ENDBERICHT

Wirksamkeitsmessung in der internationalen Zusammenarbeit: Metaevaluation DEZA/SECO/AFM

Im Auftrag der Schweizer Parlamentarischen Verwaltungskontrolle (PVK)

Autoren:
Dr. Stefan Silvestrini (Teamleiter)
Julie Ngo, M.A.

Kontakt:
Dr. Stefan Silvestrini
CEval GmbH
Dudweiler Landstrasse 5
66123 Saarbrücken

Phone +49 681 387539 73
E-Mail s.silvestrini@ceval.de
URL http://www.ceval.de

Saarbrücken, 17. April 2023

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Executive Summary

Die Geschäftsprüfungskommission (GPK) der Schweizer Bundesversammlung hat die Parlamentarische Verwaltungskontrolle (PVK) mit der Wirksamkeitsmessung in der Schweizer internationalen Zusammenarbeit (IZA) beauftragt. Hierfür soll u.a. die Qualität, der in der IZA eingesetzten Evaluationen überprüft werden. Die PVK hat diesen Teilauftrag als Mandat zur „Wirksamkeitsmessung in der internationalen Zusammenarbeit: Metaevaluation SECO/AFM“ (im Folgenden als „Metaevaluation“ bezeichnet) an die CEval GmbH vergeben. Die im Rahmen der Metaevaluation zu beantwortenden Fragen lauten wie folgt:

1. Stimmt die Qualität der Evaluationen?
2. Werden die Evaluationen für die Steuerung der IZA genutzt?
3. Sind die Evaluationskosten angemessen?

Gegenstand der Metaevaluation sind 42 vom SECO und 60 von der DEZA zufällig ausgewählte sowie alle 12 von der AFM zwischen 2018 und 2020 in Auftrag gegebenen Evaluationen. Während zur Beantwortung der Frage nach der Nutzung von Evaluationen durch deren Adressaten die PVK eigene Erhebungen und Analysen durchführen wird, befasst sich der vorliegende Bericht mit den Antworten auf die Fragen zur Qualität und den Kosten der Evaluationen.

Zur Beantwortung der Frage zur Qualität der Evaluationen wurden deren Leistungsbeschreibungen und Schlussberichte der Evaluationen sowohl einer qualitativen Inhaltsanalyse als auch einer quantitativen Analyse unterzogen. Hierbei kam ein Analyseraster zum Einsatz, mittels dessen die Qualität der Evaluationen entlang folgender Kriterien überprüft wurde: 1. **Leistungsbeschreibung**, 2. **Executive Summary**, 3. **Einleitung und Kontextanalyse**, 4. **Methodik**, 5. **Ergebnisdarstellung**, 6. **Schlussfolgerungen und Empfehlungen**. Zur Gewährleistung möglichst nachvollziehbarer und verlässlicher Untersuchungsergebnisse wurden diese Kriterien mittels Teilkriterien und Indikatoren operationalisiert. Die Ergebnisse der quantitativen Analyse wurden des Weiteren zwecks Bewertung ausgewählter Zusammenhänge zwischen den Budgets der Evaluationen, ihrer Qualität sowie ihren Ergebnissen einer Korrelationsanalyse unterzogen.

Ergebnisse

Die **Leistungsbeschreibungen** von SECO, AFM und DEZA unterscheiden sich in ihrer Qualität z.T. erheblich. Während der Anteil ausreichend dargestellter Verantwortlichkeiten, Leistungen und Zeitplan bei den drei Verwaltungseinheiten zwischen ca. 60% und 70% noch relativ nahe beieinander liegt, zeigen sich deutliche Unterschiede bei den Beschreibungen des Kontexts, in dem die Evaluation durchgeführt wird, des Evaluationsgegenstands selbst, des Hintergrunds und des Zwecks der Evaluation, ihres Gegenstandsbereichs, der dabei zu behandelnden Kriterien bzw. Fragen sowie bei den methodischen Vorgaben. Dabei erfüllen die Leistungsbeschreibungen der DEZA die Teilkriterien Kontext, Gegenstand, Fragen/Kriterien, Methodik am ehesten in zumindest zufriedenstellendem Ausmass, während Hintergrund und Zweck lediglich in den Leistungsbeschreibungen des SECO in mehr als der Hälfte der Fälle hinreichend beschrieben werden. Schliesslich gelingt es der AFM offenbar am besten, den Gegenstandsbereich der avisierten Evaluation adäquat einzugrenzen.

Die **Executive Summaries** von mindestens zwei Drittel der Berichte sind eigenständig verständlich und sie stimmen inhaltlich mit dem Hauptbericht überein. Ihr jeweiliger Aufbau unterscheidet sich jedoch mitunter wesentlich. Auffällig ist hierbei insbesondere, dass nur knapp jedes zehnte Executive Summary der AFM in zufriedenstellender Weise, die darin üblicherweise zur Verfügung gestellten Informationen enthält.

Bei der Kontextanalyse setzen die drei Verwaltungseinheiten jeweils unterschiedliche Schwerpunkte. Während die Anteile von in zumindest zufriedenstellendem Ausmass beschriebenen Hintergründen und Zwecken der jeweiligen Evaluationen mit zwischen 42 und 50 Prozent relativ wenig voneinander abweichen, unterscheidet sich die Qualität der Kontextanalyen und der Beschreibungen der Evaluationsgegenstände z.T. erheblich. Während bei der DEZA in immerhin einem guten Drittel der Berichte

die Rahmenbedingungen, unter denen das evaluierte Vorhaben operiert (hat) in ausreichendem Mass beleuchtet werden, ist dies beim SECO nur in jedem fünften Bericht der Fall, bei der AFM sogar in nur einem einzigen. Eine angemessene Beschreibung des evaluierten Vorhabens selbst findet sich hingegen am ehesten in Evaluationsberichten des SECO. Bei der DEZA gelingt dies nur in gut der Hälfte und bei der AFM in einem Drittel der Berichte.

Die Evaluationsberichte des SECO, der AFM und der DEZA weisen in ihrer **methodischen Qualität** sehr grosse Ähnlichkeiten auf. Während die meisten Berichte ausreichend detaillierte Informationen zu den Datenquellen enthalten, wird in kaum einem Drittel angemessen auf Limitationen und Herausforderungen eingegangen, mit nur geringfügigen Unterschieden zwischen den drei Verwaltungseinheiten. Auch bei der Darstellung der Datenerhebung weisen alle Berichte gleichermassen Schwächen auf. Hierzu finden sich nur in etwa jedem vierten Fall mindestens zufriedenstellende Angaben in den jeweiligen Methodenkapiteln. Tendenziell noch schlechter sieht es bei den Beschreibungen von Evaluationsdesign, Stichprobenziehung und Datenanalyse aus. Bei allen drei Teilkriterien werden die hierfür in den Berichten zur Verfügung gestellten Inhalte in mindestens drei Viertel der Fälle als unzureichend bewertet. Dabei sticht hervor, dass in kaum fünf Prozent der SECO Berichte konkrete Angaben dazu gemacht werden, wie Befragte ausgewählt wurden, und bei der AFM kein Bericht identifiziert werden konnte, in dem in angemessener Weise das Evaluationsdesign und die Datenanalyse beschrieben wurde.

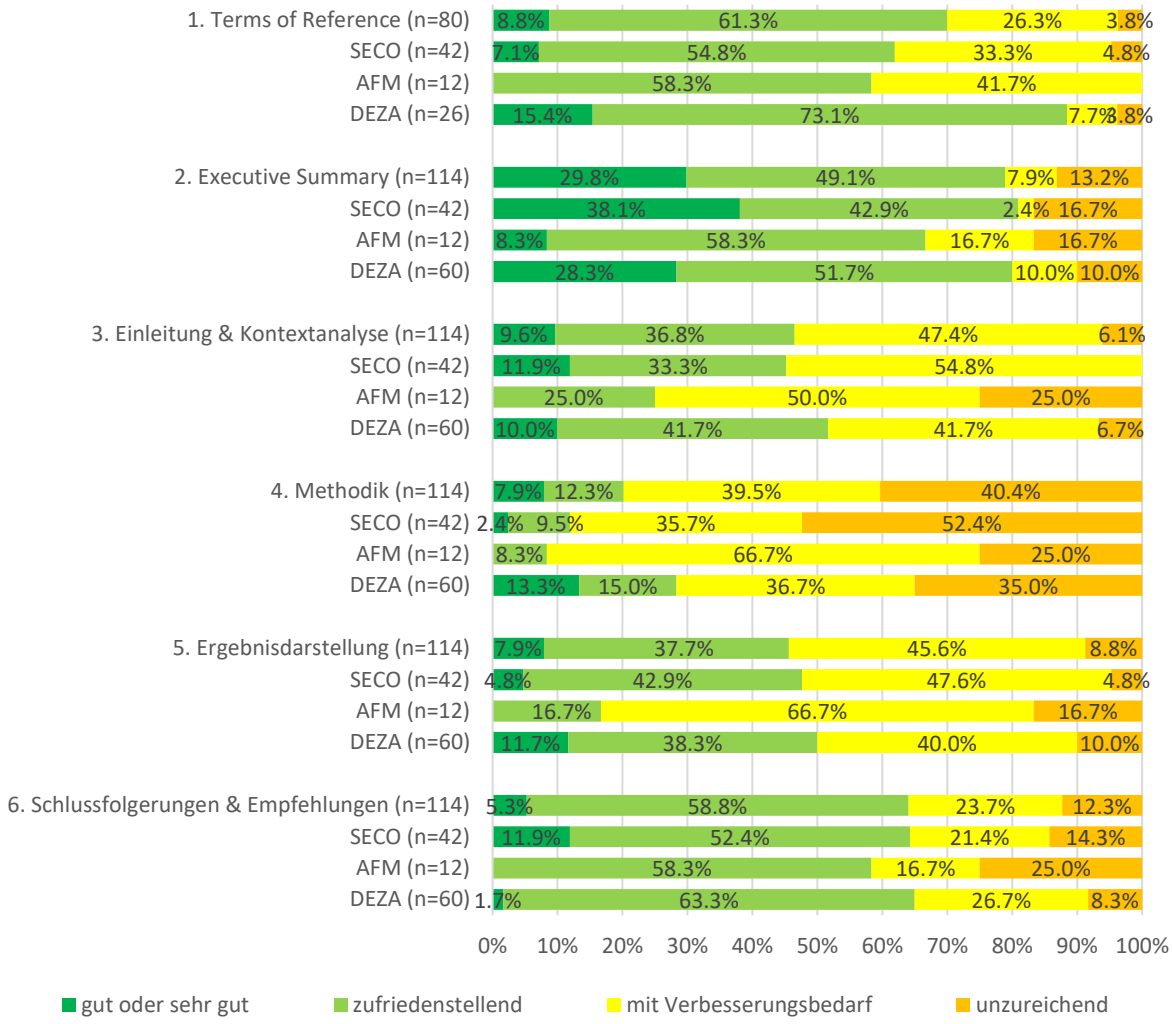
In etwa der Hälfte aller Berichte wird bei der **Ergebnisdarstellung** in mindestens zufriedenstellendem Mass auf die zugrundeliegenden Daten referenziert. Hierbei sind jedoch deutliche Unterschiede zwischen den drei Verwaltungseinheiten zu erkennen. So sind in SECO und DEZA Berichten deutlich häufiger entsprechende Verweise zu finden als in AFM Berichten. Auch die angemessene Darstellung der Relevanz des jeweils evaluierten Vorhabens gelingt in den Evaluationsberichten des SECO und der DEZA, mit in jeweils in etwa drei Viertel aller Fälle, besser als in denen der AFM (58,3%). Hinsichtlich der Beschreibung der Effektivität der Vorhaben sind hingegen keine nennenswerten Unterschiede zu erkennen. Diese wird im Schnitt in acht von zehn Fällen in ausreichender Qualität erörtert. Deutlich schlechter stellt sich die Diskussion des Impacts dar, die für die Gesamtstichprobe lediglich in einem guten Drittel der Berichte als angemessen bezeichnet werden kann, wobei dies der DEZA offenbar tendenziell häufiger gelingt als den anderen beiden Verwaltungseinheiten.[1] Bei der Qualität der Effizienzanalysen übertrifft wiederum das SECO mit annähernd zehn Prozentpunkten den Anteil der Berichte in der Gesamtstichprobe (50,0%), die in zumindest zufriedenstellender Weise die diesbezüglichen Ergebnisse darstellen. Die Nachhaltigkeitsbewertung erfolgt ebenso in etwa der Hälfte aller Berichte in ausreichendem Mass, wobei auch hier wiederum deutliche Unterschiede zwischen den drei Verwaltungseinheiten festgestellt werden können. Während sie bei der DEZA in 60 Prozent der Fälle angemessen diskutiert wird, trifft dies beim SECO nur in knapp der Hälfte aller Fälle zu und bei der AFM sogar nur bei einem Viertel.

Während in der grossen Mehrheit der Berichte **Schlussfolgerungen und Empfehlungen** logisch mit den Evaluationsergebnissen verknüpft sind, besteht erheblicher Verbesserungsbedarf hinsichtlich der Formulierung der Empfehlungen. Beide Befunde gelten ungeachtet gewisser Unterschiede im quantitativen Rating für das SECO, die AFM und die DEZA gleichermassen.

Die folgende Abbildung fasst die Ergebnisse nochmals überblicksartig zusammen:

[1] Hierbei ist zu berücksichtigen, dass bei der SECO bis 2020 der Impact kein verpflichtendes Bewertungskriterium war und er auch danach nur bei Schluss- und Ex-post Evaluationen zwingend bewertet werden musste.

Abbildung 1: Bewertung der Hauptkriterien[2]



Die Ergebnisse der Zusammenhangsanalyse stellen sich wie folgt dar:

1. Gibt es bei den externen Evaluationen von SECO und DEZA einen Zusammenhang zwischen bestimmten Qualitätskriterien und den aus den Evaluationen abgeleiteten Ratings der DAC-Kriterien?
   → Ja, die Daten weisen darauf hin, dass es bei Evaluationen des SECO einen positiven Zusammenhang zwischen der Gesamtbewertung eines Vorhabens und der Gesamtqualität des Evaluationsberichts gibt und dass bei der DEZA die Gesamtbewertung mit der Qualität der Ergebnisdarstellung zusammenhängt.
2. Gibt es einen Zusammenhang zwischen der Aussagekraft (Strength of Evidence) und der Erfolgsquote insgesamt (Overall Performance)?
   → Nein, die Untersuchungsergebnisse weisen nicht auf einen Zusammenhang zwischen der Aussagekraft der Evaluationsergebnisse und der damit ermittelten Erfolgsquote hin.
3. Gibt es einen Zusammenhang zwischen der Qualität der Diskussion der einzelnen DAC-Kriterien und deren Rating?
   → Ein direkter Zusammenhang zwischen der Qualität der Diskussion einzelner DAC-Kriterien und deren Rating konnte nur für die Nachhaltigkeitsbewertung in den Evaluationsberichten der DEZA festgestellt werden. Jedoch konnte sowohl für das SECO als auch die DEZA ein Zusammenhang zwischen der Gesamtbewertung der Vorhaben und der Qualität der Diskussion einzelner Bewertungsdimensionen identifiziert werden.

[2] Bei DEZA standen nur die ToR für 26 der 60 untersuchten Evaluationen zur Verfügung.

4. Gibt es einen Zusammenhang zwischen bestimmten Qualitätskriterien und den Kosten der Evaluationen?
   → Die Untersuchungsergebnisse legen keinen Zusammenhang zwischen der methodischen Qualität einer Evaluation und ihren Kosten nahe.
5. Gibt es einen Zusammenhang zwischen den Kosten der Evaluationen und den aus ihnen abgeleiteten DAC-Ratings?
   → Hinsichtlich des Zusammenhangs zwischen den Kosten der Evaluationen und den aus ihnen abgeleiteten DAC-Ratings sind die Untersuchungsergebnisse nicht eindeutig. Lediglich für Evaluationen der DEZA kann ein entsprechender Zusammenhang bestätigt werden.

1. Einleitung

Die Schweizer internationale Zusammenarbeit (IZA) verfolgt das Ziel, weltweit[3] einen Beitrag zur Linderung von Not und Armut, zur Achtung der Menschenrechte und zur Förderung der Demokratie, zu einem friedlichen Zusammenleben der Völker sowie zur Erhaltung der natürlichen Lebensgrundlagen zu leisten. Mit der IZA-Strategie 2021–2024 beantragt der Bundesrat fünf Rahmenkredite in der Höhe von 11,25 Milliarden Franken über vier Jahre. Themenschwerpunkte der aktuellen IZA-Strategie sind die Schaffung von menschenwürdigen Arbeitsplätzen vor Ort, der Kampf gegen den Klimawandel, die Reduktion der Ursachen von Flucht und irregulärer Migration sowie das Engagement für Rechtsstaatlichkeit.[4] Zur Umsetzung der aktuellen Strategie 2021-2024 steht den für die Umsetzung hierzu geeigneter Interventionen verantwortlichen Verwaltungseinheiten, d.h. der Direktion für Entwicklung und Zusammenarbeit (DEZA), der Abteilung Frieden und Menschenrechte (AFM) und dem Staatssekretariat für Wirtschaft (SECO), ein Budget von 11,5 Milliarden Franken zur Verfügung.

Zur Überprüfung der Wirksamkeit unterziehen alle Verwaltungseinheiten ihre Interventionen regelmässig Evaluationen. Diese Evaluationen dienen dabei neben der eigentlichen Wirksamkeitsmessung insbesondere der Schaffung empirischer Evidenzen zur Entscheidungsfindung und zur Verbesserung der Qualität der Interventionen. Weiterhin sind sie ein Kommunikationsinstrument, mittels dessen dem Parlament über den Erfolg der Schweizer IZA Bericht erstattet wird. Damit Evaluationen diese Aufgaben erfüllen können, müssen sie valide, reliable und möglichst objektive Ergebnisse liefern, was wiederum eine fachlich und methodisch korrekte Durchführung erfordert. Jüngst wurden jedoch Zweifel geäussert, ob die von den drei Einheiten vergebenen Evaluationen entsprechenden Qualitätsanforderungen genügen, ob sie tatsächlich als Steuerungsinstrument genutzt werden und ob ihre Kosten in einem angemessenen Verhältnis zu ihrem tatsächlichen Nutzen stehen.

Vor diesem Hintergrund hat die Geschäftsprüfungskommission (GPK) der Schweizer Bundesversammlung die Parlamentarische Verwaltungskontrolle (PVK) beauftragt, eine Bewertung der Wirkungsmessung in der IZA vorzunehmen. Ein Teil dieses Auftrags beinhaltet die Untersuchung der Qualität der in der IZA eingesetzten Evaluationen. Die PVK hat diesen Teilauftrag als Mandat zur „Wirksamkeitsmessung in der internationalen Zusammenarbeit: Metaevaluation SECO/AFM“ an die CEval GmbH vergeben.

Ziel der im Folgenden einfach als „Metaevaluation“ bezeichneten Untersuchung ist es entsprechend, die Qualität der zwischen 2018 und 2020 von dem SECO, der AFM und der DEZA in Auftrag gegebenen Evaluationen zu überprüfen. Als Datengrundlage dienen hierfür 42 zufällig ausgewählte der insgesamt 72 vom SECO und alle 12 von der AFM in diesem Zeitraum in Auftrag gegebenen Evaluationen, bzw. deren jeweilige Schlussberichte und Leistungsbeschreibungen (Terms of Reference, ToR). Weiterhin werden die Rohdaten der Bewertung von 60 Evaluationen aus der im vergangenen Jahr durchgeführten Metaevaluation für die DEZA[5] herangezogen.

Gemäss Pflichtenheft lauten die im Rahmen der Metaevaluation zu beantwortenden Fragen wie folgt:

4. Stimmt die Qualität der Evaluationen?
5. Werden die Evaluationen für die Steuerung der IZA genutzt?
6. Sind die Evaluationskosten angemessen?

Der vorliegende Bericht befasst sich mit den Antworten auf die Fragen 1 und 3 der Metaevaluation. Zur Frage der Nutzung von Evaluationen durch deren Adressaten zur Steuerung der IZA (Frage 2) wird

---

[3] Die dem Eidgenössischen Department für auswärtige Angelegenheiten (EDA) unterstehenden Verwaltungseinheiten DEZA und AFM fokussieren dabei ihre Arbeit auf die vier Schwerpunktregionen Nordafrika und Mittlerer Osten, Subsahara-Afrika, Asien (Zentral-, Süd- und Südostasien) und Osteuropa.

[4] https://www.eda.admin.ch/deza/de/home/aktuell/dossiers/alle-dossiers/iza-2021-2024.html

[5] Ngo, Julie; Krapp, Stefanie; Silvestrini, Stefan (2022): Quality Assessment of 60 Decentralised SDC Evaluations on behalf of the Swiss Agency for Development and Cooperation (SDC) of the Federal Department of Foreign Affairs (FDFA). Bern: Direktion für Entwicklung und Zusammenarbeit (DEZA) des Eidgenössischen Departements für auswärtige Angelegenheiten (EDA).

die PVK eigene Erhebungen und Analysen durchführen, weswegen deren Beantwortung nicht Bestandteil des Mandats ist und hier im Weiteren nicht erörtert wird.

Der Bericht ist wie folgt gegliedert: Nachdem in Kapitel 2 die Methodik der Metaevaluation erläutert wurde, widmet sich Kapitel 3 der Ergebnisdarstellung zur Qualität der Evaluationen. Die Bewertung erfolgt dabei gemäss der Struktur des in Kapitel 2 vorgestellten Analyserasters und differenziert nach den durch das SECO, die AFM und die DEZA durchgeführten Evaluationen. Anschliessend werden in Kapitel 5 die Ergebnisse der weiterhin gemäss Pflichtenheft durchzuführenden Zusammenhangsanalysen vorgestellt. Dem Bericht ist in einem separaten Dokument ein Anhang angefügt, der neben einer Liste der Evaluationen (7.1) und dem Analyseraster (7.2) die Ergebnisse der statistischen Gruppenvergleiche (7.3) sowie der Korrelationsanalysen (7.4) enthält.

2. Methodik

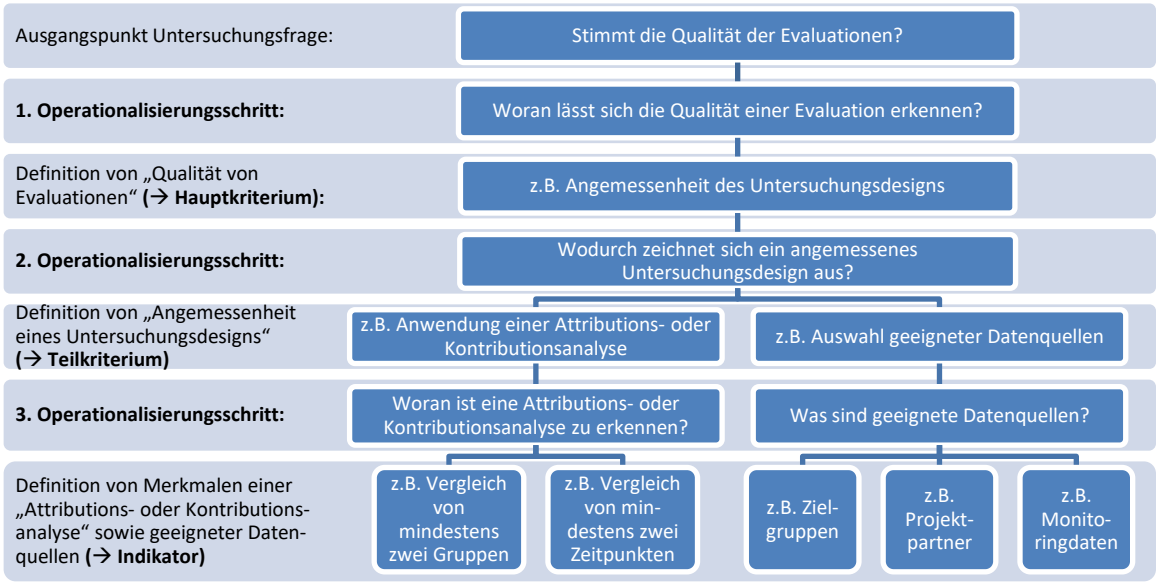
Den folgenden Ausführungen ist vorauszuschicken, dass die im Rahmen der Auswertung der SECO und AFM Evaluationen angewendete Methodik bis auf die hier nicht durchgeführe Nützlichkeitsbewertung, der Methodik der Metaevaluation im Auftrag der DEZA entspricht. Diese Vorgehensweise ist für einen Vergleich der Ergebnisse der drei Organisationen zwingend erforderlich. Insofern handelt es sich bei dieser Untersuchung um eine Replikationsstudie.

Zur Beantwortung der Frage zur Qualität der Evaluationen mussten die zur Verfügung gestellten Dokumente sowohl einer **qualitativen Inhaltsanalyse** als auch einer **quantitativen Analyse** unterzogen werden. Hierbei kam ein Analyseraster (vgl. Anhang 7.2) zum Einsatz[6], mittels dessen die Qualität der Evaluationen entlang folgender Kriterien, die der üblichen Struktur eines Evaluationsberichts, einschliesslich ihres Pflichtenhefts (Terms of Reference), entspricht, überprüft wurde:

1. Leistungsbeschreibung
2. Executive Summary
3. Einleitung und Kontextanalyse
4. Methodik
5. Ergebnisdarstellung
6. Schlussfolgerungen und Empfehlungen

Zur Gewährleistung möglichst nachvollziehbarer und verlässlicher Untersuchungsergebnisse wurden diese Kriterien, wie in der folgenden Grafik beispielhaft dargestellt, mittels Teilkriterien und Indikatoren operationalisiert.

[6] Um die Ergebnisse der Untersuchung mit denen der kürzlich durch die CEval GmbH durchgeführten Metaevaluation von Evaluationen der DEZA vergleichen zu können, wurde hier das selbe Analyseraster genutzt.



Da die Qualität der Evaluationen einer vergleichenden Analyse zu unterziehen ist, enthält das Raster weiterhin **Skalen** für die Bewertung jedes einzelnen Teilkriteriums und Indikators mit eindeutigen Definitionen für alle Bewertungsstufen. Zur Sicherstellung einer effizienten Analyse wurden hierbei für Indikatoren dichotome (ja/nein, vorhanden/nicht vorhanden) Skalen und für Teilkriterien ordinale Skalen mit den Werten: 1 = „inadequate“ („unzureichend“), 2 = „need for improvement“ („mit Verbesserungsbedarf“), 3 = „satisfactory“ („zufriedenstellend“), 4 = „good or very good“ („gut oder sehr gut“) verwendet. Zur Bewertung der Haupt- und Teilkriterien wurde schliesslich folgende **Auswertungsregel** auf Grundlage des jeweiligen Anteils der erfüllten Indikatoren festgelegt: 1 = ≤25%, 2 = >25-≤50%, 3 = >50-≤75%, 4 = >75%.

Insgesamt umfasst das Analyseraster 28 Teilkriterien und 139 Einzelindikatoren.

Datenanalyse

Die zur Verfügung gestellten Evaluationsberichte wurden computergestützt mit MaxQDA®, einer Software zur qualitativen und quantitativen Textanalyse, ausgewertet. Hierfür wurde das Analyseraster sowie die Dokumente zunächst in die Software übertragen. Die Bewertung erfolgte dann in drei Schritten: Als erstes wurden die **Berichte nach dem Raster kodiert**, indem alle relevanten Textfragmente (z.B. Textpassage in der die Datenquellen beschrieben werden) den jeweiligen Codes (z.B. Angabe zu Zielgruppenbefragung) zugeordnet. In einem zweiten Schritt wurden die für die Bewertung eines bestimmten Kriteriums (z.B. Angemessenheit des Untersuchungsdesigns) **relevanten Fragmente zusammengefasst** und gemäss der jeweiligen, vorher festgelegten Auswertungsregel bewertet. Um möglichst verlässliche Ergebnisse zu erhalten, wurde ein **Peer-Review-Verfahren** angewendet, in dem eine Reihe von Berichten von zwei Experten kodiert und analysiert wurden. Mit dieser Form der **Forschertriangulation** konnte sichergestellt werden, dass Unklarheiten in den Bewertungen aufgedeckt und damit minimiert wurden. Im dritten Schritt wurden die Bewertungen für jeden Bericht in eine Gesamtbewertungsmatrix übertragen, die schliesslich eine **vergleichende Analyse** der gesamten Stichprobe ermöglichte (z.B. im Hinblick auf signifikante Qualitätsunterschiede zwischen SECO, AFM und DEZA).

Die Daten wurden qualitativ und quantitativ ausgewertet. Während die Analyse der Dokumente dem Ansatz der **qualitativen Inhaltsanalyse** nach Mayring[7] folgte, wurden die daraus hervorgehenden

[7] Bei der qualitativen Inhaltsanalyse nach Mayring wird der Analyserahmen nicht wie bei anderen hermeneutischen Methoden, wie bspw. der Grounded Theory, aus dem Datenmaterial entwickelt, sondern a priori festgelegt. Dies ermöglicht eine zielgerichtete Untersuchung vorgegebener Fragestellungen, wie sie hier vorzunehmen ist. Vgl. Mayring, P. (1991). Qualitative Inhaltsanalyse. In U. Flick, E. v. Kardoff,

quantitativen Daten mit Hilfe **deskriptiver statistischer Methoden** ausgewertet. Die Qualitätsbewertung wird im Folgenden mittels gestapelter Balkendiagramme und Boxplots visualisiert. Für die Identifikation von Beziehungen zwischen den Bewertungen einzelner Kriterien (z.B. Qualität der ToR und Qualität der Methodik der Evaluation) wurden auf der Aggregatsebene Korrelationskoeffizienten[8] berechnet. Zur Überprüfung, inwieweit sich die Evaluationen von SECO, DEZA und AFM hinsichtlich ihrer Qualität voneinander unterscheiden, wurde schliesslich der so genannte Kruskal-Wallis-Test[9] auf die einzelnen Bewertungskriterien angewendet. Hierfür wurden die Daten auf der Ebene der Hauptkriterien in eine Normalskala, also eine Skala mit einem Werteraum zwischen 0 und 1, transformiert. Gemeinsamkeiten und Unterschiede werden mit der Lage der jeweiligen Mediane und Interquartilsabstände veranschaulicht.

Stichprobenziehung und Repräsentativität

Mit Blick auf das Ziel möglichst für die gesamte Schweizer IZA **repräsentative Ergebnisse** zu erhalten, erfolgte eine **randomisierte Stichprobenziehung** beim SECO (42 Berichte) sowie eine **Vollerhebung** bei der AFM (12 Berichte). Bei der DEZA wurden die Berichte von den dortigen Verantwortlichen ebenfalls zufällig ausgewählt. Während die Repräsentativität der Ergebnisse zur Bewertung der Qualität der von der AFM beauftragten Evaluationen aufgrund der Vollerhebung ohnehin ausser Frage steht, ist die Grösse der Stichprobe beim SECO (42 aus 72) zumindest als gleichermassen ausreichend zu erachten wie bei der DEZA (60 aus 92).

Zusammenhangsanalysen

Um herauszufinden, ob zwischen Qualitätskriterien, Ergebnisdarstellung, Evaluationskosten und Erfolgsquoten jeweils ein Zusammenhang besteht, wurden im Analyseraster die Budgets der Evaluationen ergänzt sowie, für das SECO und die DEZA, ebenfalls die Ratings der DAC-Kriterien. Somit konnten entsprechende Korrelationsanalysen durchgeführt werden.

Limitationen

Bei der folgenden Ergebnisdarstellung ist zu berücksichtigen, dass die Bewertungen lediglich auf Basis der Leistungsbeschreibungen und Evaluationsberichte einschliesslich ihrer Anhänge beruhen. Primärdaten, bspw. mittels Interviews mit Evaluationsbeteiligten (Projektmanager:innen, Evaluator:innen etc.), wurden keine erhoben. Ebenso wurden keine Rohdaten (Interviewtranskripte, Survey-Datensätze etc.) ausgewertet, die weitere Hinweise insbesondere auf die fachliche Qualität der Evaluationen hätten geben können. Dies stellt jedoch nicht das Ergebnis einer Meta-Evaluation als Ganzes in Frage. Denn weder geht üblicherweise eine gut gemachte Evaluation mit einem miserablen Evaluationsbericht einher, noch lässt sich eine schlechte durch einen hervorragend geschriebenen Evaluationsbericht kaschieren.

Weiterhin ist zu ergänzen, dass die Metaevaluation einem normativen Ansatz folgt. D.h., der Bewertungsrahmen wurde nicht auf Grundlage des Datenmaterials entwickelt, sondern er basiert auf davon unabhängigen, etablierten Standards (insb. den Standards für Evaluation der DeGEval[10]). Da die untersuchten Evaluationen nicht zwangsläufig an diesen Standards ausgerichtet waren, ist es, wie bei derartigen Untersuchungen üblich, äusserst unwahrscheinlich, dass eine Evaluation in allen Kriterien die bestmögliche Bewertung erhält.

H. Keupp, L. v. Rosenstiel, & S. Wolff (Hrsg.), Handbuch qualitative Forschung : Grundlagen, Konzepte, Methoden und Anwendungen (S. 209-213). München: Beltz - Psychologie Verl. Union. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-37278 [24.11.2022]

8 Ein Korrelationskoeffizient gibt an, wie stark eine Beziehung zwischen zwei Variablen ist, also wie sehr sich der Wert der einen Variable ändert, wenn sich der Wert der anderen ändert.

9 Der Kruskal-Wallis-Test ist ein nicht-parametrischer statistischer Test zum Gruppenvergleich von ordinalskalierten Variablen. Nicht-parametrisch bedeutet in diesem Zusammenhang, dass der Test keine Annahmen über die Werteverteilung der untersuchten Variablen erfordert. Im Vergleich zu parametrischen Tests sind sie zwar weniger sensitiv, jedoch robuster.

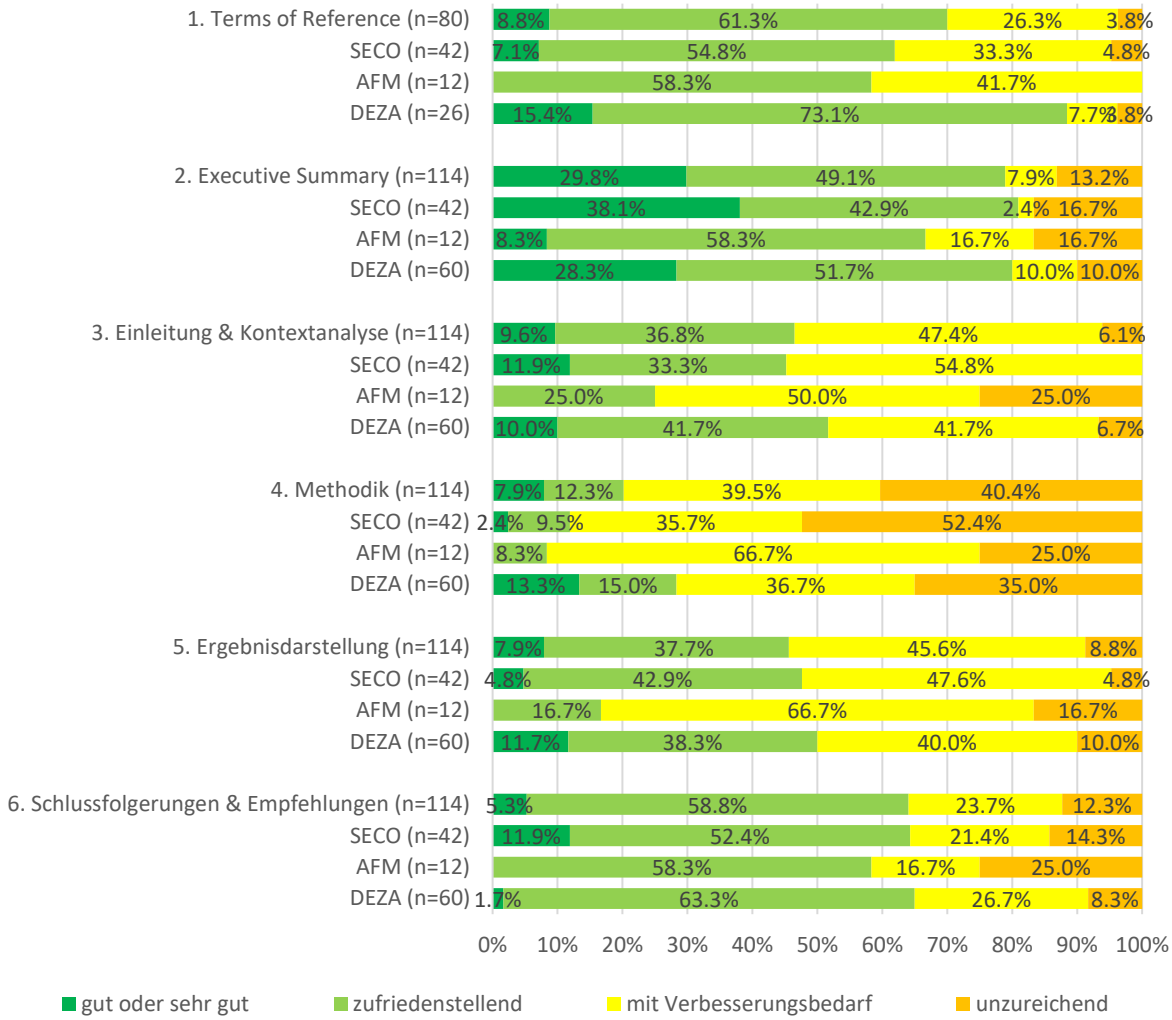
10 Vgl. https://www.degeval.org/degeval-standards/standards-fuer-evaluation/ [24.11.2022]

3. Ergebnisse

Die Ergebnisdarstellung richtet sich nach der Gliederung des im Methodenkapitel erläuterten Analyserasters, von der Bewertung der Terms of Reference bis zur Bewertung der Schlussfolgerungen und Empfehlungen (Abschnitte 3.1 bis 3.6). In den einzelnen Abschnitten erfolgt die Diskussion jeweils in fünf Schritten: Zunächst wird die Qualität der Teilkriterien vergleichend bewertet (1.). Danach werden die wesentlichsten Befunde auf der Ebene der Einzelindikatoren erörtert, wobei wiederum auf augenfällige Unterschiede zwischen den Verwaltungseinheiten eingegangen wird (2.). Im Anschluss werden diese Befunde mit den Ergebnissen der qualitativen Datenanalyse angereichert (3.). Abschliessend werden die Ergebnisse der vergleichenden statistischen Analyse (4.) sowie der Korrelationsanalysen zur Einschätzung möglicher Zusammenhänge zwischen den Bewertungskriterien (5.) vorgestellt.

Zum Einstieg gibt die folgende Abbildung einen zusammenfassenden Überblick über die Bewertungen der einzelnen Kriterien insgesamt sowie für die SECO, das AFM und die DEZA[11] im Einzelnen.

Abbildung 2: Bewertung der Hauptkriterien



Die Übersicht deutet bereits darauf hin, dass die Evaluationsberichte der AFM vergleichsweise mehr Defizite aufweisen als die des SECO und der DEZA. Weiterhin zeigt sie hinsichtlich der Qualität der Leistungsbeschreibungen (Terms of Reference, ToR) einen erkennbaren Unterschied zwischen den drei Verwaltungseinheiten zugunsten der DEZA. Hierbei ist jedoch einschränkend hinzuzufügen, dass in die Bewertung dieses Kriteriums nur 26 der insgesamt 60 Evaluationen der DEZA eingeflossen sind, da für

[11] Wie in der Einleitung des Methodenkapitels erläutert, stammen die Rohdaten hierfür aus der für die DEZA durchgeführten Metaevaluation.

die übrigen keine Leistungsbeschreibungen zur Verfügung standen. Schliesslich fällt auf, dass, ungeachtet des insgesamt vergleichsweise schlechteren Gesamtbilds, der Anteil der Berichte mit einer unzureichenden methodischen Qualität bei der AFM am geringsten ist. Bei der Interpretation dieses Befunds ist jedoch abermals die geringe Zahl der Berichte letzterer Verwaltungseinheit zu berücksichtigen.

3.1 Leitungsbeschreibung/Terms of Reference

Die Qualität der Terms of Reference der Evaluationen wird anhand von sieben Teilkriterien bewertet: der Angemessenheit der **Kontextbeschreibung**, der Beschreibung des **Evaluationsgegenstands**, der Darstellung von **Hintergrund und Zweck** der Evaluation, ihres **Gegenstandsbereichs**, der **Evaluationskriterien und -fragen**, der erwarteten **Methodik** sowie schliesslich der Darstellung der **Verantwortlichkeiten**, **Leistungen** und des **Zeitplans**. Die Teilkriterien orientieren sich an etablierten Leitlinien für Leistungsbeschreibungen für Evaluationen, wie bspw. von der United Nations Evaluation Group[12] oder der Weltbank[13] publiziert.

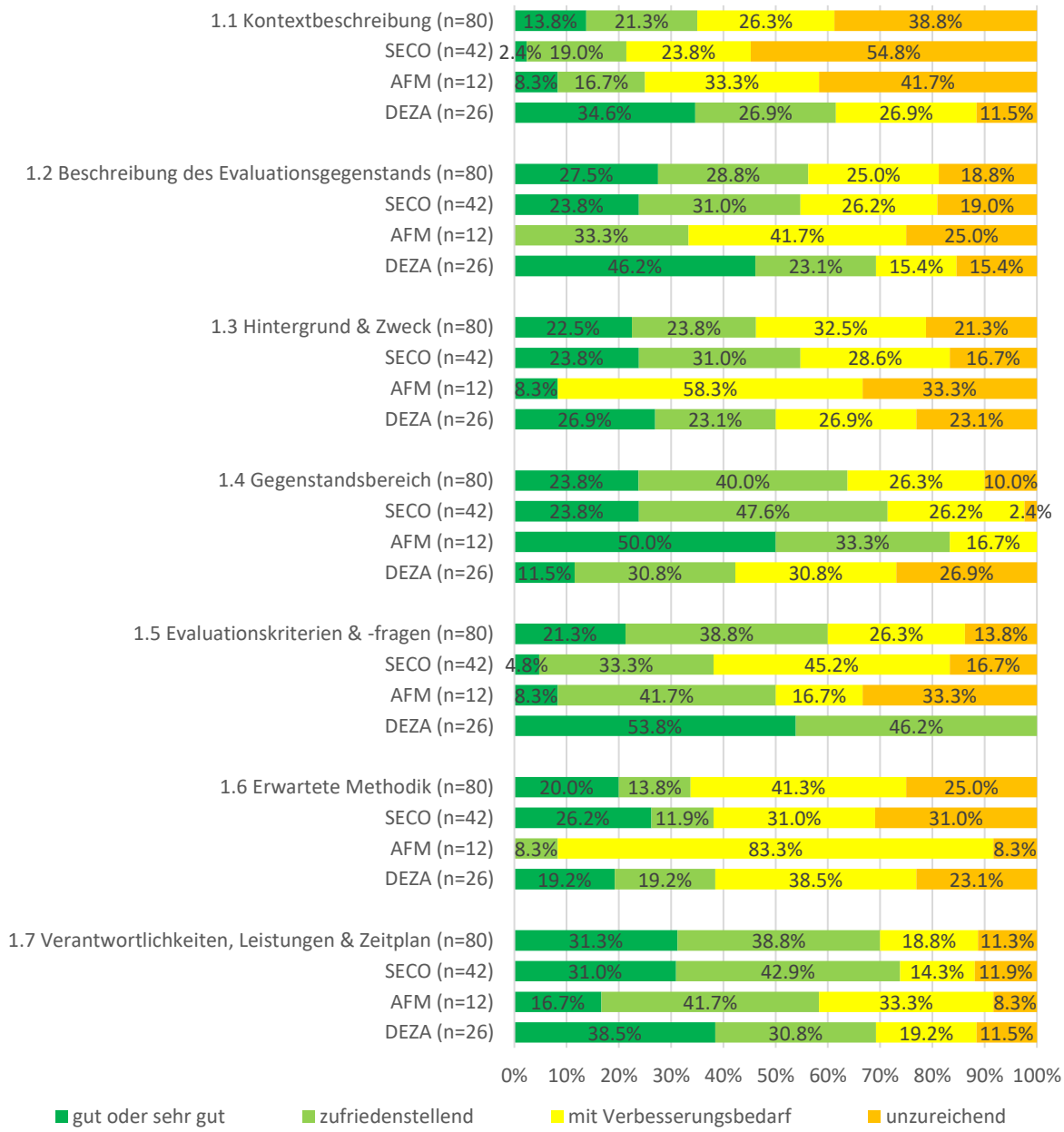
Weiterhin erfolgt eine Einschätzung der **Durchführbarkeit** der Evaluation in Anbetracht des zur Verfügung stehenden Mengengerüsts, der Komplexität des Untersuchungsgegenstands sowie der Zahl der zu beantwortenden Evaluationsfragen. Hierbei ist zu ergänzen, dass die Bewertung der Durchführbarkeit nicht auf einer rein quantitativen Auswertung dieser Indikatoren beruht, sondern im Einzelfall ebenfalls qualitativ auf Grundlage der praktischen Erfahrungen am CEval erfolgt. Daher wird auch auf eine grafische Darstellung dieses Teilkriteriums verzichtet.

Wie die folgende Abbildung veranschaulicht, unterscheiden sich die Leistungsbeschreibungen von SECO, AFM und DEZA in ihrer Qualität z.T. erheblich. Während der Anteil ausreichend dargestellter Verantwortlichkeiten, Leistungen und Zeitplan bei den drei Verwaltungseinheiten zwischen ca. 60% und 70% noch relativ nahe beieinander liegt, zeigen sich deutliche Unterschiede bei den Beschreibungen des Kontexts, in dem die Evaluation durchgeführt wird, des Evaluationsgegenstands selbst, des Hintergrunds und des Zwecks der Evaluation, ihres Gegenstandsbereichs, der dabei zu behandelnden Kriterien bzw. Fragen sowie bei den methodischen Vorgaben. Dabei erfüllen die Leistungsbeschreibungen der DEZA vier der sieben Teilkriterien am ehesten in zumindest zufriedenstellendem Ausmass (Kontext, Gegenstand, Fragen/Kriterien, Methodik), während Hintergrund und Zweck lediglich in den Leistungsbeschreibungen des SECO in mehr als der Hälfte der Fälle hinreichend beschrieben werden. Schliesslich gelingt es der AFM offenbar am besten, den Gegenstandsbereich der avisierten Evaluation adäquat einzugrenzen.

[12] Vgl. https://evaluation.iom.int/sites/g/files/tmzbdl151/files/documents/UNEG_TOR_0.pdf [19.10.2022]

[13] Vgl. https://documents1.worldbank.org/curated/en/209341599772583527/pdf/Writing-Terms-of-Reference-for-an-Evaluation-A-How-to-Guide.pdf [23.03.2023]

Abbildung 3: Bewertung der Qualität der Leistungsbeschreibungen



Schaut man sich die Indikatoren der Teilkriterien im Einzelnen an, können weitere Unterschiede festgestellt werden. Während bei der DEZA der Kontext, in dem das zu evaluierende Vorhaben operiert, in aller Regel ausreichend beschrieben wird (84,6%) und dabei auch ein Bezug zu internationalen, nationalen oder regionalen Entwicklungsstrategien hergestellt wird (76,9%), ist dies bei dem SECO lediglich in etwa vier von zehn Berichten (40,5% bzw. 35,7%), bei der AFM sogar nur in einem Sechstel bzw. einem Viertel der Berichte (16,7% bzw. 25,0%) der Fall.

Die Beschreibung des Evaluationsgegenstands erfolgt bei SECO, AFM und DEZA ebenfalls in recht unterschiedlicher Weise. Werden bei allen drei Verwaltungseinheiten in rund 80 Prozent der Leistungsbeschreibungen die Ziele des zu evaluierenden Vorhabens vorgestellt und in etwa der Hälfte der Fälle dessen Interventionsgebiet, liegt die Bandbreite der Leistungsbeschreibungen, die die Interventionslogik des Vorhabens erörtern oder dessen Zielgruppen und weitere Beteiligte nennen zwischen null (AFM) und ca. 40 bzw. 60 Prozent (DEZA). Auffällig ist auch der unterschiedliche Anteil der Leistungsbeschreibungen, die Angaben zu den Budgets der Vorhaben machen (SECO: 38,1%, AFM: 8,3%, DEZA: 61,5%).

Ähnlich heterogen stellen sich die Ergebnisse der Auswertung der einzelnen Indikatoren der übrigen Teilkriterien dar. Bei der Erörterung von Hintergrund und Zweck der Evaluation zeigen sich die grössten Unterschiede bei der Darstellung der Adressat:innen der Evaluationsergebnisse (SECO: 57,1%, AFM: 16,7%, DEZA: 30,8%) sowie bei Angaben dazu, wofür diese Ergebnisse genutzt werden sollen (SECO: 47,6%, AFM: 16,7%, DEZA: 61,5%). Die räumliche und zeitliche Abgrenzung des Gegenstandsbereichs der Evaluation sowie die Darstellung, welche Massnahmen eines Vorhabens ggf. in die Untersuchung mit einbezogen werden sollen, erfolgt ebenfalls in unterschiedlichem Ausmass, wobei die AFM hierbei offenbar den grössten Wert auf eine klare Differenzierung legt. Auffällig ist weiterhin, dass alle Leistungsbeschreibungen der DEZA Angaben zu den Kriterien machen, nach denen ein Vorhaben zu evaluieren ist (SECO: 92,9%, AFM: 75,0%) und dass dies in allen Fällen die OECD/DAC-Kriterien oder eine Auswahl davon sind. Ebenso werden in fast allen DEZA Leistungsbeschreibungen konkrete Evaluationsfragen gestellt, im Gegensatz zu den Leistungsbeschreibungen des SECO und der AFM, wo diese nur etwa in drei Viertel der Fälle spezifiziert werden. Auch Querschnittsthemen spielen bei der DEZA offenbar eine grössere Rolle (57,7%) als bei den beiden anderen Verwaltungseinheiten (SECO: 28,6%, AFM: 25,0%).

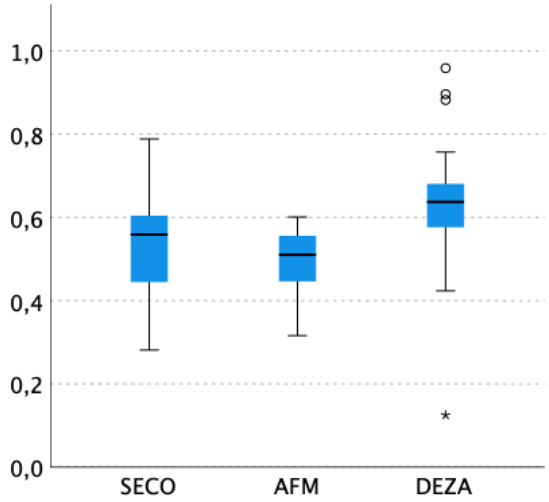
Bedenklich ist die Zahl der Evaluationsfragen beim SECO. Während diese bis auf einen Fall bei der AFM im Rahmen des üblicherweise im Rahmen einer Evaluation Machbaren liegt (je nach Budget bis ca. 15), enthalten die Leistungsbeschreibungen des SECO bis zu 58 Fragen (WEHU 175). In immerhin knapp einem Viertel der Fälle wird die Beantwortung von mehr als 20 Fragen erwartet, teilweise in einem zeitlichen Rahmen von gerade mal 20 Arbeitstagen (WEHU 183, WEIF 106, WEMU 88). In Anbetracht dieses Ergebnisses ist es umso erstaunlicher, dass dennoch die SECO Stichprobe den höchsten Anteil (69,2%) an nach Expertensicht fachlich angemessen umsetzbaren Evaluationen enthält (AFM: 50,0%, DEZA: 63,1%). Dies ist in erster Linie dem im Vergleich zur AFM und der DEZA grösseren Budget von SECO Evaluationen geschuldet. Diesem Befund ist jedoch einschränkend hinzuzufügen, dass lediglich von 26 SECO, sechs AFM und 21 DEZA Evaluationen, also weniger als der Hälfte aller Fälle ausreichende Informationen für eine valide Einschätzung ihrer Machbarkeit vorlagen.

Während in der Summe die drei Verwaltungseinheiten beim Teilkriterium Darstellung der Verantwortlichkeiten, Leistungen und des Zeitplans relativ nahe beieinander liegen, liefert ein genauerer Blick auf die einzelnen Indikatoren ein etwas differenziertes Bild. So zeigen sich hinsichtlich der Beschreibung der Rollen der an der Evaluation Beteiligten deutliche Unterschiede. Während diese bei der DEZA in etwa zwei Drittel aller Leistungsbeschreibungen spezifiziert sind, ist dies beim SECO lediglich bei einem Drittel und bei der AFM sogar nur in zwei von zwölf der Leistungsbeschreibungen der Fall. Dafür scheinen SECO und AFM tendenziell konkretere Vorstellungen von Zeitplan und Abgabefristen (SECO: 78,6% bzw. 85,7%, AFM: 58,3% bzw. 91,7%, DEZA: 57,7% bzw. 73,1%) zu haben.

Die Ergebnisse der qualitativen Analyse bestätigen die obigen Befunde insofern als vollständigere Leistungsbeschreibungen i.d.R. auch inhaltlich besser zu bewerten sind. In der aktuellen Stichprobe konnten hier vier Best Practices identifiziert werden: WEHU 168, WEIN 56, WEMU 89, DEZA 1.10. Zur erwarteten Methodik liefert weiterhin WEMU 92 klare Vorgaben. Es gibt jedoch auch einige Fälle, die diesbezüglich z.T. erhebliche inhaltliche Schwächen aufweisen bzw. in denen essenzielle Angaben dazu völlig fehlen (WEMU 79, 83, 87, AFM 1-4, 7-9, DEZA 1.23). Zu bemängeln ist weiterhin, dass einige wenige Leistungsbeschreibungen keine konkreten Fragestellungen nennen (z.B. WEHU 174).

Entsprechend der Befunde aus der deskriptiven quantitativen Analyse, weist die vergleichende statistische Auswertung auf einen Unterschied in der Qualität der Leistungsbeschreibungen der drei Verwaltungseinheiten hin. Wie die folgende Abbildung zeigt, unterscheidet sich die Verteilung ihrer jeweiligen Mediane (SECO: 0,56, AFM: 0,51, DEZA: 0,64) und Interquartilsabstände (SECO: 0,17, AFM: 0,14, DEZA: 0,12) deutlich voneinander. Weiterhin sind bei der DEZA einige sowohl positive als auch negative Ausreisser zu erkennen.

Abbildung 4: Qualität der Terms of Reference nach Verwaltungseinheit



Wie die obige Abbildung bereits erahnen lässt, bestätigt der paarweise Vergleich zwischen den drei Verwaltungseinheiten mittels Kruskal-Wallis-Test, dass die Qualität der Leistungsbeschreibungen der DEZA statistisch signifikant besser ist als die der AFM ($\alpha = 0{,}003$) und ebenso des SECO ($\alpha = 0{,}029$). Zwischen letzteren beiden ist kein signifikanter Unterschied zu erkennen. Bei der Einordnung dieses Ergebnisses muss allerdings berücksichtigt werden, dass lediglich für 26 der insgesamt 60 Evaluationen der DEZA Leistungsbeschreibungen zur Verfügung standen und es daher nur für einen Teil dieser Stichprobe (43,3%) Gültigkeit besitzt.

Wie bereits bei der Metaevaluation im Auftrag der DEZA bestätigt die Korrelationsanalyse einen statistisch signifikanten Zusammenhang zwischen der Qualität der Leistungsbeschreibungen und der Qualität der Evaluationsberichte. Für die Gesamtstichprobe zeigt sich dabei eine mässige Korrelation mit der Qualität von Einleitung und Kontextanalyse ($r = 0{,}34$), des Methodenkapitels ($r = 0{,}45$) und der Ergebnisdarstellung ($r = 0{,}40$, mit jeweils $p < 0{,}01$). Wenngleich diese Werte im Vergleich zur DEZA Teilstichprobe etwas kleiner sind, sind sie gleichermassen signifikant und auf einer grösseren Datenbasis beruhend. Zudem stehen sie in Einklang mit Befunden aus anderen Metaevaluationen, die in der Vergangenheit am CEval durchgeführt wurden.[14] Sie werden daher als äusserst valide eingestuft.

3.2 Executive Summary

Die Qualität der Executive Summaries (ES) wird anhand von zwei Teilkriterien beurteilt, nämlich der **Vollständigkeit** ihres jeweiligen Inhalts und ihrer **Verständlichkeit und Konsistenz** mit dem Hauptbericht.

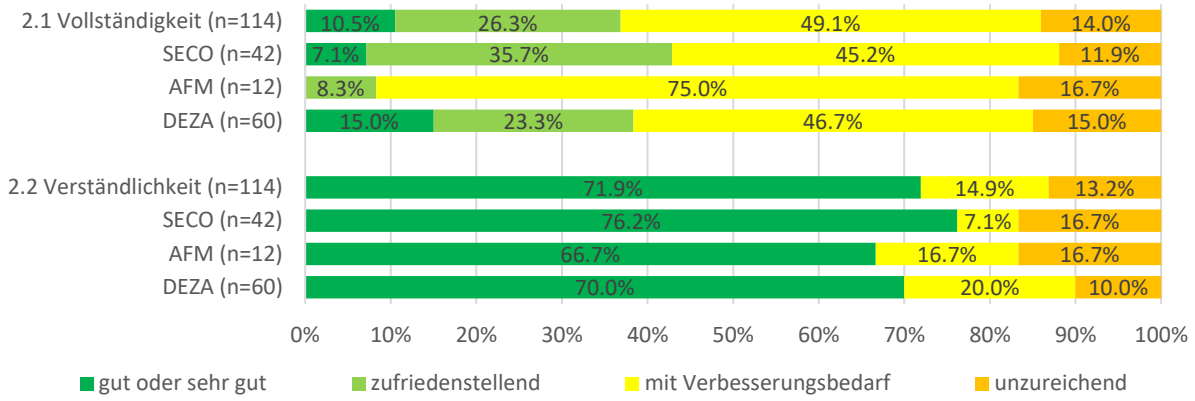
Erläuterung: Ein ES dient dazu, Entscheidungsträger:innen in Kürze die wichtigsten Informationen über eine Evaluation zu liefern. Hierfür muss es entsprechend eigenständig verständlich sein und neben den zentralen Ergebnissen, Schlussfolgerungen und Empfehlungen ebenso den Gegenstandsbereich und die Methodik der Evaluation skizzieren. Weiterhin sollte ein ES ausschliesslich auch im Hauptteil des Evaluationsberichts behandelte Befunde enthalten. Die beiden Teilkriterien wurden mit insgesamt zwölf an diesen Anforderungen ausgerichteten Indikatoren operationalisiert.

Die folgende Abbildung zeigt einen deutlichen Unterschied bei der Bewertung der beiden Teilkriterien. Während bei allen drei Verwaltungseinheiten mindestens zwei Drittel der Berichte eigenständig verständlich sind und inhaltlich mit dem Hauptbericht übereinstimmen, unterscheiden sich ihr jeweiliger Aufbau mitunter wesentlich. Auffällig ist hierbei insbesondere, dass nur knapp jedes zehnte Executive

[14] Vgl. Meta-evaluation of Project and Programme Evaluations in 2015–2017. Online Publikation: https://um.fi/documents/384998/385866/meta_evaluation_report_2018, Meta-Evaluation of ADA Project and Programme Evaluations – Executive Summary. Vienna: Austrian Development Agency. Online Publikation: https://www.entwicklung.at/fileadmin/user_upload/Dokumente/Evaluierung/Evaluierungsberichte/2019/Management_Response/Executive_Summary_ADA-Meta_Eval.pdf

Summary der AFM in zufriedenstellender Weise, die darin üblicherweise zur Verfügung gestellten Informationen enthält.

Abbildung 5: Bewertung der Qualität der Executive Summaries



Bei genauerer Betrachtung der Indikatoren des ersten Teilkriteriums zeigen sich bei dem SECO ähnliche Unterschiede, wie sie bereits bei der DEZA festgestellt werden konnten. Während vier von fünf ES eine Zusammenfassung der Empfehlungen und der zentralen Ergebnisse der Evaluation enthalten, werden ihre jeweiligen Ziele und Schlussfolgerungen nur in etwa der Hälfte und ihre Hintergründe und Zwecke in etwa einem Drittel der Fälle dargestellt. Die ES der AFM sind im Vergleich dazu insbesondere bei der Zusammenfassung der Evaluationsergebnisse und den daraus abgeleiteten Empfehlungen noch weniger aufschlussreich.

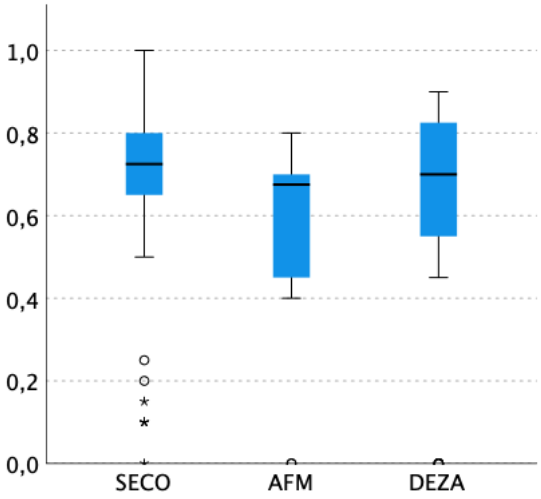
Auch bei den übrigen Indikatoren weisen die ES der AFM vergleichsweise mehr Schwächen auf. So enthalten gut drei Viertel der ES des SECO und immerhin zwei Drittel der ES der DEZA eine ausreichend detaillierte Beschreibung des evaluierten Vorhabens. Bei der AFM ist dies bei lediglich einem Drittel der Berichte der Fall. Die deutlichsten Unterschiede zeigen sich jedoch bei den Zusammenfassungen der methodischen Elemente in den ES. Hier schneiden die Berichte der DEZA mit einer jeweils angemessenen Darstellung des Gegenstandsbereichs und der Methodik in knapp der Hälfte der Fälle am besten ab, wohingegen lediglich ein gutes Drittel der SECO ES diese Informationen in zufriedenstellender Weise enthält. Bei der AFM werden in diesem Kapitel Gegenstandsbereich (16,7%) und Methodik (25,0%) noch seltener ausgeführt. Zum Evaluationsdesign werden in keinem ES der 12 AFM Berichte Angaben gemacht (DEZA: 8,3%, SECO: 14,3%).

Die qualitative Datenanalyse ergänzt die obigen Befunde um eine Reihe weiterer Hinweise, die Aufschluss über wesentliche Stärken und Schwächen der ES geben. Zunächst fällt dabei ins Auge, dass es offensichtlich bei keiner der drei Verwaltungseinheiten eindeutige Vorgaben zum Aufbau von ES gibt oder diese zumindest i.d.R. nicht eingehalten werden. Wenngleich die allermeisten Evaluationsberichte ein solches Kapitel enthalten (SECO: 41 von 42, AFM: 10 von 12, DEZA: 54 von 60), weichen diese nämlich deutlich hinsichtlich Struktur, Inhalt und Länge voneinander ab. Während die ES einiger weniger Berichte die wichtigsten Informationen enthalten (z.B. WEIF 118, WEHU 188, DEZA 1.01), weisen die meisten anderen wie bereits dargestellt z.T. erhebliche Lücken auf. Weiterhin ist bisweilen festzustellen, dass ES lediglich Ergebniszusammenfassungen enthalten, ohne ausreichende Informationen zu Gegenstand und Methodik zu liefern (z.B. WEHU 184, 189, WEIN 62, WEMU 79), dass sich die Bewertungen im ES von denen im Hauptbericht unterscheiden bzw. sich diese nicht aus den dort beschriebenen Ergebnissen ohne Weiteres ableiten lassen (z.B. WEMU 81, 92, AFM 5, 12, DEZA 1.25, 2.12) oder dass Inhalte aus dem Ergebnisteil lediglich in das ES einkopiert wurden (z.B. WEIF 106). In zwei Fällen (WEHU 183, WEIF 106) erscheinen die ES im Verhältnis zum Hauptbericht auch als zu lang für eine wesentliche Zeitersparnis.

Wie die folgende Abbildung zeigt, liefert der statistische Vergleich der ES der drei Verwaltungseinheiten ein Bild, das mit den obigen Ausführungen in Einklang steht. So liegen die jeweiligen Mediane (SECO: 0,73, AFM: 0,68, DEZA: 0,70) in etwa auf gleichem Niveau, während der Interquartilsabstand bei SECO deutlich kleiner ist als bei AFM und DEZA (SECO: 0,16, AFM: 0,28, DEZA: 0,29). Damit zeigt

sich, dass, von wenigen Ausreissern abgesehen, die Qualität der ES in den Evaluationsberichten des SECO etwas weniger streut als bei den anderen beiden Verwaltungseinheiten.

Abbildung 6: Qualität der Executive Summaries nach Verwaltungseinheit



Wenngleich der Boxplot auf geringfügige Unterschiede zwischen den drei Verwaltungseinheiten hinsichtlich der Qualität der Executive Summaries ihrer Evaluationsberichte hindeutet, zeigt der paarweise Gruppenvergleich mittels Kruskal-Wallis-Test, dass diese statistisch nicht signifikant sind.

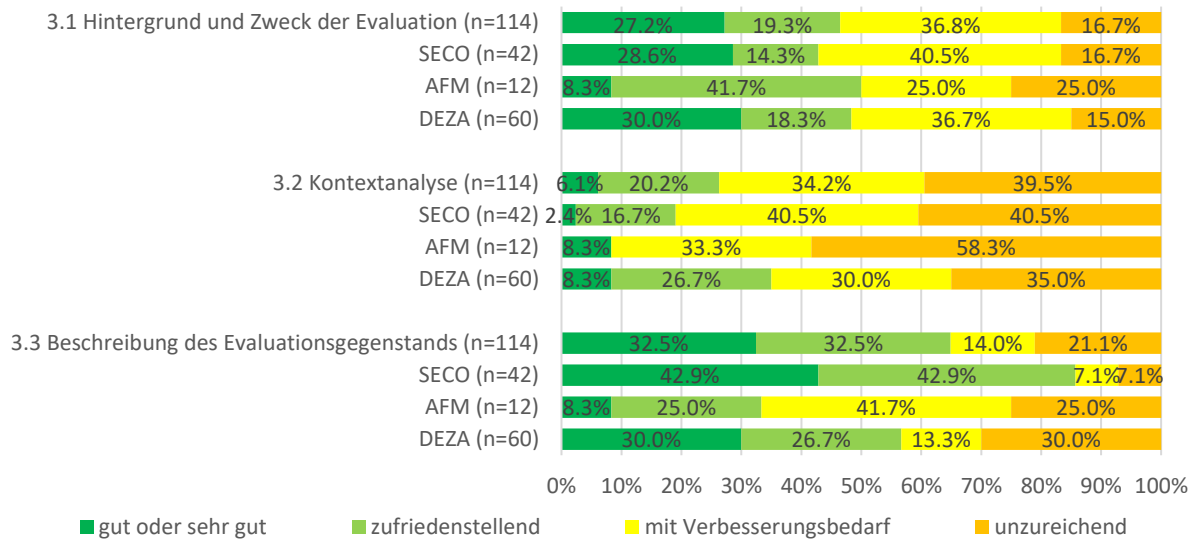
Schaut man sich als Letztes die Ergebnisse der Zusammenhangsanalysen an, so lässt sich feststellen, dass die Qualität der ES mit der der jeweiligen Methodenkapitel korreliert ($p < 0{,}044$). Da der Korrelationskoeffizient jedoch relativ klein ist ($r = 0{,}189$) und auch kein unmittelbarer inhaltlicher Zusammenhang zwischen den beiden Kapiteln hergestellt werden kann, erübrigt sich eine weitere Interpretation dieses Befunds.

3.3 Einleitung und Kontextanalyse

Die Qualität der Einleitungen und Kontextanalysen der Evaluationen wird anhand von drei Teilkriterien bewertet, die sich an den üblichen Inhalten dieses Teils von Evaluationsberichten orientiert: die Angemessenheit der Beschreibung des **Hintergrunds und des Zwecks der Evaluation**, der **Kontextanalyse** und der Beschreibung des **Evaluationsgegenstands**. Die drei Teilkriterien wurden mit insgesamt 18 Indikatoren operationalisiert.

Die folgende Abbildung legt nahe, dass SECO, AFM und DEZA hierbei jeweils unterschiedliche Schwerpunkte setzen. Während die Anteile von in zumindest zufriedenstellendem Ausmass beschriebenen Hintergründen und Zwecken der jeweiligen Evaluationen mit zwischen 42 und 50 Prozent relativ wenig voneinander abweichen, unterscheidet sich die Qualität der Kontextanalyen und der Beschreibungen der Evaluationsgegenstände zwischen den drei Verwaltungseinheiten auf jeweils verschiedenen Niveaus z.T. erheblich. Während bei der DEZA in immerhin einem guten Drittel der Berichte die Rahmenbedingungen, unter denen das evaluierte Vorhaben operiert (hat) in ausreichendem Mass beleuchtet werden, ist dies beim SECO nur in jedem fünften Bericht der Fall, bei der AFM sogar in nur einem einzigen (AFM 1). Eine angemessene Beschreibung des evaluierten Vorhabens selbst findet sich hingegen am ehesten in Evaluationsberichten des SECO. Bei der DEZA gelingt dies nur in gut der Hälfte und bei der AFM in einem Drittel der Berichte.

Abbildung 7: Bewertung der Qualität der Einleitungen und Kontextanalysen



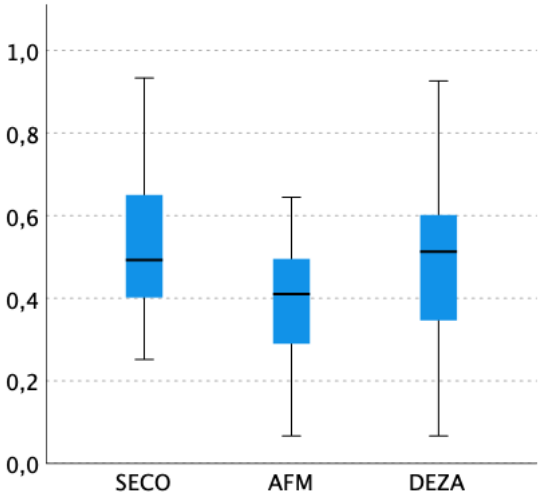
Wiederum liefert die Untersuchung der einzelnen Indikatoren der drei Teilkriterien weitere Hinweise zu den jeweiligen Stärken und Schwächen der Evaluationen von SECO, AFM und DEZA. Während die allermeisten Berichte Aufschluss über die Ziele der Evaluation geben, liefert nur etwa die Hälfte eine konkrete Begründung, wieso diese durchgeführt wurde und nur in jedem vierten Fall finden sich Informationen über die Adressat:innen der Evaluationsergebnisse. Die Unterschiede zwischen den drei Verwaltungseinheiten sind dabei verhältnismässig gering. Auffällig ist weiterhin, dass nur in einem Evaluationsbericht der AFM konkrete Angaben zur geplanten Nutzung der Ergebnisse gemacht werden. Bei dem SECO und der DEZA finden sich Informationen hierzu in immerhin knapp der Hälfte aller Berichte.

Bei den im Schnitt deutlich schlechter bewerteten Kontextanalysen fällt auf, dass diese bei der AFM vergleichsweise häufiger auf die Strategien der Schweizer Entwicklungszusammenarbeit rekurrieren als bei den anderen beiden Verwaltungseinheiten (41,7% vs. 11,9% beim SECO und 23,3% bei der DEZA). Im Gegenzug spielen internationale, nationale oder regionale Entwicklungsstrategien eher bei dem SECO und der DEZA eine Rolle (SECO: 31,0%, AFM: 8,3%, DEZA: 43,3%). Bemerkenswert ist ebenfalls, dass in weniger als zwei Drittel aller Evaluationen generell ein Bezug zwischen Entwicklungsstrategien und dem evaluierten Vorhaben selbst hergestellt wird (SECO: 59,5%, AFM: 58,3%, DEZA: 58,3%).

Wie die obige Abbildung bereits erwarten lässt, enthalten die Berichte des SECO im Schnitt die detailliertesten Beschreibungen ihres jeweiligen Evaluationsgegenstands. Erstaunlich bei diesem Befund ist, dass er sich über alle neun Einzelindikatoren dieses Teilkriteriums als konsistent darstellt. So werden Zeitraum, Budget, Umsetzungsregion, Massnahmen, Zielgruppen, Ziele, Umsetzungsstrategie und -stand des evaluierten Vorhabens sowie seine Interventionslogik (bspw. mittels Theory of Change, Wirkungsmodellen oder Logframes) in SECO Berichten am häufigsten dargestellt. Nennenswert ist weiterhin, dass in den Berichten der AFM sehr häufig Angaben zu den Zielgruppen und zur Interventionslogik des Vorhabens fehlen (in jeweils 11 von 12 Fällen).

Entsprechend der obig dargestellten heterogenen Verteilung der Bewertungen der Teilkriterien liefert auch die vergleichende statistische Analyse ein gemischtes Bild. Wie der folgenden Abbildung zu entnehmen ist, weisen alle drei Boxplots einen vergleichsweise grossen Interquartilsabstand (SECO: 0,25, AFM: 0,23, DEZA: 0,26) auf, bei nur mässigen Unterschieden in der Verteilung der Mediane (SECO: 0,49, AFM: 0,41, DEZA: 0,51).

Abbildung 8: Qualität der Einleitungen und Kontextanalysen nach Verwaltungseinheit



Die relativ grosse Streuung der Bewertungsergebnisse bei recht ähnlichen Mittelwerten bedingt auch, dass keine statistisch signifikanten Unterschiede in der Qualität der Einleitungen und Kontextanalysen zwischen den drei Verwaltungseinheiten identifiziert werden können.

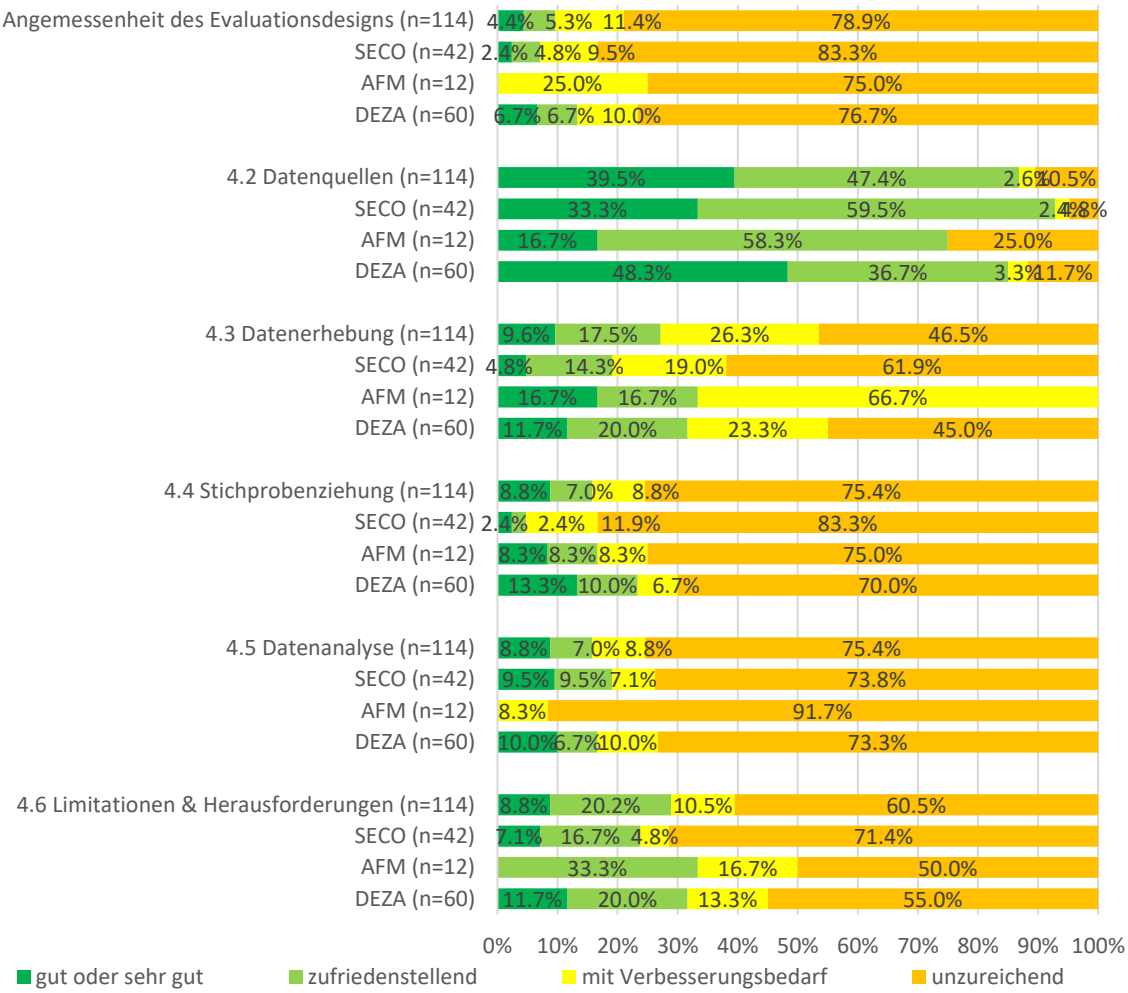
Wie in Abschnitt 3.1 bereits dargestellt, korreliert die Qualität der Einleitungen und Kontextanalysen der Evaluationsberichte mit der Qualität der Leistungsbeschreibungen . Die Korrelationsanalyse liefert weiterhin Hinweise über einen wenngleich für die Gesamtstichprobe schwachen Zusammenhang zwischen der Qualität der Einleitungen und Kontextanalysen und der Qualität der übrigen Berichtskapitel ($r = 0{,}208$ für die Methodenkapitel, $r = 0{,}222$ für die Ergebnisdarstellung sowie $r = 0{,}219$ für die Qualität der Schlussfolgerungen und Empfehlungen, jeweils mit $p < 0{,}05$). Betrachtet man die Teilpopulationen genauer, zeigt sich, dass der Zusammenhang mit dem letzteren Kapitel bei der SECO ($r = 0{,}423$, $p < 0{,}01$) deutlich stärker ausgeprägt ist als bei den anderen beiden Verwaltungseinheiten. Inwiefern diese Korrelation auf einen stärkeren inhaltlichen Zusammenhang zwischen diesen beiden Kapiteln hindeutet, kann an dieser Stelle jedoch nicht abschliessend beantwortet werden.

3.4 Methodik

Die Qualität der im Methodenkapitel der Berichte beschriebenen Methodik der Evaluationen, wird anhand von sechs Teilkriterien bewertet: der Angemessenheit des **Evaluationsdesigns**, der **Datenquellen**, der **Datenerhebung**, der **Stichprobenziehung**, der **Datenanalyse** sowie der Diskussion der **Limitationen und Herausforderungen** bei der praktischen Umsetzung der Evaluation. Die Teilkriterien sind wiederum mittels insgesamt 29 Indikatoren operationalisiert, deren Überprüfung im Wesentlichen auf die Gewährleistung objektiver, reliabler und valider Ergebnisse ausgerichtet ist.

Anhand der folgenden Abbildung wird auf Anhieb ersichtlich, dass die Evaluationsberichte des SECO, der AFM und der DEZA in ihrer methodischen Qualität sehr grosse Ähnlichkeiten aufweisen. Während die meisten Berichte ausreichend detaillierte Informationen zu den Datenquellen enthalten, wird in kaum einem Drittel angemessen auf Limitationen und Herausforderungen eingegangen, mit nur geringfügigen Unterschieden zwischen den drei Verwaltungseinheiten. Auch bei der Darstellung der Datenerhebung weisen die Berichte von SECO, AFM und DEZA gleichermassen Schwächen auf. Hierzu finden sich nur in etwa jedem vierten Fall mindestens zufriedenstellende Angaben in den jeweiligen Methodenkapiteln. Tendenziell noch schlechter sieht es bei den Beschreibungen von Evaluationsdesign, Stichprobenziehung und Datenanalyse aus. Bei allen drei Teilkriterien werden die hierfür in den Berichten zur Verfügung gestellten Inhalte in mindestens drei Viertel der Fälle als unzureichend bewertet. Dabei sticht hervor, dass in kaum fünf Prozent der SECO Berichte konkrete Angaben zur Auswahl der Befragten gemacht werden und bei der AFM kein Bericht identifiziert werden konnte, in dem in angemessener Weise das Evaluationsdesign und die Datenanalyse beschrieben wurde.

Abbildung 9: Bewertung der Qualität der Methodik der Evaluationen



Der Blick auf die einzelnen Indikatoren fügt diesem in der Summe ernüchternden Gesamtbild wieder einige wesentliche Details zu den Schwächen der Verwaltungseinheiten im Einzelnen hinzu. So ist zur Bewertung des Evaluationsdesigns festzustellen, dass lediglich in rund einem Drittel der Berichte ein Analyseraster (Evaluationsmatrix, Datenerhebungsplan) zu finden ist, mittels dessen nachvollzogen werden könnte, auf welcher empirischen und methodischen Grundlage die Evaluationsergebnisse basieren. Weiterhin werden nur vereinzelt Angaben dazu gemacht, ob eine Attributions- oder Kontributionsanalyse durchgeführt wurde.

Bei den Angaben zu den Datenquellen zeigt sich, dass bei den allermeisten Evaluationen neben empirischen Daten von den Umsetzungsverantwortlichen und weiteren Stakeholdern auch Dokumente der Vorhaben (Anträge, Fortschrittsberichte u.ä.) in die Untersuchung einbezogen werden. Das Projektmonitoring wird hingegen nur in gut einem Drittel der Berichte als Datenquelle genutzt, mit deutlichen Unterschieden zwischen den drei Verwaltungseinheiten (SECO: 35,7%, AFM: 16,7%, DEZA: 45,0%). Erstaunlich ist auch, dass in weniger als der Hälfte der Evaluationsberichte der AFM auf die Zielgruppen als Datenquelle verwiesen wird. Auch in der Nutzung weiterer, unabhängiger Datenquellen, wie bspw. nationale Statistiken oder andere Studien, unterscheiden sich die Evaluationen von SECO (21,4%), AFM (41,7%) und DEZA (60,0%) voneinander.

Die Beschreibungen der Datenerhebung in den Berichten offenbaren, dass lediglich in der Hälfte der Fälle sowohl qualitative als auch quantitative Daten erhoben wurden. Wenngleich nicht pauschal beurteilt werden kann, ob ein entsprechender Mix erforderlich ist, erscheint dieser Wert recht gering. Bei der AFM werden sogar nur in einem Drittel der Evaluationen unterschiedliche Datentypen kombiniert. Deutliche Unterschiede lassen sich auch bei der Dokumentation der Datenerhebungsinstrumente feststellen. Während elf der zwölf AFM Berichte entsprechende Angaben zu den Instrumenten

enthalten, werden diese bei der DEZA in lediglich ca. drei Viertel der Fälle genannt und beim SECO sogar in weniger als der Hälfte. Eine Diskussion der Datenqualität findet nur in den seltensten Fällen statt, noch am ehesten in Berichten der AFM.

Gleiches gilt für Informationen zur Auswahl der Datenquellen, die in immerhin zwei Drittel aller AFM Berichte enthalten sind (SECO: 28,6%, DEZA: 41,7%), jedoch in nur der Hälfte der Fälle mit Angaben zur Art der Stichprobenziehung (SECO: 9,5%, DEZA: 26,7%) und noch seltener mit einer entsprechenden Begründung der jeweiligen Vorgehensweise (SECO: 9,5%, AFM: 16,7%, DEZA: 20,0%).

Die wenigsten Evaluationsberichte enthalten eine hinreichende Beschreibung, wie die Daten ausgewertet wurden (SECO: 14,3%, AFM: 0,0%, DEZA: 21,7%). Oftmals kann dies nur aus dem Kontext der Ergebnisdarstellung geschlossen werden, wobei auch hier auf einen eher zurückhaltenden Einsatz eines adäquaten Methodenmixes geschlossen werden kann (SECO: 14,3%, AFM: 0,0%, DEZA: 26,7%).

Nur wenige nennenswerte Unterschiede bestehen zwischen den drei Verwaltungseinheiten was die Beschreibung der Limitationen und Herausforderungen betrifft, denen sich die Gutachter:innen bei der Durchführung der Evaluation stellen mussten. So werden in jeweils nur etwa einem guten Drittel der Berichte praktische Schwierigkeiten im Umsetzungsprozess und bei der Datenerhebung im Besonderen erläutert. Einschränkungen bei der Datenanalyse werden tendenziell noch seltener thematisiert (SECO: 19,0%, AFM: 33,3%, DEZA: 28,3%).

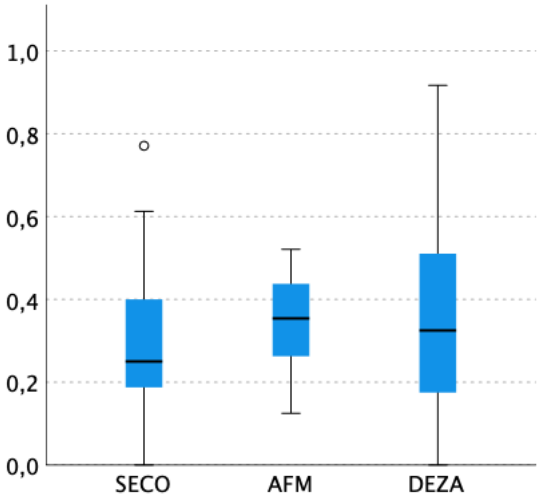
In Anbetracht der nur sehr lückenhaften Dokumentation der im Rahmen der Evaluationen angewendeten Methoden konnten den Berichten auch nur wenige qualitative Daten zur Einordnung der quantitativen Befunde entnommen werden. Dennoch lassen sich ein paar wesentliche Erkenntnisse zusammenfassen. So finden sich trotz aller Schwächen auch im Hinblick auf die Methodenbeschreibungen mehrere Good Practices, wie bspw. in der Darstellung der Datenquellen oder der Diskussion ihrer Validität und Reliabilität (WEHU 180a). Auch zwei generelle Good Practice über alle Teilkriterien hinweg, ausser der Darstellung des Evaluationsdesigns, konnten identifiziert werden (WEIN 63 & DEZA 1.25). Darüber hinaus sind vereinzelt weitere Angaben zur Methodik im Ergebnisteil (z.B. WEIF 112) zu finden. Weiterhin zeigt sich, dass, wenn Limitationen aufgezeigt werden, dies in sehr nachvollziehbarer Weise geschieht und dabei Aufschluss darüber gegeben wird, wieso welche Methoden zur Anwendung kamen bzw. auch, was nicht möglich war.

Wie bereits im Rahmen der Metaevaluation für die DEZA festgestellt wurde, herrscht bei einigen Evaluator:innen offenbar Unklarheit darüber, welche Elemente eine angemessene Methodenbeschreibung umfassen sollte. Bisweilen werden im entsprechenden Kapitel primär Angaben zum Umsetzungsprozess der Evaluation gemacht (z.B. WEHU 168, 174, WEIN 57, WEMU 86, DEZA 1.17). Wenngleich diesbezügliche Informationen für die/den Leser:in relevant zur Einordnung der Befunde sein können, sind sie üblicherweise eher ein Teil der Einleitung.

Während das methodische Design einer Evaluation nur selten beschrieben wird, wird häufig der Evaluationsansatz (z.B. theoriebasiert, partizipativ, ‚realist impact evaluation approach') genannt (z.B. WEHU 167, 173, 175, WEIN 56), wobei es oftmals im Wesentlichen bei der Nennung entsprechender Termini bleibt und Angaben dazu, wodurch sich der genannte Ansatz im jeweiligen Kontext auszeichnet (bspw. mittels welcher Verfahren welche Stakeholdergruppen zu welchen Zeitpunkten an welchen Entscheidungen partizipieren konnten oder welche Hypothesen dem theoriebasierten Ansatz zugrunde gelegt wurden) weitgehend fehlen. Weiterhin beschränken sich Angaben zur Stichprobenziehung zumeist auf die Nennung der Stichprobengrösse, während Informationen zur Auswahl der Befragten hingegen i.d.R. fehlen. Bisweilen finden sich auch Querverweise in den Berichten zu weiterführenden Informationen zur Methodik, die jedoch im Bericht ins Leere führen (z.B. WEIN 59).

Die kritische Bewertung der Methodik der Evaluationen drückt sich auch in den im Vergleich zu allen anderen Kriterien geringsten Medianen (SECO: 0,25, AFM: 0,35, DEZA: 0,33) aus, wie die folgende Abbildung darstellt. Bei der Streuung der Einzelwerte sticht hierbei die DEZA heraus, deren Interquartilsabstand gut anderthalbmal so breit ist (0,34) wie der des SECO (0,22) und der AFM (0,20), was auf durchschnittlich grössere Qualitätsunterschiede bei den Evaluationsberichten der DEZA hindeutet.

Abbildung 10: Qualität der Methodik der Evaluationen nach Verwaltungseinheit



Trotz der in der Abbildung erkennbaren Unterschiede zwischen den drei Verwaltungseinheiten ergibt der Signifikanztest ein negatives Resultat. D.h. dass sich die methodische Qualität der Evaluationen von SECO, AFM und DEZA nicht systematisch voneinander unterscheidet.

Die Korrelationsanalyse bestätigt einen recht starken signifikanten Zusammenhang zwischen der Qualität der Methodik und der Ergebnisdarstellung (Gesamtstichprobe: $r = 0{,}493$, $p < 0{,}01$; DEZA: $r = 0{,}572$, $p < 0{,}01$; SECO: $r = 0{,}331$, $p < 0{,}05$) sowie einen etwas schwächeren mit den Schlussfolgerungen und Empfehlungen (Gesamtstichprobe: $r = 0{,}251$, $p < 0{,}01$). Diese Unterschiede sind allerdings aufgrund der insgesamt kleinen Stichprobengrösse vorsichtig zu interpretieren.

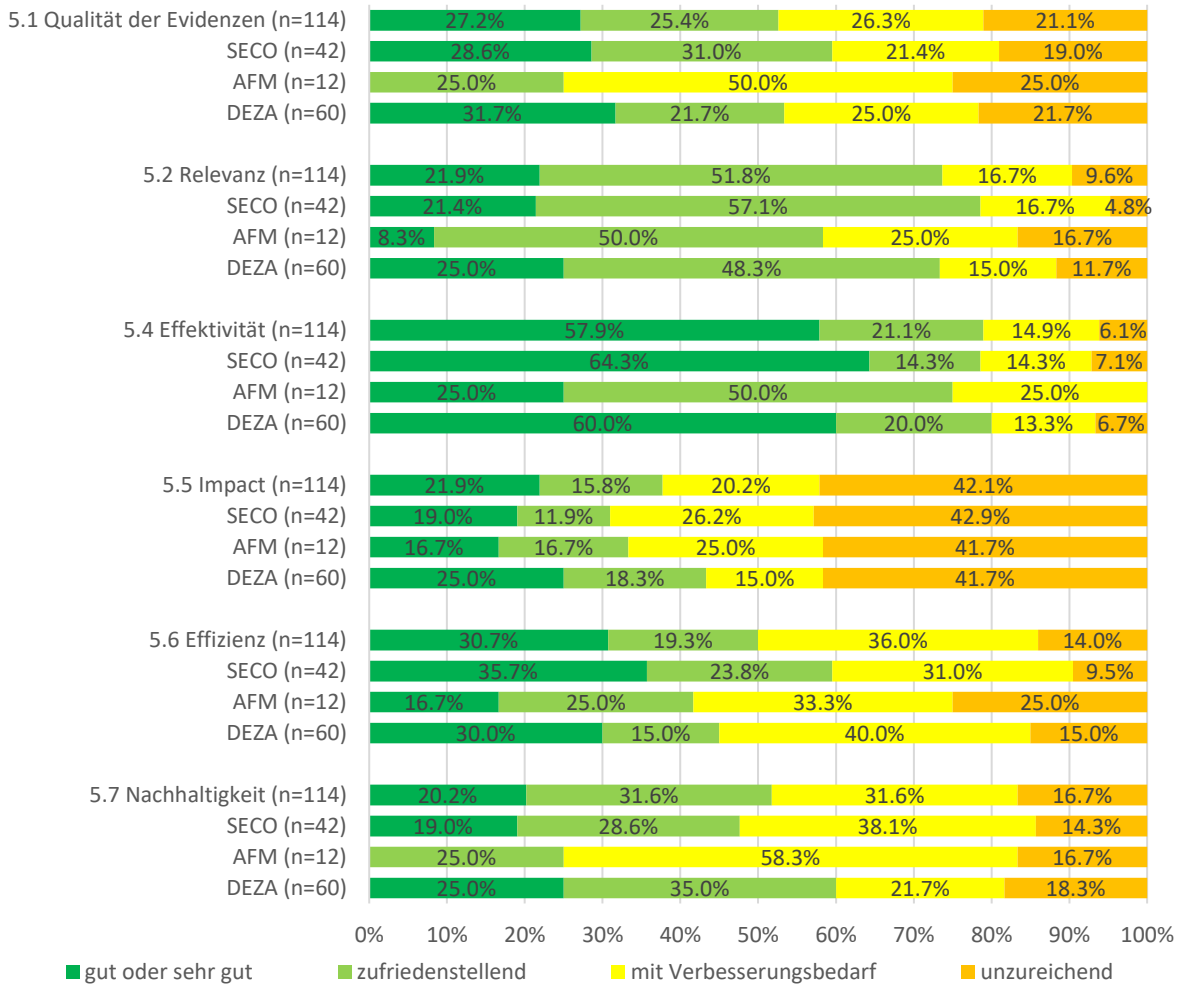
3.5 Beschreibung der Evaluationsergebnisse

Die Qualität der Beschreibung der Evaluationsergebnisse wird anhand von sieben Kriterien bewertet. Dazu gehören die Angemessenheit der Erörterung von fünf der sechs OECD/DAC-Kriterien **Relevanz**, Kohärenz[15], **Effektivität**, **Impact**, **Effizienz** und **Nachhaltigkeit** in den Berichten sowie die **Qualität der Evidenzen**, auf denen die Ergebnisse beruhen.

Wie aus der folgenden Abbildung hervorgeht, wird in etwa der Hälfte aller Berichte bei der Ergebnisdarstellung in mindestens zufriedenstellendem Mass auf die zugrundeliegenden Daten referenziert. Hierbei sind jedoch deutliche Unterschiede zwischen den drei Verwaltungseinheiten zu erkennen. So sind in SECO und DEZA Berichten deutlich häufiger entsprechende Verweise zu finden als in AFM Berichten. Auch die angemessene Darstellung der Relevanz des jeweils evaluierten Vorhabens gelingt in den Evaluationsberichten des SECO und der DEZA, mit in jeweils in etwa drei Viertel aller Fälle, besser als in denen der AFM (58,3%). Hinsichtlich der Beschreibung der Effektivität der Vorhaben sind hingegen keine nennenswerten Unterschiede zu erkennen. Diese wird im Schnitt in acht von zehn Fällen in ausreichender Qualität erörtert. Deutlich schlechter stellt sich die Diskussion des Impacts dar, die für die Gesamtstichprobe lediglich in einem guten Drittel der Berichte als angemessen bezeichnet werden kann, wobei dies der DEZA offenbar tendenziell häufiger gelingt als den anderen beiden Verwaltungseinheiten (siehe hierzu jedoch auch Fußnote 1 im Executive Summary). Bei der Qualität der Effizienzanalysen übertrifft wiederum das SECO mit annähernd zehn Prozentpunkten den Anteil der Berichte in der Gesamtstichprobe (50,0%), die in zumindest zufriedenstellender Weise die diesbezüglichen Ergebnisse darstellen. Die Nachhaltigkeitsbewertung erfolgt ebenso in etwa der Hälfte aller Berichte in ausreichendem Mass, wobei auch hier wiederum deutliche Unterschiede zwischen den drei Verwaltungseinheiten festgestellt werden können. Während sie bei der DEZA in 60 Prozent der Fälle angemessen diskutiert wird, trifft dies beim SECO nur in knapp der Hälfte aller Fälle zu und bei der AFM sogar nur bei einem Viertel.

[15] Da ein Grossteil der Evaluationen vor der Einführung des OECD/DAC-Kriteriums Kohärenz 2019 durchgeführt wurde, wird dieses Kriterium bei der Bewertung ausgeklammert.

Abbildung 11: Bewertung der Qualität der Beschreibung der Evaluationsergebnisse



Die Analyse der Indikatoren der Teilkriterien offenbart die Ursachen der teils erheblichen Unterschiede zwischen den drei Verwaltungseinheiten. Hinsichtlich der Qualität der Evidenzen zeigt sich beispielsweise, dass in AFM Berichten regelmässig eine angemessene Erörterung des kausalen Zusammenhangs zwischen den beobachteten Veränderungen und dem Vorhaben als Ursache fehlt. Wenngleich das auch bei SECO und DEZA Berichten ein häufiges Problem darstellt, finden sich dort zumindest einige Fälle (19,0% bzw. 26,7%), in denen dieser erfolgreich hergestellt wird. Auch die Darstellung der Evaluationsergebnisse entlang eines vorher definierten Analyserasters gelingt SECO und DEZA öfter als AFM. Auffällig ist ebenso, dass, während beim SECO in neun von zehn Berichten eine klare Trennung zwischen den Untersuchungsergebnissen und den daraus abgeleiteten Schlussfolgerungen und Empfehlungen erfolgt, ist dies bei AFM und DEZA nur in jeweils knapp zwei Drittel der Fälle so. Schliesslich ist festzustellen, dass lediglich in etwa einem Drittel der Berichte des SECO und der AFM ein ausreichender Bezug der Befunde zu den Datenquellen erfolgt und unterschiedliche Quellen einander gegenübergestellt werden. Bei der DEZA sind entsprechende Angaben in immerhin der Hälfte der Fälle zu finden.

Der Blick auf die Indikatoren offenbart weiterhin, dass die Diskussion der Relevanz bei den drei Verwaltungseinheiten unterschiedlich konnotiert ist. So werden beim SECO und der DEZA Zielgruppenbedarfe (81,0% bzw. 73,3%) und Partnerstrategien (78,6% bzw. 73,3%) deutlich häufiger als Bewertungsgrundlage herangezogen als bei der AFM (33,3% und 50,0%), bei der dabei eher auf das Interventionsdesign Bezug genommen wird (SECO: 66,7%, AFM: 83,3%, DEZA: 71,1%). Nationale, regionale oder internationale Strategien und Zielsetzungen (z.B. SDGs) spielen schliesslich eher bei der DEZA eine Rolle als bei den anderen beiden Verwaltungseinheiten (SECO: 31,0%, AFM: 33,3%, DEZA: 56,7%). Interessanterweise wird der Anpassungsfähigkeit des Vorhabens an sich verändernde Rahmenbedingungen als ein wesentliches Kriterium zur Bewertung von Relevanz von allen drei Verwaltungseinheiten vergleichsweise wenig Aufmerksamkeit geschenkt (SECO: 42,9%, AFM: 41,7%, DEZA: 31,7%).

Die Effektivität eines Vorhabens wird in den allermeisten Fällen auf Grundlage der erbrachten Leistungen (Outputs) sowie mutmasslich dadurch erzeugten direkten Veränderungen bei den Zielgruppen (Outcomes) bewertet, wobei auch hier wiederum festzustellen ist, dass nicht immer der kausale Zusammenhang angemessen dargestellt wird (SECO: 64,3%, AFM: 58,3%, DEZA: 60,0%). Ausser bei der DEZA werden diese Outcomes zumeist auch nicht disaggregiert nach Teilgruppen (z.B. Frauen, vulnerable Gruppen) aufgezeigt (SECO: 28,6%, AFM: 0,0%, DEZA: 50,0%).

Der Impact, also die entwicklungspolitische Wirksamkeit der Vorhaben, wird nur in zwischen 50 und 60 Prozent der Berichte dezidiert untersucht, wobei auffälliger Weise dabei relativ gesehen häufiger eine Kausalanalyse vorgenommen wird als im Effektivitätskapitel (78,4% vs. 67,9%). Zur Ermittlung einer vollständigen Wirkungsbilanz erforderliche nicht-intendierte Wirkungen werden hingegen im Schnitt nur bei jeder fünften Evaluation berücksichtigt.

Wie bereits aus der Gesamtbewertung des Teilkriteriums hervorgeht, besitzt die Effizienzbewertung beim SECO den grössten Stellenwert. Dieser Eindruck bestätigt sich auch bei der Einzelbetrachtung der Indikatoren. Während die Kosteneffizienz bei allen drei Verwaltungseinheiten gleichermassen in etwa zwei Drittel aller Berichte Berücksichtigung findet, werden die Produktionseffizienz (SECO: 59,5%, AFM: 16,7%, DEZA: 43,3%) und Allokationseffizienz (SECO: 40,5%, AFM: 16,7%, DEZA: 28,3%) deutlich öfter in Berichten des SECO beleuchtet als bei den anderen beiden. Lediglich die Steuerungseffizienz ist geringfügig häufiger Gegenstand in Evaluationen der AFM (SECO: 76,2%, AFM: 83,3%, DEZA: 73,3%).

Der Aspekt der Nachhaltigkeit wird wiederum am häufigsten in Berichten der DEZA angemessen behandelt. Hierbei ist jedoch hinzuzufügen, dass es keiner Verwaltungseinheit gelingt, alle Nachhaltigkeitsaspekte gleichermassen zu berücksichtigen. In zwei Drittel aller Fälle wird diese primär mit der Dauerhaftigkeit der durch ein Vorhaben erzeugten Wirkungen gleichgesetzt. Die übrigen Nachhaltigkeitsdimensionen werden weitaus seltener untersucht (SECO: 21,4%, AFM: 0,0%, DEZA: 28,3%). Auch eine Risikobewertung erfolgt auf die Gesamtstichprobe bezogen lediglich in etwas mehr als der Hälfte der Fälle (SECO: 52,4%, AFM: 41,7%, DEZA: 63,3%), eine Diskussion von Strategien zum Umgang mit identifizierten Risiken sogar nur in gut einem Viertel (SECO: 19,0%, AFM: 25,0%, DEZA: 33,3%).

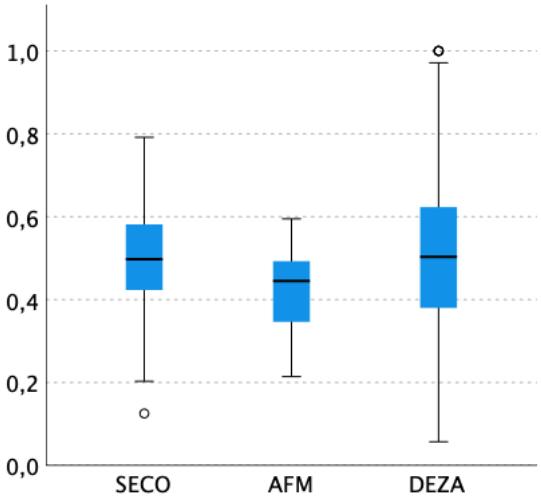
Ein weiterer Blick auf die qualitativen Daten verrät weitere Details über die Qualität der Ergebnisdarstellung in den Evaluationsberichten der drei Verwaltungseinheiten. Ein wesentlicher Befund dabei ist, dass die Struktur der Darstellung in vielen Fällen z.T. erhebliches Verbesserungspotential bietet. So werden bspw. in einem Bericht (WEHU 169) die Wirksamkeit und Nachhaltigkeit des Vorhabens im Kapitel zur Effektivität diskutiert. In anderen Berichten finden sich ebenfalls Informationen zu verschiedenen OECD/DAC Kriterien an falschen Stellen (z.B. WEHU 174, WEIN 63, WEMU 82, AFM 4, 7, DEZA 2.04). In einem Fall (WEHU 190) werden die erzielten Ergebnisse auf der Outcome-Ebene gleich auf mehrere Kapitel verteilt beschrieben. Auch bei der Differenzierung zwischen Effektivitäts- und Effizienzanalyse tun sich offenbar manche Evaluator:innen schwer (WEMU 79, 86). Jedoch lässt sich für fast jedes Ergebniskapitel eine Reihe von Good Practices identifizieren:

- ✓ Relevanz: WEHU 168, 172, 180a, WEIN 63, WEMU 86, 88, DEZA 1.02, 1.41
- ✓ Effektivität: WEHU 168, 172, 173, AFM 12, DEZA 1.03, 1.24, 1.43
- ✓ Effizienz: WEHU 174 (insb. Produktionseffizienz), WEMU 86 (insb. Managementeffizienz), AFM 8, DEZA 1.02, 1.48
- ✓ Nachhaltigkeit: WEHU 169 (zur ökologischen Nachhaltigkeit), WEHU 172, 173, DEZA 1.03, 1.48

Lediglich die Diskussion der entwicklungspolitischen Wirksamkeit ist in keinem der Berichte vollständig geglückt.

Wie die folgende Abbildung veranschaulicht, deuten die nah beieinander liegenden Mediane (SECO: 0,50, AFM: 0,45, DEZA: 0,50) darauf hin, dass keine wesentlichen Unterschiede zwischen den drei Verwaltungseinheiten hinsichtlich der Qualität der Beschreibung der Evaluationsergebnisse bestehen. Jedoch zeigt der fast doppelt so grosse Interquartilsabstand (SECO: 0,13, AFM: 0,15, DEZA: 0,25) bei der DEZA an, dass sich die Qualität ihrer Berichte bei diesem Kriterium deutlich heterogener darstellt.

Abbildung 12: Qualität der Beschreibung der Evaluationsergebnisse nach Verwaltungseinheit



Der Gruppenvergleich liefert jedoch keine Hinweise auf einen statistisch signifikanten Unterschied zwischen den drei Verwaltungseinheiten.

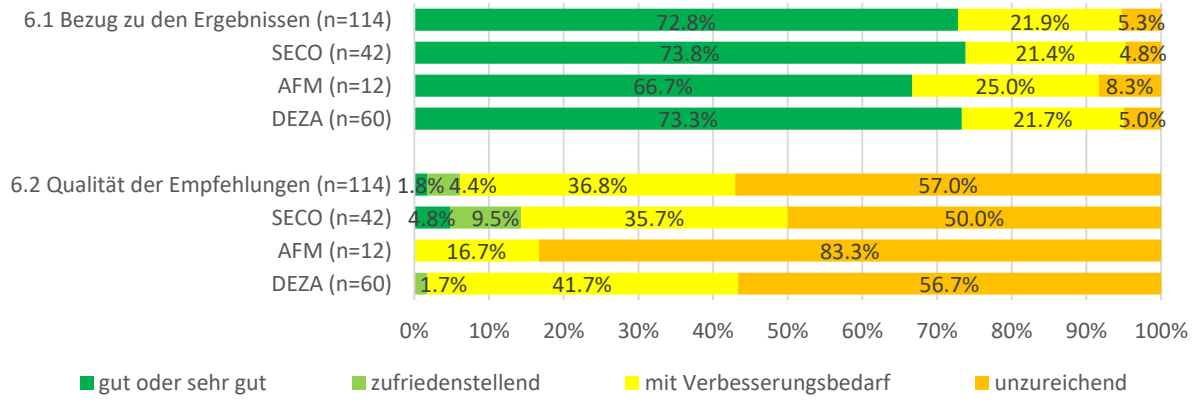
Für die Gesamtstichprobe kann eine schwache, aber signifikante Korrelation der Beschreibung der Evaluationsergebnisse mit der Qualität der Schlussfolgerungen und Empfehlungen festgestellt werden ($r = 0{,}305$, $p < 0{,}01$), die für die SECO Teilstichprobe etwas stärker ($r = 0{,}395$, $p < 0{,}01$) ausgeprägt ist. Die Korrelation kann als logischer Zusammenhang interpretiert werden, da sinnvolle Schlussfolgerungen und nützliche Empfehlungen zwingend methodisch haltbare Ergebnisse erfordern. Anders formuliert: Aus wenig validen Ergebnissen können auch keine haltbaren Schlussfolgerungen und Empfehlungen abgeleitet werden.

3.6 Schlussfolgerungen und Empfehlungen

Die Qualität der in den Evaluationsberichten enthaltenen Schlussfolgerungen und Empfehlungen wird anhand von zwei Teilkriterien bewertet: dem **Bezug** der Schlussfolgerungen **zu den Ergebnissen** und der Qualität der Empfehlungen im Sinne ihrer **Priorisierung und Handlungsorientierung**, wie in den jeweiligen Kapiteln der Evaluierungsberichte dargelegt.

Die folgende Abbildung weist auf einen deutlichen Unterschied in der Bewertung der beiden Teilkriterien hin. Während in der grossen Mehrheit der Berichte Schlussfolgerungen und Empfehlungen logisch mit den Evaluationsergebnissen verknüpft sind, besteht erheblicher Verbesserungsbedarf hinsichtlich der Formulierung der Empfehlungen. Beide Befunde gelten ungeachtet gewisser Unterschiede im quantitativen Rating für das SECO, die AFM und die DEZA gleichermassen.

Abbildung 13: Bewertung der Qualität der Schlussfolgerungen und Empfehlungen



Die den Teilkriterien zugrundeliegenden Indikatoren geben wiederum Aufschluss über die Grundlage dieses Bewertungsergebnisses. So zeigt sich, dass in drei Viertel aller Berichte in nachvollziehbarer

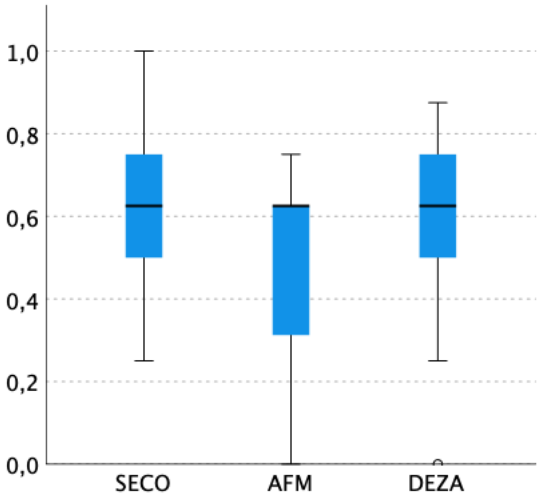
Weise ein logischer Zusammenhang zwischen den Evaluationsergebnissen und den daraus abgeleiteten Schlussfolgerungen hergestellt wird (SECO: 73,8%, AFM: 66,7%, DEZA: 78,3%). Ein Rückbezug der Empfehlungen auf Schlussfolgerungen erfolgt sogar in gut neun von zehn Fällen (SECO: 95,2%, AFM: 91,7%, DEZA: 90,0%).

Demgegenüber zeigen sich deutliche Schwächen bei der Formulierung von Empfehlungen. Nur in gut der Hälfte der Berichte sind diese an spezifische Adressaten gerichtet (bspw. Projektverantwortliche, Partner, Umsetzungsorganisation) (SECO: 54,8%, AFM: 50,0%, DEZA: 63,3%). Eine Priorisierung der Empfehlungen findet noch seltener statt (SECO: 21,4%, AFM: 0,0%, 10,0%) und ebenso eine Darstellung, in welchem zeitlichen Rahmen diese umgesetzt werden sollten (SECO: 9,5%, AFM: 0,0%, DEZA: 8,3%). Eine Zusammenfassung von Lessons Learnt scheint nur in Berichten des SECO eher üblich zu sein (61,9%), während hingegen diese bei der DEZA lediglich in der Hälfte der Berichte zu finden sind und nur in einem Drittel der AFM Berichte. Hierzu ist anhand der qualitativen Daten festzustellen, dass sich die Inhalte von Lessons Learnt z.T. recht unterschiedlich darstellen. Während in einigen Fällen auf generelle, auf andere Kontexte übertragbare Lernerfahrungen rekurriert wird, sind andere sehr konkret und handlungsleitend (bspw. im Hinblick auf ein Folgevorhaben) formuliert.

Weiterhin fällt auf, dass Schlussfolgerungen im Gegensatz zu Empfehlungen bisweilen sehr kurz und knapp formuliert sind (z.B. WEIN 62) und dadurch einen direkten Bezug zu den Evaluationsergebnissen, quasi als deren empirische Begründung, vermissen lassen (z.B. WEHU 167, 186). Schliesslich enthalten einige Berichten kein dezidiertes Kapitel für Schlussfolgerungen und Empfehlungen (z.B. WEHU 189, WEMU 89, AFM 5, 11, 12). Weiterhin finden sich Empfehlungen verstreut im Ergebnisteil bzw. lediglich im Executive Summary (WEIF 113, AFM 5) oder vermischt mit den Schlussfolgerungen (z.B. WEMU 79). Dennoch lassen sich auch für dieses Kapitel mehrere Good Practices finden, wie bspw. WEHU 181, 188, AFM 1, 6 oder DEZA 1.38.

Die folgende Abbildung lässt bereits erahnen, dass sich die Qualität der Schlussfolgerungs- und Empfehlungsteile der Evaluationsberichte der drei Verwaltungseinheiten statistisch nicht signifikant voneinander unterscheiden. So liegen die jeweiligen Mediane gleich auf (SECO: 0,63, AFM: 0,63, DEZA: 0,63). Die verhältnismässig breiten Interquartilsabstände (SECO: 0,25, AFM: 0,34, DEZA: 0,25) weisen ferner auf eine grosse Streuung in der Bewertung dieses Kriteriums hin, was ebenso auf einen nicht messbaren Unterschied schliessen lässt.

Abbildung 14: Qualität der Schlussfolgerungen und Empfehlungen nach Verwaltungseinheit



4. Zusammenhangsanalysen

Abschliessend werden nun die Ergebnisse der eingangs erwähnten Zusammenhangsanalysen vorgestellt. Gemäss Pflichtenheft gilt es im Einzelnen folgende fünf Fragen zu beantworten:

1. Gibt es bei den externen Evaluationen von SECO und DEZA einen Zusammenhang zwischen bestimmten Qualitätskriterien und den aus den Evaluationen abgeleiteten Ratings der DAC-Kriterien?
2. Gibt es einen Zusammenhang zwischen der Aussagekraft (Strength of Evidence) und der Erfolgsquote insgesamt (Overall Performance)?
3. Gibt es einen Zusammenhang zwischen der Qualität der Diskussion der einzelnen DAC-Kriterien und deren Rating?
4. Gibt es einen Zusammenhang zwischen bestimmten Qualitätskriterien und den Kosten der Evaluationen?
5. Gibt es einen Zusammenhang zwischen den Kosten der Evaluationen und den aus ihnen abgeleiteten DAC-Ratings?

4.1 Zusammenhang zwischen Qualitätskriterien und DAC-Ratings

Zur Beurteilung, ob es bei den SECO und DEZA[16] Evaluationen einen Zusammenhang zwischen bestimmten Qualitätskriterien und den aus den Evaluationen abgeleiteten Ratings der DAC-Kriterien gibt, sind die einzelnen Kriterien mit den Ratings zu korrelieren. Hierbei ist zu beachten, dass die Ratings bei dem SECO ordinal skaliert sind, während sie bei der DEZA metrisch skaliert sind[17], weswegen zwei verschiedene Koeffizienten zu berechnen sind, nämlich beim SECO der Spearman'sche Rangkorrelationskoeffizient Rho und bei der DEZA der Koeffizient nach Pearson. Die folgenden beiden Tabellen zeigen die Ergebnisse der beiden Analysen:

Tabelle 1: Korrelation zwischen Bewertungskriterien und Erfolgsratings beim SECO

SECO (Spearman-Rho)

		Gesamt-bewertung	Relevanz-bewertung	Effektivitäts-bewertung	Effizienz-bewertung	Nachhaltigkeits-bewertung
Gesamt-qualität	Spearman-Rho	**-,279***	-0,012	-0,063	-0,059	0,023
	p-Wert	**0,028**	0,928	0,610	0,641	0,851
	N	**42**	41	42	40	40
Leistungs-beschrei-bung	Spearman-Rho	-0,199	0,002	-0,111	-0,168	0,122
	p-Wert	0,119	0,990	0,374	0,189	0,328
	N	42	41	42	40	40
Executive Summary	Spearman-Rho	-0,177	-0,124	-0,132	**-,271***	-0,034
	p-Wert	0,181	0,356	0,309	**0,040**	0,790
	N	42	41	42	**40**	40
Einleitung & Kon-textana-lyse	Spearman-Rho	-0,249	0,000	-0,041	0,162	0,025
	p-Wert	0,051	1,000	0,744	0,203	0,841
	N	42	41	42	40	40
Methodik	Spearman-Rho	-0,104	-0,083	0,020	0,069	-0,006
	p-Wert	0,420	0,527	0,875	0,591	0,960
	N	42	41	42	40	40
Ergebnis-darstel-lung	Spearman-Rho	-0,146	0,088	0,065	-0,036	0,026
	p-Wert	0,252	0,494	0,601	0,777	0,831
	N	42	41	42	40	40
Schlussf. & Emp-fehlungen	Spearman-Rho	-0,166	0,139	-0,052	0,083	-0,144
	p-Wert	0,226	0,316	0,698	0,543	0,279
	N	42	41	42	40	40

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

[16] Da die AFM keine Erfolgsquoten berechnet, beruhen die folgenden Ergebnisse nur auf Daten vom SECO und der DEZA.

[17] D.h. bei der SECO werden nur diskrete Notenwerte (1, 2, 3 & 4) vergeben, bei der DEZA werden Noten auf einer kontinuierlichen Skala von 1 bis 4 vergeben.

Wie die Tabelle zeigt, lässt sich beim SECO zwar auf der Ebene der Einzelkriterien lediglich eine statistisch signifikante Korrelation zwischen der Bewertung des Executive Summaries und des Effizienzratings feststellen; ein Befund, der allerdings keine inhaltlich sinnvolle Zusammenhangsvermutung begründet und insofern eher als Scheinkorrelation zu interpretieren ist. Auf der Ebene der Gesamtbewertung der Qualität der Evaluationen findet sich jedoch ebenso eine signifikante Korrelation mit dem Overall Performance Rating.[18] Wenngleich der Zusammenhang verhältnismässig schwach ausgeprägt ist – was in Anbetracht der geringen Stichprobengrösse aber auch nicht verwundert –, verweist er auf einen durchaus schlüssigen positiven Zusammenhang zwischen der methodischen Qualität einer Evaluation und ihrem Ergebnis. Oder als Hypothese formuliert: Je besser die Qualität eines Evaluationsberichts, desto besser die darin enthaltene Bewertung der OECD-DAC Kriterien.

Das Ergebnis der Analyse der Korrelation der Qualitätskriterien mit den Performance Ratings der DEZA Berichte stellt sich etwas anders dar. Wie die folgende Tabelle zeigt, kann kein Zusammenhang zwischen der Gesamtbewertung der Qualität der Evaluation und dem Overall Performance Rating hergestellt werden. Jedoch korreliert die Darstellung der Evaluationsergebnisse signifikant mit dem Rating und der Nachhaltigkeitsbewertung. Weiterhin kann eine signifikante Korrelation zwischen der Bewertung von Einleitung und Kontextanalyse sowie der Effizienzbewertung identifiziert werden. Während letztere Korrelation wiederum sich einer inhaltlichen Begründung entzieht, erscheint zumindest ein Zusammenhang zwischen Ergebnisdarstellung und Gesamtrating schlüssig. Offenbar geht eine systematischere Ergebnisdarstellung mit einem besseren Overall Performance Rating des evaluierten Vorhabens einher.

Tabelle 2: Korrelation zwischen Bewertungskriterien und Erfolgsratings bei der DEZA[19]

DEZA (Pearson)

		Gesamtbewertung	Relevanzbewertung	Effektivitätsbewertung	Effizienzbewertung	Nachhaltigkeitsbewertung
Gesamtqualität	Pearson	-0,248	-0,206	-0,091	-0,111	-0,228
	p-Wert	0,089	0,161	0,540	0,458	0,127
	N	48	48	48	47	46
Leistungsbeschreibung	Pearson	0,057	-0,189	0,104	0,128	0,048
	p-Wert	0,702	0,199	0,483	0,392	0,752
	N	48	48	48	47	46
Executive Summary	Pearson	-0,237	-0,202	-0,239	-0,158	-0,103
	p-Wert	0,104	0,169	0,101	0,288	0,496
	N	48	48	48	47	46
Einleitung & Kontextanalyse	Pearson	0,166	-0,197	0,257	**,311***	0,093
	p-Wert	0,261	0,179	0,078	**0,034**	0,538
	N	48	48	48	**47**	46
Methodik	Pearson	-0,179	-0,043	-0,069	-0,150	-0,193
	p-Wert	0,225	0,774	0,639	0,313	0,198
	N	48	48	48	47	46
Ergebnisdarstellung	Pearson	**-,324***	-0,104	-0,168	-0,220	**-,317***
	p-Wert	**0,025**	0,480	0,253	0,137	**0,032**
	N	**48**	48	48	47	**46**
Schlussf. & Empfehlungen	Pearson	-0,155	0,000	-0,067	-0,104	-0,161
	p-Wert	0,294	1,000	0,649	0,485	0,286
	N	48	48	48	47	46

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

→ **Antwort:** Ja, die Daten weisen darauf hin, dass es bei Evaluationen des SECO einen Zusammenhang zwischen der Gesamtbewertung eines Vorhabens und der Gesamtqualität des Evaluationsberichts gibt

[18] Dass der Korrelationskoeffizient negativ ist, ist darauf zurückzuführen, dass die Skalen der Qualitätsbewertung und des Performance Ratings eine gegensätzliche Polarität aufweisen. So werden die Evaluationsberichte im Rahmen dieser Metaevaluation von 1 = „unzureichend“ bis 4 = „gut oder sehr gut“ bewertet (vgl. Kapitel 2), während die Skala der DAC-Ratings von 1 = „highly satisfactoy“ bis 4 = „highly unsatisfactory“ reicht.

[19] Aufgrund zu vieler fehlender Werte konnten zwischen den Qualitätskriterien und den Performancekriterien Kohärenz und Impact keine Korrelationen überprüft werden.

und dass bei der DEZA die Gesamtbewertung mit der Qualität der Ergebnisdarstellung zusammenhängt.

4.2 Zusammenhang zwischen der Aussagekraft und der Erfolgsquote

Um herauszufinden, ob es einen Zusammenhang zwischen der Aussagekraft (Strength of Evidence) und der Erfolgsquote insgesamt (Overall Performance) gibt, wurden die jeweiligen Bewertungen miteinander korreliert. Das Ergebnis von r = -0,070 und p = 0,512 deutet dabei für die Gesamtstichprobe auf keinen systematischen Zusammenhang hin. Auch disaggregiert nach Verwaltungseinheit können keine statistisch signifikanten Korrelationen identifiziert werden (SECO: Spearman-Rho, r = 0,067, p = 0,336, DEZA: Pearson, r = -0,154, p = 0,148).

→ **Antwort:** Nein, die Untersuchungsergebnisse weisen nicht auf einen Zusammenhang zwischen der Aussagekraft der Evaluationsergebnisse und der damit ermittelten Erfolgsquote hin.

4.3 Zusammenhang zwischen der Qualität der DAC-Diskussion und DAC-Ratings

Mittels Korrelationsanalysen konnte für die Gesamtstichprobe ebenfalls kein Zusammenhang zwischen der Qualität der Diskussion der einzelnen DAC-Kriterien und deren jeweiligen Rating bestätigt werden. Wie die folgende Tabelle zeigt, korreliert jedoch das Overall Performance Rating wenn auch nur schwach mit der Diskussion von drei der vier DAC-Kriterien, nämlich mit der der Relevanz, der Effektivität sowie der Nachhaltigkeit. Auch dieser Befund deutet auf den bereits oben formulierten Zusammenhang zwischen den Bewertungsergebnissen und der Qualität dieser Ergebnisse hin.

Tabelle 3: Korrelation zwischen Qualität der Diskussion der einzelnen DAC-Kriterien und deren jeweiligen Rating (Gesamtstichprobe)

	Qualität der…	… Diskussion der Relevanz	… Diskussion der Effektivität	… Diskussion der Effizienz	… Diskussion der Nachhaltigkeit
Gesamtbewertung	Spearman-Rho	**-,298****	**-,217***	-0,074	**-,187***
	p-Wert	**0,002**	**0,020**	0,243	**0,039**
	N	**90**	**90**	90	**90**
Relevanzbewertung	Spearman-Rho	-0,003	-0,005	-0,020	-0,060
	p-Wert	0,490	0,482	0,426	0,287
	N	89	89	89	89
Effektivitätsbewertung	Spearman-Rho	-0,133	-0,148	-0,031	-0,122
	p-Wert	0,105	0,082	0,386	0,125
	N	90	90	90	90
Effizienzbewertung	Spearman-Rho	**-,212***	**-,206***	-0,119	0,002
	p-Wert	**0,024**	**0,028**	0,136	0,494
	N	**87**	**87**	87	87
Nachhaltigkeitsbewertung	Spearman-Rho	-0,114	-0,057	-0,061	0,024
	p-Wert	0,148	0,302	0,288	0,413
	N	86	86	86	86

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Während in der SECO Teilstichprobe keine statistisch signifikanten Korrelationen identifiziert werden konnten, ist festzustellen, dass in der DEZA Teilstichprobe die Nachhaltigkeitsbewertung schwach mit der Qualität des Nachhaltigkeitskapitels korreliert (r = -0,29, p < 0,05). Wie aus der folgenden Tabelle ersichtlich wird, zeigt sich darüber hinaus ein Zusammenhang zwischen dem Overall Performance Rating und der Diskussion aller vier DAC-Kriterien; ein weiteres Indiz für den oben vermuteten Zusammenhang.

Tabelle 4: Korrelation zwischen Qualität der Diskussion der einzelnen DAC-Kriterien und deren jeweiligen Rating (DEZA)

	Qualität der…	… Diskussion der Relevanz	… Diskussion der Effektivität	… Diskussion der Effizienz	… Diskussion der Nachhaltigkeit
Gesamtbewertung	Pearson	**-,405****	**-,248***	**-,251***	**-,256***
	p-Wert	**0,002**	**0,044**	**0,043**	**0,040**

	N	**48**	**48**	**48**	**48**
Relevanz-bewertung	Pearson	-0,075	-0,128	-0,179	**-,268***
	p-Wert	0,306	0,193	0,112	**0,033**
	N	48	48	48	**48**
Effektivitäts-bewertung	Pearson	-0,232	-0,219	-0,069	-0,232
	p-Wert	0,056	0,067	0,320	0,056
	N	48	48	48	48
Effizienz-bewertung	Pearson	**-,329***	-0,201	-0,162	-0,033
	p-Wert	**0,012**	0,088	0,138	0,414
	N	**47**	47	47	47
Nachhaltig-keitsbewer-tung	Pearson	-0,236	-0,132	-0,118	**-,290***
	p-Wert	0,057	0,191	0,217	**0,025**
	N	46	46	46	**46**

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

→ **Antwort:** Ein direkter Zusammenhang zwischen der Qualität der Diskussion einzelner DAC-Kriterien und deren Rating konnte nur für die Nachhaltigkeitsbewertung in den Evaluationsberichten der DEZA festgestellt werden. Jedoch konnte sowohl für das SECO als auch die DEZA ein Zusammenhang zwischen der Gesamtbewertung der Vorhaben und der Qualität der Diskussion einzelner Bewertungsdimensionen identifiziert werden.

4.4 Zusammenhang zwischen Qualitätskriterien und Evaluationskosten

Hinsichtlich des Zusammenhangs der Kosten einer Evaluation mit ihrer methodischen Qualität konnte für die Gesamtstichprobe keine statistisch signifikante Korrelation festgestellt werden (vgl. Kapitel 7.4.1 im Anhang). Die Befunde deuten damit darauf hin, dass die Evaluationsbudgets keinen nennenswerten Einfluss auf die Qualität der Evaluationsergebnisse ausüben.

→ **Antwort:** Die Untersuchungsergebnisse legen keinen Zusammenhang zwischen der methodischen Qualität einer Evaluation und ihren Kosten nahe.

4.5 Zusammenhang zwischen Evaluationskosten und DAC-Ratings

Während die Korrelationsanalyse für die Gesamtstichprobe und die SECO-Teilstichprobe keinen Zusammenhang zwischen den Kosten der Evaluationen und den aus ihnen abgeleiteten DAC-Ratings ergab, lieferte sie für die DEZA-Teilstichprobe ein statistisch signifikantes Ergebnis ($p < 0,01$) für das Overall Performance Rating, das mit einem r von 0,401 im Vergleich zu den vorgenannten signifikanten Korrelationen sogar relativ stark ausgeprägt ist, sowie für das Sustainability Rating ($r = 0,342$, $p < 0,05$) (vgl. Kapitel 7.4.2 im Anhang). Hier liegt die Vermutung nahe, dass eine teurere Evaluation tendenziell eine schlechtere Bewertung des Vorhabens liefert als eine billigere.

→ **Antwort:** Hinsichtlich des Zusammenhangs zwischen den Kosten der Evaluationen und den aus ihnen abgeleiteten DAC-Ratings sind die Untersuchungsergebnisse nicht eindeutig. Lediglich für Evaluationen der DEZA kann ein entsprechender negativer Zusammenhang vermutet werden.

5. Schlussfolgerungen

Die in den beiden vorangegangenen Kapiteln dargestellten Ergebnisse zeichnen ein gemischtes Bild hinsichtlich der Qualität der Evaluationen des SECO, der AFM und der DEZA, zu deren gemeinsamen Stärken zweifelsohne die gute Verständlichkeit der Berichte und deren Transparenz hinsichtlich der ihnen zugrundeliegenden Daten zählen. Weiterhin gelingt es offenbar den meisten Evaluationen, ihren jeweiligen Untersuchungsgegenstand angemessen zu beschreiben, dessen Wirkungen bei den direkten Zielgruppen zu erfassen und die daraus gezogenen Schlussfolgerungen mit den empirischen Ergebnissen zu verknüpfen. Auch die Relevanzbewertungen in den Evaluationen aller drei Verwaltungseinheiten der Schweizer internationalen Zusammenarbeit sind von überwiegend zufriedenstellender Qualität.

Diesen Stärken steht jedoch eine Reihe von Schwächen gegenüber, die sich insbesondere in einer mangelnden methodischen Qualität und zumindest heterogenen Qualität der Ergebnisdarstellung in den Berichten widerspiegeln. So weisen fast alle Evaluationen z.T. erhebliche Defizite hinsichtlich der Beschreibung ihres Designs und ihrer Auswertungsverfahren auf. Auch die Darstellung von Stichprobenziehung und Erhebungsinstrumenten sowie die Diskussion von Limitationen und praktischen Herausforderungen während des Umsetzungsprozesses ist zumeist unzureichend. Die Präsentation der Evaluationsergebnisse ist insbesondere auf höheren Wirkungsebenen von methodischen Schwächen gekennzeichnet. Während im Effektivitäts-Kapitel Attributions- oder Kontributionsanalysen von ausreichender Qualität sind, sind sie im Impact-Kapitel verbesserungsbedürftig. Auch die Qualität der Empfehlungen birgt erhebliches Verbesserungspotential, insbesondere was deren Priorisierung und ihren zeitlichen Bezug anbetrifft. Dementsprechend sollte auf den Aspekt der Nützlichkeit der Empfehlungen bei der Befragung der Adressaten der Evaluationen besonderen Wert gelegt werden.

Mit Blick auf die Leistungsbeschreibungen der Evaluationen ist festzustellen, dass deren Qualität ebenfalls Steigerungspotentiale aufweist. So stellt sich in Anbetracht von Mengengerüst und Evaluationsgegenstand oftmals die Frage der Machbarkeit. Weiterhin weisen die Leistungsbeschreibungen oftmals Lücken bei der Erläuterung der methodischen Anforderungen auf, die es im Rahmen der Evaluation zu erfüllen gilt. Schliesslich fehlen vielfach Angaben zu den Aufgaben und Zielsetzung der Evaluation sowie ihres Hintergrunds und Kontextes.

Betrachtet man die Stärken und Schwächen der drei Verwaltungseinheiten im Einzelnen, lassen sich mitunter erhebliche Unterschiede erkennen. Im Schnitt stellt sich dabei die Qualität der im Auftrag der DEZA und des SECO durchgeführten Evaluationen in etwa vergleichbar dar, während die Evaluationen für die AFM deutlich mehr Mängel aufweisen. Bei genauerer Betrachtung zeigen sich jedoch zwischen allen drei Organisationen mitunter erhebliche Unterschiede hinsichtlich der einzelnen Teilkriterien. So ist die Beschreibung des Evaluationsgegenstands sowie der Datengrundlage in Berichten des SECO am detailliertesten. Die DEZA stellt ihren Evaluator:innen die besten Leistungsbeschreibungen zur Verfügung und bei Evaluationen der AFM werden Datenerhebungsinstrumente am genausten beschrieben und offenbar auch am ehesten zielgruppenadäquat eingesetzt. Wesentliche Unterschiede zeigen sich auch bei der Einbettung des Evaluationsgegenstands in seinen politischen Kontext, die in Evaluationen der DEZA am besten und AFM Evaluationen am schlechtesten erfolgt.

Bei der Darstellung der Evaluationsergebnisse zeigen sich wiederum nur geringfügige Qualitätsdifferenzen. So können nur wenige nennenswerte Unterschiede bzgl. der Qualität der Ergebnisdiskussion zwischen den OECD-DAC Kriterien ausgemacht werden. Lediglich die Nachhaltigkeitskapitel der AFM fallen durch eine vergleichsweise schwächere Darstellung auf. Es entsteht der Eindruck, dass die drei Verwaltungseinheiten gleichermassen mit den jeweiligen Anforderungen der Bewertungsdimensionen zu kämpfen haben. Bei der Qualität der Leistungsbeschreibung fällt schliesslich auf, dass es lediglich der DEZA in allen Fällen gelingt, Evaluationsfragen und -kriterien in zumindest zufriedenstellender Weise zu formulieren.