14) |
| Einbindung Stakeholder*innen | 0,20 (0,21) | 0,66 (0,35)† | −0,29 (0,26) | 0,56 (0,51) | 0,08 (0,03)* | 0,10 (0,06) |
| Remote-Datenerhebung | / | / | / | / | / | / |
| GAs-Tage/ Kosten Maßnahme | −0,44 (0,15)** | −0,18 (0,15) | −0,15 (0,17) | 0,09 (0,32) | −0,03 (0,01) | −0,06 (0,02) |
| Jahr der Evaluierung | −0,08 (0,18) | 0,56 (0,18)** | 0,15 (0,19) | 0,13 (0,29) | 0,06 (0,02)* | 0,14 (0,05)* |
| Evaluierungseinheiten absolut | 0,27 (0,17) | −0,64 (0,29)* | −0,32 (0,24) | 0,26 (0,38) | −0,03 (0,04) | −0,07 (0,06) |
| Evaluierungseinheiten/ Evaluierungen | −1,32 (1,72) | −2,80 (1,93) | −2,65 (3,55) | −4,63 (4,28) | −0,72 (0,33)† | −0,85 (0,53) |
| Evaluierungseinheiten/ Organisationen | 0,34 (0,76) | 1,50 (0,86)† | 0,86 (1,33) | 0,68 (1,55) | 0,18 (0,13) | 0,26 (0,18) |
| Evaluierungstätigkeit | −0,02 (0,03) | −0,01 (0,03) | 0,08 (0,03)* | 0,01 (0,04) | −0,00 (0,00) | −0,01 (0,01) |
| wirtschaftliche Entwicklung | 1,50 (2,33) | 5,54 (5,38) | 4,18 (2,17)† | 5,66 (1,20) | 6,69 (3,79) | 4,01 (9,82) |

| | Ergebnisse-Schlussfolgerungen | Leistungsbeschreibung | Zusammenfassung | Vorhandensein Inception Report | Summenindex Berichtslegung und Methoden | Faktorindex Berichtslegung und Methoden |
|---|---|---|---|---|---|---|
| Organisationen[a] | +* | +** | + | + | +* | +* |
| **Anzahl Beobachtungen** | **268** | **268** | **251** | **199** | **268** | **268** |
| **Pseudo-$R^2$/$R^2$ adjusted [b]** | **0,13** | **0,46** | **0,31** | **0,54** | **0,71** | **0,84** |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs = Gutachtende; IR = Inception Report. Dargestellt sind Koeffizienten logistischer (geordneter) Regressionen sowie Regressionen der kleinsten Quadrate mit den dazugehörigen Standardfehlern in Klammern. [a] Die Organisationen wurden einzeln in die Regressionen aufgenommen, aus Gründen der Anonymisierung werden nicht die Ergebnisse der einzelnen Organisationen abgebildet. [b] Bei einer linearen Regression wird das $R^2$ adjusted angegeben, bei einer nicht linearen das Pseudo-$R^2$. / = kein Zusammenhang getestet, da keine Hypothese vorhanden; Sternchen geben die statistische Signifikanz an: † $p < 0,1$; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. ologit = geordnetes logistisches Regressionsmodell; logit = logistisches Modell; ols = Regressionsmodell der kleinsten Quadrate (*ordinary least squares*)*

**Tabelle 24 Ergebnisse der Regressionsanalysen (3/3)**

| | Nützlichkeit der Empfehlungen | Kapazitätsentwicklung | Einbezug Gutachtende Partnerland | Einbindung Stakeholder*innen |
|---|---|---|---|---|
| **Regressionsmodell Nr. und Typ** | **13 (ols)** | **14 (ologit)** | **15 (logit)** | **16 (ologit)** |
| Konflikt | / | −0,35 (0,37) | 0,22 (0,45) | −0,03 (0,47) |
| Sozialkapital | / | / | / | −0,75 (1,61) |
| Jahr 2020 (Pandemie) | / | 0,55 (0,58) | −0,08 (0,51) | 0,39 (0,43) |
| Anzahl GAs | / | 0,38 (0,35) | 0,39 (0,43) | / |
| Tagessatz externe GAs | / | −0,00 (0,00) | / | −0,00 (0,00) |
| Leistungsbeschreibung | 0,02 (0,08) | / | / | / |
| Qualitätssicherung IR | 0,26 (0,11)* | −1,07 (0,48)* | / | 0,31 (0,50) |
| Einbindung Stakeholder*innen | / | / | / | / |
| Remote-Datenerhebung | / | −0,88 (0,40)* | 0,04 (0,32) | / |
| GAs-Tage/Kosten Maßnahme | / | / | / | / |
| Jahr der Evaluierung | / | / | / | / |
| Evaluierungseinheiten absolut | 0,02 (0,05) | −0,27 (0,21) | 0,10 (0,20) | −0,11 (0,26) |
| Evaluierungseinheiten/Evaluierungen | 0,36 (0,46) | −4,02 (3,44) | −3,38 (4,90) | −2,20 (3,43) |
| Evaluierungseinheiten/Organisationen | 0,28 (0,49) | 0,24 (0,90) | −0,81 (1,29) | 1,30 (1,14) |
| Evaluierungstätigkeit | −0,00 (0,01) | 0,02 (0,02) | 0,02 (0,02) | −0,02 (0,03) |
| wirtschaftliche Entwicklung | 1,81 (5,40)* | −3,63 (2,85) | −1,38 (1,65) | −4,31 (3,14) |
| Organisationen[a] | + | +† | +† | +* |
| **Anzahl Beobachtungen** | **268** | **268** | **268** | **268** |
| **Pseudo-$R^2$/$R^2$ adjusted[b]** | **0,39** | **0,43** | **0,42** | **0,41** |

*Quelle: DEval, eigene Darstellung*

*Anmerkung: GAs = Gutachtende; IR = Inception Report. Dargestellt sind Koeffizienten logistischer (geordneter) Regressionen sowie Regressionen der kleinsten Quadrate mit den dazugehörigen Standardfehlern in Klammern. [a] Die Organisationen wurden einzeln in die Regressionen aufgenommen, aus Gründen der Anonymisierung werden nicht die Ergebnisse der einzelnen Organisationen abgebildet. [b] Bei einer linearen Regression wird das $R^2$ adjusted angegeben, bei einer nicht linearen das Pseudo-$R^2$. / = kein Zusammenhang getestet, da keine Hypothese vorhanden;* Sternchen geben die statistische Signifikanz an: *† $p < 0,1$; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. ologit = geordnetes logistisches Regressionsmodell; logit = logistisches Modell; ols = Regressionsmodell der kleinsten Quadrate (ordinary least squares)*

## Tabelle 25 Odds Ratios für die signifikanten Effekte

| | | SC B & M | | | | | | | | | | Indices | | SC N | SC P, U & F | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Faktor | Proxy | Evaluierungs-gegenstand | Kontext* | Wirkungs-zusammenhänge | Erkenntnis-interesse | Informations-quellen | Darstellung Methodik | Kohärenz | Leistungs-beschreibung | Zusammenfassung GAs | Vorhandensein IR | Summenindex SC B & M | Faktorindex SC B&M | Empfehlungen* | Kapazitäts-entwicklung | GAs Partnerland | Einbindung Stakeholder*innen |
| | Regressions-modell Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Fragilität | Konflikt | – | – | –† | – | – | – | – | – | – | – | – | – | | – | – | – |
| Kultur | Sozialkapital | + | + | + | + | + | –* | + | + | + | + | –† | –† | | | | + |
| Pandemie | Jahr 2020 | – | – | – | – | – | – | – | – | – | – | – | – | | – | – | – |
| Kompetenz GAs | Anzahl GAs | + | + | + | + | + | +* | + | + | +*** | + | +† | + | | + | + | |
| | Tagessatz GAs | + | + | + | + | + | + | + | + | + | + | + | + | | + | | + |
| Qualitätssicherung | Leistungsbeschreibung | + | + | + | + | + | + | +* | | –* | + | | | + | | | |
| | Vorhandensein IR | + | + | + | + | +† 1,82 | + 1,8 | + | + | + | | | + | +* | –* | | + |
| | Einbindung Stakeholder*innen | + | +** | +† | + | +*** 1,70 | + | + | +† 1,93 | + | + | +* | + | | | | |
| Remote-Datenerhebung | | | | | | | – | | | | | | | | –* | + | |
| Kosten | GA-Tage/ Kosten Maßnahme | + | + | + | + | + | + | –** | + | + | + | + | + | | | | |
| Jahr der Evaluierung | | + | +* | +*** | +** 1,78 | –* | +* | + | +** 1,74 | + | + | +* | +* | | | | |
| Evaluierungsein-heiten/-stellen absolut | | + | + | + | + | + | – | + | –* | + | + | + | + | + | + | + | + |

| | | SC B & M | | | | | | | | | | Indices | | SC N | SC P, U & F | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Faktor | Proxy | Evaluierungs-gegenstand | Kontext* | Wirkungs-zusammenhänge | Erkenntnis-interesse | Informations-quellen | Darstellung Methodik | Kohärenz | Leistungs-beschreibung | Zusammenfassung GAs | Vorhandensein IR | Summenindex SC B & M | Faktorindex SC B&M | Empfehlungen* | Kapazitäts-entwicklung | GAs Partnerland | Einbindung Stakeholder*innen |
| | Regressions-modell Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Evaluierungsein-heiten/-stellen relativ zu Evaluierungen | | + | + | + | + | + | + | + | + | + | + | –† | + | + | + | + | + |
| Evaluierungsein-heiten/-stellen relativ zur Organisation | | + | + | + | + | + | + | + | +† 4,49 | + | + | + | + | + | + | + | + |
| Evaluierungs-tätigkeit | | + | + | + | + | + | –* | + | + | +* | + | + | + | + | + | + | + |
| Evaluierungs- und Lernkultur | Organisation[1] | + | + | + | +† | + | +* | +* | +** | + | + | +* | +* | + | +† | +† | +* |

*Quelle: DEval, eigene Darstellung*

*Anmerkungen: GAs = Gutachtende; SC = Standardcluster; B & M = Berichtslegung und Methoden; N = Nutzbarkeit; P, U & F = Partizipation Unabhängigkeit und Fairness. blau = positiver Zusammenhang; orange = negativer Zusammenhang; grau = kein statistisch signifikanter Zusammenhang; leeres Feld = nicht getestet. Sternchen geben die statistische Signifikanz an: † p < 0,1; * p < 0,05; ** p < 0,01; *** p < 0,001. Das Odds Ratio der (geordneten) logistischen Regressionen wurde als Maß für die Effektstärke herangezogen (Chen et al., 2010) und somit für alle Regressionen bis auf die Modellnummern 3, 11, 12 und 13. Die Effektstärke gilt als klein, wenn das Odds Ratio zwischen 1,68 und 3,45 liegt und als mittel, wenn das Odds Ratio zwischen 3,46 und 6,70 liegt (Chen et al., 2010). Odds Ratios, die kleiner als 1,68 sind, werden nicht abgebildet.*

# 5. LITERATUR

**Allwright, S. (2022),** „What is a good F1 score? Simply explained“, *Stephen Allwright*, *https://stephenallwright.com/good-f1-score/* (zugegriffen 13.09.2022).

**Backhaus, K. et al. (2015),** *Fortgeschrittene multivariate Analysemethoden: eine anwendungsorientierte Einführung*, Springer Gabler, Berlin, 3. Auflage.

**Brown, T. A. (2006),** *Confirmatory factor analysis for applied research*, Guilford Press, New York.

**Bundestag (2019),** *Gesetz über die Feststellung des Bundeshaushaltsplans für das Haushaltsjahr 2020 (Haushaltsgesetz)*, S. 2890–2910.

**Caspari, A. (2010),** *„Lernen aus Evaluierungen. Meta-Evaluation & Evaluationssynthese von InWEnt-Abschlussevaluierungen 2009“*, Internationale Weiterbildung und Entwicklung, Bonn.

**Caspari, A. (2011),** *„Meta-Evaluation & Evaluationssynthese 2011 - Hauptbericht“*, o.V. Frankfurt am Main.

**Chen, H. et al. (2010),** „How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies“, *Communications in Statistics - Simulation and Computation*, Vol. 39, Nr. 4, S. 860–864.

**DeGEval (2016),** *„Standards für Evaluation“*, Gesellschaft für Evaluation e.V., Mainz.

**DeGEval (2021),** *„DeGEval Beitrittsantrag“*, DeGEval – Gesellschaft für Evaluation e.V., Mainz.

**Döring, N. und J. Bortz (2016),** *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, Springer, Berlin, 5. Auflage.

**Dumm, S. (2018),** „Forschungsinfrastruktur und Forschungsdesign. Text-Mining-gestützte Inhaltsanalyse qualitativer Daten im Forschungsprojekt ‚Postdemokratie und Neoliberalismus’“, in Behnke, J., A. Blätte, K.-U. Schnapp und C. Wagemann (Hrsg.), *Computational Social Science: Die Analyse von Big Data*, Nomos, Baden-Baden, S. 191–210.

**Faulbaum, F. (2019),** „Total Survey Error“, in Baur, N. und J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung*, Springer VS, Wiesbaden, S. 505–521.

**FES (2015),** *„Metaevaluierung: Evaluierungen in der Internationalen Entwicklungszusammenarbeit der Friedrich-Ebert-Stiftung“*, Friedrich-Ebert-Stiftung, Berlin.

**Freimann, I. et al. (2016),** *„Meta-Evaluierung der Projektevaluierungen (PEV)“*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

**Freimann, I. et al. (2017),** *„Querschnittsauswertung (QSA) von Projektevaluierungen (PEV) 2016 Meta-Evaluierung“*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

**Grüne, A. (2020),** „What Is a Good F1 Score?“, *Inside GetYourGuide*, *https://inside.getyourguide.com/blog/2020/9/30/what-makes-a-good-f1-score* (zugegriffen 13.09.2022).

**Gwet, K. L. (2015),** *„On the Krippendorff ’s alpha coefficient“*, Advanced Analytics, LLC, Gaithersburg, MD.

**Hageboeck, M. et al. (2013),** *„Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012“*, United States Agency for International Development, Washington, D. C.

**HTSPE LTD. (2011),** *„Mid-term Meta Evaluation of IPA Assistance Evaluation Report“*, EU-Kommission, Brüssel.

**Koy, J. et al. (2016),** *„Meta-Evaluierung der Projektevaluierungen aus den Jahren 2014-2015“*, Misereor, Aachen.

**Krippendorff, K. (2012),** *Content analysis: an introduction to its methodology*, SAGE, Thousand Oaks, CA, 2. Auflage.

**Li, C. (2013),** „Little's Test of Missing Completely at Random", *The Stata Journal*, SAGE, Vol. 13, Nr. 4, S. 795–809.

**Lücking, K. et al. (2015),** *„Evaluierungspraxis in der deutschen Entwicklungszusammenarbeit. Umsetzungsmonitoring der letzten Systemprüfung und Charakterisierung wesentlicher Elemente"*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, Bonn.

**Maas, C. J. M. und J. J. Hox (2005),** „Sufficient Sample Sizes for Multilevel Modeling", *Methodology*, Vol. 1, Nr. 3, S. 86–92.

**Mauthofer, T. und S. Silvestrini (2018),** *„Meta-Evaluation of 33 Evaluation Reports of World Vision Germany"*, World Vision Germany, Saarbrücken.

**McKenzie, D. (2020),** *„An overview of multiple hypothesis testing commands in Stata"*, World Bank, Washington, D. C.

**Niekler, A. (2018),** *Automatisierte Verfahren für die Themenanalyse Nachrichtenorientierter Textquellen*, Methoden und Forschungslogik der Kommunikationswissenschaft, Herbert von Halem Verlag, Köln.

**Nili, A. et al. (2017),** *„A critical analysis of inter-coder reliability methods in information systems research"*, University of Tasmania, Hobart.

**Noltze, M. et al. (2018),** *„Meta-Evaluierung von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit"*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, Bonn.

**OECD DAC (2010),** *„Qualitätsstandards für die Entwicklungsevaluierung"*, Organisation for Economic Cooperation and Development, Development Assistance Committee, Paris.

**Queiroz de Souza, A. (2017),** *„Meta-Evaluation and Analysis of Project Evaluations 2016"*, Welthungerhilfe, Bielefeld.

**RKI (2009),** „Was ist eine Pandemie?", *Robert Koch-Institut*, *https://www.rki.de/SharedDocs/FAQ/Pandemie/FAQ18.html* (zugegriffen 01.08.2022).

**Silvestrini, S. und S. Bäthge (2019),** *„Meta-Evaluation of ADA Project and Programme Evaluations - Executive Summary"*, Agentur der Österreichischen Entwicklungszusammenarbeit, Saarbrücken/Wien.

**Silvestrini, S. et al. (2018),** *„Meta-evaluation of Project and Programme Evaluations in 2015–2017"*, Ministry for Foreign Affairs of Finland, Saarbrücken/Helsinki.

**UCDP (2006),** *„Definitions, sources and methods for Uppsala Conflict Data Program Battle-Death estimates"*, Uppssala Conflict Data Program.

**UNFPA (2020),** *„UNFPA Evaluation Office - Assessing the quality of developmental evaluations at UNFPA"*, United Nations Population Fund, New York.

**Väth, S. J. et al. (2022),** *„Evaluation: Metaevaluation of MFA's Project and Programme Evaluations in 2017-2020"*, Ministry for Foreign Affairs of Finland, Saarbrücken/Helsinki.

**Weltbank (2022),** „World Developement Indicators", *https://datatopics.worldbank.org/world-development-indicators/* (zugegriffen 01.08.2022).

**WVS (2022),** *„2017-2021 World Values Survey Wave 7 Master Survey Questionnaire"*, World Values Survey, Wien.

# 6. KODIERLEITFADEN

*Kurzbeschreibung: Der Kodierleitfaden enthält eine Beschreibung der Qualitätskriterien, die in der Meta-Evaluierung in den Evaluierungsdokumenten der beteiligten Organisationen untersucht wurden (Evaluierungsberichte, Berichtsanhänge, Inception Reports, Leistungsbeschreibung, sonstige Dokumente).*

## Allgemeine Informationen

Die Qualitätskriterien sind entlang der im Evaluierungsbericht abgebildeten Nummerierung aufgeführt. Im Kodierleitfaden werden für jedes Qualitätskriterium nachfolgende Punkte aufgeführt: Beschreibung des Qualitätskriteriums, Bewertungsskala, untersuchte Dokumente, Kodierregeln, Ankerbeispiele und Informationen dazu, ob das Qualitätskriterium in die Onlinebefragung integriert war.

*Bewertungsskala:* Die Einordnung in die jeweilige Bewertungsskala (entweder 2- oder 4-stufig: 1 und 4 oder 1, 2, 3, 4) wird für jedes Qualitätskriterium detailliert beschrieben.

*Dokumente:* Für die Kodierung wurden der Bericht, Berichtsanhänge, Terms of References, Inception Report und sonstige Dokumente, die nicht Teil der offiziellen Berichtsanhänge waren, herangezogen. Leistungsbeschreibungen und Inception Reports konnten Teil der regulären Berichtsanhänge sein. Ob ein Dokument Teil der regulären Berichtsanhänge ist oder nicht, war daran erkenntlich, ob die jeweiligen Anhänge im Inhaltsverzeichnis des Berichts aufgeführt wurden – ansonsten galten diese als Teil der sonstigen Dokumente. Wenn für die jeweiligen Qualitätskriterien neben den Berichtsanhängen weitere Anhänge kodiert wurden, wurde in der Spalte „Dokumente“ „sonstige Dokumente“ aufgeführt – wenn dort nur „Berichtsanhänge“ genannt wurden, wurden sonstige weitere Dokumente nicht kodiert.

*Kodierregeln:* Hier werden Details aufgelistet, die eine übereinstimmende Kodierung ermöglichen.

*Ankerbeispiele:* Ankerbeispiele wurden – wenn vorhanden – aufgenommen, um den Kodierenden Klarheit im Kodierungsprozess zu ermöglichen.

*Onlinebefragung*: Sofern nicht ausreichend Informationen zu einzelnen Qualitätskriterien in der Inter-Kodierenden-Phase[31] ermittelt werden konnten, wurden diese über eine Onlinebefragung der Verantwortlichen der Evaluierungseinheiten/-stellen der beteiligten Organisationen erfragt. Die Inter-Kodierenden-Phase stellte die erste Phase der Kodierung dar, in der jedes Qualitätskriterium in 10 Prozent der Evaluierungen (N = 30 Evaluierungen) von allen Kodierenden kodiert wurde (Döring und Bortz, 2016). In dieser Phase wurde die Anwendung eines Qualitätskriteriums mit „−99“ kodiert, wenn keine Informationen in der Evaluierung identifiziert wurden. Im Anschluss an diese Phase wurde überprüft, ob die Informationen bei mehr als 24 Evaluierungen und mindestens neun Organisationen durchgehend nicht kodiert werden konnten. Wenn dies der Fall war, wurden die Qualitätskriterien in die Onlinebefragung integriert. Die Fragen werden an der entsprechenden Stelle (in der Tabellenzeile: „Onlinebefragung“ jedes Qualitätskriteriums) aufgeführt. Entsprechend wurde bei diesen Qualitätskriterien sowohl die Operationalisierung für die Kodierung der Evaluierungsdokumente als auch die Frage in der Onlinebefragung aufgeführt.

31

### 1. Beschreibung des Evaluierungsgegenstands

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteur*innen (politische Partner*innen und/oder Träger) der EZ-Maßnahme genannt sind. | Keiner der drei Punkte ist in der Evaluierung genannt. | Einer der drei Punkte ist in der Evaluierung genannt. | Zwei der drei Punkte sind in der Evaluierung genannt. | Alle Punkte sind in der Evaluierung genannt. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | | | |
| Kodierregeln:<br>1. Einer der drei Punkte ist genannt (Bewertungsstufe 2) heißt, dass entweder Ziel(e), die Zielgruppe oder relevante Akteur*innen genannt sind.<br>2. Zielgruppe (target group): Hier geht es um die Zielgruppe der EZ-Maßnahme (zum Beispiel Landwirte, Unternehmen, Mütter, Flüchtlinge).<br>3. „Genannt“ bedeutet nicht, dass die Zielgruppe(n) oder Akteur*innen als Teil der Analyse zum Beispiel umfassend beschrieben werden müssen. | | | | |
| Ankerbeispiele:<br>• Ziel(e): „Project objective: The framework conditions for refugees and host communities in the neighbouring countries of Syria and Iraq have improved.”<br>• Zielgruppe(n): „Target group(s): The indirect target groups are Syrian refugees, internally displaced persons (IDPs).“<br>• Relevante Akteur*innen: „Implementing organisations (in the partner countries): Spanish Agency for International Development Cooperation (AECID), Expertise France (EF).“ | | | | |
| Onlinebefragung: Nein | | | | |

## 2. Beschreibung des Kontexts der Entwicklungsmaßnahme

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn der Kontext der EZ-Maßnahme beschrieben wird (zum Beispiel politische, ökonomische, soziale, [agrar-]ökologische, [agrar-]klimatische, gesetzliche, technologische, historische, kulturelle Kontextfaktoren). | Der Kontext der EZ-Maßnahme wird nicht beschrieben oder es wird ein Kontextfaktor beschrieben, aber nur mit 1 bis 50 Wörtern. | Der Kontext der EZ-Maßnahme wird nicht hinreichend beschrieben, das heißt nur ein Kontextfaktor und dieser hat 51 bis 150 Wörter. | Der Kontext der EZ-Maßnahme wird hinreichend beschrieben, das heißt mindestens zwei Kontextfaktoren mit jeweils mehr als 100 Wörtern oder ein Kontextfaktor mit mehr als 150 Wörtern. | Der Kontext der EZ-Maßnahme wird umfassend beschrieben, das heißt mindestens zwei Kontextfaktoren mit jeweils mehr als 200 Wörtern. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Hier sind explizit die Rahmenbedingungen/der Kontext der EZ-Maßnahme relevant und nicht der Evaluierung. Jedoch können Beschreibungen des Kontexts der Entwicklungsmaßnahme in der Beschreibung der Evaluierung zu finden sein.
2. Mit Kontext sind die Rahmenbedingungen gemeint, in denen die EZ-Maßnahme verortet ist, die Darstellung beziehungsweise Beschreibung der EZ-Maßnahme selbst ist nicht gefordert. Die Rahmenbedingungen können generell auf drei Ebenen dargestellt werden: der lokalen (zum Beispiel Arbeitslosigkeit und Wilderei), der regionalen/nationalen (mangelnder politischer Wille zur Wildereibekämpfung) und der globalen Ebene (globale Nachfrage nach Produkten, die aus Wilderei stammen).
3. Wenn eine Evaluierung die Relevanz der EZ-Maßnahme (beziehungsweise über das OECD-DAC-Kriterium Relevanz) untersucht, wird dabei oft der Kontext untersucht. Dies kann zur Bewertung des Qualitätskriteriums herangezogen werden.
4. Zur schnellen Zählung der Wörter kann mit einem Word-Dokument und der Funktion „Wörter zählen“ gearbeitet werden, indem die Sätze zum Kontext in das Dokument kopiert werden.
5. Die Geschichte der Organisation im Allgemeinen oder in dem spezifischen Partnerland ist keine gültige Beschreibung des Kontexts, selbst wenn diese unter der Überschrift „Kontext“ ausgeführt wird. Diese ist nur bewertungsrelevant, wenn dabei konkrete gesetzliche oder politische Faktoren aufgezählt werden, die einen Einfluss auf die EZ-Maßnahme haben. Nur Textstellen, die diese Faktoren beinhalten, werden gezählt – nicht der gesamte Textabschnitt.

Ankerbeispiel (vollständig erfüllt):

- „Family hardship experienced by communities, is borne out by not only the available data and anecdotal information, but also by what is clearly visible in the parishes of intervention. Poor infrastructure and planning have resulted in substandard transport networks, housing, roads and street lighting, as well as water and sanitation systems. The current national labour statistics indicate that the official unemployment rate is 27.6% (expanded definition takes this to 38%) – this is a 0.5% increase since the fourth quarter of 2018. The labour force has further decreased, as the number of discouraged work-seekers increase. There is a decline in employment in all sectors (Formal sector, Informal sector, Agriculture and Private households). Currently there are approximately 10.3 million persons aged 15–24 years. Of this group, the percentage of young persons who are not in employment, education or training (NEET) has increased to 33.2% (3.4 million). Of the 20.3 million young people aged 15-34 years, 40.7% are not in employment, education or training […].“

Onlinebefragung: Nein

### 3. Berücksichtigung des Kontexts bei Ergebnissen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn der Einfluss des Kontexts auf die Evaluierungsergebnisse berücksichtigt wird. | Der Einfluss des Kontexts auf keines der Ergebnisse ist beschrieben. | Der Einfluss des Kontexts auf die Ergebnisse ist nicht hinreichend beschrieben. | Der Einfluss des Kontexts auf die Ergebnisse ist hinreichend beschrieben. | Der Einfluss des Kontexts auf die Ergebnisse ist umfassend beschrieben. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Für die Betrachtung des Ergebniskapitels liegt der Fokus auf dem Bereich der Effektivität der EZ-Maßnahme (OECD-DAC-Kriterium). Effizienz, Impact, Nachhaltigkeit und Relevanz werden ausgeschlossen.
2. Wenn keine Ergebnisse zum OECD-DAC-Kriterium Effektivität vorliegen, ist dies nicht für die Bewertung des Qualitätskriteriums relevant, soweit Ergebnisse in einer anderen Form in Bezug auf die Effektivität der Zielerreichung der EZ-Maßnahme vorliegen (auch ohne die Behandlung der OECD-DAC-Kriterien). Nicht beachtet werden Empfehlungen und Schlussfolgerungen.
3. Die Ergebnisse müssen nicht quantifiziert sein, das heißt, es muss nicht für jedes Ergebnis die Effektivität bewertet werden, sondern es geht um einen allgemeinen/übergreifenden Eindruck.
4. Der Unterschied zwischen hinreichend und umfassend liegt in der Ausführlichkeit und Länge der Beschreibung des Kontextfaktors auf die Evaluierungsergebnisse. Vollständig erfüllt ist eine Berücksichtigung des Kontexts, wenn beispielsweise ein gesondertes Kapitel bei den Ergebnissen aufgeführt wird, in dem gezielt die hemmenden und fördernden externen Faktoren analysiert und/oder aufgeführt werden, die die Zielerreichung der EZ-Maßnahme beeinflusst haben.
5. Mögliche Stichwortsuche: context, factors of success or failure oder Kontext im Ergebnisteil des Berichts zu Effektivität oder auch im Abschnitt zu limitations, wenn dieser sich auf Effektivität bezieht.

Ankerbeispiele (vollständig erfüllt):

- „The country context in each case is different and mostly supportive for development of ALE as a policy segment in Ukraine, while in Belarus and Moldova there are still significant challenges to achieve the program goals, explained by lack of a clear perspective for local funding. Another challenging aspect for Moldova is selecting more objective-oriented and capable partners to focus on strategic and long-term goals in ALE.“
- „External factors [:] There were several external factors supporting and hindering the achievement of the intended results.
  External success factors included:
  - Jordan's commitment to international conventions and the existence of NBC to monitor the implementation of the NBSAP.
  - The existence of other donor initiatives to complement the project's efforts.“

Onlinebefragung: Nein

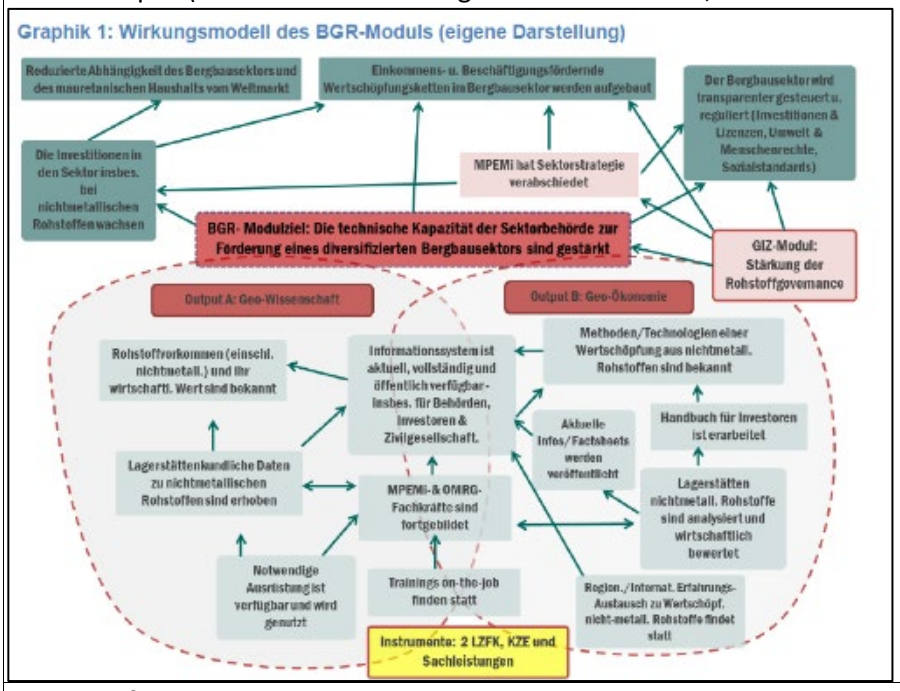
### 4. Darstellung der Wirkungszusammenhänge

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn bei der Beschreibung der intendierten Wirkungszusammenhänge der EZ-Maßnahme zwischen den Wirkungsstufen Input-Output-Outcome-Impact unterschieden wird. | Für keines der Ziele der EZ-Maßnahme werden Wirkungszusammenhänge beschrieben. | Für einen Teil oder alle der Ziele der EZ-Maßnahme werden Wirkungszusammenhänge beschrieben, aber diese sind unvollständig, das heißt, sie enthalten nicht alle Stufen der Kette Input-Output-Outcome-Impact. | Es werden für mindestens ein Ziel der EZ-Maßnahme, aber nicht für alle Ziele vollständige Wirkungszusammenhänge (Input-Output-Outcome-Impact) beschrieben. Detaillierte Wirkungshypothesen sind nicht für jede Ebene beschrieben. | Für alle Ziele der EZ-Maßnahme werden alle Wirkungszusammenhänge (von Input bis Impact) beschrieben. Außerdem wird bei mindestens einer Wirkungskette für jede Ebene eine Wirkungshypothesen beschrieben. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | | | |

Kodierregeln:

1. Schritt: Vorerst muss das Ziel (oder die Ziele) der EZ-Maßnahme festgestellt werden. Diese finden sich zumeist zu Beginn der Evaluierung oder im Ergebniskapitel „Effektivität“. Bei einzelnen Organisationen werden die Ziele der EZ-Maßnahme neben den Projektzielen auch Komponenten-, Modul- oder Programmziele genannt.
2. Schritt: Suche nach textlicher Beschreibung der Wirkungszusammenhänge beziehungsweise -stufen für diese(s) Ziel(e) der EZ-Maßnahme. Hilfreich könnte hier sein, nach Wirkmodellen zu suchen: Je nach Ansatz oder Organisation können sich die Begrifflichkeiten unterscheiden. Häufig genutzte Termini sind Wirkmodell (*results model*) oder *Theory of Change*, Ergebnismodell, Ergebnismatrix, Programmtheorie, Logisches Wirkmodell (linear – oft auf Programm- oder Projektebene).
3. Eine grafische Darstellung ist nur gültig, wenn sie textliche Beschreibungen und Erklärungen zu den Wirkungszusammenhängen beinhaltet. Eine 4 kann nur vergeben werden, wenn sowohl die vier Stufen als auch die Zusammenhänge zwischen den Stufen für jeden Input, Output, Outcome und Impact mit Text erklärt werden. Pfeile ohne Text sind nicht ausreichend.
4. Für dieses Qualitätskriterium ist die Darstellung der Wirkungslogik in der Evaluierung relevant (die in den meisten Fällen bereits vom BMZ und der Organisation zu Beginn der EZ-Maßnahme festgelegt wurde, jedoch von dem Gutachtenden-Team kommentiert werden kann), nicht die empirische Überprüfung der Wirkungslogik, die als Teil der Analyse einer Evaluierung durchgeführt wird.
5. Erklärung der Wirkungsstufen Inputs: finanzielle, personelle, materielle, administrative, organisationale und andere Mittel, die in eine EZ-Maßnahme investiert werden, um dessen Ziele zu erreichen. Outputs: Die zählbaren Dienstleistungen und Produkte eines zu evaluierenden Programms, über die Wirkungen erreicht werden sollen. Outcome: Auswirkungen des evaluierten Gegenstands auf Ebene der Zielgruppe(n) des Gegenstands. Impacts: Einwirkungen eines Evaluationsgegenstands außerhalb seiner intendierten Reichweite, also auf Objekte (Systeme, Institutionen, Organisationen, Personen, Produkte etc.), die durch den evaluierten Gegenstand mit verursacht sind.

Ankerbeispiel (Merkmal Beschreibung ist hier nicht erfüllt, da kein Fließtext zu den Pfeilen besteht):




Onlinebefragung: Nein

### 5. Beschreibung des Erkenntnisinteresses

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn 1. Zweck(e), 2. Ziel(e) und 3. Evaluierungsfrage(n) der Evaluierung genannt werden. | Keiner der drei Punkte wird genannt. | Einer der Punkte wird genannt. | Zwei der Punkte werden genannt. | Alle Punkte werden genannt. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Die Begriffe Zwecke (*purpose*) und Ziele (meist *objective*, auch *goals*) werden in den Evaluierungen häufig nicht einheitlich verwendet; zum Teil werden sie auch synonym verwendet. Wichtig für die Bewertung: Wir legen in der Meta-Evaluierung eine einheitliche Definition zugrunde, welche bei der Kodierung anzuwenden ist.
2. Zwecke beschreiben das „Warum“ einer Evaluierung (zum Beispiel Optimierung von Prozessen, Verbesserung, Entscheidungsfindung, Rechenschaftslegung oder organisationales Lernen), das heißt, sie nennen, was mit den erzeugten Evaluierungsergebnissen (ggf. übergeordnet) erreicht oder verändert werden soll.
3. Ziele beschreiben das „Was“ der Evaluierung und formulieren die direkten und unmittelbaren Ziele der Evaluierung für die EZ-Maßnahme aus (zum Beispiel „The objective is to evaluate the effectiveness of the programm“ oder „Improving Maternal and Newborn Care in South Cambodia“).
4. Hinweise zur Bewertung von Evaluierungsfragen: Evaluierungsfragen müssen entweder a) explizit als solche gekennzeichnet sein (zum Beispiel mit einem Fragezeichen) oder b) es muss einen expliziten Hinweis auf diese geben, wenn sie nicht als Fragen, sondern im Fließtext enthalten sind. Forschungsfrage(n) (*research questions*) statt Evaluierungsfrage(n) (*evaluation question[s]*) gelten auch, solange die Fragen sich auf die Evaluierung beziehen. Evaluierungsfragen finden sich häufig in der Evaluierungsmatrix.
5. Hinweise zur Bewertung von Ziel und Zweck: Allgemeine Hinweise auf Ziele und/oder Zwecke von Evaluierungen innerhalb einer Organisation sind nicht gültig (zum Beispiel „Generally, evaluations of XX cooperation projects serve for accountability towards the funding BMZ and the cooperation partners, for the institutional learning process within the project and for quality assurance and improvement of the Technical Cooperation.“). Die Nennung des Ziels einer Vormission einer Evaluierung (zum Beispiel: „Das Ziel der Mission bestand darin, das TB-Regionalprogramm entlang der fünf DAC-Kriterien zu bewerten.“) ist nicht gültig als Evaluierungsziel.

Ankerbeispiele (vollständig erfüllt)

- Zweck: „The evaluation is intended to rate the success of the current programme in the context of the unique and complex set-up of the Qudra programme as a regional multi-donor action.“
- Ziele: „The evaluation objectives are oriented on Qudra’s programme objective, ‘The framework conditions for refugees and host communities in the neighbouring countries of Syria and in Iraq have improved’.“
- Evaluierungsfrage(n): „What has the project done to ensure that the results can be sustained in the medium to long term by the partners themselves? […] In what way are advisory contents, approaches, methods or concepts of the project anchored/institutionalised in the (partner) system?“

Onlinebefragung: Nein

## 6. Nachvollziehbarkeit der Informationsquellen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die in der Evaluierung genutzten Informationsquellen ausreichend detailliert im Evaluierungsbericht dargestellt werden. | Es ist im Evaluierungsbericht nicht dargestellt, welche Informationsquellen genutzt werden. | Es ist im Evaluierungsbericht nicht hinreichend dargestellt, welche Informationsquellen genutzt werden. | Es ist im Evaluierungsbericht hinreichend dargestellt, welche Informationsquellen genutzt werden. | Es ist im Evaluierungsbericht umfassend dargestellt, welche Informationsquellen genutzt werden. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Informationsquellen sind zum Beispiel Projektdokumente, Sekundärliteratur, Sekundärdaten, befragte oder interviewte Personen (zum Beispiel im Rahmen von Surveys oder Interviews), Monitoring- oder andere Projektdaten.
2. Tipp: Zunächst sollte nach einem Kapitel zur Datengrundlage oder Methodik gesucht werden. Hierbei können auch Berichtsanhänge (*annexes*) bewertungsrelevante Informationen liefern.
3. *Nicht erfüllt*: Es ist nicht ersichtlich, welche Informationsquellen in der Evaluierung genutzt werden.
4. *Eher nicht erfüllt*: Informationsquellen werden nur rudimentär darstellt, das heißt, es werden nur Oberbegriffe für die Art der Informationsquellen verwendet (zum Beispiel: „Es wurden Sekundärdaten verwendet.“, „Es wurden Interviews durchgeführt.“).
5. *Eher erfüllt*: Informationsquellen werden mit detaillierteren Informationen ergänzt. Beispielsweise werden nicht nur Oberbegriffe wie „Sekundärdaten“ oder „Interviews“ verwendet, sondern die Informationsquellen um weitere Informationen angereichert (zum Beispiel: „Es wurden Interviews mit der Zielgruppe durchgeführt“ oder „Projektabschlussberichte der EZ-Maßnahme ausgewertet“ oder „Sekundärdaten des Landwirtschaftsministeriums in Land X ausgewertet“).
6. *Vollständig erfüllt*: Eine der beiden Bedingungen ist erfüllt:
   a) die Informationsquellen werden im Fließtext umfassend und ausführlich dargestellt (zum Beispiel „Es wurden die Baselinedaten aus dem Jahr 2017 verwendet, die von der Organisation X erhoben wurde. Außerdem wurden der Dorfvorsteher und die Leistungsempfänger in drei betroffenen Dörfern der Region X befragt.“) oder
   b) im Evaluierungsbericht oder im Annex des Evaluierungsberichtes werden Informationsquellen gelistet. Die Liste kann zum Beispiel Interviewpartner*innen mit weiterführenden Merkmalen aufführen (zum Beispiel Funktion oder Beruf) oder verwendete Projektdokumente (siehe Ankerbeispiel).
7. Aus Datenschutzgründen kann die unter 6. genannte Liste der Interviewpartner*innen oder der Dokumente in manchen Berichten oder Berichtsanhängen nicht enthalten sein. Falls ersichtlich ist, dass diese Liste vorliegt und nur aus den oben genannten Gründen nicht im Bericht oder den Berichtsanhängen enthalten ist, würde dies für die Bewertung jedoch gelten.
8. In diesem Fall kann auch der Inception Report kodiert werden, da dieser aufgrund der Nichtveröffentlichung andere Datenschutzvoraussetzungen hat.
9. Es kann der Fall auftreten, dass es mehrere Informationsquellen gibt (zum Beispiel Interviewpartner*innen und Projektdokumente) und diese unterschiedlich bewertet werden (zum Beispiel Interviewpartner*innen mit 2 und Projektdokumente mit 3. Es sollte dann der Mittelwert bestimmt und dieser für die Bewertung verwendet werden.
10. Es kann vorkommen, dass Informationsquellen zwei- oder mehrfach im Bericht dargestellt werden. In diesen Fällen wird jeweils die Textstelle mit der höchsten Bewertung für die Gesamtbewertung herangezogen.
11. Mögliche Stichwortsuche: *Interviews, Gespräche, Fokusgruppen sowie alle weiteren unter 1. genannten Informationsquellen*

Ankerbeispiel (vollständig erfüllt):

| Basic document | Is available *(Yes/No)* | Estimation of actuality and quality | Relevant for OECD/ DAC criterion: |
|---|---|---|---|
| Proposals of project and of overarching programme (28.08.2014) | Yes | Available | All OECD/DAC criteria |
| Modification offer (form for justification) (15.08.2018) | Yes | Available | All OECD/DAC criteria |
| Context, political/economic analyses/capacity assessments on social context | No | - | n/a |
| Peace and conflict assessment (PCA) (June 2015); Integr. Context and human rights analysis (PCA matrix; 7.2017); Gender analyses (6.2018, updated from 2016), environmental and climate assessments, UKP checklist (2015) | Yes | Available | Relevance, impact |
| Report: Programme Report (Oct. 2016-Sep. 2017); Special Report on Project (May 2017); Project Progress (Oct. 2017-Sep. 2018); Programme Report (27.12.2018), Progress Report Project and Programme (Jan.-Dec. 2019); Project Achievements and Lessons (Dec. 2019) | Yes | Available | All OECD/DAC criteria |
| Project evaluation report of predecessor project (March 2016) | Yes | Available | Impact, sustainability |
| Country Strategy BMZ 2018-2024 (May 2018) | Yes | Available | Relevance |
| National strategies/plans:<br>National Park Act revised 2002<br>National Environment Act 2004<br>Wildlife Policy 1998 & 2007<br>Wildlife Conservation Act 2009<br>Strategy to Combat Poaching and Illegal Wildlife Trade 2014<br>Biodiversity Strategy and Action Plan (2015)<br>Five-Year Development Plan (FYDP) (June 2016) | Yes | Available | Relevance |
| Biological Diversity Strategy (BMZ 2008)<br>BMZ/BMU Committed to Biodiversity (2014, 2016) | Yes | Available | Relevance |

Onlinebefragung: Nein

### 7. Darstellung der Angemessenheit des methodischen Vorgehens

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden. | Es sind weder Begründungen für die angewandten Methoden aufgeführt noch werden Limitationen des methodischen Vorgehens diskutiert. | Es werden entweder Begründungen für die angewandten Methoden aufgeführt oder Limitationen des methodischen Vorgehens diskutiert. | Begründungen für die angewandten Methoden werden hinreichend aufgeführt und Limitationen des methodischen Vorgehens hinreichend diskutiert. | Begründungen für die angewandten Methoden werden umfassend aufgeführt und Limitationen umfassend diskutiert. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Findet sich vorwiegend im Methoden-, Limitations- oder Diskussionsteil des Evaluierungsberichtes oder der Berichtsanhänge oder in Inception Reports – diese werden jedoch nur zur Bewertung herangezogen, wenn sie Teil der Berichtsanhänge sind.
2. Die Diskussion der Methoden kann Erhebungs- sowie Analysemethoden oder beides betreffen.
3. Für manche Evaluierungen werden Methoden in der Leistungsbeschreibung bereits vorgegeben – eine reine Vorgabe ist jedoch nicht mit einer Begründung gleichzusetzen. Enthält die Leistungsbeschreibung eine Begründung der Methodenwahl, ist dies bewertungsrelevant, aber nur, wenn die Leistungsbeschreibung Teil der Berichtsanhänge ist.
4. Die alleinige Aussage „methodisches Vorgehen ist nicht repräsentativ“ reicht nicht aus, um den Aspekt der Diskussion der Limitationen zu erfüllen.
5. Der Unterschied zwischen „eher erfüllt“ und „vollständig erfüllt“ liegt in der Ausführlichkeit der Beschreibung der Limitationen und warum bestimmte Methoden ausgewählt wurden. Wenn einer der beiden Aspekte a) Limitationen und b) Begründung für methodisches Vorgehen mit lediglich 1–2 Sätzen oder weniger als 50 Wörtern beschrieben wird, ist dieser Aspekt nur hinreichend und nicht umfassend diskutiert.
6. Als „vollständig erfüllt“ gilt das Qualitätskriterium, wenn beide Aspekte mit mehr als 1–2 Sätzen und mehr als 50 Wörtern beschrieben werden.
7. „Eher nicht erfüllt“ ist das Qualitätskriterium, wenn nur einer der beiden Aspekte beschrieben wird, ungeachtet, ob dieser Aspekt umfassend oder nur hinreichend beschrieben wird.
8. Die Begründung der Anwendung der Methoden kann zum Beispiel hinsichtlich folgender Punkte erfolgen (die Punkte sind jedoch nicht erschöpfend/vollständig, es kann auch andere Begründungszusammenhänge geben): 1. Ausgestaltung der EZ-Maßnahme, 2. Fragestellung der Evaluierung, 3. spezifische Eigenschaften der untersuchten Subjekte (zum Beispiel Geflüchtete, Frauen), 4. Zielgruppe und ihres Erkenntnisinteresses, 5. Kenntnisse der Gutachtenden über die Eignung einer Methode, 6. Verhältnisses von (Ressourcen-)Aufwand für die Anwendung einer Methode und dem Nutzen der aus der Anwendung einer Methode besteht, 7. Kompetenz und Erfahrung der Gutachtenden, 8. Möglichkeit, dass die zum Einsatz kommenden Methoden den untersuchten Gegenstand ganzheitlich erfassen und verlässliche/robuste Ergebnisse generieren.
9. Eine ausführliche Beschreibung der Methoden oder des methodischen Vorgehens ist nicht bewertungsrelevant, sondern die Ausführlichkeit der Begründung der Auswahl der Methoden.

Ankerbeispiele:

- nicht erfüllt: „Es fand ein Scoping-Treffen mit einigen Mitarbeitern des XXX statt. Der Austausch ermöglichte es, das Verständnis des Auftraggebers und der Berater bezüglich der Aufgabenstellung zu harmonisieren, Zweifel zu klären und die Methodik und den Zeitplan des Prozesses zu verfeinern. Es

| |
|---|
| war auch eine Gelegenheit, zu einem besseren Verständnis der technischen und materiellen Organisation sowie der Erwartungen des Auftraggebers im weiteren Sinne zu gelangen.“<br>• eher erfüllt: „Limitation: […] An estimation of the impact of entrepreneurship and alumni networking in the past requires a so-called tracer study, asking what happened with former graduates. Although the MTS team carried out a preparatory tracer study and provided some facts and figures, the menace of the terrorist group Boko Haram led to the temporary closure of the school from March 2014 until early 2016 and made a detailed ex-post analysis impossible. After a two-year suspension of project activities, as well the use of a ‘Before-and-After Approach’, as requested in chapter 4 of the ToR, could only be applied to a certain extent.“<br>• vollständig erfüllt: „In order to use this approach, the existence of one case with good quality data sources is sufficient. The approach is based on an existing theory for the project in question which allows the evaluator to understand the factors that cause the observed effect. As a result this approach permits an in-depth understanding of the case and its context, proving a detailed explanation of both of them […] The evaluation team chose this over other approaches to causal inference, such as the experiment/counterfactual approach, the regulatory approach or the multiple causation approach. An experiment/counterfactual approach was not deemed feasible because the project targets the whole ECOWAS Commission, for which it would hardly be feasible to identify a unit of comparison. A regulatory approach was deemed largely unsuitable because for most of the causal hypotheses of the theory of change, the number of direct beneficiaries (e.g. staff members of specific directorates of the ECOWAS Commission) is limited. It would thus hardly have been possible to draw conclusions on cause-effect relationships on the basis of a high number of observed cases for these causal hypotheses. Lastly, a multiple causation approach to causal inference was not deemed feasible due to the complexity of the project.“ |
| Onlinebefragung: Nein |

## 8. Qualitätssicherung mit Inception Report[32]

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn ein Inception Report erstellt wird. | Die Evaluierung verfügt über keinen Inception Report. | Die Evaluierung verfügt über einen Inception Report. |
| Dokumente: Inception Report | | |
| Kodierregeln:<br>1. Für dieses Qualitätskriterium wird ausschließlich betrachtet, ob ein IR vorliegt oder nicht. Wenn kein IR vorliegt, wird das Qualitätskriterium mit 1 bewertet. Bei den Organisationen haben die IRs teilweise unterschiedliche Bezeichnungen.<br>2. Es geht um den Prozess zur Qualitätssicherung der Evaluierung und nicht um die Qualitätssicherung der EZ-Maßnahme. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Welche der folgenden Komponenten umfasst einen Qualitätssicherungsprozess in Ihrer Organisation? (Formen der Qualitätssicherungsprozesse) | | |

[32] Das Vorhandensein eines Inception Reports wurde nachträglich als Proxy für das Qualitätskriterium „Qualitätssicherungsprozesse“ festgelegt. Dass der Inception eine Form der Qualitätssicherung darstellt, wurde im Rahmen einer Fokusgruppendiskussion mit den beteiligten Organisationen festgelegt und zeigt sich auch in bestehenden empirischen Untersuchungen.

### 9. Kohärenz von Daten-Ergebnissen-Schlussfolgerungen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Datenanalyse, die Ergebnisse und die Schlussfolgerungen logisch miteinander in Bezug gesetzt werden beziehungsweise aufeinander aufbauen. | Die Datenanalyse, die Ergebnisse und die Schlussfolgerungen werden bei 0 bis 24 Prozent der Hauptergebnisse logisch miteinander in Bezug gesetzt. | Die Datenanalyse, die Ergebnisse und die Schlussfolgerungen werden bei 25 bis 49 Prozent der Hauptergebnisse logisch miteinander in Bezug gesetzt. | Die Datenanalyse, die Ergebnisse und die Schlussfolgerungen werden bei 50 bis 74 Prozent der Hauptergebnisse logisch miteinander in Bezug gesetzt. | Die Datenanalyse, die Ergebnisse und die Schlussfolgerungen werden bei 75 bis 100 Prozent der Hauptergebnisse logisch miteinander in Bezug gesetzt. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Unterschied Ergebnisse (*findings*, *results*) und Schlussfolgerungen (*conclusions*): Ergebnisse haben die Daten beziehungsweise die Datenauswertung als Basis und lassen sich von ihnen ableiten. Schlussfolgerungen basieren auf den Ergebnissen und stellen diese in einen größeren Zusammenhang. Sie zeigen Konsequenzen, Folgen, Implikationen auf, die sich aus den Ergebnissen ableiten lassen.
2. Für dieses Qualitätskriterium ist es nicht erforderlich, alle Ergebnisse/Schlussfolgerungen der Evaluierung hinsichtlich der Kohärenz der Kette Analyse-Ergebnis-Schlussfolgerungen zu überprüfen. Es müssen die *main findings/main conclusions* in Bezug auf das Kriterium „Effektivität“ überprüft werden.
3. Schritt 1: Ausgangspunkt für die Bewertung sind die Hauptergebnisse der Evaluierung (*main findings*) zum Kriterium „Effektivität“. Diese müssen zunächst identifiziert werden. Die Zusammenfassung (*Executive Summary*) der Evaluierung sollte hierzu zuerst gelesen und die Hauptergebnisse für das Kriterium „Effektivität“ identifiziert werden. Finden sich dort keine Hauptergebnisse (oder gibt es keine Zusammenfassung), dient der „Ergebnisteil“ der Evaluierung zum Kriterium „Effektivität“ als Hauptorientierungspunkt. Wenn die Ergebnisse der Evaluierung nicht nach den OECD-DAC-Kriterien strukturiert sind und entsprechend das Kriterium „Effektivität“ nicht behandelt wurde, wird ein ähnlicher Abschnitt (zum Beispiel Analyse und Ergebnisse der Überprüfung der Zielerreichung) zur Bewertung herangezogen, in dem Hauptergebnisse herausgestellt werden. Wenn es keinen Ergebnisteil gibt (oder die Hauptergebnisse nicht klar erkenntlich sind), ist das Qualitätskriterium mit 1 zu bewerten.
4. Schritt 2: Für jedes dieser Hauptergebnisse zum Kriterium „Effektivität“ wird nun überprüft, ob diese mit Datenanalyse und Schlussfolgerungen logisch miteinander in Bezug gesetzt sind. Für ein effizientes Vorgehen sollte zuerst a) die Zusammenfassung, dann erst b) der Ergebnisteil zur Bewertung herangezogen werden und zuletzt gegebenenfalls c) über mehrere Abschnitte hinweg (zum Beispiel Abschnitte zu Methodik, Analyse oder Ergebnisse und Schlussfolgerungen) gelesen und bewertet werden.
5. Wenn der logische Sprung oder der textliche Abstand zwischen Analyse und Ergebnissen zu groß ist, sollte dies mit „nicht erfüllt“ bewertet werden. Dies ist zum Beispiel der Fall, wenn im Ergebnisteil keinerlei Referenz auf die Analyse angeführt wird.
6. Schritt 3: Ist bei einem von beispielsweise vier Hauptergebnissen die Kette von Analyse-Ergebnis-Schlussfolgerung nachvollziehbar, wäre das Qualitätskriterium zu einem Viertel erfüllt und würde mit einer 2 („eher nicht erfüllt“) bewertet werden.

Ankerbeispiel (vollständig erfüllt, wenn dies das einzige Ergebnis wäre):

- „Die Onlineumfrage der 17 Teilnehmer (*Analyse*) ergab, dass der Dünger häufig nicht zu den Bauen geliefert wird (*Ergebnis*). Dies stellt ein großes Risiko für die Effektivität des Projekts dar (*Schlussfolgerung*).“

Onlinebefragung: Nein

## 10. Informationsgehalt der Leistungsbeschreibung

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Leistungsbeschreibung der Evaluierung 1. Zweck, 2. Nutzende, 3. Ziele, 4. Methoden, 5. Mittel, 6. den zeitlichen Rahmen, 7. Veröffentlichungsrechte und 8. mitwirkende Personen der Evaluierung beinhaltet. | Die Leistungsbeschreibung ist nicht vorhanden oder es wird nur ein Punkt genannt. | In der Leistungsbeschreibung werden zwei bis drei der Punkte genannt. | In der Leistungsbeschreibung werden vier bis fünf der Punkte genannt. | In der Leistungsbeschreibung werden mindestens sechs der Punkte genannt. |

Dokumente: Leistungsbeschreibung

Kodierregeln:

1. Wenn keine Leistungsbeschreibung vorliegt, wird das Qualitätskriterium mit 1 bewertet.
2. Bei den Organisationen hat die Leistungsbeschreibung teilweise unterschiedliche Bezeichnungen.
3. Nutzende der Evaluierung (Punkt 2) sind zum Beispiel das Projektmanagement, politische Entscheidungsträger*innen.
4. Mittel der Evaluierung (Punkt 5) sind entweder die verfügbaren finanziellen Gesamtmittel für die Evaluierung (nicht für die EZ-Maßnahme) oder die finanziellen Mittel für Gutachtende oder die Anzahl der Gutachtendentage.
5. Zeitlicher Rahmen (Punkt 6) muss ein Anfangs- und Enddatum der Evaluierung beinhalten; eine Monat- und Jahresangabe genügt.
6. Veröffentlichungsrechte (Punkt 7) definieren, wem (zum Beispiel Organisationen/Institutionen) die Evaluierung zur Verfügung gestellt wird oder ob die Evaluierung der Öffentlichkeit zur Verfügung gestellt wird oder wer die Evaluierung veröffentlichen darf.
7. Mitwirkende Personen (Punkt 8) sind Personen oder Einheiten, die neben dem Gutachtenden-Team an der Evaluierung mitwirken (nicht im Rahmen der Analyse der Evaluierung befragt werden). Diese können beispielsweise die Auftraggebende, die zentrale Evaluierungseinheit und weitere Partner*innen/Stakeholder*innen umschließen.

Ankerbeispiel (vollständig erfüllt):

- Zweck: „The evaluation is initiated by organization XXX. […] The results will be used to improve the quality of the educational programme of XXX, to inform and plan the strategic development for the coming years and to improve existing and set up new mechanisms to productively deal with changes in the project environment.“
- Nutzende: „This […] will facilitate staff getting additional information from consultant to enable them engage meaningfully with project beneficiaries/clients. Therefore, the outcome of evaluation will provide opportunity for project staff to readjust the project approach.“
- Ziele: „The evaluation should help collect quality information and provide recommendations for the following areas of the intervention: – Development of employability courses, its products and services.“
- Methoden: „A combination of primary data collection and secondary data review is expected during the evaluation. Primary data shall be collected through a variety of methods, including field observations.“
- Mittel: /
- Zeitlicher Rahmen: „The indicative timetable for the evaluation is foreseen as follows: […].“
- Veröffentlichungsrechte: /
- Mitwirkende Personen: „The final methodological design, timetable and evaluation questions will be agreed upon between the successful applicants and the responsible staff members at organization XXX during the inception phase.“

Onlinebefragung: Nein

### 11. Veröffentlichung des Evaluierungsberichts

Der Qualitätsstandard „Zugänglichkeit“ beinhaltete drei Teilaspekte: „Veröffentlichung des Evaluierungsberichts (11)“, „Veröffentlichung der Zusammenfassung (12)“, und „Zugänglichkeit für Stakeholder*innen (13)“. Da nur der letzte Teilaspekt in den Evaluierungsdokumenten kodiert werden konnte, wurden die beiden weiteren Teilaspekte über die Onlinebefragung erfasst und keine vorläufige Operationalisierung für die Inter-Kodierenden-Phase festgelegt. Die Frage für das vorliegende Qualitätskriterium lautete: Werden die Evaluierungen Ihrer Organisation auf einer der Öffentlichkeit zugänglichen Webseite/Datenbank veröffentlicht? Auswahl: Der gesamte Bericht wird veröffentlicht (immer, überwiegend, teilweise, selten, nie).

### 12. Veröffentlichung der Zusammenfassung

Der Qualitätsstandard „Zugänglichkeit“ beinhaltete drei Teilaspekte: „Veröffentlichung des Evaluierungsberichts (11)“, „Veröffentlichung der Zusammenfassung (12)“, und „Zugänglichkeit für Stakeholder*innen (13)“. Da nur der letzte Teilaspekt in den Evaluierungsdokumenten kodiert werden konnte, wurden die beiden weiteren Teilaspekte über die Onlinebefragung erfasst und keine vorläufige Operationalisierung für die Inter-Kodierenden-Phase festgelegt. Die Frage für das vorliegende Qualitätskriterium lautete: Werden die Evaluierungen Ihrer Organisation auf einer der Öffentlichkeit zugänglichen Webseite/Datenbank veröffentlicht? Auswahl: Eine separate Zusammenfassung und/oder die Hauptergebnisse und nicht der gesamte Bericht werden veröffentlicht (immer, überwiegend, teilweise, selten, nie).

### 13. Zugänglichkeit für Stakeholder*innen

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn der Evaluierungsbericht oder eine schriftliche Zusammenfassung (auch in Form einer Ergebnispräsentation) den Stakeholder*innen der Evaluierung zugänglich gemacht wurde. | Der Evaluierungsbericht oder eine schriftliche Zusammenfassung wurde den Stakeholder*innen nicht zugänglich gemacht. | Der Evaluierungsbericht oder eine schriftliche Zusammenfassung wurde mindestens einem*r Stakeholder*in zugänglich gemacht. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, Leistungsbeschreibung | | |
| Kodierregeln:<br>1. Zumeist finden sich Informationen über die Zugänglichkeit des Evaluierungsberichts oder der Ergebnisse in der Leistungsbeschreibung. Zum Beispiel wird beschrieben, in welcher Form der Bericht veröffentlicht wird oder es wird im Zeitplan aufgeführt, dass es Präsentationen mit Stakeholder*innen zu den Ergebnissen gibt.<br>2. Auch wenn der Evaluierungsbericht oder die schriftliche Zusammenfassung veröffentlicht werden (zum Beispiel auf der Homepage) muss erkenntlich werden, dass diese (oder die Ergebnisse) gezielt mit mindestens eine*r der Stakeholder*innen geteilt werden, um als vollständig erfüllt zu gelten.<br>3. Definition Stakeholder*innen der Evaluierung: Personen, Personengruppen oder Organisationen, die je Konzeption und Umsetzung einer Evaluation etwas zu verlieren oder zu gewinnen haben. Es können z. B. Personen sein, die a) für die Planung und Umsetzung der Evaluation zuständig sind (Auftraggebende oder Geldgebende einer Evaluation), b) für das zu evaluierende Projekt operativ oder politisch zuständig sind oder c) die Evaluation nutzen können – sei es, dass sie durch die Evaluierenden aktiv in die Nutzung einbezogen werden, oder Adressierte (der Evaluation) sind oder ausgeschlossen werden. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Nein | | |

### 14. Kompetenz der Gutachtenden

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn der*die Gutachtende oder das Gutachtenden-Team der Evaluierung über die folgenden Kompetenzen verfügt: 1. Evaluierungsexpertise, 2. Fach-/Sektorexpertise und 3. Regional-/Länderexpertise. | Keine der relevanten Kompetenzen sind bei den Gutachtenden verfügbar. | Eine der relevanten Kompetenzen ist bei den Gutachtenden verfügbar. | Zwei der relevanten Kompetenzen sind bei den Gutachtenden verfügbar. | Alle drei der relevanten Kompetenzen sind bei den Gutachtenden verfügbar. |

Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report

Kodierregeln:

1. Der Fokus des Qualitätskriteriums liegt auf den tatsächlichen Kompetenzen der für die Evaluierung ausgewählten Gutachtenden. Es geht nicht um die geforderten Kompetenzen, die vorwiegend in der Leistungsbeschreibung beschrieben werden.
2. Es kann Evaluierungen geben, in denen es keine Informationen zu den Kompetenzen der für die Evaluierung ausgewählten Gutachtenden gibt, sondern nur allgemeine Kompetenzen/Expertise der Consultancy/Firma genannt werden, welche die Evaluierung durchführt. In solchen Fällen ist nicht eindeutig, ob die Gutachtenden für die Evaluierung auch die Kompetenzen/Expertise der Consultancy/Firma abdecken.
3. Evaluierungsexpertise: Erfahrung mit der Durchführung von Evaluierungen.
4. Fach-/Sektorexpertise: Erfahrungen in einem thematischen Fachbereich oder Sektor (zum Beispiel Privatwirtschafsförderung, Landwirtschaft, Rohstoffe, Mikrofinanzierung, Gesundheit, Wasser)
5. Regional-/Länderexpertise: Erfahrungen über die Region oder das Land, in dem die Evaluierung durchgeführt wird.
6. Bei Gutachtenden-Teams muss die Kompetenz nur einmal im Team vorliegen. Es müssen nicht alle Teammitglieder über alle Kompetenzen verfügen. Zum Beispiel kann es drei Gutachtende in einem Team geben; verfügt von diesen dreien nur eine*r davon über relevante Evaluierungsexpertise, ist dies ausreichend für die Verfügbarkeit dieser Expertise.

Ankerbeispiel (eher nicht erfüllt):

- „The expertise, qualifications and professional experience of the six evaluators complemented each other perfectly with regard to conducting the evaluation: Ms. […] brought broad methodological skills as evaluators.“

Onlinebefragung: Sind in Ihren Evaluierungen alle drei aufgeführten Kompetenzen bei den Gutachter*innen vertreten? (1. Evaluierungsexpertise, 2. Fach-/Sektor-expertise und 3. Regional-/Länderexpertise.)

### 15. Rechtzeitigkeit der Erkenntnisse

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Evaluierung rechtzeitig abgeschlossen wurde. | Die Evaluierung wurde nicht rechtzeitig abgeschlossen, das heißt, zum vereinbarten Zeitpunkt war sie noch nicht eingereicht. | Die Evaluierung wurde rechtzeitig abgeschlossen, das heißt, die Evaluierung wurde zum vereinbarten Zeitpunkt oder früher eingereicht. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Leistungsbeschreibung | | |
| Kodierregeln:<br>1. Schritt (Leistungsbeschreibung): Wenn in der Leistungsbeschreibung ein *submission date* (Einreichung/Abgabe-Datum oder Ähnliches) für den Endbericht (*final report*) durch das Gutachtenden-Team an den Auftraggeber aufgeführt wird, ist dies der erste Referenzzeitpunkt. Damit ist nicht das Datum der Veröffentlichung des Evaluierungsberichtes (*publication date*) gemeint.<br>2. Schritt (Evaluierungsbericht; falls das *submission date* in der Leistungsbeschreibung gefunden wurde): Suche im Evaluierungsbericht zum Beispiel nach den Wörtern *submission*, *Einreichung*, *was submitted on date XX*. Hier ist nicht das Veröffentlichungsdatum (*publication date*) gemeint, was sich häufig auf dem Titelblatt der Evaluierung befindet.<br>3. Schritt (Abgleich): Wenn ein *submission date* im Evaluierungsbericht gefunden wurde, ist ein Abgleich zwischen dem Datum der geplanten Einreichung des Endberichts (laut Leistungsbeschreibung) und der tatsächlichen Einreichung des finalen Endberichts (laut Evaluierungsbericht) vorzunehmen. Wurde der Evaluierungsbericht zum vereinbarten Zeitpunkt (oder früher) eingereicht, wird das Qualitätskriterium mit 4 bewertet. Wurde er später eingereicht, wird mit 1 bewertet. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Werden die Evaluierungen Ihrer Organisation rechtzeitig zum vereinbarten Zeitpunkt (oder früher) abgeschlossen? | | |

### 16. Adressat*innen der Empfehlungen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen der Evaluierung (alternativ: ihre Schlussfolgerungen) adressat*innen-orientiert sind. | 0 bis 24 Prozent der Empfehlungen sind adressatenorientiert. | 25 bis 49 Prozent der Empfehlungen sind adressatenorientiert. | 50 bis 74 Prozent der Empfehlungen sind adressatenorientiert. | 75 bis 100 Prozent der Empfehlungen sind adressatenorientiert. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Adressatenorientiert bedeutet, dass die Organisation (zum Beispiel *organisation XXX should …*) oder eine Abteilung/Einheit innerhalb einer Organisation (zum Beispiel *the project should …*), an die sich die Empfehlungen richten, genannt wird.
2. Schritt 1: Gibt es bei den Empfehlungen Überschriften? Beinhalten diese bereits Adressat*innen (zum Beispiel XXX: „strategic level: recommendations for XXX“ oder „zentrale Handlungsempfehlungen für das Länderprojekt Brasilien der XXX“), gilt jede einzelne dieser Empfehlungen, die unter dieser Überschrift steht, als adressat*innenorientiert.
3. Schritt 2: Gibt es keine Überschriften bei den Empfehlungen, muss für jede einzelne Empfehlung bewertet werden, ob sie adressatenorientiert ist oder nicht.
4. Schritt 3: Eine Bewertung hinsichtlich der Adressat*innenorientierung („erfüllt“ oder „nicht erfüllt“) wird für jede Empfehlung vorgenommen und anschließend ermittelt, wie viel Prozent der Empfehlungen eine Adressat*innenorientierung aufweisen.
5. Falls keine Empfehlungen in den Evaluierungsberichten vorliegen, wird die Adressat*innenorientierung der Schlussfolgerungen bewertet.
6. Liegen weder Empfehlungen noch Schlussfolgerungen vor, wird mit 1 bewertet.

Ankerbeispiel (diese Empfehlung gilt als „erfüllt“ – siehe Kodierregel 4):

- „INMETRO needs to strengthen internal cooperation between departments, especially by improving communication with the different reference laboratories. This also applies to relevant calibration laboratories, research institutes, industrial associations (ABSOLAR, ABEEólica, CNI, FIESP, Chambers of Commerce, and others), as well as industrial companies. This refers especially to INMETRO’s reference laboratories, as only some of them have regular contact with their clients.“

Onlinebefragung: Nein

### 17. Handlungsorientierung der Empfehlungen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Empfehlungen (alternativ: Schlussfolgerungen) der Evaluierung handlungsorientiert sind. | 0 bis 24 Prozent der Empfehlungen sind handlungsorientiert. | 25 bis 49 Prozent der Empfehlungen sind handlungsorientiert. | 50 bis 74 Prozent der Empfehlungen sind handlungsorientiert. | 75 bis 100 Prozent der Empfehlungen sind handlungsorientiert. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Handlungsorientiert bedeutet, dass die Empfehlung spezifisch formuliert ist und es klar wird, wie sie umgesetzt werden soll. Leitfrage für die Bewertung ist: Werden in der Evaluierung konkrete Handlungsschritte beschrieben, die zeigen, wann und wie die Empfehlung(en) umgesetzt werden sollen?
2. Eine Bewertung hinsichtlich der Handlungsorientierung („erfüllt“ oder „nicht erfüllt“) wird für jede Empfehlung vorgenommen und anschließend ermittelt, wie viel Prozent der Empfehlungen eine Handlungsorientierung aufweisen.
3. Falls keine Empfehlungen in den Evaluierungsberichten vorliegen, wird die Handlungsorientierung der Schlussfolgerungen bewertet.
4. Liegen weder Empfehlungen noch Schlussfolgerungen vor, wird mit 1 bewertet.

Ankerbeispiel (für diese Empfehlung „nicht erfüllt“ – siehe Kodierregel 2):

- „This refers especially to INMETRO’s reference laboratories, as only some of them have regular contact with their clients. For this, a special technological and innovation service within INMETRO could help improving the relationship between the institute, industry, and other users of metrological services.“

Onlinebefragung: Nein

### 18. Evaluierungseffizienz

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn während der Evaluierung Überlegungen dargestellt werden, ob und wie die aufgewendeten Ressourcen einer Evaluierung (in Bezug auf Zeit, Personal, finanzielle Mittel und Aufwand für Partner*innen/Beteiligte) im Verhältnis zu ihrem Nutzen stehen. | Überlegungen zum Aufwand-Nutzen-Verhältnis der Evaluierung werden nicht dargestellt. | Überlegungen zum Aufwand-Nutzen-Verhältnis der Evaluierung werden dargestellt. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, Leistungsbeschreibung, sonstige Dokumente | | |
| Kodierregeln:<br>1. Das Qualitätskriterium zielt auf Effizienzüberlegungen der Evaluierung ab, nicht auf die Überprüfung der Effizienz der EZ-Maßnahme.<br>2. Es handelt sich nur um Überlegungen zur Effizienz; es müssen keine konkreten Berechnungen beziehungsweise Effizienzanalysen in den Evaluierungsdokumenten gefunden werden.<br>3. Mögliche Suchbegriffe: *Efficiency, efficient, financial means, financial resources, resources, expense, expenditure, costs, time plan, schedule, changes in the time table, budget, funds*. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Wird in Ihrer Organisation sichergestellt, dass die Effizienz (Aufwand-Nutzen-Verhältnis) einer Evaluierung reflektiert wird? | | |

### 19. Transparenz von Meinungsverschiedenheiten

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn Meinungsverschiedenheiten der Stakeholder*innen und des Gutachtenden-Teams hinsichtlich bestimmter Inhalte des Evaluierungsberichts (zum Beispiel Bewertungen, Schlussfolgerungen oder Empfehlungen) im Evaluierungsbericht transparent gemacht werden. | Meinungsverschiedenheiten der Stakeholder*innen und der Gutachtenden der Evaluierung werden im Evaluierungsbericht nicht transparent gemacht. | Meinungsverschiedenheiten der Stakeholder*innen und der Gutachtenden der Evaluierung werden im Evaluierungsbericht transparent gemacht. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Bei diesem Qualitätskriterium geht es nicht um Meinungsverschiedenheiten, die zum Beispiel durch Interviews mit verschiedenen Personen zum Evaluierungsgegenstand deutlich werden. Es geht um Meinungsverschiedenheiten zu bestimmten Inhalten des Evaluierungsberichts.
2. Es wird mit 4 bewertet, falls das Thema/der Gegenstand der Meinungsverschiedenheit im Evaluierungsbericht transparent gemacht wird und dies somit für die Kodierenden ersichtlich ist. Das heißt, für die Kodierenden muss erkennbar sein, worin die Meinungsverschiedenheit besteht.
3. Es wird mit 1 bewertet, wenn es nur einen allgemeinen Hinweis auf auftretende Meinungsverschiedenheiten gibt, diese jedoch nicht weiter erläutert beziehungsweise konkretisiert werden.
4. Hinweise darauf, dass a) Stakeholder*innen die Möglichkeit haben, den Evaluierungsbericht zu kommentieren oder b) dass erwähnt wird, dass Meinungsverschiedenheiten im Evaluierungsbericht transparent oder berücksichtigt werden sollen, ist nicht ausreichend für eine Bewertung mit 4.
5. Tipp: Bewertungsrelevante Informationen finden sich zum Beispiel im Methodenkapitel, im Abschnitt zu Limitationen oder auch bei der Diskussion der Ergebnisse.
6. Definition Stakeholder*innen der Evaluierung: Personen, Personengruppen oder Organisationen, die je Konzeption und Umsetzung einer Evaluation etwas zu verlieren oder zu gewinnen haben. Es können z. B. Personen sein, die a) für die Planung und Umsetzung der Evaluation zuständig sind (Auftraggebende oder Geldgebende einer Evaluation), b) für das zu evaluierende Projekt operativ oder politisch zuständig sind oder c) die Evaluation nutzen können – sei es, dass sie durch die Evaluierenden aktiv in die Nutzung einbezogen werden, oder Adressierte (der Evaluation) sind oder ausgeschlossen werden.
7. Mögliche Suchbegriffe: *disagreement (Uneinigkeit, Unstimmigkeit), dissent (Dissens, Meinungsverschiedenheit), conflict of opinions, difference in opinions, comment (Kommentar, Bemerkung).*

Ankerbeispiel: /

Onlinebefragung: Werden Gutachter*innen darauf hingewiesen, Meinungsverschiedenheiten, wenn es welche gibt, im Evaluierungsbericht transparent zu machen?

### 20. Einbindung der internen und externen Stakeholder*innen

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn Stakeholder*innen der Evaluierung in den Evaluierungsprozess (1. Konzeption/Planung, 2. Durchführung, 3. Berichtslegung) einbezogen werden. | Keine der Stakeholder*in wird in den Evaluierungsprozesses einbezogen. | Mindestens eine Stakeholder*in wird in einer der drei genannten Evaluierungsphasen einbezogen. | Mindestens eine Stakeholder*in wird in zwei der drei genannten Evaluierungsphasen einbezogen. | Mindestens eine Stakeholder*in wird in allen drei genannten Evaluierungsphasen einbezogen. |

Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, Leistungsbeschreibung

Kodierregeln:

1. In der Evaluierungsphase „Durchführung“ geht es nicht um die Einbindung als Informationsquelle für die Datenerhebung, beispielsweise als Teilnehmende*r einer Fokusgruppendiskussion.
2. Die 1. Phase „Konzeption/Planung“ umfasst beispielsweise erste Briefings oder Austausche zum Projekt oder den Methoden. Die 2. Phase „Durchführung“ beginnt, sobald die tatsächliche Evaluierungsarbeit beginnt und schon erste Interviews und Befragungen durchgeführt werden – damit zählt auch die Erstellung des Inception Reports zu dieser Phase. Außerdem kann die Durchführungsphase zum Beispiel die Einbindung von Stakeholder*innen im Rahmen von Referenzgruppen, die Kommentierungsmöglichkeiten von Instrumenten oder Präsentationen von IR mit Feedback seitens der Stakeholder*innen umschließen. Die 3. Phase der Berichtslegung umschließt beispielsweise Kommentierungsschleifen zum Berichtsentwurf und finalen Bericht.
3. Es wird unterschieden zwischen einem „Informieren“ von Stakeholder*innen und der Möglichkeit, dass diese dazu auch eine Rückmeldung abgeben können, die von den Gutachtenden berücksichtigt wird. Ein „Informieren“ könnte aus einer E-Mail bestehen, bei der nicht gewährleistet ist, dass sie gelesen wird. Daher wird ein Informieren nur als Einbindungsform gewertet, wenn auf das Informieren eine „Rückmeldung“ möglich ist. Beides muss aus dem Evaluierungsbericht hervorgehen.
4. Definition Stakeholder*innen der Evaluierung: Personen, Personengruppen oder Organisationen, die je Konzeption und Umsetzung einer Evaluation etwas zu verlieren oder zu gewinnen haben. Es können z. B. Personen sein, die a) für die Planung und Umsetzung der Evaluation zuständig sind (Auftraggebende oder Geldgebende einer Evaluation), b) für das zu evaluierende Projekt operativ oder politisch zuständig sind oder c) die Evaluation nutzen können – sei es, dass sie durch die Evaluierenden aktiv in die Nutzung einbezogen werden, oder Adressierte (der Evaluation) sind oder ausgeschlossen werden.
5. Mögliche Suchbegriffe: *participatory, participation, consultation, involvement, comment, discussion of results, stakeholder.*

Ankerbeispiel (vollständig erfüllt):

- „Participation should promote a learning and growth process, action-reflection-action, which generates changes and adjustments, not judgment. It is a qualitative and quantitative assessment of what was found (retrospective), and for the formulation of proposals (prospective). It is considered as a space for meeting and dialogue between various stakeholders and the multidisciplinary external consulting team. Direct participation took place at three moments: at the beginning, with contributions to the methodological proposal; during the execution, through dialogue in the focus group; and finally, in the presentation of the final results of the evaluation. It should be noted that there was a lot of interest in the evaluation on the part of all those involved. The commitment made it possible to develop the proposed participatory moments and fulfill the planned activities, in the given time.“

Onlinebefragung: Nein

### 21. Einbindung der Partner*innen

Da Stakeholder*innen nach Definition der DeGEval sowohl externe (zum Beispiel Projektpartner*innen) als auch interne Stakeholder*innen (zum Beispiel die Bereichsleitung innerhalb der Organisation, die die Evaluierung in Auftrag gegeben hat und zentral [mit-]verantwortet) mit einschließt, wurde auf Evaluierungsebene kein Unterschied zwischen „internen“ und „externen“ Stakeholder*innen gemacht. Um Informationen zu der Einbindung von ausschließlich externen Stakeholder*innen zu erheben, wurde dies zusätzlich über die Onlinebefragung erfasst. Frage: Binden Sie externe Stakeholder*innen in die Konzeption, Durchführung und Berichtslegung der Evaluierungen Ihrer Organisation ein?

### 22. Evaluierungsethik

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Evaluierung Vorgaben zur Sicherheit und zu den Rechten der Evaluierungsteilnehmenden trifft und/oder über entsprechende Richt- und Leitlinien verfügt. | Die Evaluierung macht keine oder nur implizite Vorgaben zur Sicherheit und zu den Rechten der Evaluierungsteilnehmenden und verfügt über keine entsprechenden Richt- und Leitlinien. | Die Evaluierung macht explizite Vorgaben zur Sicherheit und zu den Rechten der Evaluierungsteilnehmenden und/oder verfügt über entsprechende Richt- und Leitlinien. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, Leistungsbeschreibung | | |
| Kodierregeln:<br>1. Oft findet sich in der Leistungsbeschreibung ein Abschnitt zum Schutz der Evaluierungsteilnehmenden oder dem Umgang mit sensiblen Daten.<br>2. Datenschutz ist nur ein Element des Qualitätskriteriums und nicht ausreichend zur Erfüllung.<br>3. Das Qualitätskriterium zielt nicht auf die Überprüfung der Ethik und Menschrechte der EZ-Maßnahme ab, sondern auf die Einhaltung von ethischen Grundsätzen in der Durchführung der Evaluierung.<br>4. Mögliche Suchbegriffe: *ethic, ethics, ethical guidelines, safety, security, dignity, anonymous, anonymity, (human) rights*. | | |
| Ankerbeispiele (vollständig erfüllt):<br>• „To make the evaluation as useful as possible for decision-making processes, the following standards should be met […] Protection of individual rights: evaluations should be planned and carried out so that the security, dignity and rights of the persons included in an evaluation are protected.“<br>• „The evaluator/s should take all reasonable steps to ensure that the evaluation is designed and conducted to respect and protect the rights and welfare of people and the communities of which they are members. […] Therefore, the evaluator/s should adhere to the evaluation standards of the IFRC.“ | | |
| Onlinebefragung: Werden Vorgaben zu Sicherheit und Rechten der Evaluierungsteilnehmenden gemacht? | | |

### 23. Darstellung organisationale Unabhängigkeit der Gutachtenden

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Gutachtenden der Evaluierung eine ausreichende Distanz zur EZ-Maßnahme haben, das heißt, wenn sie weder 1) politisch 2) operativ 3) und beratend an der EZ-Maßnahme beteiligt sind und 4) nicht zur Zielgruppe der EZ-Maßnahme gehören. | Einer der vier Fälle trifft für mindestens eine*n Gutachtende*n/das Gutachtenden-Team zu. | Keiner der vier Fälle trifft für eine*n Gutachtende*n/das Gutachtenden-Team zu. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report | | |
| Kodierregeln:<br><br>1. Es wird mit 4 bewertet, wenn eine der zwei Punkte erfüllt ist:<br>a) Auf dem Titelblatt findet sich der Name oder die Kontaktdaten einer Consultancy/Firma/Universität oder einer anderen Institution. Ist dies der Fall, werden die Gutachtenden als strukturell unabhängig betrachtet und somit mit einer 4 bewertet.<br>b) Eine weitere hilfreiche Quelle zur Bewertung ist eine *declaration of independence* (Unabhängigkeitserklärung). Ist diese unterschrieben und erfüllt sie alle oben genannten vier Fälle von Unabhängigkeit, wird auch mit 4 bewertet.<br>2. Mit 1 wird bewertet, wenn es explizite Hinweise darauf gibt, dass zumindest ein*e Gutachtende (des Teams) strukturell abhängig ist. Dies ist der Fall, wenn<br>a) es Hinweise darauf gibt, dass ein Gutachtende (des Teams) 1) politisch oder 2) operativ oder 3) beratend der EZ-Maßnahme beteiligt ist oder 4) zur Zielgruppe der EZ-Maßnahme gehört<br>b) eine *declaration of independence* vorliegt, die zwar auf die strukturelle Unabhängigkeit der Gutachtenden eingeht, aber nicht von den Gutachtenden unterschrieben wurde. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Nein | | |

### 24. Darstellung Unvoreingenommenheit der Gutachtenden

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn der*die Gutachtende/das Gutachtenden-Team dem Evaluierungsgegenstand gegenüber unvoreingenommen ist. | Es gibt keinen Hinweis darauf, dass eine *declaration of independence* vorliegt, die auf die persönliche Unvoreingenommenheit des*der Gutachtenden/des Gutachtenden-Teams abzielt. | Es liegt eine *declaration of independence* vor, die auf die persönliche Unvoreingenommenheit des*der Gutachtenden/des Gutachtenden-Teams abzielt, und die von der*dem Gutachtenden oder den Gutachtenden unterschrieben wurde. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report | | |
| Kodierregeln:<br>1. Hier geht es um das „Mindset“ und die „persönliche Einstellung“ des Gutachtenden-Teams zur Evaluierung beziehungsweise zum Evaluierungsgegenstand und grenzt sich dadurch vom Qualitätskriterium „organisationale Unabhängigkeit“ ab.<br>2. Eine Annäherung zur Messung dieses Qualitätskriteriums bietet die *declaration of independence* (Unabhängigkeitserklärung). Damit mit 4 bewertet wird, müssen folgende Bedingungen erfüllt sein: Die *declaration of independence* liegt vor, enthält einen expliziten Hinweis zur Unvoreingenommenheit (*impartiality* oder *unbiasedness*) und ist von den Gutachtenden unterschrieben worden. | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Wird für Evaluierungen Ihrer Organisation von allen Gutachter*innen eine „Erklärung zur Unabhängigkeit“ unterschrieben, die Aspekte zur persönlichen Unvoreingenommenheit beinhaltet? | | |

## 25. Informationsgehalt der Zusammenfassung

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn eine Zusammenfassung vorliegt und diese 1. Ergebnisse und entweder 2. Schlussfolgerungen oder 3. Empfehlungen der Evaluierung nennt. | a) Eine Zusammenfassung liegt nicht vor oder b) eine Zusammenfassung liegt vor, enthält jedoch nur Schlussfolgerungen oder Empfehlungen und keine Ergebnisse, oder c) eine Zusammenfassung liegt vor, enthält jedoch nur Ergebnisse und keine Schlussfolgerungen oder Empfehlungen. | Die Zusammenfassung enthält Ergebnisse und zusätzlich entweder Schlussfolgerungen oder Empfehlungen. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Die Zusammenfassung ist innerhalb des Evaluierungsberichts zu finden und meistens durch eine Überschrift abgegrenzt (zum Beispiel *Executive Summary, Summary*).
2. Erklärung des Unterschiedes zwischen Ergebnissen (zum Beispiel *findings*, *results*) und Schlussfolgerungen (*conclusions*): Ergebnisse haben die Daten beziehungsweise die Datenauswertung als Basis und lassen sich von ihnen ableiten. Schlussfolgerungen basieren auf den Ergebnissen und stellen diese in einen größeren Zusammenhang. Sie zeigen Konsequenzen, Folgen, Implikationen auf, die sich aus den Ergebnissen ableiten lassen.
3. Eine Kurzbeschreibung der EZ-Maßnahme gilt nicht als Zusammenfassung der Evaluierung.

Ankerbeispiele (vollständig erfüllt):

- Ergebnisse: „Wichtige Ergebnisse: Das Vorhaben entfaltete hohe entwicklungspolitische Relevanz und Wirksamkeit. Die Förderung und Stärkung der Bank ermöglichte die signifikante Verbesserung des Zugangs zu Finanzierungsangeboten für die Zielgruppe im KKMU-Sektor.“
- Schlussfolgerungen: „Der mangelnde Zugang zu Finanzdienstleistungen von KKMU sowie die Finanzierungslücke für das ursprünglich anvisierte Kundensegment besteht auch heute weiterhin fort. Das verhältnismäßig kleine EZ-Engagement limitierte die umfassende Verbesserung des Zugangs zu Finanzdienstleistungen sowie die Entwicklung und das Wachstum des Finanzsektors.“
- Empfehlungen: „Für eine langfristige nachhaltige Verbesserung des Zugangs zu Finanzdienstleistungen für KKMU wäre eine größere Investition sinnvoll gewesen, die die Erreichung deutlich größerer Marktanteile ermöglicht hätte.“

Onlinebefragung: Nein

### 26. Berücksichtigung Kapazitätsentwicklung

| *Beschreibung* | *nicht erfüllt (1)* | *eher nicht erfüllt (2)* | *eher erfüllt (3)* | *vollständig erfüllt (4)* |
|---|---|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn im Evaluierungsprozess Aktivitäten zur Entwicklung von Evaluierungskapazitäten der Akteur*innen im Partnerland durchgeführt werden (1. Verbesserung individueller Evaluierungskenntnisse, 2. Stärkung des Evaluierungsmanagements, 3. Förderung der Nutzung von Evaluierungsergebnissen, 4. Unterstützung eines Klimas/Umfelds von Rechenschaftspflicht und Lernen in den Partnerländern). | Keine der vier Aktivitäten wird im Evaluierungsprozess durchgeführt. | Eine der vier Aktivitäten wird im Evaluierungsprozess durchgeführt. | Zwei der vier Aktivitäten werden im Evaluierungsprozess durchgeführt. | Drei oder vier der vier Aktivitäten werden im Evaluierungsprozess durchgeführt. |

Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, Leistungsbeschreibung

Kodierregeln:

1. Aktivität 1: Verbesserung individueller Evaluierungskenntnisse: Mindestens ein*e lokale*r/nationale*r Gutachtende wird in Entscheidungsprozesse der Evaluierung einbezogen. Eine Unterstützung an der Datenerhebung zum Beispiel mittels Übersetzungen ist nicht ausreichend. Eventuell gibt es auch ausschließlich lokale/nationale Gutachtende bei einer Evaluierung. Diese sind dann automatisch in allen Entscheidungsprozessen involviert.
2. Aktivität 2: Stärkung des Evaluierungsmanagements: Mindestens eine Partnerorganisation wird in den Planungs- und Konzeptionsprozess der Evaluierung einbezogen (zum Beispiel durch gemeinsame Erarbeitung der Leistungsbeschreibung oder der Evaluierungsfragen). Stichwortsuche: zum Beispiel *partner organis(z)ation*, *contracted*, *involved*, *partner agencies*. Tipp: Es ist auch hilfreich im *preface* (Vorwort) oder in *acknowledgements* (Danksagung) zu suchen.
3. Aktivität 3: Förderung der Nutzung von Evaluierungsergebnissen: Vorstellung und gemeinsame Diskussion der Ergebnisse der Evaluierung mit den Partnerorganisationen. Auch eine Kommentierung des Evaluierungsberichtes durch die Partnerorganisationen ist hier bewertungsrelevant. Wird der Evaluierungsbericht lediglich an die Partnerorganisationen versendet, muss der Bericht in die lokale Sprache übersetzt sein, damit diese Aktivität erfüllt ist. Stichwortsuche: *early* oder *preliminary finding*, *debriefing*, Reiseverlauf, *time plan*, *time schedule.*
4. Aktivität 4: Unterstützung eines Klimas/Umfelds von Rechenschaftspflicht und Lernen in den Partnerländern: Es besteht eine Kommunikation/ein Austausch über die Evaluierung mit relevanten Ministerien in den Partnerländern wie dem Planungs-, Entwicklungs- oder Landwirtschaftsministerium. Werden Repräsentant*innen von Ministerien ausschließlich als Gesprächs- beziehungsweise Interviewpartner*innen genannt, ist dies nicht ausreichend. Stichwortsuche: *ministry.*
5. Mögliche Suchbegriffe: *local/national consultant*; *nationale*r/lokale*r Gutachter*in/Gutachtende.*

Ankerbeispiel: /

Onlinebefragung: Nein

### 27. Berücksichtigung von Gemeinschaftsevaluierungen

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn Geber und Partnerregierung die Möglichkeit einer Gemeinschaftsevaluierung prüfen. | Geber und Partnerregierung haben die Möglichkeit einer Gemeinschaftsevaluierung nicht geprüft. | Geber und Partnerregierung haben nachweisbar die Möglichkeit einer Gemeinschaftsevaluierung geprüft. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Leistungsbeschreibung, Inception Report | | |
| Kodierregeln:<br>1. Eine Gemeinschaftsevaluierung liegt dann vor, wenn mindestens ein Geber und eine Partnerregierung a) gemeinsam ein Evaluierungskonzept erarbeiten, b) die Evaluierung gemeinsam steuern, c) einen gemeinsamen Evaluierungsbericht schreiben und d) die Evaluierung gemeinsam finanzieren.<br>2. Das Qualitätskriterium ist erfüllt (mit 4 zu bewerten), sobald die Möglichkeit (*options*) einer Gemeinschaftsevaluierung geprüft wird. Die tatsächliche Durchführung der Gemeinschaftsevaluierung ist für die Erfüllung des Qualitätskriteriums nicht notwendig. Wurde eine Evaluierung als Gemeinschaftsevaluierung durchgeführt, gilt das Qualitätskriterium als erfüllt.<br>3. Das Qualitätskriterium meint nicht die gemeinsame Implementierung einer EZ-Maßnahme mit anderen Gebern/Partnerregierungen.<br>4. Mögliche Suchbegriffe: *Gemeinschaftsevaluierung, gemeinschaftliche Evaluierung (joint evaluation), Steuerung (steering/management), Kofinanzierung (co-financing oder co-funding), BMZ(-finanziert), Geber (donor), partner.* | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Prüfen Sie bei den Evaluierungen Ihrer Organisation, ob diese als Gemeinschaftsevaluierung mit Gebern und Partnerregierungen durchgeführt werden können? | | |

### 28. Berücksichtigung partnerschaftlicher Ansätze

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Evaluierung nationale oder lokale Evaluierungsaktivitäten und Evaluierungspläne/-politiken berücksichtigt. | a) Weder Evaluierungsaktivitäten noch Evaluierungspläne/-politiken werden in der Evaluierung berücksichtigt.<br>oder<br>b) Es wird einer der beiden Bereiche in der Evaluierung berücksichtigt. | Sowohl Evaluierungsaktivitäten als auch Evaluierungspläne/-politiken werden in der Evaluierung berücksichtigt. |

Dokumente: Evaluierungsbericht und Berichtsanhänge, Leistungsbeschreibung, Inception Report

Kodierregeln:

1. Evaluierungsaktivitäten bedeutet, dass sich die Evaluierung mit bereits vor Ort laufenden Evaluierungen koordiniert. Es sollte erklärt werden, wie genau die Koordination/Abstimmung stattfindet und zu welchem Thema.
2. Evaluierungspläne/-politiken: Dazu gehören zum Beispiel *national evaluation guidelines* (Leitfaden), *manuals* (Handbücher), *plans* oder *policies* der jeweiligen Partnerländer. Darüber hinaus können auch nationale (aber auch regionale, zum Beispiel Lateinamerika) Evaluierungsstandards dazu gehören. Berücksichtigt bedeutet, dass die Evaluierung sich in ihrer Konzeption und Durchführung an nationalen Evaluierungsplänen oder -politiken orientiert. Es sollte erklärt werden, wie genau sich die Evaluierung an den Plänen oder Politiken orientiert.
3. Für die Suche ist ein Blick in das Inhaltsverzeichnis hilfreich. Gegebenenfalls gibt es dort einen Abschnitt zu Evaluierungsablauf beziehungsweise -prozess, Vorgehensweise oder *evaluation objective*. Auch ein Blick in das Literaturverzeichnis kann hilfreich sein.
4. Mögliche Suchbegriffe: *ongoing evaluation, coordination, Evaluierungspläne/-politiken: (national) guidelines, manuals, plans, policies, standard, AEG (African Evaluation Guidelines), Evaluation Standards for Latin America and the Caribbean (RELAC).*

Ankerbeispiel: /

Onlinebefragung: Berücksichtigen Evaluierungen Ihrer Organisation nationale Evaluierungsaktivitäten und Evaluierungspläne/ -politiken der Partnerländer, in denen die Evaluierung durchgeführt wird?

### 29. Genderbalance im Team

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn bei der Zusammenstellung des Gutachtenden-Teams auf einen ausgewogenen Anteil von Männern und Frauen geachtet wird. | Bei der Zusammenstellung des Gutachtenden-Teams wird auf keinen ausgewogenen Anteil von Männern und Frauen geachtet. | Bei der Zusammenstellung des Gutachtenden-Teams wird auf einen ausgewogenen Anteil von Männern und Frauen geachtet. |

Dokumente: Leistungsbeschreibung

Kodierregeln:

1. Das Qualitätskriterium wird mit 4 bewertet, wenn in der Leistungsbeschreibung erwähnt wird, dass bei der Auswahl des Gutachtenden-Teams auf einen ausgewogenen Anteil von Männern und Frauen geachtet wird, also intendiert wird, dies zu erreichen. Es geht nicht um das tatsächliche Verhältnis zwischen Frauen und Männern im Gutachtenden-Team, das für die Durchführung der Evaluierung verantwortlich ist.
2. Informationen zu diesem Qualitätskriterium lassen sich in der Leistungsbeschreibung bei den geforderten Qualifikationen für Gutachtende oder Evaluator*innen finden.
3. Wenn keine Leistungsbeschreibung vorliegt, wird das Qualitätskriterium mit 1 bewertet.
4. Ein einfacher Hinweis in der Leistungsbeschreibung auf den Begriff *diversity* (Diversität) ist nicht ausreichend für eine Bewertung mit 4, da *diversity* verschiedene Aspekte (Nationalität, Kulturen etc.) beinhalten kann. Es muss explizit Aussagen zu *gender diversity* oder *balance* geben.
5. Mögliche Suchbegriffe: *geforderte Qualifikation, qualifications, required qualifications, diversity.*

Ankerbeispiel (nicht erfüllt):

- „Qualifikationen der Gutachter/innen. Die Evaluierung wird von zwei Personen durchgeführt, die über fundierte Erfahrungen und fachlichen Hintergrund in der Arbeit in internationalen Projekten, umfassenden Methodenkenntnissen sowie soweit möglich einem Bezug zu zivilgesellschaftlichen Prozessen in Israel und Kenntnisse über die Region verfügen. Die Gutachter/innen sollten eine berufliche Distanz zum Auftraggeber haben. Die Gutachter/innen sind federführend für die Einsatzplanung, das Vorgehen und die Abfassung des Berichts verantwortlich.“

Onlinebefragung: Wird in Evaluierungen, die von mindestens zwei Gutachter*innen durchgeführt wird, auf einen ausgewogenen Anteil von Männer und Frauen im Gutachter*innen-Team geachtet?

### 30. Einbezug von Gutachtenden aus dem Partnerland

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn sich im Gutachtenden-Team Fachleute aus den betroffenen Partnerländern befinden. | Das Gutachtenden-Team umfasst kein Teammitglied aus den betroffenen Partnerländern. | Das Gutachtenden-Team umfasst mindestens ein Teammitglied aus den betroffenen Partnerländern. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge, Inception Report, sonstige Dokumente | | |
| Kodierregeln:<br>1. Sogenannte „Enumerator“ (Angestellte, die ausschließlich die [oder Teile der] Datenerhebung/Befragung durchführen), die im Partnerland angestellt werden, zählen nicht als lokale/nationale Gutachtende.<br>2. Wenn lokale/nationale Gutachtende aus benachbarten Ländern und Regionen kommen, ist das Qualitätskriterium nicht erfüllt.<br>3. Wenn EZ-Maßnahmen in mehreren Partnerländern evaluiert werden, lokale/nationale Gutachtende aber nur aus einem der Partnerländer kommen, ist das Qualitätskriterium nicht erfüllt.<br>4. Mögliche Suchbegriffe: *evaluation team, team, regional expertise, local evaluator, national (consultant, evaluator).* | | |
| Ankerbeispiel (vollständig erfüllt):<br>• “The evaluation team was composed of an international and a local evaluator. […] while the local evaluator from Mexico […]. Together, both evaluators reflected the findings from the documents and interviews in light of the specific country context of Mexico.” | | |
| Onlinebefragung: Nein | | |

### 31. Ausreichende Ressourcen vorhanden

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn die Ressourcen (zum Beispiel in Bezug auf Finanzierung und/oder Personalausstattung) ausreichend sind, um die Ziele der Evaluierung zu erreichen. | Die Ressourcen sind unzureichend, um die Ziele der Evaluierung zu erreichen. | Die Ressourcen sind ausreichend, um die Ziele der Evaluierung zu erreichen. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | |
| Kodierregeln:<br>1. Relevante Aussagen darüber, ob die Ressourcen ausreichend sind, um die Evaluierungsziele zu erreichen, finden sich gegebenenfalls in Abschnitten zu den Limitationen der Evaluierung oder im Ergebnisteil. Es geht nicht um Ressourcen der EZ-Maßnahme, sondern um die Ressourcen der Evaluierung.<br>2. Gegebenenfalls finden sich Informationen über Ressourcen/Personalausstattung in einem Abschnitt zu *self-evaluation*“, in der das Gutachtenden-Team die eigene Evaluierung bewertet.<br>3. Es muss explizit Hinweise darauf geben, dass aufgrund nicht vorhandener Ressourcen das übergeordnete Evaluierungsziel nicht erreicht werden konnte (zum Beispiel „assess the impact of the project on poverty reduction“).<br>4. Vereinzelte Hinweise darauf, dass aufgrund von nicht ausreichenden Ressourcen bestimmte Arbeitsschritte (zum Beispiel Durchführung von Fokusgruppen) nicht durchgeführt werden konnten, sind nicht ausreichend für „vollständig erfüllt.“ Es muss klar und nachvollziehbar sein, dass in der Folge das Evaluierungsziel nicht erreicht werden konnte.<br>5. Mögliche Suchbegriffe: *financial resources*, *human resources*, *time resources*, *funds*, *budget* | | |
| Ankerbeispiel: / | | |
| Onlinebefragung: Sind die für die Evaluierungen zur Verfügung stehenden Ressourcen ausreichend, um die Ziele der Evaluierungen zu erreichen? | | |

### 32. Vorhandensein einer Management-Response

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn eine formelle Stellungnahme („Management-Response“) verfasst wird. | Es gibt keine Hinweise, dass eine formelle Stellung-nahme verfasst wird. | Es gibt Hinweise, dass eine formelle Stellungnahme verfasst wird. |
| Dokumente: Bericht und Berichtsanhänge, Leistungsbeschreibung, Inception Report, sonstige Dokumente | | |
| Kodierregeln:<br>1. Liegt eine Management-Response zur Evaluierung unter den Evaluierungsdokumenten vor, ist das Qualitätskriterium mit 4 zu bewerten.<br>2. Zum Zeitpunkt der Veröffentlichung/Publikation des Evaluierungsberichts liegt zumeist noch keine Management-Response vor. | | |
| Ankerbeispiel (vollständig erfüllt):<br>• „This evaluation is order by the evaluation unit and performed by external evaluator/s. The report is written by external evaluator/s. It is the opinion and the valuation of the evaluator/s. Organization X wrote the statement and the management response.“ | | |
| Onlinebefragung: Wird bei den Evaluierungen Ihrer Organisation eine „Management Response“ verfasst? | | |

### 33. Anwendung OECD-DAC-Kriterium – Relevanz

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn das OECD-DAC-Kriterium Relevanz angemessen bearbeitet wurde. | In der Evaluierung wurde keine Prüffrage laut BMZ-Evaluierungskriterien bearbeitet. | Mindestens eine Prüffrage laut BMZ-Evaluierungskriterien wurde bearbeitet. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | |
| Kodierregeln:<br>1. Bewertungsrelevante Hinweise finden sich vor allem im Ergebnisteil des Evaluierungsberichts. Dort sind die Ergebnisse häufig nach OECD-DAC-Kriterien strukturiert und die Bearbeitung des entsprechenden Kriteriums kann überprüft werden. Ist der Ergebnisteil nicht nach OECD-DAC-Kriterien unterteilt, empfiehlt es sich, über die Stichwortsuche oder Evaluierungsfragen im Ergebnisteil nach Aspekten des jeweiligen Kriteriums zu suchen.<br>2. Im Evaluierungsbericht muss es keinen expliziten Verweis auf die Prüffragen des BMZ geben, solange sich die Ergebnisse inhaltlich mit dem Kern der Prüffragen beschäftigen. Ebenso kann es vorkommen, dass die Prüffragen in leicht abgeänderter Form behandelt werden. Wichtig ist hierbei, dass die inhaltliche Substanz mit der jeweiligen Prüffrage in Zusammenhang gebracht werden kann.<br>3. Es kann vorkommen, dass Aspekte des OECD-DAC-Kriteriums Relevanz (zum Beispiel zur Frage der Ausrichtung der Projektziele an den Bedarfen der Zielgruppe) in der Evaluierung bearbeitet werden, ohne dass dabei explizit Bezug auf das Kriterium genommen wird. Hierbei ist das Qualitätskriterium auch erfüllt, solange der inhaltliche Kern der Prüffragen bearbeitet wird.<br>4. Die Prüffragen finden sich im BMZ-Dokument „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit“ von 2006. | | |
| Ankerbeispiel (vollständig erfüllt):<br>• „Assessment dimension: Relevance. 1.1. The project concept* is in line with the relevant strategic reference frameworks. 1. Which strategic reference frameworks exist for the project? (e.g. national strategies incl. National implementation strategy for 2030 agenda, regional and international strategies, sectoral, […].“ | | |
| Onlinebefragung: Nein | | |

## 34. Anwendung OECD-DAC-Kriterium – Effektivität

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn das OECD-DAC-Kriterium Effektivität angemessen bearbeitet wurde. | In der Evaluierung wurde keine Prüffrage laut BMZ-Evaluierungskriterien bearbeitet. | Mindestens eine Prüffrage laut BMZ-Evaluierungskriterien wurde bearbeitet. |

Dokumente: Evaluierungsbericht und Berichtsanhänge

Kodierregeln:

1. Bewertungsrelevante Hinweise finden sich vor allem im Ergebnisteil des Evaluierungsberichts. Dort sind die Ergebnisse häufig nach OECD-DAC-Kriterien strukturiert und die Bearbeitung des entsprechenden Kriteriums kann überprüft werden. Ist der Ergebnisteil nicht nach OECD-DAC-Kriterien unterteilt, empfiehlt es sich, über die Stichwortsuche oder Evaluierungsfragen im Ergebnisteil nach Aspekten des jeweiligen Kriteriums zu suchen.
2. Im Evaluierungsbericht muss es keinen expliziten Verweis auf die Prüffragen des BMZ geben, solange sich die Ergebnisse inhaltlich mit dem Kern der Prüffragen beschäftigen. Ebenso kann es vorkommen, dass die Prüffragen in leicht abgeänderter Form behandelt werden. Wichtig ist hierbei, dass die inhaltliche Substanz mit der jeweiligen Prüffrage in Zusammenhang gebracht werden kann.
3. Es kann vorkommen, dass Aspekte des OECD-DAC-Kriteriums Effektivität in der Evaluierung bearbeitet werden, ohne dass dabei explizit Bezug auf das Kriterium genommen wird. Hierbei ist das Qualitätskriterium auch erfüllt, solange der inhaltliche Kern der Prüffragen bearbeitet wird.
4. Die Prüffragen finden sich im BMZ-Dokument „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit“ von 2006.

Ankerbeispiel (vollständig erfüllt):

- „Inwieweit erscheinen die Ziele [der Jahre 2012 – 2017] des Projekts im Rückblick realistisch und erreichbar? Wurden die Ziele erreicht?“

Onlinebefragung: Nein

## 35. Anwendung OECD-DAC-Kriterium – Effizienz

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn das OECD-DAC-Kriterium Effizienz angemessen bearbeitet wurde. | In der Evaluierung wurde keine Prüffrage laut BMZ-Evaluierungskriterien bearbeitet. | Mindestens eine Prüffrage laut BMZ-Evaluierungskriterien wurde bearbeitet. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | |
| Kodierregeln:<br>1. Bewertungsrelevante Hinweise finden sich vor allem im Ergebnisteil des Berichts. Dort sind die Ergebnisse häufig nach OECD-DAC-Kriterien strukturiert und die Bearbeitung des entsprechenden Kriteriums kann gemäß dem Qualitätskriterium überprüft werden. Ist der Ergebnisteil nicht nach OECD-DAC-Kriterien unterteilt, empfiehlt es sich, über die Stichwortsuche oder Evaluierungsfragen im Ergebnisteil nach Aspekten des jeweiligen Evaluierungskriteriums zu suchen.<br>2. Im Evaluierungsbericht muss es keinen expliziten Verweis auf die Prüffragen des BMZ geben, solange sich die Ergebnisse inhaltlich mit dem Kern der Prüffragen beschäftigen. Ebenso kann es vorkommen, dass die Prüffragen in leicht abgeänderter Form behandelt werden. Wichtig ist hierbei, dass die inhaltliche Substanz mit der jeweiligen Prüffrage in Zusammenhang gebracht werden kann.<br>3. Es kann vorkommen, dass Aspekte des OECD-DAC-Kriteriums Effizienz (zum Beispiel inwieweit Kosten und Nutzen in einem angemessenen Verhältnis zueinander stehen) in der Evaluierung bearbeitet werden, ohne dass dabei explizit Bezug auf das Kriterium genommen wird. Hierbei ist das Qualitätskriterium auch erfüllt, solange der inhaltliche Kern der Prüffragen bearbeitet wird.<br>4. Die Prüffragen finden sich im BMZ-Dokument „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit“ von 2006. | | |
| Ankerbeispiel (vollständig erfüllt):<br>• „At the time of the most recent audit report, June 2017, the project, which is planned to end on 30 June 2018, had run 71% of its full course. At this point the project had spent 47% of the initial total budget, indicating some significant under spending. This level of under spending is a significant concern.“ | | |
| Onlinebefragung: Nein | | |

## 36. Anwendung OECD-DAC-Kriterium – Wirkungen

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn das OECD-DAC-Kriterium Wirkungen angemessen bearbeitet wurde. | In der Evaluierung wurde keine Prüffrage laut BMZ-Evaluierungskriterien bearbeitet. | Mindestens eine Prüffrage laut BMZ-Evaluierungskriterien wurde bearbeitet. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | |
| Kodierregeln:<br>1. Bewertungsrelevante Hinweise finden sich vor allem im Ergebnisteil des Berichts. Dort sind die Ergebnisse häufig nach OECD-DAC-Kriterien strukturiert und die Bearbeitung des entsprechenden Kriteriums kann gemäß dem Qualitätskriterium überprüft werden. Ist der Ergebnisteil nicht nach OECD-DAC-Kriterien unterteilt, empfiehlt es sich, über die Stichwortsuche oder Evaluierungsfragen im Ergebnisteil nach Aspekten des jeweiligen Evaluierungskriteriums zu suchen.<br>2. Im Evaluierungsbericht muss es keinen expliziten Verweis auf die Prüffragen des BMZ geben, solange sich die Ergebnisse inhaltlich mit dem Kern der Prüffragen beschäftigen. Ebenso kann es vorkommen, dass die Prüffragen in leicht abgeänderter Form behandelt werden. Wichtig ist hierbei, dass die inhaltliche Substanz mit der jeweiligen Prüffrage in Zusammenhang gebracht werden kann.<br>3. Es kann vorkommen, dass Aspekte des OECD-DAC-Kriteriums Wirkungen (zum Beispiel inwieweit übergeordnete entwicklungspolitische Wirkungen erreicht wurden) in der Evaluierung bearbeitet werden, ohne dass dabei explizit Bezug auf das Kriterium genommen wird. Hierbei ist das Qualitätskriterium auch erfüllt, solange der inhaltliche Kern der Prüffragen bearbeitet wird.<br>4. Die Prüffragen finden sich im BMZ-Dokument „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit“ (2006). | | |
| Ankerbeispiel (vollständig erfüllt):<br>• „Wie sind die festgestellten übergeordneten Wirkungen des Projektes unter Abwägung positiver und negativer Effekte zusammenfassend zu bewerten?“ | | |
| Onlinebefragung: Nein | | |

## 37. Anwendung OECD-DAC-Kriterium – Nachhaltigkeit

| *Beschreibung* | *nicht erfüllt (1)* | *vollständig erfüllt (4)* |
|---|---|---|
| Das Qualitätskriterium ist erfüllt, wenn das OECD-DAC-Kriterium Nachhaltigkeit angemessen bearbeitet wurde. | In der Evaluierung wurde keine Prüffrage laut BMZ-Evaluierungskriterien bearbeitet. | Mindestens eine Prüffrage laut BMZ-Evaluierungskriterien wurde bearbeitet. |
| Dokumente: Evaluierungsbericht und Berichtsanhänge | | |
| Kodierregeln:<br>1. Bewertungsrelevante Hinweise finden sich vor allem im Ergebnisteil des Berichts. Dort sind die Ergebnisse häufig nach OECD-DAC-Kriterien strukturiert und die Bearbeitung des entsprechenden Kriteriums kann gemäß dem Qualitätskriterium überprüft werden. Ist der Ergebnisteil nicht nach OECD-DAC-Kriterien unterteilt, empfiehlt es sich, über die Stichwortsuche oder Evaluierungsfragen im Ergebnisteil nach Aspekten des jeweiligen Evaluierungskriteriums zu suchen.<br>2. Im Evaluierungsbericht muss es keinen expliziten Verweis auf die Prüffragen des BMZ geben, solange sich die Ergebnisse inhaltlich mit dem Kern der Prüffragen beschäftigen. Ebenso kann es vorkommen, dass die Prüffragen in leicht abgeänderter Form behandelt werden. Wichtig ist hierbei, dass die inhaltliche Substanz mit der jeweiligen Prüffrage in Zusammenhang gebracht werden kann.<br>3. Es kann vorkommen, dass Aspekte des OECD-DAC-Kriteriums Nachhaltigkeit (zum Beispiel inwieweit die Wirkungen von Dauer sind) in der Evaluierung bearbeitet werden, ohne dass dabei explizit Bezug auf das Kriterium genommen wird. Hierbei ist das Qualitätskriterium auch erfüllt, solange der inhaltliche Kern der Prüffragen bearbeitet wird.<br>4. Die Prüffragen finden sich im BMZ-Dokument „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit“ von 2006. | | |
| Ankerbeispiel (nicht erfüllt):<br>• Überschrift: „Zusammenarbeit mit Partner/innen und Nachhaltigkeit.“ Dazu untersuchte Fragestellung: „Wie ist die Auswahl von und Zusammenarbeit mit Projektpartner/innen in Bezug auf die strategische Fortentwicklung der Stiftungsarbeit anhand der formulierten Ziele zu bewerten?“ | | |
| Onlinebefragung: Nein | | |