Stabsstelle Evaluierung Querschnittauswertung

WISSEN WAS WIRKT

Meta-Evaluierung und statistische Auswertung der Projektevaluierungen 2017 / 2018

Teil I: Meta-Evaluierung
Im Auftrag der GIZ durchgeführt durch externe Evaluator/innen

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Impressum

Als Bundesunternehmen unterstützt die GIZ die deutsche Bundesregierung bei der Erreichung ihrer Ziele in der Internationalen Zusammenarbeit für nachhaltige Entwicklung.

Als Stabsstelle Evaluierung der GIZ untersteht sie organisatorisch direkt dem Vorstand und ist vom operativen Geschäft getrennt. Diese Organisationsstruktur stärkt ihre Unabhängigkeit. Die Stabsstelle Evaluierung ist mandatiert, zur Entscheidungsfindung evidenzbasierte Ergebnisse und Empfehlungen zu generieren, einen glaubwürdigen Wirkungsnachweis zu erbringen und die Transparenz zu den Ergebnissen zu erhöhen.

Diese Evaluierung wurde im Auftrag der Stabsstelle Evaluierung von externen Evaluator/innen durchgeführt und der Evaluierungsbericht von externen Evaluator/innen verfasst. Er gibt ausschließlich deren Meinung und Wertung wieder. Die GIZ hat eine Stellungnahme zu den Ergebnissen und eine Management Response zu den Empfehlungen verfasst.

Evaluator/innen:
Matías Krämer, Syspons GmbH
Olga Almqvist, Syspons GmbH
Lennart Raetzell, Syspopns GmbH
Birgit Alber, Syspons GmbH
Lukas Verfürden, Syspons GmbH

Autor/innen des Evaluierungsberichts:
Matías Krämer, Syspons GmbH
Olga Almqvist, Syspons GmbH

Consulting:
Syspons GmbH
Prinzenstraße 84, Aufgang 1
10969 Berlin

Konzeption, Koordination und Management
Dr. Annette Backhaus, GIZ Stabsstelle Evaluierung, Gruppenleiterin Unternehmensstrategische Evaluierungen

Michael Florian, GIZ Stabsstelle Evaluierung Senior-Fachkonzeptionist Unternehmensstrategische Evaluierungen

Christoph Mairesse, GIZ Stabsstelle Evaluierung, Senior-Fachkonzeptionist Unternehmensstrategische Evaluierungen

Lucas Jacobs, GIZ Stabsstelle Evaluierung, Fachkonzeptionist Unternehmensstrategische Evaluierungen

Verantwortlich:
Dr. Ricardo Gomez, GIZ, Leiter Stabsstelle Evaluierung

Herausgeberin
Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Sitz der Gesellschaft
Bonn und Eschborn

Friedrich-Ebert-Allee 36 + 40
53113 Bonn, Deutschland
T +49 228 4460-0
F +49 228 4460 - 1766

E evaluierung@giz.de
I www.giz.de/evaluierung
www.youtube.com/user/GIZonlineTV
www.facebook.com/gizprofile
https://twitter.com/giz_gmbh

Design
DITHO Design GmbH, Köln

Bonn 2019

Das vorliegende Dokument ist auf der GIZ-Website als pdf-Download verfügbar unter www.giz.de/wissenwaswirkt
Anfragen nach einer gedruckten Ausgabe richten Sie bitte an evaluierung@giz.de

Inhalt

Verzeichnis der Darstellung der Ergebnisse im Evaluierungsstandard Genauigkeit nach Bewertungskriterien

Abbildungen

Abkürzungsverzeichnis

AIZ	Akademie für Internationale Zusammenarbeit
AV	Auftragsverantwortliche/r
B 1	Regionalbereich Afrika
B 2	Regionalbereich Asien/Pazifik, Lateinamerika/Karibik
B 3	Regionalbereich Mittelmeer, Europa und Zentralasien
BMZ	Bundesministerium für Wirtschaftliche Zusammenarbeit und Entwicklung
D	Durchführbarkeit (Evaluierungsstandard)
DAC	Development Assistance Committee
DeGEval	Deutsche Gesellschaft für Evaluation e.V.
EZ	Entwicklungszusammenarbeit
F	Fairness (Evaluierungsstandard)
FMB	Fach- und Methodenbereich
FZ	Finanzielle Zusammenarbeit
G	Genauigkeit (Evaluierungsstandard)
GIZ	Gesellschaft für Internationale Zusammenarbeit GmbH
GloBe	Sektor- und Globalvorhaben
GVR	Gemeinsame Verfahrensreform
IATI	International Aid Transparency Initiative
InWEnt	Internationale Weiterbildung und Entwicklung gGmbH
JCSEE	Joint Committee on Standards for Educational Evaluation
M	Mittelwert
N	Nützlichkeit (Evaluierungsstandard)
OECD	Organisation for Economic Cooperation and Development
PEV	Projektevaluierung
PFK	Projektfortschrittskontrolle
SD	Standardabweichung
SMART	Specific, measurable, accepted, realistic, time-bound
ToR	Terms of Reference
TZ	Technische Zusammenarbeit
UE	Unabhängige Evaluierung
VGK	Verwaltungsgemeinkosten
ZAS	Zeitaufschriebe

1 Vorwort

Die Stabstelle Evaluierung der deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) führt seit 2009 Meta-Evaluierungen von Evaluationsberichten durch. Die vorliegende Meta-Evaluierung analysiert die methodische Qualität aller zwischen Oktober 2016 und September 2018 fertiggestellten Projektevaluierungen (PEV) und bestimmt Trends in der methodischen Qualität auf Basis der Ergebnisse aus den Meta-Evaluierungen der PEV aus den Jahren 2016 und 2017.

Die PEV waren im April 2014 in der GIZ als verbindliches Instrument im Auftragsmanagement für das BMZ-Geschäft eingeführt worden. Diese waren flächendeckend und dezentral gesteuert und verbanden die Evaluierung mit der Prüfung der Folgemaßnahmen in einem Prozess. Die PEV stellten seinerzeit zudem eine Zusammenführung der früheren Projektfortschrittskontrollen (PFK) und der Unabhängigen Evaluierungen (UE) dar und fusionierten die eher internen Beratungs- und Lernfunktion der PFK mit den Funktionen der Transparenz und der Rechenschaftslegung der UE. Damit wurden zugleich die evaluatorischen Qualitätsansprüche an die PEV im Vergleich zu den PFK deutlich erhöht.

Die erste META der PEV aus dem Jahr 2016, die nicht nur die methodische Qualität, sondern auch die Prozessqualität und Nutzung untersuchte, zeigte, dass die PEV zwar die Standards eher erfüllten, insbesondere aber im Hinblick auf die methodische Qualität deutliche Defizite aufwiesen. Diese waren teilweise bedingt in der Multifunktionalität der PEV und ihrer Eingebundenheit in das Auftragsmanagement mit engen prozessualen Vorgaben. Da zudem die Anforderungen an Evaluierungen wiederum gestiegen waren, unter anderem durch die Agenda 2030, eine zunehmende Diversifizierung von Auftragstypen, die gestiegenen Anforderungen an die Unabhängigkeit von Evaluierungen und neue Erkenntnisinteressen der Politik im Hinblick auf Wirkungsnachweise, hat die GIZ im Nachgang zu der META eine umfassende Reform der Projektevaluierungen beschlossen.

Die wesentlichen Neuerungen der Reform waren folgende: Einführung zentraler Projektevaluierungen (ZPE), Trennung von Prüfung und Evaluierung, Einführung von Schlussevaluierungen, Verzicht auf flächendeckende Evaluierungen zugunsten einer Stichprobe, die sukzessive bis zu 50% der Vorhaben beinhaltet. Mit der Reform sollen unterschiedliche Ziele erreicht werden:

- Die methodische und Prozessqualität der Projektevaluierungen verbessern.
- Die Aussagekraft zur Wirkungserzielung der Projekte erhöhen.
- Die Glaubwürdigkeit der Evaluierungsergebnisse verstärken.
- Das Evaluierungssystem stärker an die sich verändernde Auftragsrealität anpassen.
- Die Evaluierungen flexibler an spezifischen Erkenntnisinteressen ausrichten.
- Das Evaluierungssystem auf neue Herausforderungen ausrichten.
- Das Prüfungs- und Evaluierungssystem optimieren und zukunftsfähig machen

Die Reform ist seit Anfang 2017 in der Umsetzung und befindet sich nach einer Aufbau- und Pilotierungsphase, in der die inhaltlich-methodischen und die prozessualen Grundlagen für das neue System geschaffen wurden, seit Mitte 2018 im Regelbetrieb. Eine erste Zwischenbilanz zeigt deutlich, dass die Ziele der Reform weitgehend erreicht werden können, was zukünftige Meta-Evaluierungen genauer zu untersuchen haben. Die Reform stößt auch auf große Anerkennung im BMZ und der Evaluierungscommunity und hat zudem wichtige Impulse gesetzt, wie zum Beispiel in der Anpassung der OECD-DAC- Kriterien im Rahmen der Diskussion über die Evaluierungskriterien der deutschen EZ.

Obwohl die PEV nach der Entscheidung zur Reform im Jahr 2016 ein „Auslaufmodell“ waren hat die GIZ beschlossen, noch die Meta-Evaluierungen 2017 und die jetzt vorliegende Meta-Evaluierung auch für die letzten Jahrgänge der PEV zu machen. Meta-Evaluierungen dienen nicht nur der Qualitätssicherung und Reflektion des Instruments, sondern auch der Identifikation methodisch angemessener Evaluierungen für Evaluierungssynthesen, die die GIZ regelmäßig zu unterschiedlichen Themen macht. Darüber hinaus sind sie Grundlage für die Analyse, ob es Unterschiede in der Bewertungspraxis zwischen methodisch angemessenen und nicht angemessenen Evaluierungen gibt. Hierüber informiert die GIZ in den Evaluierungsberichten des Unternehmens, die alle zwei Jahre veröffentlicht werden. (Link zum GIZ-Evaluierungsbericht 2017 **hier**[1].)

Zudem wurden im Rahmen dieser Meta-Evaluierung zusätzliche statistische Auswertungen vorgenommen, die wichtige Informationen geben, zum Beispiel zur regionalen und sektoralen Differenzierung der Evaluierungen und der Bewertungen. (Link zur statistischen Auswertung **hier**[2].) Auswertungen dieser Art werden auch bei künftigen Metaevaluierungen erfolgen. Auch für die Reform der Projektevaluierung hat die vorliegende Meta-Evaluierung noch wichtige Anregungen.

Auch die ZPE werden künftig regelmäßig durch Meta-Evaluierungen überprüft werden.

[1] www.giz.de/wissenwaswirkt -> Evaluierungsbericht 2017

[2] www.giz.de/wissenwaswirkt -> QUERSCHNITTAUSWERTUNGEN -> PROJEKTEVALUIERUNGEN

2 Evaluierungsbericht

Matías Krämer
Olga Almqvist
Syspons GmbH

2.1 Executive Summary

Die Stabstelle Evaluierung der deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) führt seit 2009 Meta-Evaluierungen von Evaluationsberichten durch. Die Meta-Evaluierungen dienen der Qualitätssicherung und Reflektion des Instruments, und leisten so einen wichtigen Beitrag zu den in der Evaluierungspolicy der GIZ definierten Grundfunktionen von Evaluierung b: i) Unterstützung evidenzbasierter Entscheidungen, ii) Transparenz und Rechenschaftslegung und iii) organisationales Lernen. Die Meta-Evaluierungen dienen zum anderen der Identifikation methodisch angemessener Evaluierungen für Evaluierungssynthesen. Darüber hinaus sind sie Grundlage für die Analyse, ob es Unterschiede in der Bewertungspraxis zwischen methodisch angemessenen und nicht angemessenen Evaluierungen gibt. Letztes ist für die statistische Auswertung für den GIZ-Evaluierungsberichte von Bedeutung.

Die vorliegende Meta-Evaluierung analysiert die methodische Qualität aller zwischen Oktober 2016 und September 2018 fertiggestellten Projektevaluierungen (PEV) und bestimmt Trends in der methodischen Qualität auf Basis der Ergebnisse aus den Meta-Evaluierungen 2015 und 2016. Dabei handelt es sich ausschließlich um dezentrale Projektevaluierungen. Diese mussten von allen Vorhaben im BMZ-Geschäft mit einem Auftragsvolumen von über einer Million Euro und einer Mindestlaufzeit von drei Jahren durchgeführt werden.

Zielsetzung und methodische Grundlagen der Meta-Evaluierung

Das **Ziel der Meta-Evaluierung** ist es, die Qualität der dezentralen PEV zu bewerten. Dabei lag der zentrale Fokus der Meta-Evaluierung 2017 / 2018 auf der methodischen Qualität. Diese wurde auf der Grundlage anerkannter (inter-)nationaler Evaluierungsstandards geprüft, insbesondere auf Grundlage des Evaluierungsstandards Genauigkeit der deutschen Gesellschaft für Evaluation (DeGEval). Die Bewertung wurde anhand von Textanalysen der PEV-Berichte vorgenommen.

Für die Meta-Evaluierung 2017 / 2018 wurde ein Analyseraster verwendet, das in weiten Teilen dem Analyseraster der vergangenen Jahre für den Evaluierungsstandard Genauigkeit entspricht. Gegenüber den vergangenen Jahren wurden lediglich sechs neue Indikatoren hinzugefügt. Diese wurden gesondert ausgewertet, um die Vergleichbarkeit gegenüber den Meta-Evaluierungen der beiden letzten Jahre zu gewährleisten. Die Ergebnisse der Meta-Evaluierung 2017 / 2018 konnten somit den Ergebnissen der beiden letzten Jahre gegenübergestellt werden.

Die Bewertungskriterien zur methodischen Qualität sollen sicherstellen, *„dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt“* (DeGEval 2008). Dafür wurden in der Meta-Evaluierung die DeGEval-Kriterien zur Genauigkeit überprüft.

Die Meta-Evaluierung 2017 / 2018 basiert methodisch auf der **Textanalyse** von 176 Berichten (Vollerhebung), die zwischen Oktober 2016 und September 2018 finalisiert wurden. Dem verwendeten Bewertungssystem zur Prüfung der PEV-Qualität liegt der **Anspruch** einer vollständigen Zielerreichung zugrunde. Dies bedeutet, dass es Anspruch der GIZ ist, alle gängigen (inter-)nationalen Qualitätsstandards für Evaluierungen vollständig zu erfüllen. Um darauf aufbauend die Bewertungen einheitlich und nachvollziehbar zu gestalten, wird die **Erfüllung von Standards** wie folgt angegeben: Erfüllung zu 90–100 % (größtenteils bis vollständig erfüllt); Erfüllung zu 80–89 % (zumeist erfüllt); Erfüllung zu 60–79 % (eher erfüllt); Erfüllung geringer als 60 % (bedingt erfüllt).

Zusammenfassung der wesentlichen Ergebnisse, Bewertungen und Schlussfolgerungen

Mit einer Erfüllung von 66 % im Durchschnitt im Evaluierungsstandard methodische Genauigkeit **kommen die PEV der Jahre 2017 / 2018 den (inter-)nationalen Qualitätsstandards eher nach**. Dieses Ergebnis lässt sich ausdifferenzieren in einen Durchschnittswert von 66 % für die 125 PEV-Berichte die im Jahr 2017 finalisiert wurden, und 67 % für die 51 Berichte, die im Jahr 2018 finalisiert wurden. Damit ergibt sich gegenüber den Meta-Evaluierungen 2015 (59 %, n=70) und 2016 (62 %, n=100) ein positiver Trend in der methodischen Qualität.

Die Meta-Evaluierung zeigt, dass die Stärken und

Schwächen der PEV zwischen Oktober 2016 und September 2018 ähnlich ausfallen wie in den Vorjahren. Hierbei liegen die **Stärken** wie in der Vergangenheit auch für 2017 / 2018 vor allem in der Erfüllung formaler und deskriptiver Anforderungen. Hinsichtlich der methodischen Anforderungen lassen sich zwar Verbesserungen verzeichnen, es bestehen jedoch weiterhin auch Herausforderungen.

Die formalen und deskriptiven Anforderungen, bei denen die PEV vergleichsweise gut abschneiden, sind auch in dieser Meta-Evaluierung die graphische oder textliche Darstellung des Wirkungsmodells sowie die Darstellung des Kontextes, in den sich die untersuchten Vorhaben eingliedern. Darüber hinaus geht mit der ausführlichen Beschreibung des Kontextes auch eine vergleichsweise umfassende methodische Bearbeitung des OECD-Kriteriums Relevanz in den PEV-Berichten einher. Eine weitere Stärke liegt in der allgemeinen Beschreibung der angewandten Methoden. Weiterhin sind auch die Datentriangulation sowie eine Triangulation von Ergebnissen mit Beteiligten und Betroffenen weitgehend gegeben.

Die methodischen **Schwächen** der dezentralen PEV sind in den gleichen Bereichen zu verorten wie in den vergangenen Meta-Evaluierungen. Diesbezüglich ist anzumerken, dass es im Untersuchungszeitraum keine Veränderungen hinsichtlich der Vorgaben an die PEV-Teams gegeben hat. Wie in den vergangenen Jahren ist demnach zu berücksichtigen, dass nicht alle in der Meta-Evaluierung untersuchten Aspekte Bestandteil der Vorgaben für die Durchführung von PEV waren. Mit der Umstellung des Evaluierungssystems der GIZ auf zentrale Projektevaluierungen lag der Schwerpunkt der Stabsstelle Evaluierung auf der Einführung höherer methodischer Standards für das neue Instrument.

Eine zentrale Schwäche der PEV ist aus Sicht des Meta-Evaluierungsteams auch in 2017 / 2018 eine unzureichende Reflektion der methodischen Vorgehensweise. So hat es sich noch nicht durchgesetzt, in den Berichten darzustellen, warum bzw. zu welchem Sachverhalt welche Quellen ausgewertet wurden, und zu erläutern, unter welchen Gesichtspunkten die Auswahl von Gesprächspartnern/innen erfolgte. Auch werden die Stärken und Schwächen der gewählten Methodik nur wenig thematisiert. Darüber hinaus fällt auf, dass die Auseinandersetzung mit der Qualität der zugrunde gelegten Daten ausbaufähig bleibt. So gehen die Berichte zwar in der Regel kurz auf die allgemeine Qualität des wirkungsorientierten Monitorings ein, sie setzen sich jedoch nur bedingt explizit mit der Belastbarkeit von Baseline-Daten auseinander.

Mit Blick auf die methodische Bearbeitung der OECD-DAC-Kriterien lässt sich folgendes feststellen:

- Die Bearbeitung des **Effizienz**-Kriteriums weist nach wie vor die größten Herausforderungen auf. Das entsprechende Kapitel beschränkt sich zumeist auf die Analyse der Implementierungseffizienz. Diesbezüglich ist anzumerken, dass vor der gemeinsamen Verfahrensreform von BMZ und GIZ geplante Vorhaben in der Regel keine Zuordnung von Kosten zu Outputs oder Outcomes vorgenommen haben, was die Analyse von Produktions- und Allokationseffizienz erschwert. Allerdings wird in den untersuchten PEV die Auswahl von Verfahren und Methoden der Effizienzmessung nur in Ausnahmefällen begründet. Insgesamt ist das Effizienz-Kapitel darüber hinaus in der Regel sehr kurz gehalten und die Analysen fallen dementsprechend knapp aus.

- Für die Bearbeitung der **Effektivität** werden zwar fast flächendeckend Modulzielindikatoren herangezogen, diese entsprechen jedoch nicht durchgehend den SMART-Qualitätskriterien. Zwar hat sich dieser Aspekt gegenüber den Vorjahren gebessert, aber angesichts des zentralen Stellenwerts dieser Anforderung besteht hier noch Optimierungspotenzial. In der Regel werden Mängel hinsichtlich der Erfüllung der SMART-Kriterien von den PEV-Prüfteams korrekt identifiziert, aber es werden dann keine angepassten oder neuen Indikatoren formuliert. Dies zeigte sich insbesondere bei Sektor- oder Globalvorhaben, für die in mehreren PEV-Berichten darauf verwiesen wurde, dass die Nutzung von erarbeiten Strategien innerhalb der Laufzeit des Vorhabens nicht messbar sei.

- Die Qualität **kontributionsanalytischer Aspekte** kommt sowohl bei der Bearbeitung der Effektivität als auch bei der Bearbeitung

des Impacts zum Tragen. Hierbei lässt sich feststellen, dass eine Auseinandersetzung mit der Kausalität zwischen den Maßnahmen des Vorhabens und den beobachteten Veränderungen eher für die Impact- als für die Outcome-Ebene erfolgt. Allerdings erfolgt diese Auseinandersetzung auch für die übergeordneten Wirkungen oft nur implizit durch Formulierungen wie „das Vorhaben leistet einen Beitrag“. Eine differenzierte Darlegung anderer Einflussfaktoren auf die beschriebenen Veränderungen erfolgt sowohl im Effektivitäts- als auch im Impact-Kapitel nur selten.

- Erstmalig untersucht wurde in diesem Jahr die **Nachvollziehbarkeit** der Bewertung für jedes der OECD-DAC-Kriterien. Auch wenn die Nachvollziehbarkeit größtenteils gegeben ist, zeigt sich, dass in bis zu einem von zehn Berichten entweder Punkte vergeben werden für Aspekte, auf die im Text nicht eingegangen wird, oder die Bewertung nach Einschätzung des Meta-Evaluierungsteams angesichts im Berichtstext beschriebener Schwächen des evaluierten Vorhabens zu positiv ausfällt.

Übergeordnet zeigt sich, dass die **Benotung der Vorhaben** durch die PEV-Teams ausgesprochen positiv ausfällt. So erhalten 85 % der Vorhaben die Gesamtnote „sehr erfolgreich“ oder „erfolgreich“. Besonders ausgeprägt ist die positive Benotung hinsichtlich des Relevanz-Kriteriums, für das 99 % aller Vorhaben als „sehr erfolgreich“ bzw. „erfolgreich“ eingestuft werden. Dies wirft die Frage auf, inwiefern die Evaluator*innen sich angemessen kritisch mit den Vorhaben auseinandersetzen. Dieser Qualitätsaspekt ist im Hinblick auf das Potential von Evaluationen für Lernen und Entscheidungsfindung von Bedeutung. Diesbezüglich stellt die Umstellung innerhalb der GIZ von dezentral gesteuerten PEV auf zentral gesteuerte Projektevaluierungen (ZPE) eine Chance dar. Mit der Umstellung werden Projektevaluierungen von der Stabstelle für Evaluierungen beauftragt und nicht mehr durch das Vorhaben selbst, wodurch die evaluatorische Unabhängigkeit gestärkt wird. Gleichzeitig werden die Projektevaluierungen nunmehr entkoppelt von der Planung von Folgevorhaben umgesetzt, wodurch die Fokussierung der Gutachterteams auf ihre evaluatorische Rolle gewährleistet wird. Darüber hinaus wurden für die zentral gesteuerten Projektevaluierungen methodische Vorgaben eingeführt, die weit über die Vorgaben für die PEV hinausgehen.

Hinsichtlich der **Einflussfaktoren** auf die methodische Qualität der PEV zeigt die statistische Auswertung mehrere zentrale Stellschrauben auf, die sich bereits in den vorhergehenden Meta-Evaluierungen als signifikante Zusammenhänge erwiesen hatten:

1) Ein signifikanter positiver Zusammenhang besteht zwischen dem Umfang der Methoden und der methodischen Qualität (Erfüllung des Genauigkeitsstandards). Besonders auffallend ist hierbei der Qualitätssprung zwischen PEV, die bis zu zwei Methoden einsetzen, und PEV, in denen drei oder mehr Methoden angewandt werden. Wie in den vergangenen Jahren ist allerdings der Methodenmix recht einseitig. Die PEV-Prüfteams arbeiten vornehmlich mit Interviews und Dokumentenauswertungen. Darüber hinaus erfolgt zum Teil eine eigene Auswertung von Monitoringdaten durch die PEV-Teams, und in einigen Fällen wird mit Fokusgruppen gearbeitet.

2) Die zweite Stellschraube für die methodische Genauigkeit ist das Mengengerüst für die Gutachter. Hier weisen die Daten darauf hin, dass die methodische Qualität mit steigender Anzahl an Gutachtertagen im Mittel zunächst ansteigt, der Zusammenhang dann jedoch abflacht. Es lässt sich vor allem ein positiver Zusammenhang zwischen der Anzahl der Gutachtertage für die Vorbereitung und Durchführung der PEV und deren methodischer Qualität ausmachen. Allerdings ist hierbei zu berücksichtigen, dass in den PEV-Berichten in der Regel nicht differenziert wird zwischen Anzahl der Gutachtertage für die Evaluierung und ggf. Anzahl der Tage für die Prüfung eines Folgevorhabens.

3) Weiterhin lässt sich ein positiver Zusammenhang ausmachen zwischen einer klaren Darstellung des Evaluierungsgegenstands sowie der Auseinandersetzung mit dessen Wirkungslogik und der methodischen Quali-

tät der Evaluierung. Darüber hinaus zeichnen sich PEV, die eine klare Trennung zwischen dem Dreischnitt Beschreibung, Analyse und Bewertung vornehmen, durch eine vergleichsweise höhere Genauigkeit aus. Dieser Sachverhalt entspricht dem allgemeinen Evaluierungsverständnis, dass intendierte Ziele und Wege zur Zielerreichung nachvollziehbar darstellt sein müssen, um eine differenzierte Auseinandersetzung mit den Ergebnissen zu gewährleisten.

Mit Blick auf die **Trends in der Qualität der dezentralen PEV** über vier Jahre hinweg lässt sich eine Verbesserung der methodischen Qualität für die meisten untersuchten Bewertungskriterien ausmachen. Von den 13 Bewertungskriterien, die in allen Meta-Evaluierungen erfasst wurden, ist dabei für sieben Kriterien zwischen 2015 und 2018 eine durchschnittliche Steigerung von zehn Prozentpunkten und mehr auszumachen.
Die stärkste Verbesserung ist für das Bewertungskriterium zu validen und reliablen Informationen auszumachen, sowie für das Bewertungskriterium zur Beschreibung des Evaluierungsgestands. Auf Indikatorenebene sind diese Entwicklungen vor allem auf Verbesserungen hinsichtlich der Datentriangulation und der Darstellung relevanter Wirkungshypothesen zurückzuführen. Eine rückläufige Tendenz zeigt sich hingegen für die Durchschnittswerte hinsichtlich des Bewertungskriteriums zur systematischen Fehlerprüfung. Nur minimale Veränderungen gab es hinsichtlich der Kriterien zu Beschreibung von Zweck und Vorgehen und zur Bearbeitung der Nachhaltigkeit.

In der Gesamtschau zeigt sich jährlich eine leichte, aber **über die Jahre hinweg stetig positive Entwicklung der Qualität der PEV**, ohne dass sich die formalen Vorgaben an die dezentralen Evaluierungen im Untersuchungszeitraum verändert hätten. Aus Sicht des Meta-Evaluierungsteams sind mögliche Erklärungsfaktoren hierfür die Rückmeldungen der Stabsstelle Evaluierung an die PEV-Prüfteams z.B. im Rahmen von Qualitätschecks[3], aber auch eine allgemeine Sensibilisierung innerhalb der GIZ für Qualitätsaspekte wie bspw. die Auseinandersetzung mit dem Wirkungsmodell oder die Berücksichtigung von SMART-Qualitätskriterien für Indikatoren. Trotz der insgesamt positiven Entwicklungen wurden jedoch auch in den Jahren 2017 und 2018 zum Teil grundlegende Evaluierungsstandards in den PEV nicht eingehalten. Es ist daher zu begrüßen, dass mit der Umstellung von dezentralen auf zentralen Evaluierungen im GIZ Evaluierungssystem höhere methodische Anforderungen in den Vorgaben verankert und in der Abnahme der Berichte nachgehalten werden.

[3] In der Meta-Evaluierung 2017 / 2018 ließen sich keine signifikanten Qualitätsunterschiede feststellen zwischen PEV, für die ein PEV-Check durchgeführt wurde, und solche, für die kein PEV-Check durchgeführt wurde. Dies schließt jedoch nicht aus, dass Evaluatoren, die in der Vergangenheit Rückmeldungen im Rahmen von PEV-Checks bekommen haben und später an der Umsetzung weiterer PEV beteiligt waren, Feedback von der Stabsstelle berücksichtigt haben.

2.2 Einleitung

Die Stabstelle Evaluierung der deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) führt seit 2009 Meta-Evaluierungen von Evaluationsberichten durch. Die Meta-Evaluierungen dienen der Qualitätssicherung und Reflektion des Instruments, und leisten so einen wichtigen Beitrag zu den in der Evaluierungspolicy der GIZ definierten Grundfunktionen von Evaluierung b: i) Unterstützung evidenzbasierter Entscheidungen, ii) Transparenz und Rechenschaftslegung und iii) organisationales Lernen.. Die Meta-Evaluierungen dienen zum anderen der Identifikation methodisch angemessener Evaluierungen für Evaluierungssynthesen. Darüber hinaus sind sie Grundlage für die Analyse, ob es Unterschiede in der Bewertungspraxis zwischen methodisch angemessenen und nicht angemessenen Evaluierungen gibt. Letztes ist für die statistische Auswertung für den GIZ-Evaluierungsberichte von Bedeutung.

Die vorliegende Meta-Evaluierung analysiert die methodische Qualität aller zwischen Oktober 2016 und September 2018 fertiggestellten Projektevaluierungen und bestimmt Trends in der methodischen Qualität auf Basis der Ergebnisse aus den Meta-Evaluierungen 2015 und 2016. Dabei handelt es sich ausschließlich um dezentrale Projektevaluierungen (PEV). Diese mussten seit April 2014 von allen Vorhaben im BMZ-Geschäft mit einem Auftragsvolumen von über einer Million Euro und einer Mindestlaufzeit von drei Jahren durchgeführt werden. Mitte 2017 begann die Umstellung des Evaluierungssystems der GIZ. In dem neuen System werden Projektevaluierungen zentral von der GIZ Stabsstelle Evaluierung gesteuert. Zur Gewährleistung eines reibungslosen Übergangs vom dezentralen System hin zum System der zentral gesteuerten Projektevaluierungen wurden in der „Übergangsphase“ sowohl dezentrale als auch zentral gesteuerte Projektevaluierungen durchgeführt (GIZ 2018). Die ersten zentralen Projektevaluierungen wurden nach dem Erhebungszeitraum für die vorliegende Meta-Evaluierung fertiggestellt.

Zielsetzung und methodische Grundlagen der Meta-Evaluierung
Das **Ziel der Meta-Evaluierung** ist es, die Qualität der dezentralen PEV zu bewerten. Dabei lag der zentrale Fokus der Meta-Evaluierung 2017 / 2018 auf der methodischen Qualität. Diese wurde auf der Grundlage anerkannter (inter-)nationaler Evaluierungsstandards geprüft, insbesondere auf Grundlage des Evaluierungsstandards Genauigkeit der deutschen Gesellschaft für Evaluation (DeGEval). Die Bewertung wurde anhand von Textanalysen der PEV-Berichte vorgenommen.

Für die Meta-Evaluierung 2017 / 2018 wurde ein Analyseraster verwendet, das in weiten Teilen dem Analyseraster der vergangenen Jahre für den Evaluierungsstandard Genauigkeit entspricht. Gegenüber den vergangenen Jahren wurden lediglich sechs neue Indikatoren hinzugefügt. Diese wurden gesondert ausgewertet, um die Vergleichbarkeit gegenüber den Meta-Evaluierungen der beiden letzten Jahre zu gewährleisten. Die Ergebnisse der Meta-Evaluierung 2017 / 2018 konnten somit den Ergebnissen der beiden letzten Jahre gegenübergestellt werden.

Die Bewertungskriterien zur methodischen Qualität sollen sicherstellen*, „dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt“* (DeGEval 2008). Dafür wurden in der Meta-Evaluierung die DeGEval-Kriterien zur Genauigkeit überprüft.

Die Meta-Evaluierung 2017 / 2018 basiert methodisch auf der **Textanalyse** von 176 Berichten (Vollerhebung), die zwischen Oktober 2016 und September 2018 finalisiert wurden. Dem verwendeten Bewertungssystem zur Prüfung der PEV-Qualität liegt der **Anspruch** einer vollständigen Zielerreichung zugrunde. Dies bedeutet, dass es Anspruch der GIZ ist, alle gängigen (inter-)nationalen Qualitätsstandards für Evaluierungen vollständig zu erfüllen. Um darauf aufbauend die Bewertungen einheitlich und nachvollziehbar zu gestalten, wird die **Erfüllung von Standards** wie folgt angegeben: Erfüllung zu 90–100 % (größtenteils bis vollständig erfüllt); Erfüllung zu 80–89 % (zumeist erfüllt); Erfüllung zu 60–79 % (eher erfüllt); Erfüllung geringer als 60 % (bedingt erfüllt).

2.3 Methodik

Für die Meta-Evaluierung 2015[4] wurde das Analyseraster ausgearbeitet, welches 2016 sowie in diesem Jahr punktuell weiterentwickelt wurde. Bei der (Weiter-)Entwicklung des Analyserasters wurde darauf geachtet, dass die **folgenden Anforderungen** an die Bewertungsinstrumente von Meta-Evaluierungen erfüllt werden, die aus der einschlägigen Literatur abgeleitet wurden:

Anforderungen nach Widmer (1996, S. 9/10)	Umsetzung der Anforderungen in der Meta-Evaluierung
Methodische Offenheit: Offenheit gegenüber Evaluierungen unterschiedlicher methodologischer Ausrichtungen und Evaluierungsansätze	Die Bewertungskriterien und Indikatoren sind bewusst so definiert, dass sie sich nicht auf spezifische Theorien, methodische Vorgehensweisen oder Evaluierungsmethoden stützen. Vielmehr befassen sie sich damit, inwieweit die PEV ihre methodischen Ansätze begründen und sich daran ausrichten. Die gewählten Vorgehensweisen und Evaluierungsmethoden werden dann deskriptiv durch Checklisten abgefragt.
Umfassende Bewertung: Bewertung der gesamten Evaluierung und nicht lediglich von Teilaspekten einer Evaluierung	Das erweiterte Qualitätsverständnis der Evaluierung umfasst die Evaluierungsstandards Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit und bewertet somit die Qualität der gesamten Evaluierung. Die Meta-Evaluierungen 2015 und 2016 haben sowohl die methodische Qualität als auch die Prozessqualität und die Nützlichkeit untersucht. Die Meta-Evaluierung 2017 / 2018 beschränkt sich vor dem Hintergrund des Auslaufens des Instruments der dezentralen PEV aus Kosten-Nutzen-Erwägungen auf den Aspekt der Genauigkeit.
Verbreitung, Anerkennung und Akzeptanz: Hohe Akzeptanz im Evaluierungsfeld, Anerkennung durch relevante Fachorganisationen und einschlägige Literatur	Das Analyseraster wurde 2015 auf Grundlage nationaler und internationaler Evaluierungsstandards entwickelt. Hierbei wurde abgeglichen, welche Evaluierungsstandards die GIZ in ihren internen Qualitätsinstrumenten (noch nicht) berücksichtigt. International sind verschiedene Kataloge von Evaluierungsstandards entwickelt worden, von denen insbesondere die Standards des Joint Committee on Educational Evaluation (JCSEE) (2000; 2006) eine breite Beachtung fanden. Fachorganisationen wie die DeGEval haben sich den Standards des JCSEE angeschlossen, um einen internationalen Erfahrungsaustausch zu vereinfachen (z. B. DeGEval 2008, S. 20; 44). Auch die DeGEval hat ihre Standards zuletzt weiterentwickelt. Die Weiterentwicklung umfasst jedoch keine grundlegenden Anpassungen, sondern ist auf einzelne inhaltliche und sprachliche Konkretisierungen fokussiert. Schließlich nimmt das Analyseraster Bezug auf die Qualitätsstandards der OECD-DAC (1991; 2010), um ggf. Besonderheiten von Evaluierungen in der Entwicklungszusammenarbeit (EZ) zu berücksichtigen.

Abbildung 1: Anforderungen an die Bewertungsinstrumente von Meta-Evaluierungen (Quelle: Syspons 2018)

Unter Beachtung dieser Anforderungen ist das Bewertungssystem der Meta-Evaluierung methodisch entlang der folgenden drei Ebenen strukturiert (siehe auch Abbildung 2):

- Auf der ersten Ebene steht der **DeGEval-Evaluierungsstandard** Genauigkeit (methodische Qualität).[5]

[4] Freimann, I., Krämer, M. (2016): Querschnittsauswertung (QSA) von Projektevaluierungen (PEV) 2015 – Meta-Evaluierung. Bonn und Eschborn. Herausgeber: Deutsche Gesellschaft für Internationale Zusammenarbeit. Online unter: https://www.giz.de/de/downloads/giz2016-de-Finaler_Bericht_Metaevaluierung_der_PEV.pdf . (zuletzt abgerufen: 03.11.2017)

[5] In den vorherigen Meta-Evaluierungen wurden zudem die Evaluierungsstandards Nützlichkeit, Durchführbarkeit und Fairness bewertet.

- Die **Bewertungskriterien**[6], welche ebenfalls weitgehend von der DeGEval übernommen wurden, bilden die zweite Ebene des Analyserasters. Dem Evaluierungsstandard Genauigkeit werden 13 Bewertungskriterien zugeordnet.
- Auf der dritten Ebene werden jedem Bewertungskriterium **Indikatoren** zugeordnet, die beschreiben, welche Aspekte gegeben sein müssen, damit ein Kriterium als erfüllt gilt. Die Anzahl der einem Bewertungskriterium zugeordneten Indikatoren schwankt zwischen einem und sechs Indikatoren.

Die nachfolgende Abbildung gibt einen graphischen Überblick über die 13 Bewertungskriterien zum DeGEval-Standard Genauigkeit, die in den vergangenen Jahren und in dieser Meta-Evaluierung identisch erhoben wurden und zu deren Entwicklung sich daher Trends darstellen lassen:

Abbildung 2: Schematische Darstellung des Analyserasters

(Quelle: Syspons 2018)

Zusätzlich zu den bereits in den beiden vergangenen Meta-Evaluierungen erhobenen Indikatoren wurden dieses Jahr in Abstimmung mit der GIZ sechs neue Indikatoren erhoben. Diese erfassen kontributionsanalytische Aspekte und die Nachvollziehbarkeit der Bewertung. Um die Vergleichbarkeit der Ergebnisse gegenüber den Meta-Evaluierungen 2015 und 2016 zu gewährleisten, erfolgte eine gesonderte Auswertung zu kontributionsanalytischen Aspekten (hierbei wurde ein Index gebildet aus alten und einem neuen Indikator) und der Nachvollziehbarkeit der Bewertung (Index aus fünf neuen Indikatoren).

[6] Die DeGEval bezeichnet diese als „Standards“. Die DeGEval definiert neun Genauigkeitsstandards (wobei in der Meta-Evaluierung der siebte Genauigkeitsstandard weiter ausdifferenziert und der neunte Genauigkeitsstandard „Meta-Evaluation“ nicht erhoben wurde). Die DeGEval definiert drei Durchführbarkeits- und fünf Fairnessstandards (wie in der Meta-Evaluierung) sowie acht Nützlichkeitsstandards (in dieser Meta-Evaluierung wurde der achte Standard in Nützlichkeit und Nutzung ausdifferenziert).

Kurzdarstellung der methodischen Vorgehensweise

Die methodische Vorgehensweise gliederte sich in folgende Schritte:

In einem **Auftragsklärungsgespräch** mit der Stabsstelle Evaluierung wurden die Zielsetzungen der diesjährigen Meta-Evaluierung festgelegt und Fragen zur Anpassung des Analyserasters diskutiert.

Das in 2015 entwickelte **Analyseraster** wurde **an wenigen Stellen erweitert**. Dabei wurden die Ergebnisse der letzten Meta-Evaluierung, in 2016, als Ausgangspunkt für Anpassungen genutzt, um die Relevanz des Systems weiter zu erhöhen und gleichzeitig die Vergleichbarkeit zu den Meta-Evaluierungen von 2015 und 2016 sicherzustellen. Das Analyseraster wurde um sechs neue Indikatoren[7] ergänzt, diese werden jedoch gesondert ausgewertet, um die Vergleichbarkeit zu den Vorjahren zu gewährleisten (für eine vollständige Darstellung aller Indikatoren siehe Analyseraster in Anlage 2).

Das **Analyseraster** wurde 2015 ursprünglich aufbauend auf **acht explorativen Interviews** mit aktuellen und ehemaligen Mitarbeiter/innen der Stabsstelle Evaluierung sowie einem mehrstufigen Desk Research entwickelt. Die damaligen Interviewpartner/innen wurden ausgewählt, wenn sie (1) in die Entwicklung der PEV-Unterstützungsmaterialien involviert waren und somit den Einführungsprozess der PEV begleitet hatten, (2) Erfahrungen mit der Durchführung von Qualitätschecks der PEV-Berichte hatten und/ oder (3) bereits Meta-Evaluierungen gesteuert hatten. Zudem wurde 2015 ein mehrstufiger **Desk Research** durchgeführt, in welchem anhand der wissenschaftlichen Literatur Anforderungen an Meta-Evaluierungen ausgearbeitet wurden. Darauf aufbauend wurde entlang der Evaluierungsstandards der DeGEval durch eine Textanalyse abgeglichen, welche Bewertungsmaßstäbe sich in anderen internationalen Standards und in den GIZ-internen Qualitätssicherungsinstrumenten wiederfinden. Das Analyseraster (siehe Anlage 2) dokumentiert die Ergebnisse dieses Vergleichs.

Die Durchführung der Meta-Evaluierung besteht in diesem Jahr ausschließlich aus einer **Textanalyse von PEV-Berichten**. Die AV-Befragung, die in den zwei vorangegangenen Meta-Evaluierungen durchgeführt wurde, fand in diesem Jahr nicht statt. Grund hierfür ist erstens, dass die PEV mit der Einführung der ZPE auslaufen, sodass eine erneute Befassung der AV mit diesem Instrument nicht prioritär erschien. Zweitens hat sich in den vorhergehenden Meta-Evaluierungen gezeigt, dass insbesondere im Bereich der methodischen Genauigkeit Optimierungsbedarf besteht. Da die methodische Genauigkeit der Evaluationen über die Text-Analyse erfasst wird, beschränkt sich die diesjährige Meta-Evaluierung auf diese Analyse.

Die Grundgesamtheit der Textanalyse der PEV-Berichte umfasst eine Vollerhebung aller 176 zwischen 01. Oktober 2016 und 30. September 2018 finalisierten PEV-Berichte. Die Auswertung der PEV-Berichte liefert somit ein umfassendes Bild im Betrachtungszeitraum hinsichtlich der zugrundeliegenden Bewertungskriterien. Für jeden Bericht wurde ein Auswertungsblatt angelegt, welches die Rahmendaten aufführt und die qualitative und quantitative Bewertung der Indikatoren beinhaltet.[8] Um die Bewertung jedes Indikators transparent und nachvollziehbar zu machen, wurden die Grundlagen der Beurteilung erläutert. Negativ bewertete Indikatoren wurden textlich begründet. Bei einer positiven Bewertung wurde auf die entsprechende Seite im Bericht verwiesen.

[7] Z 7 (B) Die Bewertung des OECD/DAC Kriteriums "Relevanz" ist nachvollziehbar; Z 7 (C) Die Bewertung des OECD/DAC Kriteriums "Effektivität" ist nachvollziehbar; Z 7 (D) Die Bewertung des OECD/DAC Kriteriums "Effizienz" ist nachvollziehbar; Z 7 (E) Die Bewertung des OECD/DAC Kriteriums "Impact" ist nachvollziehbar; Z 7 (F) Die Bewertung des OECD/DAC Kriteriums "Nachhaltigkeit" ist nachvollziehbar; Z 7 (F) Die Bewertung des OECD/DAC Kriteriums "Nachhaltigkeit" ist nachvollziehbar

[8] Auswertungsblätter wurden zudem auch für fünf Berichte erstellt, die vor Oktober 2016 finalisiert wurden. Diese Berichte wurden jedoch nicht in die Analyse für die Meta-Evaluierung einbezogen.

Die Auswertung der PEV-Berichte basierte auf dem folgenden Bewertungssystem:

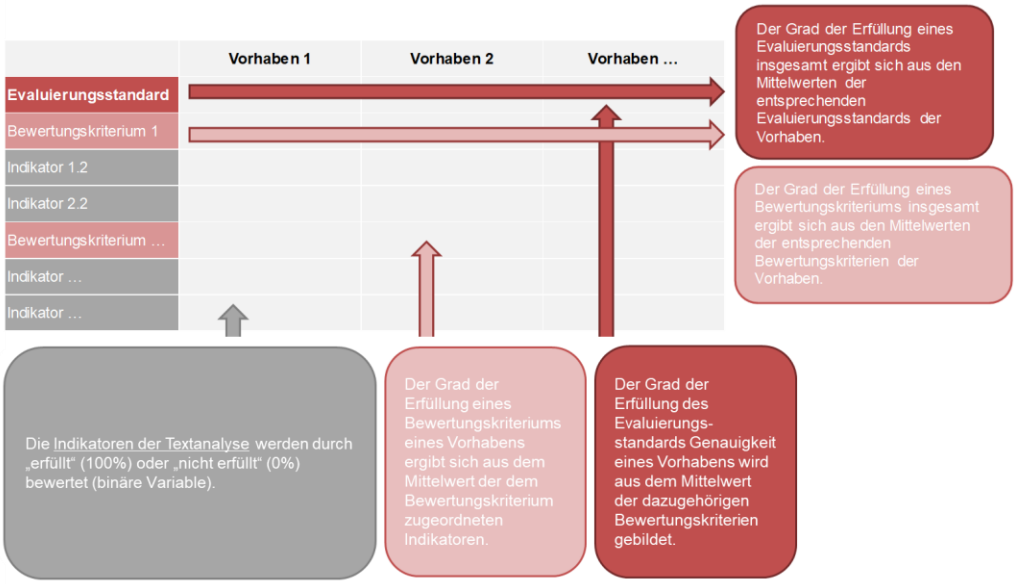


Abbildung 3: Bewertungssystem (Quelle: Syspons 2018)

Dem Bewertungssystem liegt der Anspruch einer vollständigen Zielerreichung zugrunde. Um darauf aufbauend die Bewertungen auf der **Ebene des Evaluierungsstandards Genauigkeit und der Bewertungskriterien** einheitlich und nachvollziehbar zu gestalten, werden diese in diesem Bericht, angelehnt an die Skala der jährlichen Externen Qualitätskontrolle der GIZ[9], wie folgt angegeben:

- 90–100 % (größtenteils bis vollständig erfüllt)
- 80–89 % (zumeist erfüllt)
- 60–79 % (eher erfüllt)
- < 60 % (bedingt erfüllt).

Neben der Auswertung der PEV-Berichte anhand des Analyserasters wurden in einem zweiten Schritt mögliche allgemeine Zusammenhänge zwischen einzelnen Qualitätsaspekten mithilfe von bi- und multivariaten Auswertungsmethoden untersucht. Das Ziel war es hierbei, mögliche Einflussfaktoren auf die methodische Qualität zu identifizieren. Dabei wurden auch spezifischere Zusammenhänge auf den dahinterliegenden Ebenen der Bewertungskriterien, Indikatoren und Befragungsitems untersucht. Hierbei wurden deduktiv zunächst **Hypothesen zu allen theoretisch erwarteten Zusammenhängen** formuliert und in Indizes gruppiert, die verschiedene Befragungsitems und Indikatoren zusammenfassen. Die verwendeten Hypothesen sind das Ergebnis der explorativen Interviews von 2015 sowie des Austauschs zwischen dem Meta-Evaluierungsteam und der Stabsstelle Evaluierung in 2015 und 2016, sowie in diesem Jahr.

Die theoretisch erwarteten Zusammenhänge wurden **bi- und multivariat analysiert**. Hierzu wurde auf Kreuztabellen, Korrelationen und multivariate Regressionen zurückgegriffen. Mit Hilfe von statistischen Tests wurden die untersuchten Zusammenhänge zudem auf ihre Signifikanz untersucht. Die identifizierten Einflussfaktoren oder „Stellschrauben" wurden anschließend näher untersucht, um weitere Wirkungszusammenhänge zu identifizieren. In diesem Bericht werden nur ausgewählte – das heißt in der Regel signifikante – Zusammenhänge erläutert und analysiert, die einen systematischen Einfluss auf die PEV-Qualität haben können. Dabei wurden Zusammenhänge dann als signifikant bewertet, wenn die entsprechenden Tests einen p-Wert kleiner als 0,05 aufweisen. Zur Einschätzung der Größe der identifizierten Zusammenhänge wurden neben Korrelationen nach Pearson auch Effektstärken nach Cohen berechnet. Die verschiedenen Maßzahlen der Größe der Zusammenhänge wurden durchgehend in die gängigen Kategorien nach Cohen[10] eingeteilt und in den Fußnoten des Berichts angegeben.

[9] Stern, T., Scheller, O., Freimann, I. (2015). Externe Qualitätskontrolle der GIZ. Ergebnisbericht 2015.

[10] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

Es liegen der diesjährigen Meta-Evaluierung alle PEV-Berichte des Zeitraums Oktober 2016 bis September 2018 zugrunde. Aus Sicht des Meta-Evaluierungsteams ist die Anwendung von Signifikanztests in der bi- und multivariaten Analyse jedoch trotz „Vollerhebung“ sinnvoll, da auch bei Vollerhebungen die Daten stochastischen Prozessen unterliegen können.[11]

Um den Einfluss persönlicher Bewertungstendenzen zu verringern, haben im Rahmen einer **Forschertriangulation** insgesamt drei Personen die Bewertungen der PEV-Berichte vorgenommen, sodass der Einfluss persönlicher Bewertungstendenzen verringert werden konnte. Zugleich wurde, um eine vergleichbare Bewertung sicherzustellen, eine zufällige Stichprobe von Berichten (8 % aller Berichte) durch zwei Evaluator/innen doppelt geprüft. Hierdurch wurde eine Kalibrierung vorgenommen. Die Ergebnisse der Doppelprüfungen (z. B. abweichende Interpretationen von Indikatoren) wurden anschließend im Team diskutiert und reflektiert. Somit wurde durch den individuellen Austausch über die Bewertungsmaßstäbe die Qualität der Auswertungen erhöht. Zudem wurden der Mittelwert (M) und die Standardabweichung (SD), also die durchschnittliche Abweichung der Eingaben einzelner Evaluator/innen vom Mittelwert eines Bewertungskriteriums, ausgewertet, um die Verlässlichkeit der Ergebnisse zu reflektieren.

Während eine **Forschertriangulation** sichergestellt wurde, war eine **Methodentriangulation** nicht möglich. Die Bewertung eines Indikators unseres Bewertungssystems bezieht sich jeweils nur auf die Datenerhebungsmethode der Textanalyse. Um dieser Limitierung des methodischen Vorgehens zu begegnen, reflektiert das Meta-Evaluierungsteam die Aussagekraft der Ergebnisse in der Analyse und der Bewertung der Evaluierungsstandards entsprechend.

2.4 Evaluierungsgegenstand und Datengrundlage

Evaluierungsgegenstand

Seit April 2014 mussten PEV in allen Vorhaben im Auftrag des Bundesministeriums für Wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) durchgeführt werden, die ein Mindestvolumen von einer Millionen Euro sowie eine Mindestlaufzeit von drei Jahren aufwiesen. Mitte 2017 begann die Umstellung des Evaluierungssystems der GIZ. In dem neuen System werden Projektevaluierungen zentral von der GIZ Stabsstelle Evaluierung gesteuert. Zur Gewährleistung eines reibungslosen Übergangs vom dezentralen System hin zum System der zentral gesteuerten Projektevaluierungen wurden in der „Übergangsphase“ sowohl dezentrale als auch zentral gesteuerte Projektevaluierungen durchgeführt (GIZ 2018). Die ersten zentralen Projektevaluierungen wurden im letzten Quartal 2018 fertiggestellt. Die vorliegende Meta-Evaluierung fokussiert aufgrund dessen ausschließlich auf die methodische Qualität der zwischen Oktober 2016 und September 2018 fertiggestellten dezentralen PEV.

PEV wurden ein Jahr bis sechs Monate vor Ende der Vorhabenslaufzeit durchgeführt. Waren sie als Abschlussevaluierung ohne Folgemaßnahme konzipiert, konnten sie auch später erfolgen. Falls eine Folgemaßnahme geplant war, wurden PEV in den Prüfprozess von Neuvorhaben integriert (GIZ 2015a).

PEV sollten einem iterativen Vorgehen entlang von **acht Schritten** folgen. (1) In der Vorbereitungsphase wurde der Evaluierungsgegenstand erfasst, wozu der AV vorbereitende Unterlagen bereitstellt. (2) Abgeleitet aus dem Evaluierungsgegenstand und den OECD-DAC-Kriterien wurde das Evaluierungsdesign entwickelt und (3) der Ablauf vor Ort geplant. (4) Die Durchführung der PEV umfasste einen Auftaktworkshop, (5) die Datenerhebung und -auswertung sowie (6) ein GIZ-internes Debriefing und (7) einen Abschlussworkshop mit den Partnern, in welchem die Evaluierungsergebnisse vorgestellt wurden. (8) In der Nachbereitungsphase wurde der Evaluierungsbericht mit seinen Anhängen – inklusive der internen Managementanlage und des Kurzberichts – erstellt (GIZ 2015a).

Für die PEVs lagen **verschiedene Instrumente der Qualitätssicherung** vor. Hierzu gehörten:

[11] Zur Verwendung von Inferenzstatistik in sog. Vollerhebungen vgl. beispielsweise Andreas Broscheid / Thomas Gschwendn (2005). Zur statistischen Analyse von Vollerhebungen. Politische Vierteljahresschrift, 46. Jg. (2005), Heft 1, S. O-16–O-26.

- eine **Handreichung**, welche als Anlage zum Leitfaden für Programmvorschläge (PV) das Vorgehen für PEV in Kombination mit der Prüfung von Folgemaßnahmen darstellt (GIZ 2015a);
- weitere **Unterstützungsmaterialien**:
    - ein Fahrplan zur PEV, welcher die notwendigen Aufgaben zur Planung, Terminabstimmung und Klärung im Evaluierungsprozess strukturieren soll (GIZ 2015b);
    - ein Muster der Terms of Reference (ToR) für die PEV (GIZ);
    - eine Checkliste zur Selbsteinschätzung der Capacity-WORKS-Erfolgsfaktoren, welche sich mit dem Kooperationsmanagement befasst (GIZ o. D. b);
    - eine Checkliste, in welcher Bewertungsdimensionen mit Analysefragen für die Bewertung der OECD-DAC-Kriterien zusammengefasst werden. Diese Checkliste zeigt zudem auf, wie die Gesamtnote des evaluierten Vorhabens zu berechnen ist (GIZ 2015c);
    - jeweils eine (annotierte) Berichtsgliederung für PEV-Berichte und PEV-Kurzberichte, welche die wesentlichen Inhalte, Analyseaspekte und Leitfragen zu den einzelnen Kapiteln des Berichts auflistet (GIZ o. D. c; 2015d; o. D. d; 2015e);
    - eine (annotierte) Berichtsgliederung zur internen Managementanlage (GIZ o. D. f; 2015e);
    - ein Kommentierungsblatt mit Qualitätsstandards (Mindeststandards, weitere Standards), mit dem die Stabsstelle Evaluierung die methodische Qualität der PEV-Berichte prüft und das den PEV-Teams als Hilfsmittel für die Berichtslegung zur Verfügung steht (GIZ 2015g);
    - eine Methoden-Toolbox, in welcher Anleitungen und Arbeitshilfen zu analytischen Grundlagen, Evaluierungsdesigns, Evaluierungsmethoden (Interviews, Befragungen, Fokusgruppen, Fallstudien) sowie zu Qualitätskriterien von Indikatoren etc. gesammelt werden (GIZ 2016h);
    - Arbeitshilfen zu Indikatoren (GIZ 2014) und zum Wirkungsmodell der GIZ (GIZ 2015f);
- zum Ende des Prozesses: ein **Qualitätscheck**[12] der PEV-Langberichte und Kurzberichte durch die Stabsstelle Evaluierung, wobei nur die Kommentare im Kurzbericht verbindlich zu überarbeiten waren.

Beschreibung der Datengrundlage

Vorhabensdaten

176 Evaluierungsberichte, die zwischen 01. Oktober 2016 und 30. September 2018 abgenommen wurden, bildeten die **Datengrundlage der Textanalyse**. 125 dieser Berichte wurden im Zeitraum 01. Oktober 2016 bis 30. September 2017 abgenommen, während 51 dieser Berichte im Zeitraum 01. Oktober bis 30. September 2018 abgenommen wurden. Im Folgenden wird eine regionale, fachliche, finanzielle und zeitliche **Einordnung der evaluierten Vorhaben** gegeben.

Die Vorhaben, die in den PEV evaluiert wurden, sind **in allen Regionalbereichen** der GIZ angesiedelt. Ein knappes Viertel der Vorhaben sind im Regionalbereich Afrika angesiedelt, und jeweils ein gutes Viertel im Regionalbereich APLAK (Asien, Lateinamerika, Karibik) sowie im Regionalbereich EMZ (Europa, Mittelmeer, Zentralasien). Die restlichen Vorhaben sind im Bereich der Sektor- und Globalvorhaben (GloBe) verortet. Im Vergleich zur letzten Meta-Evaluierung ist damit der Anteil der Vorhaben aus dem Bereich Afrika und APLAK gesunken, während der Anteil der Vorhaben aus dem Bereich EMZ und GloBe gestiegen ist.

Alle Partnerländer der GIZ werden hinsichtlich ihrer Fragilität eingeordnet. Die Einordnung der Länder mit Risikopotenzial für die Lieferfähigkeit der GIZ erfolgt folgendermaßen:

- 23,3 % der Vorhaben (n=41) wurden in Ländern, in denen die GIZ krisenbedingt ein hohes Risiko für die Lieferfähigkeit (**Kategorie I**[13]) sieht, umgesetzt; 5,7 % (n=10) der Vorhaben wurden in Ländern mit erhöhter Gefährdung für die Lieferfähigkeit (**Kategorie II**) umgesetzt); während 31,3 % (n=55) der Vorhaben in Ländern mit geringer Gefährdung der Lieferfähigkeit (**Kategorie III**) angesiedelt waren. Zu

[12] PEV Checks wurden bis Februar 2017 durchgeführt

[13] Einstufung durch die GIZ in der Liste aller PEV mit Rahmendaten, die dem Evaluationsteam übermittelt wurde. Die Einstufung basiert auf einem internen Bewertungssystem mit drei Kategorien: Kategorie 3 zeigt höchstes Risikopotenzial an.

39,8% der Vorhaben (n=70) gab es keine Angaben hinsichtlich des Risikopotenzials für die Lieferfähigkeit.

Die evaluierten Vorhaben decken hinsichtlich der **fachlichen Gruppierung des Fach- und Methodenbereichs (FMB)** ein breites Portfolio ab. Dabei fallen die meisten Vorhaben (42,2 %) in den Themenbereich **Klima, ländliche Entwicklung und Infrastruktur** (hierzu gehören: Klima und Umweltpolitik; Wald, Biodiversität, Landwirtschaft; Ländliche Entwicklung, Ernährungssicherung; Wasser, Abwasser, Abfall; Energie und Verkehr). Ähnlich viele Vorhaben (37,3 %) sind dem Themenbereich Wirtschaft, Beschäftigung und soziale Entwicklung (Bildung; Berufliche Bildung, Arbeitsmarkt; Finanzsystementwicklung, Versicherungen; Gesundheit und Soziale Sicherung; Wirtschaftspolitik und Privatwirtschaftsförderung; neue gesellschaftspolitische Themen) zugeordnet. Der am wenigsten stark repräsentierte Themenbereich ist mit 20.5 % der Bereich Governance und Konflikt (Rechtsstaat und Sicherheit; öffentliche Finanzen und Verwaltung; Demokratie, Politikdialog, Stadt; Frieden und Nothilfe). Abbildung 4 stellt die Verteilung nach Regionalbereichen und fachlichen Schwerpunkten der evaluierten Vorhaben dar.

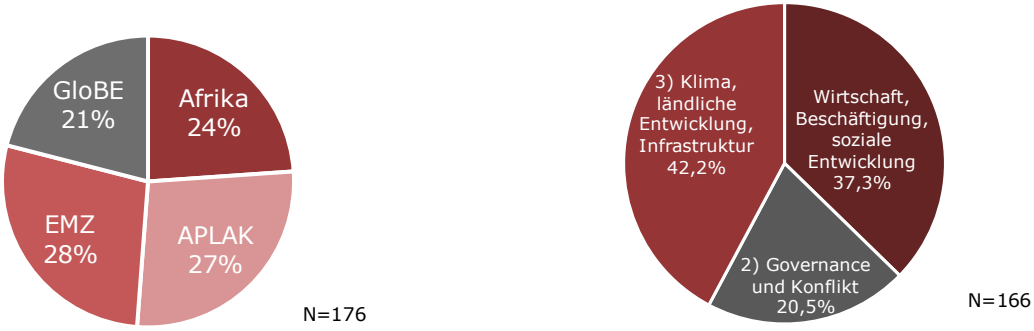


Abbildung 4: Regionalbereiche und fachliche Schwerpunkte der evaluierten Vorhaben.
(Quelle: Syspons 2018, Textanalyse der PEV-Berichte)
Das unterschiedliche n ist darauf zurückzuführen, dass eine Einordnung in die fachliche Gruppierung des Fach- und Methodenbereichs für einige wenige Vorhaben nicht möglich war

Auch hinsichtlich des Auftragswerts und der Laufzeit bilden die evaluierten Vorhaben ein breites Spektrum ab. Das durchschnittliche **Auftragsvolumen** beträgt 8,6 Mio. Euro, wobei das kleinste Auftragsvolumen 1,5 Mio. Euro beträgt und das größte Auftragsvolumen bei 33 Mio. Euro liegt. Die durchschnittliche Laufzeit der Vorhaben beträgt knapp 4 Jahre (45 Monate). Dabei ist die kürzeste Laufzeit etwas über ein Jahr (15 Monate), während die längste Laufzeit gut 8 Jahre beträgt (100 Monate).

In der Gesamtschau lässt sich nur für ein knappes Drittel anhand der Berichte die **Modulphase** der Berichte ableiten. Hiervon waren 27,1 % (n=13) in der ersten Phase, 31,3 % (n=15) in der zweiten Phase, 33,3 % (n= 16) in der dritten Phase, und 8,3 % (n= 4) in der vierten Phase.

Art der PEV

PEV sollten einen kritischen Rückblick auf der Grundlage der Bewertung der OECD-DAC-Kriterien ermöglichen. PEV, die zugleich eine **Folgemaßnahme** prüfen, sollten zusätzlich zur Planung der Weiterführung des Vorhabens dienen. Von den Berichten, zu denen hierzu Angaben verfügbar waren, zählen 35,2 % (n=62) zu PEV, die zugleich eine Folgemaßnahme prüfen. Wenn keine Folgemaßnahme geplant war, wurden PEV als **Abschlussevaluierung** durchgeführt. Dies trifft auf 63,1 % (n= 111) der Berichte der Textanalyse zu. Demnach befinden sich im Vergleich zur letzten Meta-Evaluierung mehr Abschlussevaluierungen unter den analysierten PEV.

Rahmendaten der PEV

In den Terms of Reference (ToR) wird die maximale Anzahl der für die PEV zur Verfügung stehenden Gutachtertage aufgelistet. Anhand der ToR lässt sich allerdings nicht immer ableiten, wie viele Gutachtertage für die Evaluierung und wie viele Tage für die Planung eines Folgevorhabens eingesetzt werden. Die ToR der evaluierten Vorhaben veranschlagen durchschnittlich 8,9 Tage für die Vorbereitung, 29,2 Tage für die Durchführung und 18,4 Tage für die Nachbereitung. Im Durchschnitt werden somit für eine PEV 56,5 Tage veranschlagt, wobei die Werte zwischen einzelnen Vorhaben stark variieren. Die Auswertungen zeigen, dass in der Tendenz ein höherer Auftragswert eines Vorhabens mit einer höheren Anzahl an Gutachtertagen einhergeht, wenngleich es einzelne PEVs gibt, die dieser Tendenz nicht entsprechen. Abbildung 5 zeigt eine Kreuzung des Auftragswertes der Vorhaben mit der Anzahl der Gutachtertage. Sie stützt die Hypothese, dass mit einem höheren Auftragswert auch eine höhere Komplexität der Vorhabensarchitektur einhergeht, die dann in der Evaluierung tendenziell mit mehr Gutachtertagen untersucht wird.

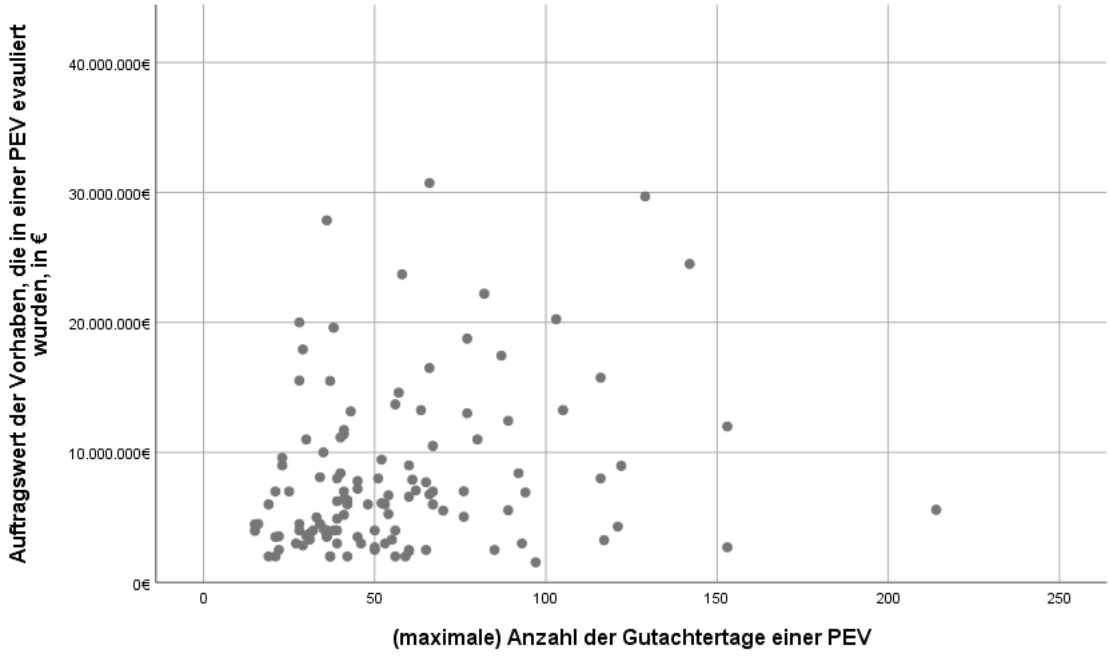


Abbildung 5: Kreuzung des gesamten Auftragswerts der Vorhaben, die in einer PEV evaluiert wurden, mit der Anzahl der Gutachtertage (Quelle: Syspons 2018, Textanalyse)

Zugleich gibt es Unterschiede bei der Anzahl der Gutachtertage zwischen PEV mit oder ohne Folgemaßnahme. Während für PEV ohne Folgemaßnahme in den ToR durchschnittlich 51,3 Tage veranschlagt werden, werden für PEV mit Folgemaßnahmen 67,4 Tage veranschlagt. Der größte Unterschied besteht dabei in der Anzahl der Tage, die für die Nachbereitung vorgesehen sind. Diese beträgt für Schlussevaluierungen durchschnittlich 15,7 Tage, während sie für PEV mit Folgevorhaben 23,9 Tage beträgt. Die Unterschiede im Mengengerüst für die Vorbereitung (8,2, bzw. 10,3 Tage) und die Durchführung (27,4, bzw. 33,4 Tage) sind demgegenüber weniger ausgeprägt. Insgesamt ist der Unterschied zwischen dem Mengengerüst für PEV mit oder ohne Folgemaßnahme weniger ausgeprägt als in der letzten Meta-Evaluierung

von 2016, aber immer noch beachtlich.

Ein Blick auf die Zahl der Gutachter/innen einer PEV zeigt, dass PEV zumeist von zwei bis fünf Gutachter/innen durchgeführt wurden. Nur in Ausnahmen sind mehr oder weniger Gutachter/innen beteiligt (siehe Abbildung 6). Größtenteils bestehen die **PEV-Teams sowohl aus internen als auch aus externen Gutachter/innen** (87,7 %, n= 136). Ausschließlich externe Gutachter/innen wurden vergleichsweise selten (9,7 %, n= 15) eingesetzt, und nur in Ausnahmen wurden lediglich interne Gutachter/innen eingesetzt (2,6 %, n=4).

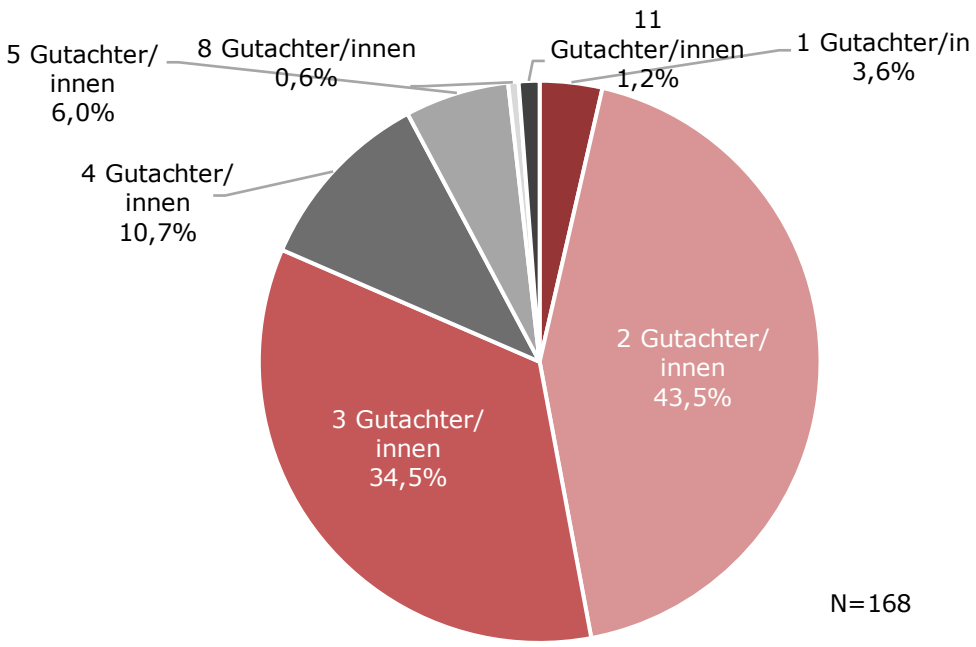


Abbildung 6: Anzahl der Gutachter/innen (Quelle: Syspons 2018, Textanalyse)

Ebenfalls untersucht wurde, wie die Gutachter/innen die evaluierten Vorhaben hinsichtlich der **OECD-DAC-Kriterien** bewerten. Insgesamt betrachtet werden die evaluierten Vorhaben seitens der Gutachter/innen zumeist sehr positiv bewertet. Von den insgesamt 176 Vorhaben erhielten 150 (85,2 %)die Gesamtbenotung „sehr erfolgreich“ oder „erfolgreich“.

Ein detaillierter Blick auf die OECD-DAC Kriterien verdeutlicht, dass das Kriterium der Relevanz insgesamt am besten beurteilt wurde. Die Mehrheit der Vorhaben schneiden hier mit „sehr erfolgreich“ ab. Hinsichtlich der weiteren Kriterien (Effizienz, Impact und Nachhaltigkeit) erhalten die meisten Vorhaben die Bewertung „erfolgreich“. Bei den Kriterien Impact und Nachhaltigkeit weist die Verteilung ein größeres Spektrum auf und verhältnismäßig mehr Vorhaben erhalten die Bewertung „eher erfolgreich“ oder „eher unbefriedigend“.

	Stufe 1 = sehr erfolgreich	Stufe 2 = erfolgreich	Stufe 3 = eher erfolgreich	Stufe 4 = eher unbefriedigend	Stufe 5 = unbefriedigend	Stufe 6 = sehr unbefriegigend
Relevanz	157	18	1	0	0	0
Effektivität	36	104	28	6	2	0
Effizienz	43	96	27	9	1	0
Impact	28	89	42	17	0	0
Nachhaltigkeit	25	74	56	18	3	0
Gesamtbenotung	42	108	24	2	0	0

Abbildung 7: OECD-DAC-Bewertungen der Vorhaben nach Schulnoten (n=176) durch die Gutachter/innen (Quelle: Syspons 2018, Textanalyse)

Somit fällt die durchschnittliche Benotung entsprechend der in den PEV genutzten Skala von 0 bis 16 Punkten für das Relevanz-Kriterium mit 14,9 Punkten am besten aus. Die schwächste durchschnittliche Benotung durch die PEV-Gutachter liegt mit einer durchschnittlichen Punktzahl von 11,2 im Kriterium Nachhaltigkeit vor.

	Durchschnittliche Bewertung
Relevanz	14,9
Effektivität	12,8
Effizienz	12,7
Impact	12,0
Nachhaltigkeit	11,2
Gesamtbenotung	12,7

Abbildung 8: Durchschnittliche Bewertung der Vorhaben nach Punkten in den OECD-DAC-Kriterien (n=176) durch die Gutachter/innen (Quelle: Syspons 2018, Textanalyse)

2.5 Ergebnisse zur methodischen Qualität (Genauigkeit)

Die Meta-Evaluierung der PEV hinsichtlich der methodischen Qualität ist wie folgt aufgebaut:

Im Kapitel 2.4 werden die Ergebnisse entlang der 13 Bewertungskriterien dargestellt, aus denen sich der Evaluierungsstandard Genauigkeit (G) zusammensetzt. Zur Feststellung der PEV-Qualität bezüglich der 13 Bewertungskriterien wurde auf Indikatoren der Textanalyse zurückgegriffen. Die Indikatoren werden mit „erfüllt" (100%) oder „nicht erfüllt" (0%) bewertet. In diesem Bericht wird der Erfüllungsgrad der Bewertungskriterien und dazugehörigen Indikatoren im Durchschnitt aller 176 PEV-Berichte aus den Jahren 2017 / 2018 gemeinsam dargestellt. Nach der Darstellung der Ergebnisse für 2017 / 2018 folgt eine Analyse der Trends in der Qualität der dezentralen PEV 2015 – 2018 unter Einbezug der Ergebnisse der vorangegangenen Meta-Evaluierungen.

Eine tabellarische Darstellung der Ergebnisse für 2017 und 2018, einmal aufgeschlüsselt nach Jahren und einmal zusammengefasst, findet sich im Anlage 1. Die Ergebnisse für die Jahre 2015 – 2018 sind in der digitalen Anlage 3 verfügbar. **Für weitere Informationen zum Bewertungssystem siehe auch Abbildung 3 in Kapitel 2.2.**

Kapitel 2.5 stellt die in diesem Jahr erstmalig gesondert vorgenommene Betrachtung von kontributionsanalytischen Qualitätsaspekten und die Ergebnisse zu den erstmalig erhobenen Indikatoren zur Nachvollziehbarkeit dar.

Kapitel 2.6 stellt die Einflussfaktoren auf die methodische Qualität dar, welche auf Grundlage der vorliegenden Daten identifiziert werden konnten.

Kapitel 2.7 enthält schließlich die Bewertung und Schlussfolgerung zur methodischen Qualität durch das Meta-Evaluierungsteam. Dies beinhaltet eine Abwägung der identifizierten Stärken,

Beschreibung und Analyse des Evaluierungsstandards Genauigkeit

Der Evaluierungsstandard Genauigkeit befasst sich mit der **methodischen Qualität** einer Evaluierung. Die DeGEval definiert den Evaluierungsstandard wie folgt:

> *„Die Genauigkeitsstandards sollen sicherstellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt" (DeGEval 2008).*

Dafür formuliert die DeGEval neun Bewertungskriterien, deren Einhaltung die methodische Qualität einer Evaluierung sicherstellt. Die vorliegende Meta-Evaluierung differenziert das siebte Kriterium weiter aus, um der methodischen Bewertung der OECD-DAC-Kriterien ausreichend Raum zu geben, und erhebt das neunte Kriterium „Meta-Evaluation" der DeGEval nicht. Die folgende Abbildung fasst die 13 Bewertungskriterien die in dieser und in den vergangenen Jahren in der Meta-Evaluierung erhoben wurden übersichtlich zusammen und

zeigt deren durchschnittliche Erfüllung für die Jahre 2017 / 2018, wie sie sich in den jeweils analysierten Indikatoren in den evaluierten PEV darstellt, in Prozent auf. Darüber hinaus wurden dieses Jahr erstmals zusätzliche Indikatoren zur Nachvollziehbarkeit der Bewertung erhoben, die separat ausgewertet wurden und in der Abbildung sowie im Text gesondert dargestellt werden, da hierfür keine Vergleichbarkeit gegenüber den Meta-Evaluierungen 2015 und 2016 gegeben ist.

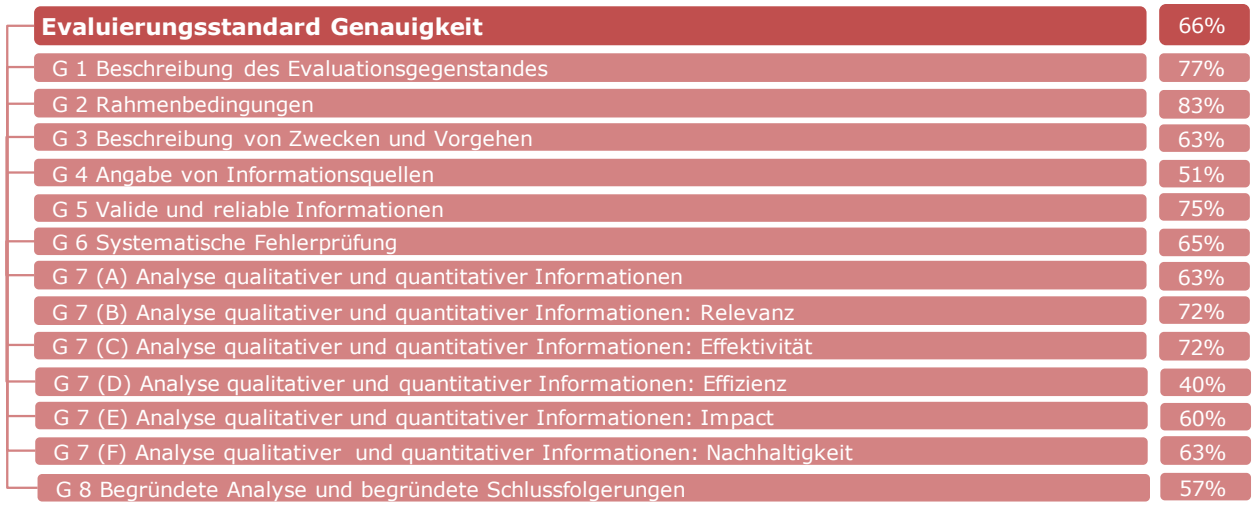


Abbildung 9: Erfüllungsgrad der Bewertungskriterien im Evaluierungsstandard Genauigkeit (Quelle: Syspons 2018, Textanalyse)

Der Evaluierungsstandard Genauigkeit wird im Durchschnitt aller 176 Berichte **eher[14] erfüllt** (66 %). Wie die Abbildung 9 zeigt, gibt es hierbei deutliche Unterschiede hinsichtlich des durchschnittlichen Erfüllungsgrades der Bewertungskriterien. Am besten schneidet hierbei mit 83 % das Kriterium G 2 ab, das erfasst, inwiefern des untersuchten Vorhabens ausreichend detailliert untersucht und analysiert werden. Am schwächsten schneidet das Kriterium G 7 C ab, welches die methodische Auseinandersetzung mit der Effizienz untersucht.

Von den 176 Berichten erfüllen 51 Berichte den Evaluierungsstandard Genauigkeit bedingt, d.h. zu weniger als 60 %. Für 109 Berichte ist der Evaluierungsstandard eher erfüllt (60-79%), und für 15 Vorhaben ist er zumeist erfüllt (80-89%). Ein Vorhaben erfüllt den Standard größtenteils bis vollständig (90-100%).

Erfüllung des Evaluierungsstandards Genauigkeit	Anzahl PEV
bedingt erfüllt (<60%)	51
eher erfüllt (60-79%)	109
zumeist erfüllt (80-89%)	15
größtenteils bis vollständig erfüllt (90-100%)	1
Gesamt	176

Abbildung 10: Anzahl von Berichten, die den Evaluierungsstandard bedingt, eher, zumeist bzw. größtenteils bis vollständig erfüllen (Quelle: Syspons 2018, Textanalyse)

Übergeordnete methodische Aspekte in den PEV 2017 / 2018

Die ersten drei Bewertungskriterien und zugeordnete Indikatoren zur methodischen Qualität befassen sich mit übergeordneten methodischen Aspekten (siehe Abbildung 11).

[14] Um die Bewertungen auf Ebene der Evaluierungsstandards und Bewertungskriterien einheitlich und nachvollziehbar zu gestalten, werden diese wie folgt angegeben: 90–100 % (größtenteils bis vollständig erfüllt); 80–89 % (zumeist erfüllt); 60–79 % (eher erfüllt); < 60 % (bedingt erfüllt). Siehe hierzu auch Kapitel 3.2

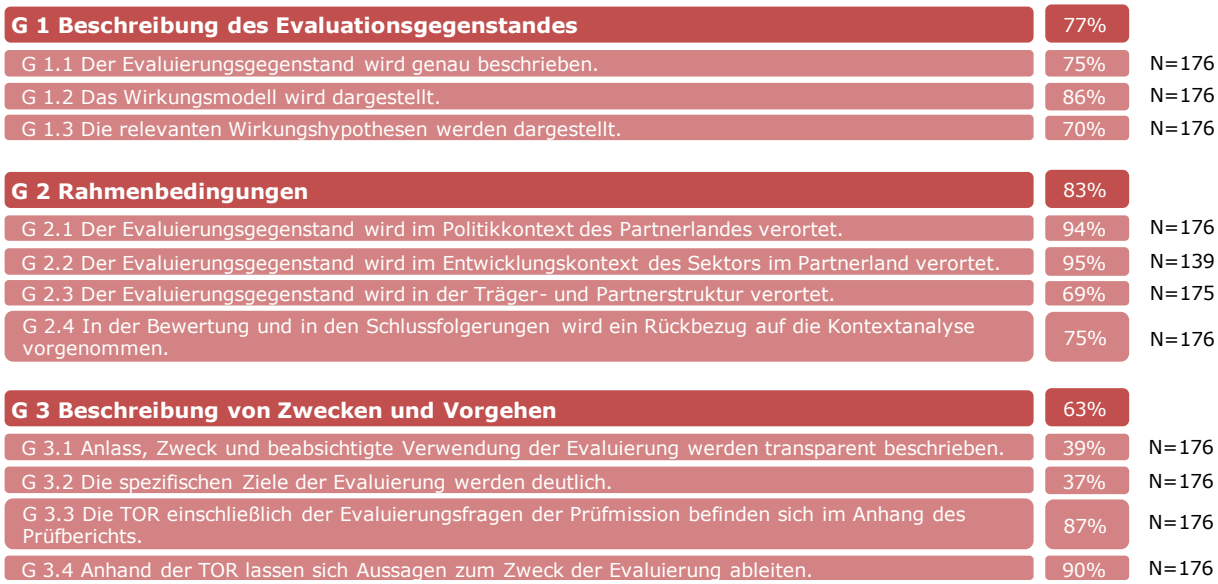


Abbildung 11: Durchschnittlicher Erfüllungsgrad der Indikatoren in den Bewertungskriterien G1 - G3 in den PEV 2017 / 2018 (Quelle: Syspons 2018, Textanalyse). Das N beschreibt Anzahl der Berichte, für die der jeweilige Indikator bewertet wurde. [15]

G1 Beschreibung des Evaluationsgegenstandes in den Berichten 2017 / 2018: Das erste Evaluationskriterium zeigt, ob das evaluierte Vorhaben in den PEV-Berichten klar und genau beschrieben wird. Hierzu wurde anhand von drei Indikatoren erfasst, ob eine Definition und Abgrenzung des Evaluationsgegenstandes erfolgt (G 1.1), ob das Wirkungsmodell dargestellt wird (G 1.2), und ob die Wirkungshypothesen ausgeführt werden (G 1.3).

Wie die Analyse der Berichte zeigt, sind diese drei Indikatoren im Mittel zu 77 % erfüllt. Damit ist das Kriterium G1 **eher erfüllt** (siehe Abbildung 11). In drei Viertel der Berichte (75 %, n = 132) wird der Evaluierungsgegenstand genau beschrieben (G 1.1). Dabei ist der Indikator dann erfüllt, wenn mindestens vier der folgenden sechs Aspekte gegeben sind: zeitliche Abgrenzung, Darstellung des Vorhabensbudgets, regionale Abgrenzung, sowie Einordnung des Mehrebenenansatzes, der Systemgrenze und der Capacity Development-Ebenen. Ein vertiefter Blick in die Ergebnisse zeigt, dass hohe Werte vor allem bei der zeitlichen Abgrenzung (86 %, n= 152), der regionalen Abgrenzung (84 %, n= 148) sowie der Darstellung der Systemgrenze gegeben sind (72 %, n= 127). Eine inhaltliche Abgrenzung hinsichtlich des Mehrebenenansatzes (65 %, n=115) und der Capacity Development Ebenen (61 %, n=108) erfolgt immer noch für deutlich über die Hälfte der Vorhaben, während das Vorhabensbudget in knapp weniger als der Hälfte der Berichte (47 %, n= 83) aufgeführt wird.

Die Darstellung des Wirkungsmodells ist in 86 % (n=151) der Berichte enthalten (G 1.2). Darüber hinaus formulieren 70 % der Berichte (n=123) die relevanten Wirkungshypothesen narrativ aus (G.1.3). Hierbei handelt es sich um eine zentrale Voraussetzung für das Verständnis der intendierten Ziele und Wege zur Zielerreichung des Vorhabens, die für Evaluationen einen großen Stellenwert hat. Der Indikator G 1.2 wurde dann positiv bewertet, wenn entweder eine graphische Darstellung oder eine Beschreibung als Text vorlag. Der Indikator G 1.3 ist dann erfüllt, wenn mindestens die zentralen Wirkungszusammenhänge hinsichtlich der Erreichung des Modulziels dargestellt sind.

G2 Rahmenbedingungen in den Berichten 2017 / 2018: Das zweite Bewertungskriterium prüft, ob in den Evaluierungsberichten ausreichend detailliert auf den Kontext der evaluierten Vorhaben eingegangen wird. Diesem Anspruch liegt zugrunde, dass eine fundierte Analyse der Rahmenbedingungen eine Grundvoraussetzung für die Interpretation der Ergebnisse und die Lernfunktion einer Evaluierung ist. Das Bewertungskriterium

[15] G 2.2. wurde für GloBE-Vorhaben als unpassend bewertet, folglich verändert sich die Grundgesamtheit. G.2.3 wurde für ein Vorhaben als unpassend bewertet, das keinen politischen Träger hat.

setzt sich aus vier Indikatoren zusammen. Die ersten drei Indikatoren erfassen, ob die PEV Berichte das evaluierte Vorhaben im Politikkontext verorten (G 2.1) sowie im Sektorkontext des Partnerlandes (G 2.2) und der Träger- und Partnerstruktur (G 2.3). Der vierte Indikator untersucht, ob in der Bewertung und in den Schlussfolgerungen ein Rückbezug auf die Kontextanalyse vorgenommen wird (G 2.4).

Das Bewertungskriterium G2 wird im Mittel in 83 % der Berichte zumeist erfüllt (siehe Abbildung 11). Die Berichte beschreiben fast durchgängig (94 %, n= 132) den Politikkontext, in dem das Vorhaben verortet ist. Auch die Verortung im Entwicklungskontext des Sektors auf sozio-ökonomischer, politischer / oder kultureller Ebene erfolgt nahezu flächendeckend (95 %, n= 132[16]). Die Verortung in der Träger- und Partnerstruktur erfolgt demgegenüber weniger häufig (69 %, n= 121). Schließlich zeigt sich, dass die Kontextfaktoren in drei Vierteln der PEV-Berichte in der Bewertung und in den Schlussfolgerungen wieder aufgegriffen werden (75 %, n= 132).

G3 Beschreibung von Zweck und Vorgehen in den Berichten 2017 / 2018: Das dritte Bewertungskriterium untersucht, ob Gegenstand, Zwecke und Fragestellungen der Evaluation dokumentiert sind. Dieser Aspekt ist wichtig, um festzustellen, inwiefern die PEV den jeweiligen Zielen gerecht werden. Das Bewertungskriterium setzt sich aus vier Indikatoren zusammen. Diese bilden ab, ob in den PEV-Berichten Anlass, Zweck und beabsichtige Verwendung der Evaluierung dargestellt sind (G 3.1) und ob die spezifischen Ziele deutlich werden (G 3.2). Weiterhin messen sie, ob die ToR einschließlich der Evaluierungsfragen dem Bericht angehängt sind (G 3.3) und ob sich anhand dieser Aussagen zum Zweck der Evaluierung ableiten lassen (G 3.4).

Insgesamt wird das Bewertungskriterium G3 im Durchschnitt aller Berichte mit 63 % eher erfüllt (siehe Abbildung 11). Dabei zeigen sich deutliche Unterschiede zwischen der Darstellung von Zweck im PEV-Bericht selbst gegenüber der Darstellung in den Terms of Reference. In den Berichten selbst werden Anlass, Zweck und beabsichtigte Verwendung der Evaluierung lediglich in 39 % der Fälle transparent beschrieben (n=69) (G 3.1). Auch die spezifischen Ziele der Evaluierung werden aus den PEV mit 37 % nur bedingt deutlich (n=65) (G 3.2). Allerdings befinden sich die ToR fast flächendeckend im Anhang des Berichts und geben Aufschluss über die Evaluierungsfragen (87 %, n=153). Ebenso lassen sich für fast alle Berichte anhand der ToR Aussagen zum Zweck der Evaluierung ableiten (90%, n=159). Die fehlende Darstellung von Zweck und Zielen der Evaluierung in den Berichten selbst macht es jedoch für den Leser z.T. schwer nachzuvollziehen, worin die Ziele bestehen und inwiefern die Prüfmission diesen gerecht geworden ist.

[16] Sektor- und Globalvorhaben wurden im Indikator G 2.2 mit „unpassend“ bewertet. Eine entsprechende Kontextualisierung dieser Vorhaben wurde bereits bei Indikator G.2.1 erwartet. Folglich verringert sich die Grundgesamtheit.

Trends hinsichtlich übergeordneter methodischer Aspekte

Trends in der Qualität der dezentralen PEV 2015 – 2018 für die Bewertungskriterien G1 – G3[17]: Das Abschneiden der ersten beiden Bewertungskriterien zu übergeordneten methodischen Aspekten weist im Zeitverlauf über die durchgeführten PEV 2015 – 2018 eine positive Entwicklung auf, während das dritte Bewertungskriterium relativ konstant bleibt (siehe Abbildung 12).

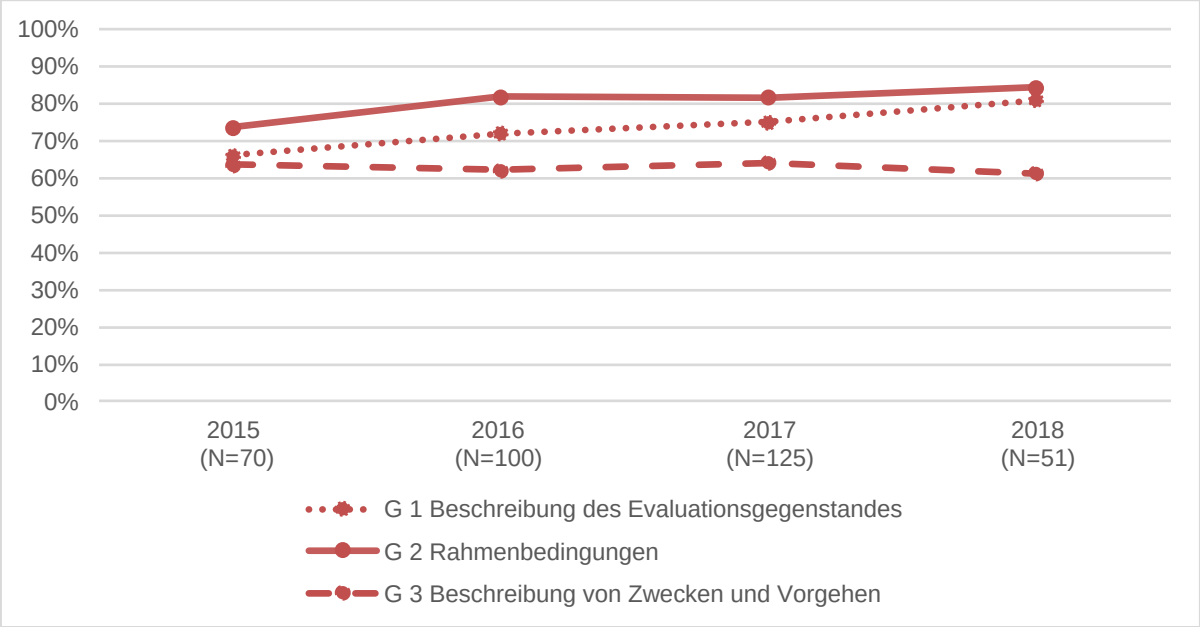


Abbildung 12: Durchschnittlicher Erfüllungsgrad der Indikatoren in den Bewertungskriterien G1 - G3 über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Das erste Bewertungskriterium, G1, welches die Beschreibung des Untersuchungsgegenstandes erfasst, hat sich dabei jedes Jahr positiv entwickelt. In der Gesamtschau konnte sich das Kriterium um 15 Prozentpunkte steigern, von einem Ausgangswert von 66 % in den PEVs 2015 auf einen Wert von 81 % in den PEV 2018 (siehe Abbildung 12). Verbesserungen wurden dabei vor allem hinsichtlich der Darstellung des Evaluierungsgegenstandes und der Beschreibung der Wirkungshypothesen festgestellt.

Das Abschneiden hinsichtlich der Bewertung des zweiten Kriteriums, G2, welches die Darstellung der Rahmenbedingungen untersucht, hat sich zwischen 2015 und 2018 um 10 Prozentpunkte verbessert. Auch hier ist ein kontinuierlicher positiver Trend festzustellen, mit durchschnittlichen Werten von 74 % in 2015, 82 % in 2016 und 2017, und 84 % in 2018 (siehe Abbildung 12). Hierbei haben sich die Indikatoren zur Verortung des Evaluierungsgegenstands im Politikkontext des Partnerlandes und im Entwicklungskontext des Sektors sowie zum Rückbezug auf die Kontextanalyse in der Darstellung von Bewertung und Schlussfolgerungen kontinuierlich positiv entwickelt. Schwankungen sind auf Ebene des Indikators zur Verortung des Evaluierungsgegenstands in Träger- und Partnerstruktur festzustellen.

Das Abschneiden des dritten Bewertungskriteriums G3, Beschreibung von Zweck und Vorgehen, ist im Zeitverlauf relativ konstant geblieben. Je nach Jahr schwankt der durchschnittliche Wert zwischen mindestens 61 % (2018) und höchstens 64 % (2015) (siehe Abbildung 12). In allen Jahren schneidet die Darstellung von Zweck und Zielen der Evaluierung im Berichtstext selbst deutlich schwächer ab als die entsprechende Darstellung in den Terms of Reference.

[17] Während für die Darstellung der Ergebnisse dieser Meta-Evaluierung die Ergebnisse gebündelt für 2017 / 2018 dargestellt wird, erfolgt die Trendanalyse aufgeschlüsselt nach Jahren.

Qualität in der methodischen Durchführung in den PEV 2017 / 2018

Die Qualität in der methodischen Durchführung wird in den nächsten drei Bewertungskriterien und zugeordneten Indikatoren erhoben (siehe Abbildung 13).

G4 Angabe von Informationsquellen in den Berichten 2017 / 2018: Das vierte Bewertungskriterium erfasst, inwiefern die in den PEVs genutzten Informationsquellen in angemessenem Maße dargestellt sind. Das Kriterium setzt sich aus sechs Indikatoren zusammen. Diese stellen fest, ob eine vollständige Liste der Gesprächspartner/innen und sonstigen Informationsquellen vorliegt (G 4.1), und ob die Systematik der Dokumentenauswertung (G 4.2) und die Kriterien für die Auswahl von Gesprächspartnern bzw. für die Ziehung von Stichproben (G 4.3) deutlich wird. Darüber hinaus erfassen die Indikatoren, ob die PEV-Berichte sich mit der Verlässlichkeit der Daten des wirkungsorientierten Monitorings (WOM) auseinandersetzen (G 4.4), und ob sie hierbei die Belastbarkeit von Baseline-Daten (G 4.5) und Partnerdaten (G 4.6) reflektieren.

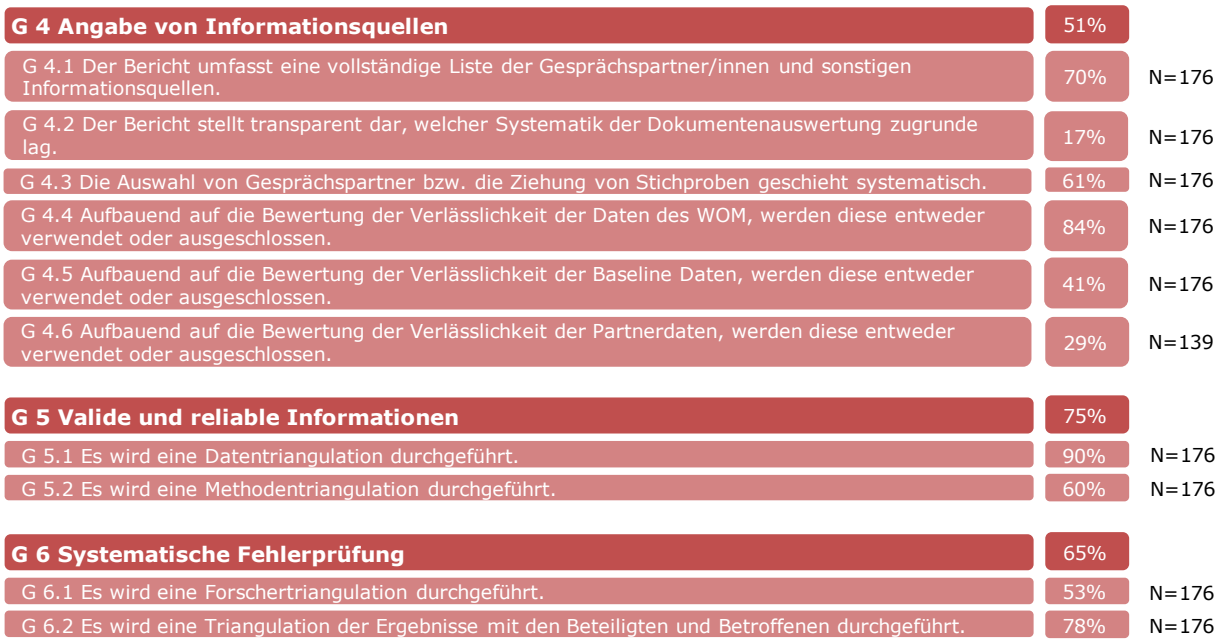


Abbildung 13: Durchschnittlicher Erfüllungsgrad der Indikatoren in den Bewertungskriterien G 4- G6 in den PEV 2017 / 2018 (Quelle:Syspons, Textanalyse). Das N beschreibt die Anzahl der Berichte, für die der jeweilige Indikator bewertet wurde.[18]

Mit einem durchschnittlichen Erfüllungsgrad von 51 % wird das Bewertungskriterium G4 **bedingt erfüllt** (siehe Abbildung 13). Während zwei Drittel der Berichte (70 %, n= 123) darstellen, welche Gesprächspartner/innen und sonstige Informationsquellen einbezogen worden sind, wird seltener darauf eingegangen, anhand welcher Systematik die Dokumentenauswertung (17 %, n=30) erfolgt ist. Hingegen wird in über der Hälfte der Berichte (61 %, n= 107) dargestellt, wie und warum die Auswahl von Gesprächspartner/innen bzw. die Ziehung von Stichproben erfolgt ist. Mit Blick auf die Darstellung der ausgewerteten Dokumente zeigt sich, dass die PEV in über der Hälfte der Fälle Bezug nehmen auf die Programmvorschläge (PV) (60 %, n= 106), den Operationsplan (64 %, n=113), die Capacity Development Strategie (54 %, n= 95), die Akteursanalyse (67 %, n= 118) sowie das Wirkungsmodell (66 %, n=116). Deutlich seltener wird mit in 37 % der Fälle (n=65) Bezug genommen auf die Steuerungsstruktur. Diesbezüglich ist zu berücksichtigen, dass eine reine Nennung der zu Rate gezogenen Dokumente nicht ausreichte, um den Indikator hinsichtlich der Darstellung der Systematik der Dokumentenauswertung zu erfüllen. Für die Erfüllung des Indikators musste darüber hinaus deutlich werden, wie die jeweiligen Dokumente in die Analyse eingeflossen sind.

Hinsichtlich der Auseinandersetzung mit der Verlässlichkeit der Daten des Wirkungsorientierten Monitorings

[18] Indikator G 4.6 wurde für GloBE-Vorhaben als unpassend bewertet, folglich verändert sich die Grundgesamtheit.

(WOM) in den PEV-Berichten konnte eine übergeordnete Auseinandersetzung in 85 % der Fälle (n=142) festgestellt werden. Voraussetzung für die Erfüllung des Indikators (G 4.4) war, dass aufbauend auf einer Bewertung der Verlässlichkeit der Daten diese entweder von der Prüfmission verwendet oder ausgeschlossen wurden. Eine entsprechende explizite Befassung mit der Belastbarkeit der Baseline-Daten erfolgte deutlich seltener (40 %, n= 67). Diesbezüglich ist anzumerken, dass es eine Vorgabe für die GIZ-Vorhaben ist, eine Baseline für die Modulzielindikatoren festzuhalten. Eine fehlende Auseinandersetzung der PEV-Teams mit der Qualität der Baseline-Daten lässt sich demnach nicht dadurch erklären, dass keine Baseline-Werte vorliegen, sondern eher dadurch, dass diese nicht hinterfragt werden. Hinsichtlich der Baseline-Daten fällt auf, dass die Baseline in GIZ-Angeboten insbesondere, wenn es um die Erarbeitung von Strategien oder die Einführung von Prozessen geht, oft mit „0“ angegeben ist. Auch diese Werte kritisch zu hinterfragen ist geboten, um die Belastbarkeit von Kontributionsanalysen zu gewährleisten. In weniger als einem Drittel der Fälle (29 %, n= 40) erfolgte eine Auseinandersetzung mit der Verlässlichkeit der Partnerdaten. Dies ist ggf. teilweise damit zu erklären, dass nur für einen Teil der Vorhaben Partnerdaten vorliegen. Da zwei Drittel der PEV-Berichte Partnerdaten gar nicht adressieren, lässt sich für diese jedoch nicht nachvollziehen, ob die Evaluatoren sich damit auseinandergesetzt haben, und ob Partnerdaten vorlagen. Insgesamt ist somit die Reflektion mit der Belastbarkeit der Daten des wirkungsorientierten Monitorings nur in Teilen gegeben.

G5 Valide und verlässliche Informationen in den Berichten 2017 / 2018: Die Gewährleistung der Zuverlässigkeit und Gültigkeit der durch die Evaluation gewonnen Daten nach Qualitätskriterien der Sozialforschung ist maßgeblich für die Belastbarkeit der Evaluationsergebnisse. Inwiefern die PEVs diesem Anspruch gerecht werden, wurde im fünften Bewertungskriterium durch zwei Indikatoren operationalisiert. Der erste Indikator untersucht die Datentriangulation (G 5.1), während der zweite Indikator die Methodentriangulation (G 5.2) analysiert[19].

Im Mittel wird das Bewertungskriterium mit 75 % **eher erfüllt** (siehe Abbildung 13). In der Textanalyse wurden vornehmlich in den Kapiteln zur Darstellung der methodischen Vorgehensweise oder in den Kapiteln zu Effektivität und Impact Anhaltspunkte für die Triangulation durch die Prüfmission identifiziert. Hierbei wurden in neun von zehn Berichten (90 %, n=159) Daten / Informationen zu einem gleichen Sachverhalt durch die Einbeziehung verschiedener Akteure erhoben (G 5.1). Eine Methodentriangulation (G 5.2), also die Erfassung von Informationen zum gleichen Sachverhalt durch unterschiedliche Methoden, ist demgegenüber für einen geringeren Anteil der Berichte (60%, n= 106) gegeben.

G6 Systematische Fehlerprüfung in den Berichten 2017 / 2018: Das sechste Bewertungskriterium erfasst, ebenfalls durch Ansätze der Triangulation, inwiefern die durch die PEV-Prüfteams erhobenen Informationen systematisch auf Fehler hin untersucht werden. Um hierüber Aussagen zu treffen, wurden dem sechsten Bewertungskriterium zwei Indikatoren zugrunde gelegt. Der erste Indikator untersucht, ob eine Forschertriangulation durchgeführt wurde (G 6.1)[20], während der zweite Indikator erhebt, ob eine Triangulation der Ergebnisse mit den Beteiligten und Betroffenen durchgeführt wurde (G 6.2). Über diese beiden Sachverhalte wird abgebildet, inwiefern die PEV-Prüfteams Mechanismen entwickelt haben, um mit widersprüchlichen Ergebnissen umzugehen und fehlerhafte Aussagen zu vermeiden.

Im Durchschnitt der Berichte wird das Kriterium G 6 mit 65 % **eher erfüllt** (siehe Abbildung 13). In knapp über der Hälfte der Berichte (53 %, n= 93) gibt es Hinweise auf eine Forschertriangulation. In der Textanalyse waren Anhaltspunkte für die systematische Überprüfung der gesammelten Informationen innerhalb des PEV-Teams beispielsweise Synthesetreffen, die dokumentiert wurden, oder die Darstellung von unterschiedlichen Wahr-

[19] Einer Datentriangulation liegt in dieser Meta-Evaluierung der Anspruch zugrunde, dass mindestens drei Daten zu einem Sachverhalt herangezogen werden müssen; einer Methodentriangulation, dass mindestens drei Methoden zu einem Sachverhalt herangezogen werden müssen.

[20] Einer Forschertriangulation liegt in dieser Meta-Evaluierung der Anspruch zugrunde, dass mindestens die Perspektiven von drei Forscher/innen zu einem Sachverhalt herangezogen werden müssen.

nehmungen innerhalb des Prüfteams, die reflektiert wurden. Demgegenüber konnte eine Triangulation der Ergebnisse mit Partnern, Zielgruppen oder Auftraggebern in 78 % der Fälle (n= 137) festgestellt werden. Als Anhaltspunkte wurden hierfür in der Textanalyse beispielsweise Debriefings oder die Kommentierung des Berichts durch Beteiligte und Betroffene genutzt. Sowohl qualitative als auch quantitative Analysen erfordern diesen iterativen Prozess, in welchem die eigenen Interpretationen und Folgerungen einer Überprüfung unterzogen werden müssen.

Trends in der Qualität der methodischen Durchführung

Trends in der Qualität der dezentralen PEV 2015 – 2018 für die Bewertungskriterien G4 – G6: Die Entwicklung der drei Bewertungskriterien hinsichtlich der Qualität in der methodischen Durchführung ist durchwachsen, aber überwiegend positiv. Das Kriterium G4 hat sich leicht positiv entwickelt und das Kriterium G5 stärker positiv, während es hinsichtlich des Kriteriums G6 leichte Rückschritte gab (siehe Abbildung 14).

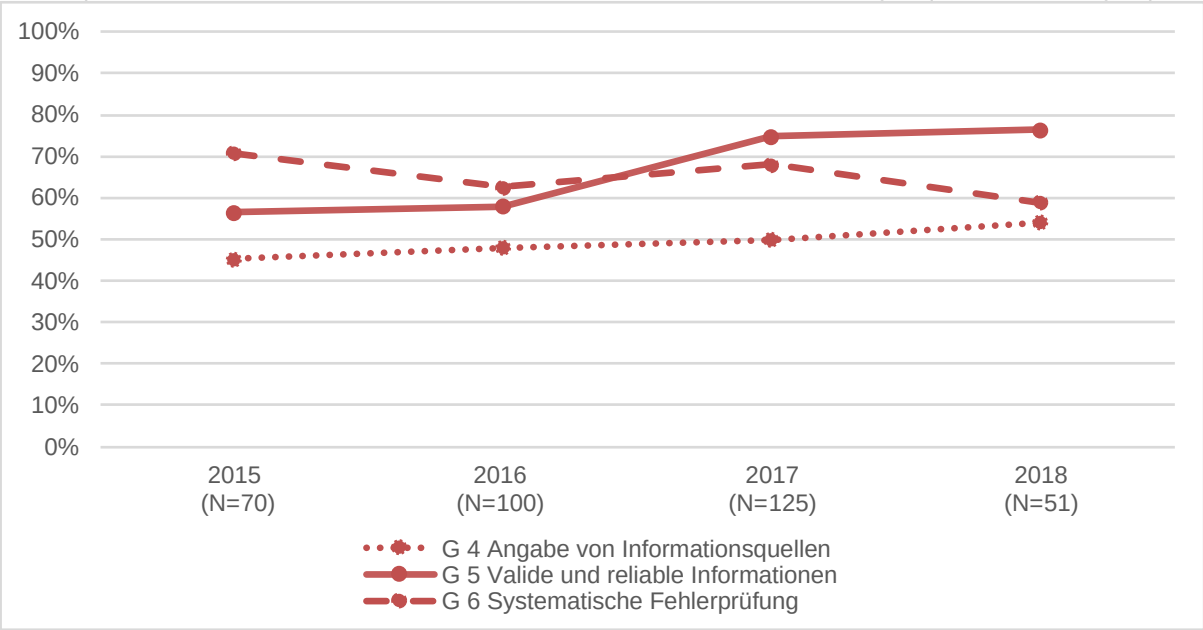


Abbildung 14: Durchschnittlicher Erfüllungsgrad der Indikatoren in den Bewertungskriterien G4 - G6 über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Das Kriterium G4, welches sich auf die Angabe von Informationsquellen bezieht, hat sich jedes Jahr positiv entwickelt, und macht in der Gesamtschau eine Entwicklung vom Basiswert 45 % in 2015 zu einem Höchstwert von 54 % in 2018 durch (siehe Abbildung 14). Diese Veränderungen sind vor allem auf Verbesserungen hinsichtlich der Existenz von vollständigen Listen der Gesprächspartner/innen und sonstigen Informationsquellen, der Darstellung der Systematik der Informationsquellen und der allgemeinen Darstellung der Verlässlichkeit der Daten des WoM zurückzuführen. Weniger Veränderungen gab es demgegenüber hinsichtlich der Darstellung der Kriterien für die Auswahl der Gesprächspartner, bzw. der Kriterien für die Ziehung der Stichproben.

Das Kriterium G5 zur Validität und Reliabilität der Informationen hat sich von einem durchschnittlichen Wert von 56 % in 2015 auf einen durchschnittlichen Wert von 76 % in 2018 entwickelt. Diesbezüglich fällt auf, dass die Werte von 2015 und 2016 relativ nah beieinanderliegen (56 % und 58%), und dann ein Sprung erfolgt, und die Werte für 2017 und 2018 wieder vergleichsweise nah beieinander liegen (75 % bzw. 76 %) (siehe Abbildung 14). Hierbei ergibt sich dieser Trend aus der Entwicklung für beide Indikatoren, die dem Bewertungskriterium zugrunde liegen. Sowohl die Datentriangulation als auch die Methodentriangulation hat sich deutlich positiv entwickelt. Hierbei hat in allen Jahren mehr Datentriangulation als Methodentriangulation stattgefunden.

Das Kriterium G6 zur systematischen Fehlerprüfung weist eine leicht rückläufige Tendenz auf, mit Schwankungen zwischen den Jahren. Der durchschnittliche Wert lag in 2015 bei 71 %, in 2016 bei 63 %, in 2017 bei 68 %, und in 2018 bei 59 % (siehe Abbildung 14). Schwankungen sind dabei bei beiden Indikatoren, die dem Bewertungskriterium zugrunde liegen, zu verzeichnen. Sowohl der Anteil der PEVs, für die eine Forschertriangulation stattgefunden hat, als auch der Anteil der PEVS, in denen eine Triangulation von Ergebnissen mit den Beteiligten und Betroffenen durchgeführt wurde, schwankt im Zeitverlauf. Die Schwankungen hinsichtlich der Forschertriangulation lassen sich darauf zurückführen, dass die Anzahl der eingesetzten Gutachter/innen zwischen den Jahren schwankt, und für die Erfüllung des Indikators zur Forschertriangulation in dieser Meta-Evaluierung der Einsatz von mindestens drei Gutachter/innen vorausgesetzt wurde. Für die Erfüllung der Indikatoren zur Triangulation von Ergebnissen wurden in der Textanalyse Anhaltspunkte z.B. auf Synthesetreffen zugrunde gelegt.

Angemessenheit der Analyse und Bewertung

Die Angemessenheit der Analyse und Bewertung in den PEV-Berichten wird mithilfe der nachfolgenden sieben Bewertungskriterien und zugeordneten Indikatoren erhoben (siehe Abbildung 15).

G 7 (A) Analyse qualitativer und quantitativer Informationen	63%	
G 7.1 (A) Die methodische Vorgehensweise beantwortet die Frage nach der Zuordnungs- und/ oder Beitragsanalyse.	24%	N=176
G 7.2 (A) Die Evaluierungsmethoden werden dargestellt.	96%	N=176
G 7.3 (A) Es gibt eine Methodenvielfalt.	100%	N=176
G 7.4 (A) Die Vor- und Nachteile der gewählten Methoden werden dargestellt.	32%	N=176
G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz	72%	
G 7.5 (B) Die Nachvollziehbarkeit der zugrunde gelegten Rahmenbedingungen und Kernprobleme der Maßnahme ist gegeben.	91%	N=176
G 7.6 (B) Die Mehrdimensionalität der Relevanz wird analysiert.	93%	N=176
G 7.7 (B) Die strategische Ausrichtung der Maßnahme an veränderten Rahmenbedingungen werden analysiert.	33%	N=176
G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität	72%	
G 7.8 (C) Die Kausalität zwischen Maßnahmen und Wirkungen wird differenziert analysiert und eingeschätzt.	48%	N=176
G 7.9 (C) Die Zielerreichung wird anhand von Modulindikatoren bewertet.	99%	N=176
G 7.10 (C) Die verwendeten Indikatoren zur Messung und Beurteilung der Zielerreichung sind SMART (spezifisch, messbar, erreichbar, relevant, zeitgebunden).	70%	N=176
G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz	40%	
G 7.11 (D) Die verschiedenen Ebenen der Effizienz einer Maßnahme werden analysiert.	65%	N=176
G 7.12 (D) Die Auswahl von Methoden und Verfahren der Effizienzmessung wird begründet.	12%	N=176
G 7.13 (D) Die Bearbeitung der Effizienz ermöglicht die Identifikation von Potenzialen zur Effizienzsteigerung.	44%	N=176
G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact	60%	
G 7.14 (E) In der Analyse und Beurteilung der übergeordneten Wirkungen (Impacts) wird die Attributionslücke analysiert.	68%	N=176
G 7.15 (E) Die Plausibilität der Hypothesen zu den intendierten langfristigen Wirkungen (Impacts) wird bewertet.	54%	N=176
G 7.16 (E) Bewertungsmaßstäbe zur Analyse und Beurteilung des Beitrages der Maßnahme zu übergeordneten Wirkungen (Impacts) sind transparent dargestellt.	60%	N=176
G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit	63%	
G 7.17 (F) Die Grenzen der Messung der Nachhaltigkeit werden beschrieben.	34%	N=176
G 7.18 (F) Die angelegten Ansätze zur Schaffung von Nachhaltigkeit werden analysiert.	75%	N=176
G 7.19 (F) Es werden mindestens zwei Ebenen der Nachhaltigkeit im Rahmen einer Prognose analysiert.	79%	N=176
G 8 Begründete Analyse und begründete Schlussfolgerung	57%	
G 8.1 Es wird zwischen Beschreibung, Analyse und Bewertung unterschieden.	50%	N=176
G 8.2 Beschreibungen und Analysen werden belegt.	45%	N=176
G 8.3 Empfehlungen werden aus der Analyse abgeleitet und sind spezifisch, realistisch und termingebunden. Sie richten sich an die wesentlichen Nutzer und ihre Umsetzung ist messbar.	46%	N=176
G 8.4 Die positiven und/ oder negativen nicht-intendierten Wirkungen der Maßnahme werden beschrieben.	85%	N=176

Abbildung 15: Durchschnittlicher Erfüllungsgrad der Indikatoren in den Bewertungskriterien G7 (A-F) - G8 in den PEV 2017 / 2018 (Quelle: Syspons, Textanalyse). Das N beschreibt die Anzahl der Berichte, für die der jeweilige Indikator bewertet wurde.

Analyse qualitativer und quantitativer Informationen in den PEV 2017 / 2018

G7 (A) Analyse qualitativer und quantitativer Informationen in den Berichten 2017 / 2018: Das siebte Bewertungskriterium befasst sich mit der Frage, inwieweit die methodische Vorgehensweise angemessen ausgewählt wurde. Dieses Kriterium bildet damit den Ausgangspunkt für eine robuste Evidenz zur Beantwortung der Evaluierungsfragen und wird durch vier Indikatoren abgebildet. Der erste Indikator erhebt, inwieweit die methodische Vorgehensweise in der PEV die Frage nach der Zuordnungs- und / oder Beitragsanalyse beantwortet (G 7.1). Weitere Voraussetzungen für eine nach fachlichen Maßstäben angemessene und systematische Analyse der Informationen sind die Darstellung der Evaluierungsmethoden (G 7.2), die Nutzung einer Methodenvielfalt (G 7.3) und die Darstellung der Vor- und Nachteile der gewählten Methoden (G.7.4).

Mit einem durchschnittlichen Erfüllungsgrad von 63 % wird das Bewertungskriterium G7 (a) **eher erfüllt** (siehe Abbildung 15). Der erste zugrundeliegende Indikator erfasst, ob in der Erläuterung des methodischen Designs adressiert wird, ob ein theoriebasierter Ansatz, eine Kontributionsanalyse oder ein experimentelles oder quasi-experimentelles Evaluationsdesign angewandt wurde (G 7.1). Dies ist in 24 % der Berichte (n= 43) der Fall. Von den PEV für die dies gegeben ist, wenden 20 % (n= 36) eine Kontributionsanalyse an, während 4 % (n= 7) einen theoriebasierten Ansatz verfolgen. In keinem der Berichte wird auf eine experimentelles oder quasi-experimentelles Design verwiesen.

Nahezu alle PEV-Berichte (96 %, n= 169) stellen die verwendeten Evaluierungsmethoden dar. Dies beschränkt sich jedoch meist auf eine allgemeine Auflistung der Methoden. In etwa einem Fünftel der Berichte (23 %, n= 41) wird darüber hinaus deutlich, wie die Methoden im Prozess der Datenerhebung eingesetzt wurden. In mehreren Fällen wurde bspw. tabellarisch dargestellt, welche Quellen für welches Evaluierungskriterium herangezogen wurden. Nur vereinzelt gehen die Berichte hingegen darauf ein, wie die erhobenen Daten dokumentiert wurden (2 %, n= 4) oder wie erhobene qualitative Daten ausgewertet wurden (2 %, n= 3). Eine Darstellung der Vorgehensweise zur Auswertung quantitativer Daten findet sich in keinem der Berichte (0%, n=0).

Eine Methodenvielfalt, also die Anwendung von mindestens zwei Methoden, ist in allen geprüften Berichten gegeben (100 %, n= 176). Am häufigsten kommen hierbei die Dokumentenauswertung (99 %, n= 174) sowie Interviews (100 %, n= 176) zur Anwendung. In etwas weniger als der Hälfte der Fälle (43 %, n= 76) erfolgt eine eigene Auswertung von Monitoring-Daten durch die Prüfmission, und in einem Fünftel der Fälle (21 %, n= 37) wird mit Fokusgruppen gearbeitet. Eine Befragung hingegen kommt nur in vergleichsweise wenigen PEVs (7%, n=13) zum Einsatz. Hinsichtlich der methodischen Vorgehensweise legen 32 % (n= 56) die Vor- und Nachteile der gewählten Methoden dar.

Trends hinsichtlich der Analyse qualitativer und quantitativer Informationen

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (A): Das Bewertungskriterium zur Analyse qualitativer und quantitativer Informationen hat sich im Zeitverlauf fast jedes Jahr minimal positiv entwickelt. In der Gesamtschau über vier Jahre hinweg lässt sich eine positive Entwicklung von 5 Prozentpunkten feststellen, von einem Durchschnittswert von 58 % in 2015 auf einen Durchschnittswert von 63 % in 2018 (siehe Abbildung 16).

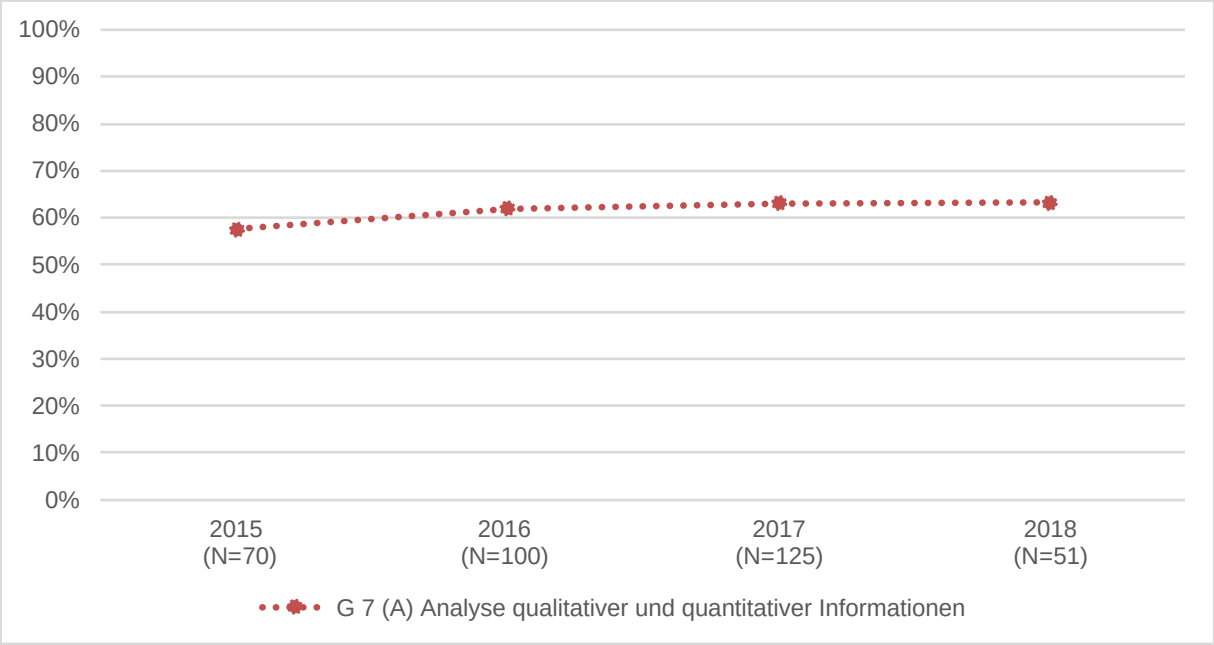


Abbildung 16: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (A) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Auf Indikatorenebene lassen sich die größten Entwicklungen hinsichtlich der Auseinandersetzung mit der Frage nach der Zuordnungs- und Beitragsanalyse und hinsichtlich der Darstellung der Evaluierungsmethoden festmachen. Kaum Veränderungen gab es hinsichtlich der Methodenvielfalt; in allen Jahren wurden nahezu flächendeckend mindestens zwei Methoden angewandt. Die Werte hinsichtlich der Darstellung der Vor- und Nachteile der gewählten Methoden schwanken von Jahr zu Jahr.

Methodische Bearbeitung der Relevanz in den PEV 2017 / 2018

G7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz in den Berichten 2017 / 2018: Das erste der Bewertungskriterien zur methodischen Bearbeitung der OECD-DAC Kriterien befasst sich mit der Relevanz. Auf Ebene der Indikatoren wurde hierzu zunächst erhoben, ob die Nachvollziehbarkeit der zugrunde gelegten Rahmenbedingungen und Kernprobleme gegeben ist (G 7.5). Weiterhin wurde erfasst, ob die Mehrdimensionalität der Relevanz erfasst wird (G 7.6) und ob die strategische Ausrichtung an veränderten Rahmenbedingungen in den PEV-Berichten analysiert wurden (G 7.7).

Im Durchschnitt aller Berichte erreicht das Bewertungskriterium G 7 (B) 72 % und ist damit **eher erfüllt** (siehe Abbildung 15). Das Relevanzkriterium schneidet damit, zusammen mit dem Effektivitätskriterium, am besten ab bei der Bearbeitung der OECD-DAC-Kriterien. Die beiden ersten Aspekte der Relevanz, die Nachvollziehbarkeit der dargestellten Rahmenbedingungen und die Berücksichtigung der Mehrdimensionalität der Relevanz, sind in neun von zehn Berichten gegeben. Die Auseinandersetzung mit den Rahmenbedingungen (G 7.5 B, 91 %, n= 161)) ist dabei von Bedeutung, um zu analysieren, ob ein Vorhaben sich mehrere Jahre nach seiner ursprünglichen Konzeption noch an aktuellen entwicklungspolitischen Fragestellungen orientiert. Hinsichtlich der Mehrdimensionalität der Relevanz (G 7.6 B, 93 %, n= 163) wurde untersucht, ob bei der Bearbeitung dieses OECD-DAC Kriteriums sowohl die Kernbedarfe von Zielgruppe und Partner als auch des Auftraggebers berücksichtigt werden. Dies erfolgt in der Regel durch einen Abgleich der Ziele und Strategien des Vorhabens mit länder- und regionalspezifischen Vorgaben des BMZ sowie mit Politik- und Strategiedokumenten der Partnerländer wie z.B. Armutsstrategien. Den Anspruch des dritten Indikators, die Ausrichtung des Vorhabens an Veränderungen in den Rahmenbedingungen zu untersuchen (G 7.7 B), erfüllen ein Drittel der geprüften PEV-Berichte (33 %, n= 58).

Trends hinsichtlich der methodischen Bearbeitung der Relevanz

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (B): Das Bewertungskriterium zur methodischen Bearbeitung der Relevanz hat sich zwischen den Jahren 2015 und 2018 jedes Jahr positiv entwickelt. Ausgehend von einem durchschnittlichen Wert von 65 % in der ersten Metaevaluierung 2015 erfolgte bis zur letzten Metaevaluierung 2018 eine Steigerung auf einen Durchschnittwert von 73 % (siehe Abbildung 17).

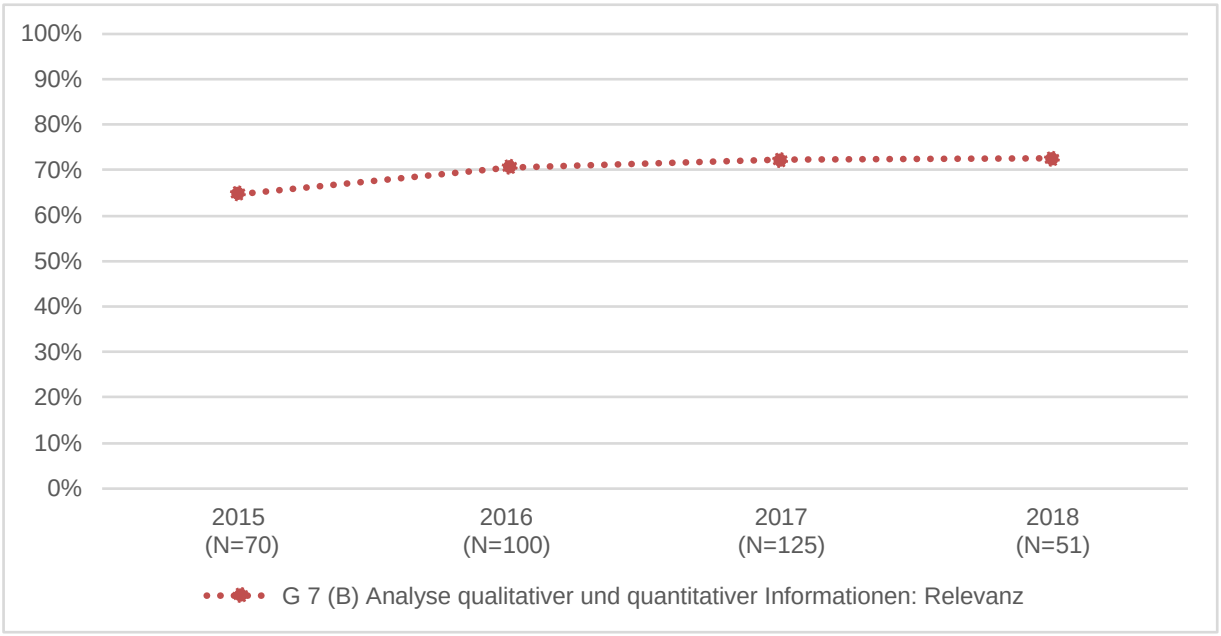


Abbildung 17: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (B) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Blickt man auf die drei Indikatoren zum Bewertungskriterium, so lässt sich dieser Trend vor allem auf eine positive Entwicklung des ersten Indikators zurückführen. Die Nachvollziehbarkeit der zugrunde gelegten Rahmenbedingungen und Kernprobleme der Maßnahme hat sich über die Jahre stetig verbessert.

Die Benotung des Kriteriums Relevanz zeigt kaum Veränderungen im Vergleich zu vergangenen Jahren und verbleibt auf einem sehr hohen Niveau. Dies lässt aus Sicht des Gutachterteams die Vermutung zu, dass ggf. die Fragen, mit denen die Relevanz von Vorhaben in den PEV analysiert und bewertet werden, keine ausreichend kritische Auseinandersetzung ermöglichen. Es ist zu vermuten, dass mit der gemeinsamen Verfahrensreform von BMZ und GIZ (GVR) eine kritischere Auseinandersetzung mit dem Relevanzkriterium möglich sein könnte, da in den in der GVR eingeführten Formaten deutlicher herausgearbeitet wird, wann und warum ein Vorhaben entwicklungspolitisch relevant ist. Mit der GVR ergibt sich zum Beispiel die Relevanz eines Vorhabens in erster Linie durch die Beiträge des Moduls zum übergeordneten strategischen Bezugsrahmen (insbesondere den Programmindikatoren sowie die Komplementarität und die Synergien mit anderen Vorhaben im Programm). Zudem ist nachgeordnet zur entwicklungspolitischen Relevanz für das BMZ die Relevanz auch auf Grundlage des spezifischen entwicklungspolitischen Problems zu analysieren, wobei hier ein Fokus für die Relevanz auf folgende Aspekte liegt: (1) die Veränderbarkeit des entwicklungspolitischen Problems, (2) das Aufsetzen auf bereits erreichten Outcomes sowie (3) die Fähigkeit der Vorhaben bei verändertem Kontext umzusteuern.

Methodische Bearbeitung der Effektivität in den PEV 2017 / 2018

G7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität in den Berichten 2017 / 2018: Ausgehend von den Anforderungen an eine methodisch fundierte PEV sollen diese eine Analyse und Bewertung der Effektivität der Vorhaben vornehmen. Operationalisiert in drei Indikatoren wurde zunächst geprüft, ob die Kausalität zwischen Maßnahme und Wirkung differenziert analysiert wurde (G 7.8 C). Weiterhin wurde mit zwei Indikatoren gemessen, ob die Zielerreichung anhand von Modulzielindikatoren gemessen wurde (G 7.9 C) und die verwendeten Indikatoren die SMART-Qualitätskriterien erfüllen (G 7.10 C).

Über alle Berichte hinweg wird das Bewertungskriterium G 7 (C) durchschnittlich mit 72 % **eher erfüllt** (siehe Abbildung 15). Hierbei analysieren knapp die Hälfte der Berichte (48 %, n= 84) explizit oder implizit, ob sich beobachtete Veränderungen auf das evaluierte Vorhaben zurückführen lassen. Die Auseinandersetzung mit der Kausalität und der Ausschluss von Drittvariablen (G 7.8 C) erfolgt demnach nicht flächendeckend. Hingegen zeigen die Ergebnisse zum zweiten Indikator, dass die Bewertung der Zielerreichung nahezu flächendeckend (99 %, n= 174) anhand der Modulzielindikatoren erfolgt. Weniger durchgesetzt hat sich demgegenüber die Auseinandersetzung der Prüfer mit der Qualität der Indikatoren (G 7.10 C), sie erfolgt in zwei Dritteln der PEV-Berichte (70 %, n= 124). Ein Anspruch für die Erfüllung des entsprechenden Indikators der Metaevaluierung war, dass dort wo die Evaluatoren Qualitätsmängel zu den Indikatoren feststellen, eine Neuformulierung oder Anpassung der Indikatoren entsprechend der SMART-Qualitätskriterien erfolgt. Die Bearbeitung der Effektivität weist somit sowohl hinsichtlich der stringenten Verwendung von SMARTen Indikatoren sowie mit Blick auf eine konsequente Auseinandersetzung mit der Kausalität zwischen Maßnahmen und Wirkungen Optimierungspotenziale auf.

Trends hinsichtlich der methodischen Bearbeitung der Effektivität

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (C): Das Bewertungskriterium zur methodischen Bearbeitung der Effektivität hat sich insgesamt positiv entwickelt, allerdings gab es zwischen 2017 und 2018 eine leicht rückläufige Entwicklung. Ausgehend von einem durchschnittlichen Wert von 56 % in 2015 steigert sich das Abschneiden der PEV in der Metaevaluierung in diesem Kriterium in 2016 auf 69 % in 2016 und auf 74 % in 2017, um dann in 2018 wieder bei 69 % zu liegen (siehe Abbildung 18).

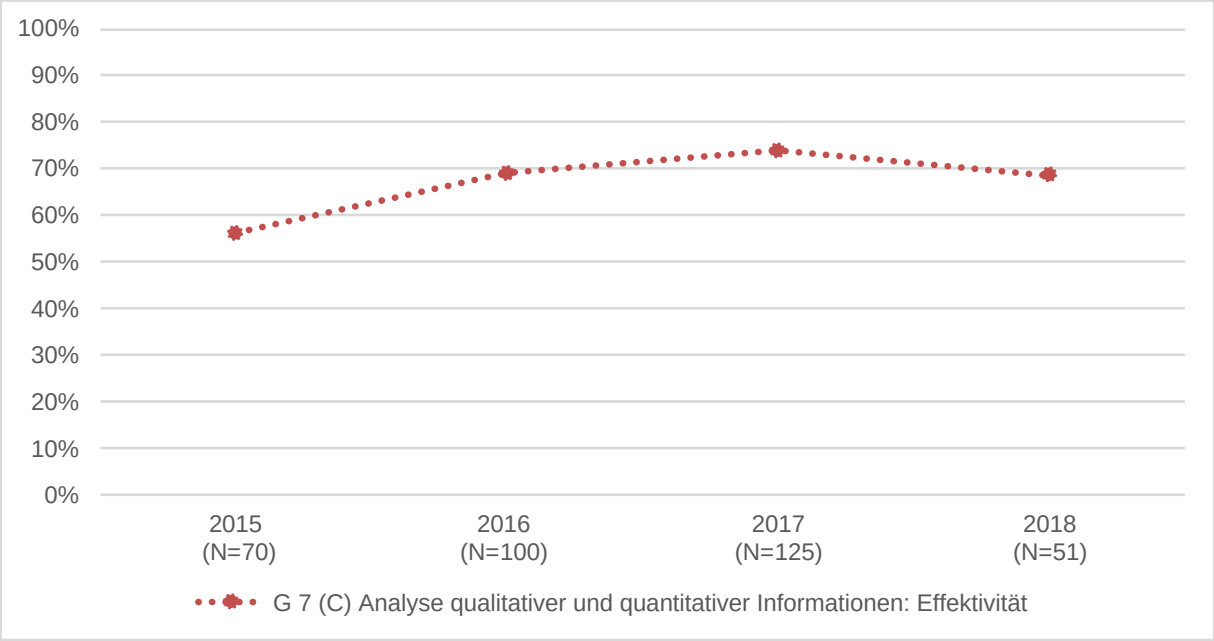


Abbildung 18: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (C) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Ein vertiefter Blick in die Ergebnisse zeigt, dass sich die Auseinandersetzung mit der Kausalität zwischen Maßnahmen und Wirkungen insgesamt leicht positiv entwickelt hat, mit einer leicht rückläufigen Entwicklung in 2018. Je nach Jahr erfüllen zwischen vier und fünf von zehn Berichten diesen methodischen Anspruch. Die Bewertung der Zielerreichung anhand von Modulzielindikatoren ist in allen Jahren nahezu flächendeckend gegeben. Hinsichtlich der Einhaltung von SMART-Qualitätskriterien für die Indikatoren, die zur Beurteilung der Zielerreichung herangezogen werden, konnte zwischen der Metaevaluierung 2015 und 2016 ein bedeutender Qualitätssprung festgestellt werden. Der Durchschnittswert ist hier innerhalb eines Jahres von 36 % auf 68 % gestiegen. In den beiden darauffolgenden Jahren gibt es nur noch minimale Veränderungen gegenüber dem Wert von 2016.

Methodische Bearbeitung der Effizienz in den PEV 2017 / 2018

G7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz in den Berichten 2017 / 2018: Dieses Bewertungskriterium erfasst die methodische Angemessenheit der Effizienzbewertung. Hierfür werden drei Indikatoren herangezogen. Zunächst wird untersucht, ob mindestens eine Ebene der Effizienz (Implementierungs-, Produktions- oder Allokationseffizienz) analysiert (G 7.11 D) und die Auswahl von Methoden und Verfahren der Effizienzmessung begründet (G 7.12 D) wurde. Darüber hinaus wurde erfasst, ob die Bearbeitung der Effizienz die Identifikation von Potenzialen zur Effizienzsteigerung ermöglicht (G 7.13).

Insgesamt wird das Bewertungskriterium G 7 (D) **bedingt erfüllt**; im Durchschnitt sind 40 % der Indikatoren zu diesem Kriterium erfüllt (siehe Abbildung 15). Damit ist die Bearbeitung der Effizienz das OECD-DAC-Kriterium, dessen Bearbeitung in der Metaevaluierung am schwächsten abschneidet. Innerhalb des Kriteriums schneidet der erste Indikator, der misst, ob mindestens eine Ebene der Effizienz analysiert wurde, am besten ab. Er ist im Mittel zu 65 % (n= 114) erfüllt. Ein differenzierter Blick in die Auswertung zeigt, dass zumeist die Implementierungseffizienz analysiert wird (57 %, n= 101). Die Produktionseffizienz wird demgegenüber deutlich weniger betrachtet (14 %, n= 25), und die Allokationseffizienz am wenigsten häufig (4 %, n= 7). Dort wo die Produktions- und Allokationseffizienz untersucht wurde, beschränkt sich die Analyse auf eine gutachterliche Einschätzung zum Verhältnis von eingesetzten Mitteln zu Outputs bzw. Wirkungen. Eine zahlengestützte Analyse, bei der ein Vergleich zu anderen Vorhaben erfolgt, ist in keinem der Berichte gegeben.

Der zweite Indikator, der erfasst, ob die Auswahl von Methoden und Verfahren der Effizienzmessung begründet wurde (G 7.14), ist lediglich für 12 % (n= 21) der Berichte erfüllt. In diesen Fällen wird zumeist erläutert, warum vornehmlich auf deskriptive Methoden (11 %, n= 19) zurückgegriffen wurde. In wenigen Fällen (5%, n= 8) wird zusätzlich der Einsatz von Level 1-Methoden erläutert. Level 2-Methoden kommen hingegen in keiner der geprüften PEVS zum Einsatz (0 %, n= 0). Die Identifikation von Potenzialen zur Effizienzsteigerung ist hingegen in knapp der Hälfte der Berichte (44 %, n= 78) gegeben. Hierbei handelt es sich in der Regel um gutachterliche Einschätzungen, in denen auf Potenziale zur Verbesserung der Implementierungseffizienz hingewiesen wird. Vereinzelt wird in den gutachterlichen Einschätzungen auch formuliert, dass sich durch ein verstärktes Augenmerk auf bestimmte Handlungsfelder, durch die Akquise von Kofinanzierungen oder durch die Anpassung des Personalkonzepts das Verhältnis von Inputs zu erbrachten Leistungen bzw. erzielten Wirkungen steigern ließe.

Übergeordnet lässt sich feststellen, dass die Bearbeitung der Effizienz durch die PEV-Teams weitestgehend auf Basis subjektiver Einschätzungen erfolgt. Auch die wenigen Berichte, die den Anspruch haben, über eine Analyse der Implementierungseffizienz hinauszugehen, tun dies nur auf deskriptiver Ebene. In einigen wenigen Berichten werden Zahlen dazu aufgeführt, wie viele Mittel in bestimmte Handlungsfelder fließen. Hierzu erfolgt in diesen Fällen aber entweder gar keine Analyse, oder eine gutachterliche Einschätzung der Angemessenheit der Kosten ohne Einordnung zu Vergleichsgrößen oder Kosten für alternative Handlungsoptionen In den Berichten werden vereinzelt die Stichwort Allokations- oder Produktionseffizienz genannt, bzw. es wird auf das Verhältnis von Input zu Output / Outcome eingegangen, In diesen Fällen stehen hierzu jedoch lediglich 1 – 3 Sätze Text, die eine gutachterliche Einschätzung ohne quantitative Datengrundlage darstellen. Als Begründungen für die Einschätzungen werden Erfahrungswerte der Gutachter oder schlüssige Darlegungen des Projektteams genannt. In wenigen „good practice“ Fällen werden konkrete Ansätze des Vorhabens analysiert, bspw. Scaling Up-Strategien oder das Instrumentenkonzept, die sich nach gutachterlicher Einschätzung auf das Verhältnis zwischen eingesetzten Mitteln und erbrachten Leistungen bzw. erzielten Wirkungen auswirken.

Trends hinsichtlich der methodischen Bearbeitung der Effizienz

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (D): Das Bewertungskriterium zur methodischen Bearbeitung der Effizienz weist eine positive Entwicklung auf. Das durchschnittliche Abschneiden in der Metaevaluierung hat sich von einem Wert von 28 % in 2015 auf einen Wert von 39 % gesteigert. In der Gesamtschau hat zeigt sich demnach einer Verbesserung von 11 Prozentpunkten, auch wenn es zwischen 2017 und 2018 eine minimal rückläufige Entwicklung von 2 Prozentpunkten gab (siehe Abbildung 19). Unabhängig von dieser positiven Entwicklung schneidet die Bearbeitung der Effizienz von allen Kriterien der Metaevaluierung, die sich auf die Bearbeitung der OECD-DAC-Kriterien beziehen, in allen Jahren am schwächsten ab.

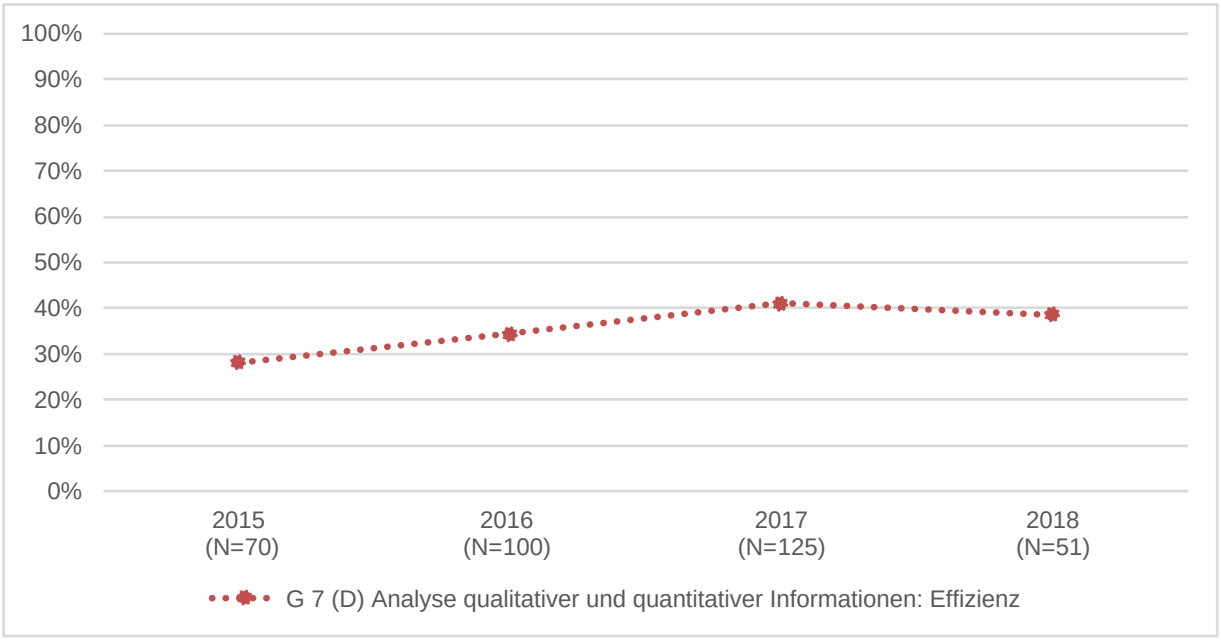


Abbildung 19: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (D) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Die Verbesserung im Kriterium zur methodischen Bearbeitung der Effizienz lässt sich vor allem auf Verbesserungen hinsichtlich der Identifikation von Potenzialen zur Effizienzsteigerung zurückführen. Hinsichtlich der Bearbeitung von verschiedenen Ebenen der Effizienz und der Begründung der genutzten Methoden und Verfahren lassen sich nur unwesentliche Entwicklungen feststellen.

Mit Blick auf die Beschränkung der PEV auf deskriptive Methoden ist festzustellen, dass die Vorhaben der GIZ den PEV-Teams bisher in der Regel keine quantitativen Daten für eine Zuordnung von Kosten zu Outputs zur Verfügung stellen konnten. Durch die GVR ändert sich dies und zukünftig werden Evaluatoren Informationen darüber bereitgestellt werden können, wie viel Budget in welche Outputs fließt. Bei größeren Budgetabweichungen werden auch Informationen darüber vorliegen, wie diese Abweichungen zustande gekommen sind. Sollte ein entsprechendes Wissensmanagement durch die GIZ hierzu betrieben werden, stünden den Evaluatoren zudem ggf. quantitative Daten aus Vorhaben mit ähnlichen Handlungsbereichen zur Verfügung, um einen Vergleich von Kosten für die Erbringung bestimmter Leistungen anzustellen. Hierdurch könnten sich die Benotungen in Zukunft ggf. weniger auf gutachterliche Einschätzungen und verstärkt auf quantitative und vergleichende Effizienzanalysen stützen.

Methodische Bearbeitung des Impacts in den PEV 2017 / 2018

G7 (E) Analyse qualitativer und quantitativer Informationen: Impact in den Berichten 2017 / 2018: Das Bewertungssystem, welches untersucht, inwiefern in den PEV-Berichten eine methodisch fundierte Bewertung des Impacts der evaluierten Vorhaben erfolgt, wird ebenfalls durch drei Indikatoren operationalisiert. Der erste Indikator untersucht, ob bei der Analyse und Bewertung des Impacts die Attributionslücke (auch “Zuordnungslücke” oder “Systemgrenze”) adressiert wurde (G 7.14). Die anderen beiden Indikatoren messen, ob in der Berichterstattung die Plausibilität der Wirkungshypothesen bewertet wird (G 7.15) und ob die Bewertungsmaßstäbe zur Analyse und Beurteilung des Impacts transparent dargestellt sind.

Im Mittel der geprüften PEV-Berichte ist das Bewertungskriterium G 7 (E) mit 60 % **eher erfüllt** (siehe Abbildung 15). Die methodische Bearbeitung des Impacts schneidet somit schwächer ab als die Bearbeitung von Relevanz und Effektivität, aber besser als die Bearbeitung von Effizienz. Im Folgenden wird auf das Abschneiden der einzelnen Indikatoren, die dem Bewertungskriterium zum Impact zugrunde liegen, eingegangen. 68 % (n= 119) der Berichte adressieren die Attributionslücke zwischen den Maßnahmen des Vorhabens und den übergeordneten Wirkungen (G 7.14) implizit oder explizit. Als impliziter Verweis auf die Zuordnungslücke wurden dabei in der Textanalyse Formulierungen gewertet, die darauf verweisen, dass ein Vorhaben lediglich einen Beitrag zu den beschriebenen Veränderungen leistet. Etwas über die Hälfte der Berichte (54 %, n= 95) setzt sich mit der Nachvollziehbarkeit der Hypothesen zu den langfristigen Wirkungen auseinander (G 7.15). Darüber hinaus stellen sechs von zehn (60%, n= 105) der Berichte die Bewertungsmaßstäbe für die Beurteilung des Impacts dar (G 7.16). Dies können beispielsweise die Programmindikatoren sein, der Beitrag zu SDGs oder zu einem entwicklungspolitischen Bezugsrahmen auf nationaler Ebene. Zum Teil erfolgt hierbei anstatt einer quantitativen Analyse von Indikatoren eine deskriptive Einordnung, die dadurch begründet wird, dass zum Zeitpunkt der Evaluierung noch keine Wirkungen gemessen werden können. Insgesamt ist mit Blick auf das Abschneiden der drei Indikatoren noch Potenzial für die Verbesserung der methodischen Bearbeitung des Impact-Kriteriums durch von der GIZ beauftragte Evaluatoren gegeben.

Trends hinsichtlich der methodischen Bearbeitung des Impacts

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (E): Das Bewertungskriterium zur methodischen Bearbeitung des Impacts hat sich positiv entwickelt, zwischen der ersten Metaevaluierung von 2015 und der letzten Metaevaluierung von 2018 hat sich das durchschnittliche Abschneiden der geprüften Berichte von 49 % auf 59 % gesteigert. Blickt man auf die Entwicklung in allen vier Jahren, so lässt sich der größte Sprung zwischen der Metaevaluierung 2015 und der Metaevaluierung 2016 feststellen. Zwischen diesen beiden Jahren ist eine Verbesserung von 8 Prozentpunkten zu verzeichnen, von 49 % auf 57 % (siehe Abbildung 20).

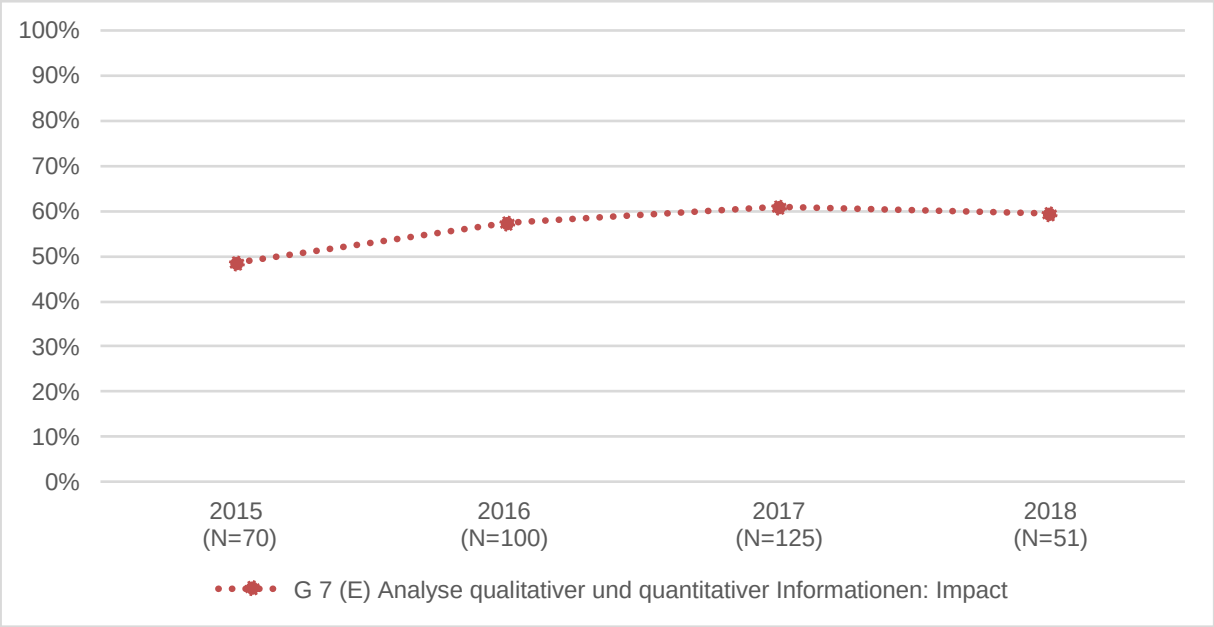


Abbildung 20: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (E) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Analysiert man die Durchschnittswerte für die zugrundeliegenden Indikatoren, so zeigt sich, dass der positive Trend vor allem auf die Entwicklung bei einem der drei Indikatoren zurückzuführen ist. Hinsichtlich der Darstellung der Bewertungsmaßstäbe zur Analyse und Beurteilung des Beitrags der Maßnahme zu den übergeordneten Wirkungen (G7.16) lässt sich eine deutliche Verbesserung feststellen. Am ausgeprägtesten fällt diese zwischen den Jahren 2015 und 2016 aus. Hinsichtlich der Bewertung der Plausibilität der Wirkungshypothesen sind nur leichte Veränderungen zu verzeichnen. Mit Blick auf die Auseinandersetzung mit der Zuordnungslücke schließlich sind die Veränderungen unwesentlich.

Methodische Bearbeitung der Nachhaltigkeit in den PEV 2017 / 2018

G7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit in den Berichten 2017 / 2018: Das fünfte Kapitel zur Bearbeitung der OECD-DAC-Kriterien untersucht die methodische Qualität der Nachhaltigkeitskapitel. Hierfür werden drei Indikatoren herangezogen. Zunächst wurde untersucht, ob sich die Prüfteams mit den Grenzen der Messung der Nachhaltigkeit auseinandersetzen (G 7.17 F). Weiterhin wurde erfasst, ob die Berichte die angelegten Ansätze zur Schaffung von Nachhaltigkeit darlegen (G 7.18 F), und ob mindestens zwei Ebenen der Nachhaltigkeit analysiert wurden (G 7.19 F). Als mögliche Analyseebenen wurden hierbei die finanzielle, institutionelle, personelle, soziale, technologische und ökologische Nachhaltigkeit in Betracht gezogen.

Das Kriterium G 8 wird im Durchschnitt der Berichte zu 63 % und damit **eher erfüllt** (siehe Abbildung 15). Der erste Indikator schneidet hierbei mit Abstand am schwächsten ab. In knapp einem Drittel der Berichte (34 %,

n= 60) werden die Grenzen der Messung der Nachhaltigkeit beschrieben. Dies ist gegeben, wenn der Bericht sich damit auseinandersetzt, inwiefern zum Zeitpunkt der Evaluierung eine ergebnisorientierte Analyse und Beurteilung von Nachhaltigkeit überhaupt möglich ist (G 7.17). Die anderen beiden Indikatoren zur Bearbeitung der Nachhaltigkeit sind in knapp drei Viertel der Berichte erfüllt, sie schneiden somit wesentlich besser ab. So werden in 75 % der Berichte (n= 132) die angelegten Ansätze zur Schaffung der Nachhaltigkeit analysiert, indem bspw. auf Exit- oder Nachhaltigkeitsstrategien des Vorhabens eingegangen wird oder reflektiert wird, wie auf die Anschlussfähigkeit der Maßnahmen und die Verantwortungsübergabe an Partner hingewirkt wird (G 7.18 F). Darüber hinaus werden in 79 % (n= 139) der PEVs für mindestens zwei Ebenen der Nachhaltigkeit eine Prognose formuliert (G 7.19 F).

Trends hinsichtlich der methodischen Bearbeitung der Nachhaltigkeit

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G7 (F): Das Bewertungskriterium zur methodischen Bearbeitung der Nachhaltigkeit hat sich minimal positiv entwickelt. Es hat sich von einem durchschnittlichen Wert von 61 % in den Metaevaluierungen 2015 und 2016 auf 63 % in den Metaevaluierungen 2017 und 2018 gesteigert (siehe Abbildung 21).

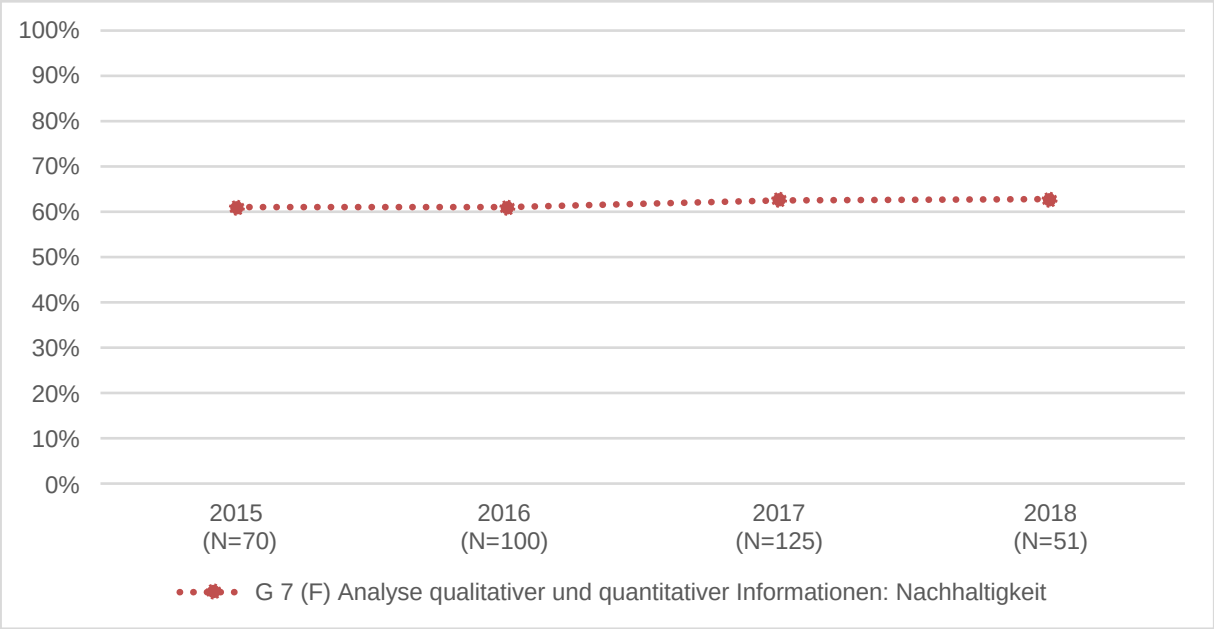


Abbildung 21: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G7 (F) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

Ein Blick auf die Indikatoren zeigt, dass es hinsichtlich der Beschreibung der Grenzen der Nachhaltigkeit in den PEV-Berichten insgesamt eine leicht rückläufige Tendenz gibt. Für die anderen beiden Indikatoren, die Analyse angelegter Ansätze zur Schaffung von Nachhaltigkeit sowie die Formulierung von Prognosen zur Nachhaltigkeit, gibt es eine minimal positive Tendenz.

Begründete Analyse und Schlussfolgerungen in den PEV 2017 / 2018

G8: Begründete Analyse und Schlussfolgerungen in den Berichten 2017 / 2018: Das Bewertungskriterium G8 untersucht, inwiefern Analysen und Schlussfolgerungen der PEV nachvollziehbar begründet werden. Dies wird anhand von vier Indikatoren erfasst. Der erste Indikator prüft, ob zwischen Beschreibung, Analyse und Bewertung unterschieden wurde (G 8.1). Weiterhin wurde untersucht, ob Beschreibungen und Analysen belegt werden (G 8.2). Ebenfalls erfasst wurde, ob die Empfehlungen den methodischen Anforderungen gerecht wer-

den (G 8.3) und ob nicht-intendierte (positive oder negative) Wirkungen beschrieben oder zumindest ausgeschlossen werden (G 8.4).

Das Bewertungskriterium G 8 wird im Durchschnitt der Berichte zu 57 % und damit **bedingt erfüllt** (siehe Abbildung 15). Die ersten drei zugrundeliegenden Indikatoren schneiden ähnlich ab und sind ca. zur Hälfte erfüllt, während der letzte Indikator deutlich besser abschneidet. So wird in 50 % der Berichte (n= 88) der Dreischritt zwischen Beschreibung, Analyse und Bewertung eingehalten (G 8.1). Für die Nachvollziehbarkeit des zugrunde gelegten Wirkungsmodells sowie der Ergebnisse der Evaluation ist es wichtig, dass Beschreibung, Analyse und Bewertung nicht vermengt werden. Zugleich sollen Beschreibungen und Analysen durch Quellen belegt werden, um die Belastbarkeit zu gewährleisten. Diesem Anspruch (G 8.2) kommen 45 % der PEVs (n= 79) weitestgehend nach. Die Ergebnisse des dritten Indikators zeigen auf, dass 46 % der PEV-Berichte (n= 81) Empfehlungen aus der Analyse ableiten, diese spezifisch und realistisch formulieren sowie die Adressaten der Empfehlungen benennen (G 8.3). Der letzte Indikator zum Bewertungskriterium G 8 schließlich ist in 85 % der Fälle (n= 150) erfüllt. Er erfasst, ob auf die positiven und / oder negativen nicht-intendierten Wirkungen der Maßnahme eingegangen wurde. In den meisten Berichten, die diesen Indikator erfüllen, wurde dabei explizit darauf verwiesen, dass keine nicht-intendierten negativen Wirkungen festgestellt wurden.

Trends hinsichtlich begründeter Analyse und Schlussfolgerungen

Trends in der Qualität der dezentralen PEV 2015 – 2018 für das Bewertungskriterium G8: Das Bewertungskriterium, das prüft, inwiefern begründete Analysen und Schlussfolgerungen vorliegen, hat sich zwischen den Jahren 2015 und 2018 in der Metaevaluierung kontinuierlich positiv entwickelt. In der Gesamtschau lässt sich eine Verbesserung von 11 Prozentpunkten verzeichnen, von einem durchschnittlichen Wert von 50 % in 2015 auf einen durchschnittlichen Wert von 61 % in 2018.

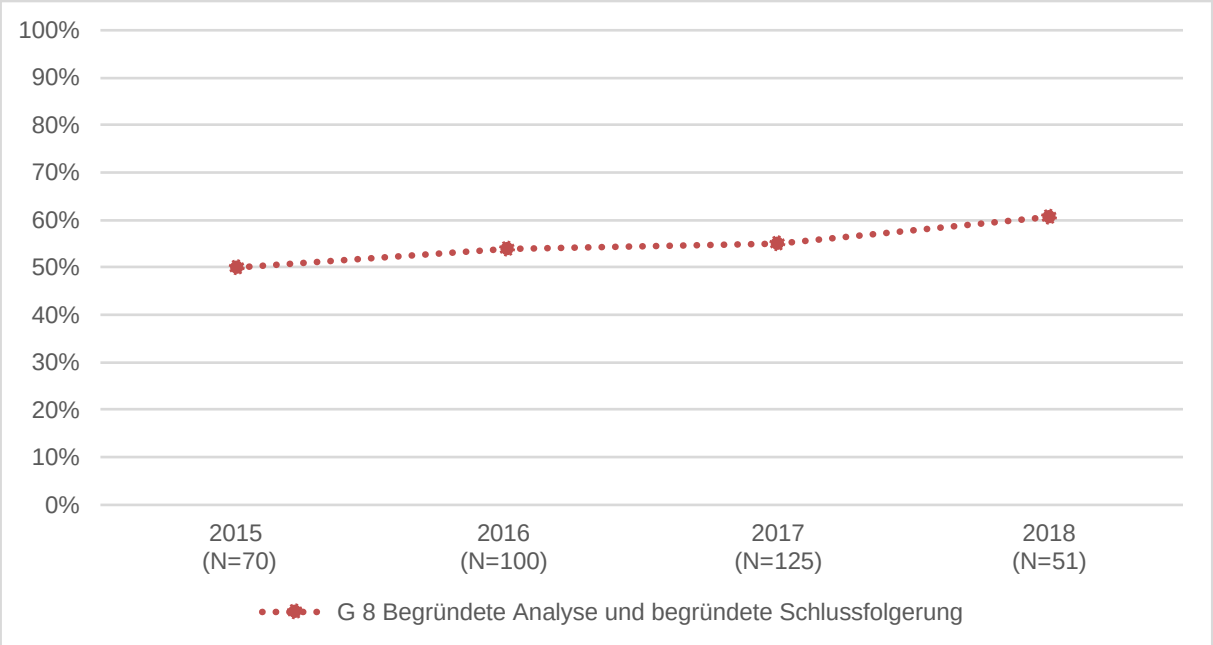


Abbildung 22: Durchschnittlicher Erfüllungsgrad der Indikatoren im Bewertungskriterium G8 über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Grundgesamtheit im jeweiligen Zeitraum

Ein vertiefter Blick in die Ergebnisse zeigt, dass die positive Entwicklung vor allem auf Verbesserungen bei zwei der vier Indikatoren zum Bewertungskriterium zurückzuführen ist. Die Meta-Evaluierungen zeigen, dass sich vor allem die Unterscheidung zwischen Beschreibung, Analyse und Schlussfolgerungen verbessert hat, sowie die Angaben von Quellen zu Beschreibungen und Analysen. Weniger Veränderungen gab es demgegenüber hinsichtlich der methodischen Qualität der Empfehlungen und der Beschreibung von positiven und / oder negativen nicht-intendierten Wirkungen.

2.6 Sonderauswertung kontributionsanalytische Qualitätsaspekte und Nachvollziehbarkeit

Ein besonderes Erkenntnisinteresse liegt in der diesjährigen Meta-Evaluierung auf kontributionsanalytischen Qualitätsaspekten und der Nachvollziehbarkeit von Ergebnissen. Für die Darstellung des ersten Aspekts wird in diesem Kapitel zunächst gesondert auf einzelne Indikatoren eingegangen, von denen drei bereits in den vergangenen Meta-Evaluierungen erfasst wurden, und von denen einer 2017 / 2018 erstmalig erhoben wurde. Für die Indikatoren, die bereits in den vergangenen Jahren identisch erhoben wurden, werden Trends dargestellt. Darüber hinaus werden zum Aspekt der Nachvollziehbarkeit der Ergebnisse die Werte zu fünf neuen Indikatoren dargestellt, die in der diesjährigen Meta-Evaluierung erstmalig erfasst wurden.

Kontributionsanalytische Qualitätsaspekte

Die Berücksichtigung kontributionsanalytischer Qualitätsaspekte wurde in den Meta-Evaluierungen 2015 – 2018 durch fünf Indikatoren erfasst, die sich jeweils auf die Kapitel zur Beschreibung des Evaluierungsgegenstands (G 1.2 und G 1.3), zur methodischen Vorgehensweise (G 7.1(A)), zur Effektivität (G 7.8 (C)) und zu den Impacts (G 7.14 (E) beziehen. Erstmalig erfasst wurde 2017 / 2018 ein Indikator zur Bewertung der untersuchten Zusammenhänge und Beiträge des Vorhabens in den Schlussfolgerungen (Z 8). Abbildung 23 zeigt die Werte für die Indikatoren differenziert nach Jahren.

Diese Indikatoren wurden aufgrund ihres Stellenwerts für die Kontributionsanalyse ausgewählt. Hierbei wird zu Grunde gelegt, dass ein Verständnis der intendierten Ziele und Wege zur Zielerreichung (G 1.2 und G 1.3) eine Voraussetzung dafür ist, dass untersucht werden kann, inwiefern Ziele wie geplant erreicht werden konnten. Weiterhin sollten Evaluatoren sich übergeordnet damit auseinandersetzen, inwiefern es möglich ist, angesichts der Rahmenbedingungen des Vorhabens und mit dem gewählten Evaluationsdesign eingetretene Veränderungen ursächlich auf das Vorhaben zurückzuführen (G 7.1 (A)). In der Analyse sollte dann sowohl für die Auseinandersetzung mit der Erreichung des Modulziels ((G 7.8 (C)) als auch mit den übergeordneten entwicklungspolitischen Wirkungen (G 7.14 (E)) differenziert analysiert werden, wie das Vorhaben zu den beschriebenen Veränderungen beigetragen hat. Eine entsprechende Differenzierung sollte schließlich auch in der Bewertung und in den Schlussfolgerungen noch einmal vorgenommen werden. (Z 8)[21].

Bei der Auswahl der Indikatoren zur Sonderauswertung zu kontributionsanalytischen Qualitätsaspekten wurde darauf geachtet, weitgehend auf Indikatoren zurückzugreifen, die seit der ersten Meta-Evaluation erfasst werden, um Entwicklungen im Zeitverlauf untersuchen zu können.

[21] Auch wenn ein logischer Zusammenhang zwischen den genannten Indikatoren besteht, ist die Erfüllung dieser Indikatoren n nicht in jedem Bericht korreliert. Beispielweise gibt es PEV, die in der Beschreibung des Evaluierungsgegenstands das Wirkungsmodell und die Wirkungshypothesen nicht angemessen darstellen, aber dennoch im Kapitel zur Effektivität reflektieren, ob die beschriebenen Veränderungen auf das Vorhaben zurückzuführen sind.

Indikatoren zu kontributionsanalytischen Qualitätsaspekten	2015 (N=70)	2016 (N=100)	2017 (N=125)	2018 N=(51)	2017 / 2018 (N=176)
G 1.2 Das Wirkungsmodell wird dargestellt.	89%	86%	84%	90%	86%
G 1.3 Die relevanten Wirkungshypothesen werden dargestellt.	41%	65%	70%	69%	70%
G 7.1 (A) Die methodische Vorgehensweise beantwortet die Frage nach der Zuordnungs- und/ oder Beitragsanalyse.	19%	19%	23%	27%	24%
G 7.8 (C) Die Kausalität zwischen Maßnahmen und Wirkungen wird differenziert analysiert und eingeschätzt.	41%	44%	50%	43%	48%
G 7.14 (E) In der Analyse und Beurteilung der übergeordneten Wirkungen (Impacts) wird die Attributionslücke analysiert.	67%	64%	68%	67%	68%
Z 8 In den Schlussfolgerungen werden die untersuchten Zusammenhänge und Beiträge des Vorhabens differenziert bewertet.			68%	67%	60%
Qualität des kontributionsanalytischen Ansatzes			59%	61%	60%

Abbildung 23: Durchschnittlicher Erfüllungsgrad der Indikatoren zu kontributionsanalytischen Qualitätsaspekten (G 1.2, G 1.3, G7.1 (A), G7.8 (C), G.7.14 (E) und Z8) über die in den Meta-Evaluationen untersuchten Jahreszeiträume hinweg (Quelle: Syspons 2018, Textanalyse. Das N beschreibt die Grundgesamtheit der untersuchten Berichte im jeweiligen Zeitraum

G7.1 Trends hinsichtlich der Darstellung des Wirkungsmodells in den dezentralen PEV 2015 – 2018: Der Indikator, der erfasst, ob die Wirkungslogik des aktuellen Moduls graphisch dargestellt oder textlich beschrieben wird, hat sich im Zeitverlauf kaum verändert. Der Erfüllungsgrad für diesen Indikator schwankt je nach Jahr zwischen 85 % und 90 %, wobei der Wert bereits in der ersten Meta-Evaluierung 2015 bei 89 % lag. Somit sind Werte für diesen Indikator durchgängig hoch, was darauf zurückzuführen sein kann, dass es für die Vorhaben der GIZ seit einigen Jahren eine verbindliche Vorgabe ist, eine graphische Darstellung ihres Wirkungsmodells zu erarbeiten. Die PEV können somit auf bestehende Formate zurückgreifen.

G7.2 Trends hinsichtlich der Darstellung der relevanten Wirkungshypothesen in den dezentralen PEV 2015 – 2018: Der Indikator, der erfasst, ob die Wirkungslogik eines Vorhabens anhand eines Narratives dargestellt wird, hat sich stark positiv entwickelt, von 41 % in der ersten Meta-Evaluierung 2015 auf 69 bzw 70 % für 2017 und 2018. Die größte Entwicklung hat hierbei zwischen 2015 und 2016 stattgefunden: hier erfolgte eine Verbesserung von 24 Prozentpunkten, von 41 % auf 65 %. Eine mögliche Erklärung für den positiven Trend hinsichtlich dieses Indikators könnte die Sensibilisierung für die Bedeutung von Wirkungsmodellen- und Hypothesen in der Evaluierungs-Community im Allgemeinen sowie durch die GIZ gegenüber ihren Mitarbeitenden und externen Gutachtern sein.

G7.1 (A) Trends hinsichtlich der Auseinandersetzung mit der Frage nach der Zuordnungs- und Beitragsanalyse in den dezentralen PEV 2015 – 2018: Der Indikator, der erfasst, ob die PEVs bei der Darlegung der methodischen Vorgehensweise die Frage nach der Zuordnungs- und / oder Beitragsanalyse adressieren,

hat sich im Zeitverlauf leicht positiv entwickelt. Ausgehend von einem Ausgangswert von 19 % in der Meta-Evaluierung 2015 blieb dieser Wert im darauffolgenden Jahr zunächst stabil, um sich dann in der Meta-Evaluierung 2017 auf 23 % zu steigern und in 2018 schließlich einen Wert von 27 % zu erreichen. Somit ergibt sich eine Verbesserung von insgesamt 8 Prozentpunkten im Zeitverlauf, aber auch 2018 setzen sich weniger als ein Drittel der PEV-Berichte im Methodik-Kapitel mit der Frage nach der Zuordnungs- und Beitragsanalyse auseinander.

G7.8 (C) Trends hinsichtlich der differenzierten Analyse und Einschätzung zwischen Maßnahmen und Wirkungen in den dezentralen PEV 2015 – 2018: Der Indikator, der erfasst, ob die PEVs in der methodischen Bearbeitung der Effektivität explizit oder implizit hinterfragen, wie tatsächliche kausale Wirkungen von anderweitig verursachten Zusammenhängen unterschieden werden können, hat sich im Zeitverlauf positiv entwickelt. Ausgehend von einem Ausgangswert von 41% in der Meta-Evaluierung 2015 schwankte dieser Wert in den darauffolgenden Jahren zwischen 43 % und 50 %. Zwischen den Durchschnittswerten in der ersten Meta-Evaluierung 2015 und der letzten Meta-Evaluierung 2018 liegt ein Unterschied von zwei Prozentpunkten. 2018 setzen sich weniger als die Hälfte (43%) der PEV-Berichte im Effektivitätskapitel damit auseinander, ob die beschriebenen Veränderungen auf das Vorhaben zurückzuführen sind.

G7.14 (E) Trends hinsichtlich der Analyse der Attributionslücke in den PEV 2015 – 2018: Für den Indikator, der untersucht, ob die PEVs in der methodischen Bearbeitung des Impact-Kriteriums die Attributionslücke adressieren, lassen sich nur kleinere Schwankungen und kein Trend feststellen. Ausgehend von einem Ausgangswert von 67 % in 2015 ist der Indikator im darauffolgenden Jahr um drei Prozentpunkte gesunken (64 % in 2016), ein Jahr später wieder um zwei Prozentpunkte gestiegen (68 % in 2017), um sich für das Jahr 2018 wieder bei dem Ausgangswert von 67 % einzupendeln. Einerseits zeigt sich, dass keine nennenswerten Veränderungen hinsichtlich kontributionsanalytischer Qualitätsaspekte im Impact-Kapitel zu verzeichnen sind. Andererseits sind kontributionsanalytische Qualitätsaspekte im Impact-Kapitel deutlich stärker berücksichtigt als im Effektivitätskapitel und im Kapitel zur methodischen Vorgehensweise.

Differenzierte Bewertung der untersuchten Zusammenhänge und Beiträge des Vorhabens in den Schlussfolgerungen (Z 8) in den Berichten 2017 / 2018: Der Indikator, der ein Aufgreifen kontributionsanalytische Qualitätsaspekte in den Schlussfolgerungen untersucht, ist mit 60 % **eher erfüllt**. Da in den PEV-Berichten kein eigenes Kapitel für Schlussfolgerungen vorgesehen ist, wurden die entsprechenden Elemente entweder im Bewertungsteil des Effektivitätskapitels oder – häufiger – des Impact-Kapitels identifiziert.

Nachvollziehbarkeit

In der Meta-Evaluierung erstmalig untersucht wurden in diesem Jahr anhand neuer Indikatoren die Nachvollziehbarkeit der Bewertung in den Kapiteln Relevanz (Z 7 (B)), Effektivität (Z 7 (C)), Effizienz (Z 7 (D)), Impact (Z 7 (E)) und Nachhaltigkeit (Z 7 (F)). Voraussetzung für die Erfüllung dieser Indikatoren war, dass die Bewertung des jeweiligen Kriteriums anhand der Analyseergebnisse begründet wurde, und dass Punktevergabe und die Bewertung zueinander passten. Es wurde demnach nicht die Qualität der Analyse bewertet, sondern lediglich, ob die Bewertung kohärent zu den dargestellten Ergebnissen war. Nachfolgend wird die Erfüllung dieser Indikatoren zusammengefasst für 2017 / 2018 dargestellt (siehe Abbildung 24). Eine nach den Jahren 2017 und 2018 aufgeschlüsselte Darstellung findet sich in Anlage 1.

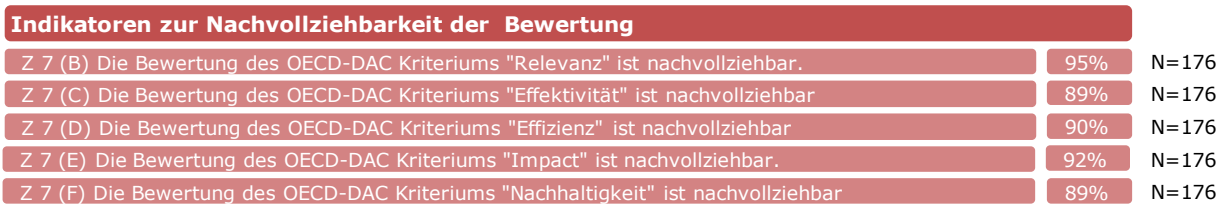


Abbildung 24: Durchschnittlicher Erfüllungsgrad der Indikatoren zur Nachvollziehbarkeit (Z 7 (B), Z 7 (C), Z 7 (D), Z 7 (E), Z 7 (F) in den PEV

2017 / 2018 (Quelle: Syspons 2018, Textanalyse). Das N beschreibt die Anzahl der Berichte, für die der Indikator bewertet wurde.

Nachvollziehbarkeit in der Bewertung der OECD-DAC-Kriterien in den Berichten 2017 / 2018: Die Nachvollziehbarkeit in der Bewertung der Relevanz ist mit 95 % **größtenteils bis vollständig erfüllt** (Z 7 (B)). Die Relevanz ist damit das OECD-Kriterium, dessen Bewertung am besten nachvollziehbar ist. Auch die Nachvollziehbarkeit der Bewertung des Kriteriums Effizienz ((Z 7 (D)) ist in 90 % der Fälle gegeben und der Indikator größtenteils bis vollständig erfüllt. Gleiches gilt für die Bewertung des Impact-Kriteriums, wo der entsprechende Indikator in 92 % der Fälle erreicht ist. **Zumeist erfüllt** sind die Indikatoren, die die Nachvollziehbarkeit der Bewertung der Effektivität (89%) und der Nachhaltigkeit (89%) erfassen. In den Fällen, in denen die entsprechenden Indikatoren negativ bewertet wurden, wurden in der Regel von den Gutachtern entweder Punkte vergeben für Aspekte, die im Berichtstext der PEVs nicht adressiert wurden, oder die Punktevergabe erschien angesichts im Berichtstext formulierter Schwächen des Vorhabens zu positiv.

2.7 Einflussfaktoren auf die methodische Qualität

Zusätzlich zu den im letzten Kapitel dargestellten deskriptiven Ergebnissen zur methodischen Qualität werden in diesem Kapitel anhand statistischer Analysen mögliche Einflussfaktoren auf den Evaluierungsstandard Genauigkeit (G) – und damit die methodische Qualität der PEV-Berichte – identifiziert. Hierbei wird zunächst der Einfluss der Rahmendaten und einzelner methodischer Aspekte analysiert. Anschließend werden potenzielle Stellschrauben identifiziert, die innerhalb des Evaluierungsstandards Genauigkeit die methodische Qualität beeinflussen. Zuletzt wird der Zusammenhang zwischen der Bewertung durch die PEV Teams und der methodischen Genauigkeit analysiert, um die Hypothese zu überprüfen, ob methodisch bessere PEV im Mittel kritischer von Gutachter/innen beurteilt werden.

Rahmendaten und ausgewählte methodische Aspekte

Für die Analyse der Rahmendaten werden zunächst bivariate Zusammenhänge der Variablen mit der methodischen Qualität untersucht, um mögliche Einflussfaktoren zu identifizieren. Hierbei werden für die **Anzahl der in der Evaluation eingesetzten Methoden**, die **Anzahl der Gutachter/innen** sowie die **Zahl der Gutachtertage** positive Effekte auf die methodische Genauigkeit festgestellt. Darüber hinaus stellt sich heraus, dass **längere Vorhaben** im Durchschnitt schlechtere Bewertungen des Genauigkeitsstandards aufweisen. Die so identifizierten Zusammenhänge werden anschließend in einer multiplen Analyse gemeinsam untersucht. Dabei ergeben sich zwei zentrale Stellschrauben für die methodische Genauigkeit. Erstens führt eine Erhöhung der Methodenvielfalt (insb. von bis zu zwei auf über zwei Methoden) im Mittel zu einer höheren methodischen Genauigkeit.[22] Zweitens wirkt sich die Anzahl an Gutachtertagen grundsätzlich positiv auf den Bewertungsmaßstab Genauigkeit aus. Es zeigt sich jedoch auch, dass eine weitere Erhöhung der Gutachtertage ausgehend von einer bereits hohen Anzahl nur noch eine geringe Verbesserung des Genauigkeitsstandards mit sich bringt. Für sehr hohe Anzahlen an Gutachtertagen sinkt die methodische Qualität der PEV Berichte im Mittel sogar wieder. Hierbei sollte jedoch berücksichtigt werden, dass dieser nicht lineare Zusammenhang stark von einzelnen PEV abhängt. Wie Abbildung 26 zeigt, sind in der Vollerhebung zwei PEV mit auffallend hohen Anzahlen an Gutachtertagen und geringer methodischer Qualität

Im Folgenden werden die Ergebnisse der bi- und multivariaten Analysen für die PEV 2017 / 18 detailliert beschrieben.

[22] Zur Messung der Methodenvielfalt wurde die Anzahl der eingesetzten Methoden (Indikator G 7.3 (A)) verwendet.

Einflussfaktoren		N	M	SD
Umfang der Methoden	2 oder weniger	72	60%	0,11
	mehr als 2	104	71%	0,10
Anzahl der Gutachter/innen	2 oder weniger	79	64%	0,11
	mehr als 2	89	69%	0,12
Laufzeit	bis zu 36 Monaten	75	69%	0,11
	mehr als 36 Monate	101	65%	0,12

Abbildung 25: Einflussfaktoren Umfang der Methoden, Anzahl der Gutachter/innnen, Laufzeit (Quelle: Syspons 2018, Textanalyse)

N= Anzahl der Berichte, M= Mittelwert der Berichte, SD= Standardabweichung

Der signifikante, positive Zusammenhang[23] zwischen dem **Umfang der Methoden** und der methodischen Qualität ergibt sich in der bivariaten Analyse aus dem Vergleich von PEV, die bis zu zwei Methoden einsetzen, und PEV, die auf mehr als zwei Methoden zurückgreifen. So erfüllen PEV, die mehr als zwei Methoden anwenden, den Genauigkeitsstandard zu durchschnittlich 71%. PEV, die lediglich bis zu zwei Methoden einsetzen, kommen mit 60% hingegen auf einen niedrigeren Mittelwert (siehe Abbildung 25).

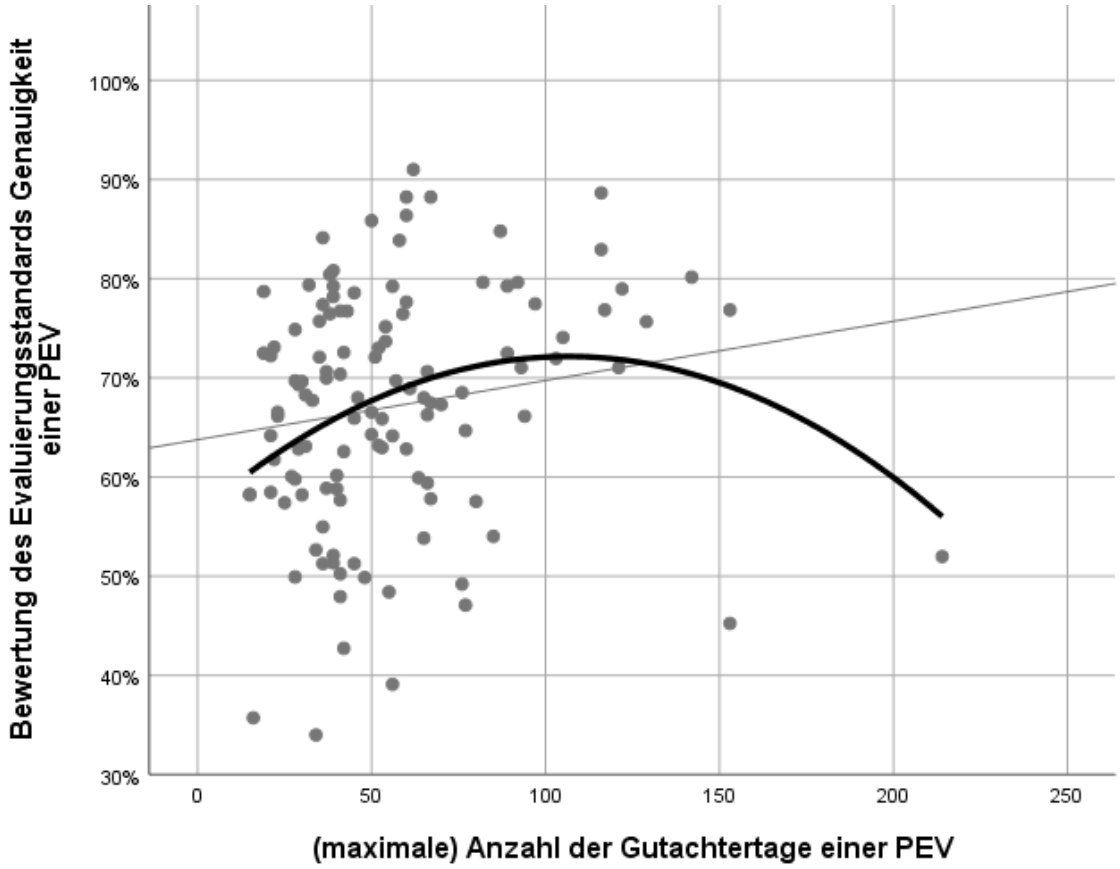
Ebenso wird in der bivariaten Analyse ein signifikanter Unterschied[24] zwischen PEV mit **weniger und mehr Gutachter/innen** deutlich. So erfüllen PEV, die von bis zu zwei Gutachter/innen durchgeführt werden den Genauigkeitsstandard im Durchschnitt zu 64%, wohingegen PEV mit mehr als zwei Gutachter/innen den Standard durchschnittlich zu 69% erfüllen (siehe Abbildung 25). Keine signifikanten Unterschiede im Genauigkeitsstandard lassen sich zwischen PEV mit 3, 4 oder 5 und mehr Gutachter/innen feststellen.

Zur Untersuchung des Zusammenhangs zwischen der Anzahl der Gutachtertage und der methodischen Genauigkeit wird zunächst ein entsprechendes Streudiagramm herangezogen. Das Streudiagramm in Abbildung 26 stellt den Zusammenhang zwischen der Anzahl der Gutachtertage und der Erfüllung des Genauigkeitsstandards dar. Grundsätzlich fällt auf, dass für kleine Anzahlen an Gutachtertagen sehr viele PEV mit einer hohen Varianz der methodischen Genauigkeit vorliegen. Für größere Werte an Gutachtertagen liegen dagegen weniger PEV vor.

Abbildung 26 enthält sowohl eine lineare Anpassungslinie als auch eine nicht lineare (quadratische) Anpassungskurve. Dabei scheint die kurvenförmige Anpassungslinie den Zusammenhang über die gesamte Spannbreite an Gutachtertagen etwas besser abzubilden. Mit zunehmender Anzahl der Gutachtertage steigt der Genauigkeitsstandard der PEV im Mittel zunächst an, flacht jedoch im weiteren Verlauf ab und wird für besonders hohe Anzahlen von Gutachtertagen sogar kleiner. Diesbezüglich gilt es jedoch zu berücksichtigen, dass die Fallzahlen für PEV mit einer hohen Anzahl an Gutachtertagen verhältnismäßig klein sind, sodass einzelne Vorhaben die Werte die Krümmung der Kurve leicht verändern können. Insbesondere fallen zwei Vorhaben mit besonders hohen Werten an Gutachtertagen auf, die gleichzeitig sehr niedrige Werte im Genauigkeitsstandard aufweisen und so entscheidend zur relativ starken Krümmung der Anpassungskurve beitragen. Da keine Anzeichen dafür vorliegen, dass die Angaben dieser PEV zur Anzahl der Gutachtertage sowie ihre Bewertung des Genauigkeitsstandards fehlerhaft sind, liegt aus statistischer Sicht kein Grund vor, diese aus der weiteren Analyse auszuschließen. Es wird jedoch im Rahmen einer Sensitivitätsanalyse untersucht, wie sich der Zusammenhang der Gutachtertage und der methodischen Genauigkeit verändert, wenn diese PEV aus der Analyse ausgeschlossen werden.

[23] $t_{(174)} = -6{,}801$, $p = 0{,}000$, $d = 1{,}035$ (große Effektgröße). Die Einordnung der Effektgrößen in klein (0,1 bis 0,3), mittel (0,3 bis 0,5) und groß (< 0,5) erfolgt nach Cohen (1988). .

[24] $t_{(166)} = -3{,}151$, $p = 0{,}002$, $d = 0{,}488$ (mittlere Effektgröße).



s

Abbildung 26: Kreuzung des Erfüllungsgrades des Evaluierungsstandards Genauigkeit und der Anzahl der Gutachtertage (Quelle: Syspons 2018, Textanalyse)

Um die Beobachtung aus dem Streudiagramm (dass unter Einschluss *aller* PEV tendenziell kein linearer, sondern eher ein nicht linearer Zusammenhang vorliegt), auf ihre Signifikanz zu testen, werden zwei Regressionsmodelle verwendet. Beide Modelle verwenden den Genauigkeitsstandard als abhängige Variable. Das erste Modell testet, ob ein linearer Zusammenhang zwischen dem Genauigkeitsstandard und der Anzahl der Gutachtertage besteht, ob also mehr Gutachtertage den Genauigkeitsstandard im Mittel konstant erhöhen. Das Modell enthält die Gutachtertage als einzige erklärende Variable. Das zweite Modell überprüft den kurvenförmigen Zusammenhang auf seine Signifikanz. Dies erfolgt, indem neben den Gutachtertagen auch die Gutachtertage zum Quadrat als erklärende Variable aufgenommen werden. Abbildung 27 stellt die Ergebnisse der beiden Regressionsmodelle dar.[25]

[25] Beide Modelle enthalten zudem eine Konstante.

Modell 1

	Standardisierte Koeffizienten	T-Test	Signifikanz
(Konstante)		29,790	0,000
Gutachtertage ingesamt	0,166	1,825	0,071

Modell 2

	Standardisierte Koeffizienten	T-Test	Signifikanz
(Konstante)		15,564	0,000
Gutachtertage ingesamt	0,828	2,984	0,003
Gutachtertage ingesamt zum Quadrat	-0,699	-2,518	0,013

Abbildung 27: Prüfung der Annahmen durch Regressionsmodelle (Quelle: Syspons 2018)
Abhängige Variable: Erfüllung des Evaluierungsstandards Genauigkeit (G)

Im ersten Modell weisen die Gutachtertage insgesamt *keinen* signifikanten Effekt für die Beurteilung des Genauigkeitsstandards auf (p-Wert größer als 0,05). Dies bedeutet, dass unter Einschluss aller PEV ein Modell mit den Gutachtertagen als einzig erklärender Variablen den Zusammenhang nicht ausreichend erklärt. Die stetige Erhöhung der Gutachtertage führt also im Mittel nicht zu einer konstanten Verbesserung des Genauigkeitsstandards. Im zweiten Modell sind dagegen alle Effekte signifikant.[26] Dies bedeutet, dass unter Einschluss aller PEV eine Anpassungskurve wie in Abbildung 26 dargestellt eine bessere Annäherung an den Zusammenhang zwischen den beiden Variablen darstellt als eine Gerade. : Laut dem Modell hat eine Erhöhung der Gutachtertage zunächst einen positiven Effekt auf die methodische Qualität der PEV. Jedoch bewirkt eine stetige Erhöhung der Gutachtertage keinen konstanten Anstieg des Genauigkeitsstandards. Für höhere Anzahlen von Gutachtertagen wird der Effekt kleiner und ab einem bestimmten Level sogar negativ. Die Erhöhung der Gutachtertage über einen bestimmten Punkt hinaus bewirkt so laut Modell keine große Verbesserung im Genauigkeitsstandard mehr und kann sich sogar negativ auf diesen auswirken. Hierbei ist zu berücksichtigen, dass dieser Effekt wie oben dargestellt stark von zwei PEV mit hohen Anzahlen an Gutachtertagen und geringer methodischer Qualität beeinflusst wird.

Im Rahmen einer Sensitivitätsanalyse wird der Zusammenhang zwischen der Anzahl der Gutachtertage und der methodischen Qualität auch ohne die beiden PEV untersucht, die besonders hohe Werte an Gutachtertagen und gleichzeitig eine niedrige methodische Qualität aufweisen. Werden diese beiden PEV aus der Analyse ausgeschlossen, zeigt sich, dass eher ein *linearer*, positiver Zusammenhang zwischen den beiden Variablen besteht. Eine Erhöhung der Gutachtertage um einen Tag führt dann im Mittel zu einer konstanten Erhöhung der methodischen Qualität um 0,001 Prozentpunkte. Dies bedeutet, dass der kurvenförmige Zusammenhang zwischen den Gutachtertagen und der methodischen Qualität, der in der Analyse der Gesamtheit der PEV deutlich wird, stark von den beiden extremen PEV abhängig ist. Für die weitere Analyse werden die beiden PEV trotzdem nicht aus der Datengrundlage ausgeschlossen, da dafür wie oben erläutert aus statistischer Sicht kein Grund vorliegt. Es sollte jedoch berücksichtigt werden, dass für eine robuste Analyse des Zusammenhangs eine größere Zahl an PEV mit hohen Anzahlen von Gutachtertagen nötig wären.

Zuletzt ergibt die bivariate statistische Analyse einen signifikanten, negativen Zusammenhang der methodischen Qualität mit der **Laufzeit des Vorhabens**. Die PEV längerer Vorhaben weisen im Mittel einen niedrigeren Genauigkeitsstandard auf.[27] So erfüllen PEV von Vorhaben mit einer Laufzeit von bis zu 36 Monaten den Genauigkeitsstandard im Mittel zu 69% während PEV von Vorhaben mit einer längeren Laufzeit den Standard im Mittel nur zu 65% erfüllen (siehe Abbildung 25).

Um zu untersuchen, ob die identifizierten Einflussfaktoren unabhängig voneinander einen Einfluss auf die methodische Qualität haben oder redundante Informationen beinhalten, werden sie gemeinsam in einem **multiplen Regressionsmodell** analysiert (siehe Abbildung 27). Dieses Analyseverfahren untersucht den Einfluss einzelner Faktoren (hier: Umfang der Methoden, Anzahl der Gutachter/innen, Anzahl der Gutachtertage (inkl.

26 Die Effekte wurden zudem auf ihre gemeinsame Signifikanz überprüft.

27 r = -0.241 (kleine Effektgröße), n=176, p=0,001.

Quadrat der Gutachtertage), Laufzeit des Vorhabens) auf ein Kriterium (hier: methodische Qualität), bei gleichzeitiger Kontrolle der jeweils anderen Faktoren. Speziell wurde die Methode der *schrittweisen Regression* angewendet. Hierbei wird zunächst ein Modell geschätzt, dass alle zuvor in bivariaten Analysen identifizierten Einflussfaktoren enthält. Anschließend wird der am wenigsten signifikante Einflussfaktor (der Einflussfaktor mit dem höchsten p-Wert) entfernt und das reduzierte Modell erneut geschätzt. Dieser Vorgang wird solange wiederholt, bis das Modell nur noch signifikante Einflussfaktoren enthält. Abbildung 28 zeigt das finale Regressionsmodell, in dem nur noch signifikante Einflussfaktoren enthalten sind.

	Standardisierte Koeffizienten	T-Test	Signifikanz
(Konstante)		15,543	0,000
Umfang der Methoden: 2 oder weniger, mehr als 2	0,474	6,034	0,000
Gutachtertage ingesamt	0,674	2,760	0,007
Gutachtertage ingesamt zum Quadrat	-0,547	-2,238	0,027

Abbildung 28: Prüfung der Annahmen durch ein multiples Regressionsmodell (Quelle: Syspons 2018, Textanalyse)
Abhängige Variable: Erfüllung des Evaluierungsstandards Genauigkeit (G)

Im multiplen Regressionsmodell bleiben die Faktoren **Umfang der Methoden** sowie **Anzahl der Gutachtertage** (inkl. des quadratischen Terms) weiterhin signifikant.[28] Diese Faktoren haben demnach auch bei gleichzeitiger Kontrolle des jeweilig anderen Faktors einen signifikanten Einfluss auf den Genauigkeitsstandard. Die Ergebnisse zeigen damit, dass es zur Verbesserung der methodischen Genauigkeit der PEV zwei zentrale Stellschrauben gibt: Zum einen kann die Erhöhung der Methodenvielfalt (insbesondere von bis zu zwei auf mehr als zwei Methoden) zu einem höheren Genauigkeitsstandard der PEV beitragen. Zum anderen wirkt sich eine höhere Anzahl an Gutachtertagen im Mittel positiv auf den Genauigkeitsstandard aus. Für hohe Werte der Gutachtertage wird der Effekt einer weiteren Erhöhung jedoch kleiner und für sehr große Werte sogar negativ, wobei dieser Effekt wie oben dargestellt stark von zwei PEV mit hohen Anzahlen an Gutachtertagen und geringer methodischer Qualität beeinflusst wird. Auf Grundlage des oben geschätzten Regressionsmodells ergibt sich ein Wert von *circa 92 Tagen*, ab dem bei einer weiteren Erhöhung der Gutachtertage der Genauigkeitsstandard im Mittel wieder sinkt. Praktisch bedeutet dies, dass bei einer Erhöhung der Gutachtertage über diesen Wert hinaus keine weitere Verbesserung des Genauigkeitsstandards zu erwarten ist.

Die Faktoren **Anzahl der Gutachter/innen** (in Form der geclusterten Variable bis zu zwei Gutachter/innen / mehr als zwei Gutachterinnen) sowie die **Laufzeit des Vorhabens** sind bei gleichzeitiger Kontrolle der jeweils anderen Faktoren nicht länger signifikant. Dies kann dadurch erklärt werden, dass beide Faktoren Zusammenhänge zu den signifikanten Faktoren Methodenumfang und Anzahl der Gutachtertage aufweisen. Sie enthalten daher zumindest teilweise redundante Informationen, die im multiplen Regressionsmodell bei gleichzeitiger Kontrolle der anderen Faktoren keinen zusätzlichen Mehrwert für die Erklärung der abhängigen Variable (Genauigkeitsstandard) bringen.

Insbesondere **der Zusammenhang zwischen Gutachtertagen und Anzahl der Gutachter/innen** sollte noch einmal genauer beleuchtet werden. So ergibt eine bivariate Analyse einen signifikanten, positiven Zusammenhang zwischen der Anzahl der Gutachtertage und der Anzahl der Gutachter/innen (geclustert in bis zu zwei Gutachter/innen und mehr als zwei Gutachter/innen).[29] Prüfungen, bei denen bis zu zwei Gutachter/innen eingesetzt werden haben demnach im Durchschnitt 39 Gutachtertage. PEV mit mehr als zwei Gutachter/innen kommen dagegen im Mittel auf 77 Gutachtertage. Das Ergebnis der multiplen Regressionsanalyse zeigt, dass die Anzahl der Gutachter/innen keinen „eigenen" Effekt auf die Genauigkeit der PEV hat, wenn gleichzeitig die Anzahl der Gutachtertage einbezogen wird. Insgesamt lässt sich demnach schlussfolgern, dass die Anzahl der Gutachter/innen eine weniger entscheidende Rolle für die Erfüllung des Genauigkeitsstandards spielt als die

[28] Die beiden Terme für Anzahl der Gutachtertage wurden auch auf ihre gemeinsame Signifikanz getestet.

[29] $t_{(117)}$= -7,588, p-Wert=0,000, d= 1,361.

den Gutachter/innen zur Verfügung stehenden Tage. Die unterdurchschnittlichen Ergebnisse des Genauigkeitsstandards für die wenigen PEV mit nur einem/einer Gutachter/in[30] deuten jedoch darauf hin, dass der Einsatz vom mind. zwei Gutachter/innen sinnvoll ist. Dies wäre ohnehin im Sinne der Forschertriangulation zu fördern.

Die Erkenntnisse zum Einfluss des Mengengerüsts können weiter ausdifferenziert werden. Dies zeigt ein Blick auf die **Gutachtertage in der Vorbereitung, Durchführung und Nachbereitung** einer PEV. Dabei können für die Anzahl der Gutachtertage in der Vorbereitung sowie in der Durchführung ähnliche nicht linearer Zusammenhänge zur Erfüllung des Genauigkeitsstandards festgestellt werden wie für die Gutachtertage insgesamt. Auch hier gilt also, dass eine Erhöhung der Gutachtertage im Mittel zu einer Erhöhung des Genauigkeitsstandards führt. Der Effekt der Gutachtertage auf den Genauigkeitsstandard wird jedoch für große Anzahlen an Gutachtertagen kleiner und ab einem bestimmten Level sogar negativ. Auch hier gilt jedoch, dass diese Zusammenhänge stark von einzelnen PEV mit hohen Werten an Gutachtertagen bei gleichzeitig geringer methodischer Qualität abhängen. Hingegen gibt es in dieser Meta-Evaluierung keine Hinweise auf einen linearen oder nicht linearen signifikanten Zusammenhang der Gutachtertage in der Nachbereitung einer PEV mit der methodischen Qualität. Dies galt zwar auch für die Untersuchung im Rahmen der Meta-PEV 2016, ist jedoch trotzdem überraschend, da die Bewertung der Genauigkeit auf den PEV-Berichten basiert, welche in der Nachbereitungsphase verfasst werden. Bei der Analyse des Mengengerüstes gilt es insgesamt zu berücksichtigen, dass in den ToR der PEV die Anzahl der Gutachtertage für die Evaluierung und für die Prüfung eines Folgevorhabens **gemeinsam angegeben** werden. Informationen zur Verteilung der Gutachtertage zwischen Evaluierung und Prüfung lagen der Meta-Evaluierung folglich nicht vor.

Keine signifikanten Unterschiede: Die Ergebnisse der Meta-Evaluierung geben **keine Anhaltspunkte** dafür, dass es signifikante Unterschiede im Evaluierungsstandard Genauigkeit gibt zwischen PEV, für die ein PEV-Check durchgeführt wurde und solchen, für die kein PEV-Check durchgeführt wurde.[31] Ebenfalls ergeben sich keine signifikanten Unterschiede zwischen einzelnen Regionen, den länderbezogenen Regionalbereichen B1, B2 und B3, sowie zwischen bilateralen und nicht-bilateralen Vorhaben. Auch zwischen den unterschiedlichen Schwerpunkten des FMB (Themenbereich Klima, Ländliche Entwicklung und Infrastruktur; Themenbereich Wirtschaft, Beschäftigung, Soziales; Themenbereich Governance und Konflikte), sowie zwischen PEV in Ländern mit krisenbedingt hohem Risiko für die Lieferfähigkeit der GIZ und Ländern mit geringer oder erhöhter Gefährdung gibt es keine Zusammenhänge mit dem Genauigkeitsstandard. Keine signifikanten Effekte lassen sich zudem für die Art der PEV (PEV mit oder ohne Folgevorhaben) identifizieren.

Methodische Aspekte innerhalb des Evaluierungsstandards Genauigkeit

Zusätzlich zu den Rahmendaten wurden ausgewählte methodische Aspekte untersucht, die innerhalb des Evaluierungsstandards Genauigkeit die methodische Qualität beeinflussen. Hierbei können zunächst die gleichen signifikanten Stellschrauben wie in den beiden vorangegangenen Meta-Evaluierungen identifiziert werden. So besteht ein signifikanter Zusammenhang zwischen dem Evaluierungsstandard Genauigkeit und einem klaren **Verständnis des Vorhabens** sowie der Auseinandersetzung mit dessen Wirkungslogik (Index aus: G 1.2, G 1.3, G 8.1 1 (erster Punkt der Checkliste))[32]. Ebenso kann ein signifikanter Zusammenhang zwischen dem Evaluierungsstandard Genauigkeit und der **Trennung zwischen Beschreibung, Analyse und Bewertung** (G 8.1) identifiziert werden.[33] Diese Erkenntnisse legen die Annahme nahe, dass ein fundiertes Verständnis des Evaluierungsgegenstands sowie eine klare Trennung des Dreischritts Beschreibung, Analyse und Bewertung zu einer höheren methodischen Qualität der PEV führen. Diese Annahme entspricht auch dem allgemeinen Evaluierungsverständnis – nur wenn die Wirkungslogik eines Vorhabens seitens des PEV-Teams verstanden wird und Beschreibung, Analyse und Bewertung klar getrennt sind, kann eine Evaluierung methodisch nachvollziehbar sein.

[30] Die 6 PEV mit nur einem / einer Gutachter/in haben einen durchschnittlichen Wert im Genauigkeitsstandard von 54%.

[31] PEV Checks wurden bis Februar 2017 durchgeführt.

[32] r = 0,395 (mittlere Effektgröße), n = 176, p = 0,000. Die entsprechenden korrelativen Indikatoren wurden aus G herausgerechnet. Dies bedeutet, dass der Genauigkeitsstandard G noch einmal neu berechnet wurde, wobei die Indikatoren zur Messung des Verständnisses des Vorhabens nicht miteinbezogen wurden.

[33] r = 0,379 (mittlere Effektgröße), n = 176, p = 0,000. Der entsprechende korrelative Indikator wurde aus G herausgerechnet.

In diesem Jahr wird erstmals auch der Zusammenhang zwischen der **Qualität des kontributionsanalytischen Ansatzes** (Index aus G 1.2, G 1.3, G 7.1, G 7.8, G7.14 und Z 8[34]) und der Erfüllung des Genauigkeitsstandard untersucht. Auch hier wird ein signifikanter, positiver Zusammenhang deutlich, der ähnlich stark ausgeprägt ist, wie der Zusammenhang zwischen der methodischen Genauigkeit und dem Verständnis des Vorhabens sowie der Trennung zwischen Beschreibung, Analyse und Bewertung.[35] Eine vertiefte Analyse ergibt, dass vor allem die Angemessenheit der Analyse und Bewertung (G 7.1 bis G 8.4 ausgenommen der für den Index zugrunde gelegten Indikatoren) mit der Qualität des kontributionsanalytischen Ansatzes korreliert sind.

Zusätzlich wird der Zusammenhang zwischen fünf der sechs neuen Indikatoren (Z 7 (B), Z 7 (C), Z 7 (D), Z 7 (E) und Z 7 (F)), die die *Nachvollziehbarkeit* der Bewertung beurteilen und der Erfüllung des Genauigkeitsstandards untersucht. Dabei kann ein signifikanter, positiver Zusammenhang identifiziert werden.[36] Dies erscheint naheliegend, da eine nachvollziehbare Beschreibung maßgeblich zur methodischen Genauigkeit der Evaluation beiträgt.

Benotungen durch die PEV-Teams

Zuletzt wird geprüft, inwieweit ein Zusammenhang zwischen den Benotungen durch die PEV-Teams (OECD DAC Kriterien) und der methodischen Qualität der PEV besteht. Dem unterliegt die Annahme, dass die methodische Qualität einer PEV Einfluss auf die Gesamtbenotung der PEV haben könnte. Die statistische Analyse liefert keine signifikanten Hinweise, dass PEV, die im gesamten Evaluierungsstandard Genauigkeit besser abschneiden, durchschnittlich schlechtere OECD-DAC-Gesamtbenotungen der Vorhaben vergeben. Dies wird beispielsweise am Vergleich von PEV, die im Mittel den Bewertungsmaßstab Genauigkeit nur bedingt erfüllen **(Mittelwert < 60 %)**, und methodisch besseren PEV **(Mittelwert > 60 %)** deutlich. So haben PEV, die im Mittel den Bewertungsmaßstab Genauigkeit nur bedingt erfüllen eine durchschnittliche OECD DAC Gesamtbenotung von 13 Punkten, während methodisch bessere PEV im Mittel eine Punktzahl von 12,6 erhalten. Der Unterschied ist klein und nicht signifikant.[37]

Es zeigt sich jedoch, dass die methodische Qualität einer PEV dann signifikant negativ mit der Benotung korreliert ist, wenn man einzelne Bewertungskriterien bzw. **Cluster von Bewertungskriterien** untersucht. Die Ergebnisse dieser Analyse sind in Abbildung 29 in Form von Korrelationen einzelner Cluster von Bewertungskriterien mit der Gesamtbenotung der Vorhaben (Punktzahl) dargestellt. Es zeigt sich, dass ein signifikant negativer Zusammenhang zwischen den Bewertungskriterien zur Analyse qualitativer und quantitativer Informationen (Index: G 7 B bis G 7 E) mit der Gesamtbenotung der der Vorhaben vorliegt (kleine Effektgröße). Die Kriterien zur Analyse qualitativer und quantitativer Informationen befassen sich mit der Qualität der methodischen Bewertung der OECD-DAC-Kriterien. Dies bedeutet: Ist die methodische Qualität bei der Bewertung anhand der OECD-DAC-Kriterien besser, werden die entsprechenden Vorhaben kritischer und damit durchschnittlich schlechter benotet. Zu anderen Clustern von Bewertungskriterien bestehen keine signifikanten Zusammenhänge.

[34] Indikatoren G 1.2 Das Wirkungsmodell wird dargestellt.; G 1.3 Die relevanten Wirkungshypothesen werden dargestellt.; G 7.1 (A) Die methodische Vorgehensweise beantwortet die Frage nach der Zuordnungs- und/ oder Beitragsanalyse; G 7.8 (C) Die Kausalität zwischen Maßnahmen und Wirkungen wird differenziert analysiert und eingeschätzt; G 7.14 (E) In der Analyse und Beurteilung der übergeordneten Wirkungen (Impacts) wird die Attributionslücke analysiert; Z8 In den Schlussfolgerungen werden die untersuchten Zusammenhänge und Beiträge des Vorhabens differenziert bewertet.

[35] r =0,380 (mittlere Effektgröße), n = 176, p = 0,000. Die entsprechenden korrelativen Indikatoren wurden aus G herausgerechnet.

[36] r =0,416 (große Effektgröße), n = 176, p = 0,000.

[37] $t_{(174)} = 1{,}73$, p = 0,085.

		Evaluierungs-standard Genauigkeit	Cluster von Bewertungskriterien		
			G 1.1 – G3.4 Übergeordnete methodische Aspekte	G 4.1 – G6.2 Reflektion der Qualität in der methodischen Durchführung	G 7.5 – G7.19 OECD DAC Kriterien
OECD DAC Gesamtbenotung des Vorhabens - Punktzahl	Korrelation Pearson	-0,058	0,115	-0,089	**-,166***
	Signifikanz (2-seitig)	0,444	0,127	0,238	0,028
	N	176	176	176	176

Abbildung 29: Korrelationen einzelner Cluster von Bewertungskriterien (Quelle: Syspons 2018, Textanalyse)

Darüber hinaus wird untersucht, ob und wie die **methodische Auseinandersetzung innerhalb der einzelnen OECD-DAC-Kriterien** die Benotung der Vorhaben im jeweiligen Kriterium beeinflusst. Hier wird deutlich, dass eine gute methodische Auseinandersetzung innerhalb des Kriteriums Effizienz tendenziell mit einer schlechteren Benotung in dem Kriterium Effizienz einhergeht.[38] Weitere signifikante Zusammenhänge können im Gegensatz zur letzten Meta-PEV jedoch nicht identifiziert werden.

[38] r = -0,208, n = 176, p = 0,006.

2.8 Bewertung und Schlussfolgerungen

Die Meta-Evaluierung zeigt, dass die Stärken und Schwächen der PEV zwischen Oktober 2016 und September 2018 ähnlich ausfallen wie in den Vorjahren. Gleichzeitig gibt es einen insgesamt leicht positiven Trend hinsichtlich der Qualität der dezentralen PEV über die in allen Meta-Evaluierungen erfassten Jahreszeiträume hinweg. Der Durchschnittswert aller geprüften PEV für den Evaluierungsstandard Genauigkeit hat sich jährlich etwas verbessert, von 59 % (n= 70) in 2015 auf 62 % (n= 100) in 2016, hin zu 66 % (n= 125) für 2017 und 67 % (n= 51) in 2018. Während der Evaluierungsstandard Genauigkeit demnach in der Meta-Evaluierung 2015 bedingt erfüllt war, ist er in den darauffolgenden Jahren eher erfüllt.

Unabhängig vom insgesamt leicht positiven Trend in der Entwicklung der Qualität der PEV über die Jahre hinweg liegen die **Stärken** wie in der Vergangenheit auch für 2017 / 2018 vor allem in der Erfüllung formaler und deskriptiver Anforderungen. Hinsichtlich der methodischen Anforderungen lassen sich zwar Verbesserungen verzeichnen, es bestehen jedoch weiterhin auch Herausforderungen.

Die formalen und deskriptiven Anforderungen, bei denen die PEV vergleichsweise gut abschneiden, sind auch in dieser Meta-Evaluierung die graphische oder textliche Darstellung des Wirkungsmodells sowie die Darstellung des Kontextes, in den sich die untersuchten Vorhaben eingliedern. Darüber hinaus geht mit der ausführlichen Beschreibung des Kontextes auch eine vergleichsweise umfassende methodische Bearbeitung des OECD-Kriteriums Relevanz in den PEV-Berichten einher. Eine weitere Stärke liegt in der allgemeinen Beschreibung der angewandten Methoden. Darüber hinaus sind auch die Datentriangulation sowie eine Triangulation von Ergebnissen mit Beteiligten und Betroffenen weitgehend gegeben.

Die methodischen **Schwächen** der dezentralen PEV sind in den gleichen Bereichen zu verorten wie in den vergangenen Meta-Evaluierungen. Diesbezüglich ist anzumerken, dass es im Untersuchungszeitraum keine Veränderungen hinsichtlich der Vorgaben an die PEV-Teams gegeben hat. Mit der Umstellung des Evaluierungssystems der GIZ lag der Schwerpunkt der Stabsstelle Evaluierung auf der Einführung höherer methodischer Standards für das neue Instrument der zentral gesteuerten Projektevaluierungen. Wie in den vergangenen Jahren ist demnach zu berücksichtigen, dass nicht alle in der Meta-Evaluierung untersuchten Aspekte Bestandteil der Vorgaben für die Durchführung von PEVs waren.

Eine zentrale Schwäche der PEVs ist aus Sicht des Meta-Evaluierungsteams weiterhin eine unzureichende Reflektion der methodischen Vorgehensweise. So hat es sich noch nicht durchgesetzt, in den Berichten darzustellen, warum, bzw. zu welchem Sachverhalt welche Quellen ausgewertet wurden, und zu erläutern unter welchen Gesichtspunkten die Auswahl von Gesprächspartner/innen erfolgte. Auch werden die Stärken und Schwächen der gewählten Methodik nur wenig thematisiert. Darüber hinaus fällt auf, dass die Auseinandersetzung mit der Qualität der zugrunde gelegten Daten ausbaufähig bleibt. So gehen die Berichte zwar in der Regel kurz auf die allgemeine Qualität des wirkungsorientierten Monitorings ein, sie setzen sich jedoch nur bedingt explizit mit der Belastbarkeit der Baseline-Daten auseinander.

Mit Blick auf die methodische Bearbeitung der OECD-DAC-Kriterien lässt sich folgendes feststellen:

- Die Bearbeitung des Effizienz-Kriteriums weist nach wie vor die größten Herausforderungen auf. Das entsprechende Kapitel beschränkt sich zumeist auf die Analyse der Implementierungseffizienz. Diesbezüglich ist anzumerken, dass vor der gemeinsamen Verfahrensreform von BMZ und GIZ geplante Vorhaben in der Regel keine Zuordnung von Kosten zu Outputs oder Outcomes vorgenommen haben, was die Analyse von Produktions- und Allokationseffizienz erschwert. Allerdings wird in den untersuchten PEVs die Auswahl von Verfahren und Methoden der Effizienzmessung nur in Ausnahmefällen begründet. Insgesamt ist das Effizienz-Kapitel darüber hinaus in der Regel sehr kurz gehalten und die Analysen fallen dementsprechend oberflächlich aus.

- Für die Bearbeitung der Effektivität werden zwar fast flächendeckend Modulzielindikatoren herangezogen, diese entsprechen jedoch nicht durchgehend den SMART-Qualitätskriterien. Zwar hat sich dieser Aspekt gegenüber den Vorjahren gebessert, aber angesichts des zentralen Stellenwerts dieser Anforderung besteht hier noch Optimierungspotenzial. In der Regel werden Mängel hinsichtlich der Erfüllung der SMART-Kriterien von den PEV-Prüfteams korrekt identifiziert, aber es werden dann keine angepassten oder neuen Indikatoren formuliert. Dies zeigte sich insbesondere bei Sektor- oder Globalvorhaben, für die in mehreren PEV-Berichten darauf verwiesen wurde, dass die Nutzung von erarbeiten Strategien innerhalb der Laufzeit des Vorhabens nicht messbar sei.

- Die Qualität kontributionsanalytischer Aspekte kommt sowohl bei der Bearbeitung der Effektivität als auch bei der Bearbeitung des Impacts zum Tragen. Hierbei lässt sich feststellen, dass eine Auseinandersetzung mit der Kausalität zwischen den Maßnahmen des Vorhabens und den beobachteten Veränderungen eher für die Impact- als für die Outcome-Ebene erfolgt. Allerdings erfolgt diese Auseinandersetzung auch für die übergeordneten Wirkungen oft nur implizit durch Formulierungen wie „das Vorhaben leistet einen Beitrag“. Eine differenzierte Darlegung anderer Einflussfaktoren auf die beschriebenen Veränderungen erfolgt sowohl im Effektivitäts- als auch im Impact-Kapitel nur selten.

- Erstmalig untersucht wurde in diesem Jahr die Nachvollziehbarkeit der Bewertung für jedes der OECD-DAC-Kriterien. Auch wenn die Nachvollziehbarkeit größtenteils gegeben ist, zeigt sich, dass in bis zu einem von zehn Berichten entweder Punkte vergeben werden für Aspekte, auf die im Text nicht eingegangen wird, oder die Bewertung nach Einschätzung des Meta-Evaluierungsteams zu positiv ausfällt angesichts im Berichtstext beschriebener Schwächen des evaluierten Vorhabens.

Übergeordnet zeigt sich, dass die Benotung der Vorhaben durch die PEV-Teams ausgesprochen positiv ausfällt. So erhalten 85 % der Vorhaben die Gesamtnote „sehr erfolgreich“ oder „erfolgreich“. Besonders ausgeprägt ist die positive Benotung hinsichtlich des Relevanz-Kriteriums, für das 99 % aller Vorhaben als „sehr erfolgreich“ bzw. „erfolgreich“ eingestuft werden. Dies wirft die Frage auf, inwiefern die Evaluator*innen sich angemessen kritisch mit den Vorhaben auseinandersetzen. Dieser Qualitätsaspekt ist im Hinblick auf das Potential von Evaluationen für Lernen und Entscheidungsfindung von Bedeutung. Diesbezüglich stellt die Umstellung innerhalb der GIZ von dezentral gesteuerten PEV auf zentral gesteuerte Projektevaluierungen (ZPE) eine Chance dar. Mit der Umstellung werden Projektevaluierungen von der Stabstelle Evaluierung beauftragt und nicht mehr durch das Vorhaben selbst, wodurch die evaluatorische Unabhängigkeit gestärkt wird. Gleichzeitig werden die Projektevaluierungen nunmehr entkoppelt von der Planung von Folgevorhaben umgesetzt, wodurch die Fokussierung der Gutachterteams auf ihre evaluatorische Rolle gewährleistet wird. Darüber hinaus wurden für die zentral gesteuerten Projektevaluierungen methodische Vorgaben eingeführt, die weit über die Vorgaben für die PEV hinausgehen.

Hinsichtlich der **Einflussfaktoren** auf die methodische Qualität der PEV zeigt die statistische Auswertung mehrere zentrale Stellschrauben auf, die sich bereits in den vorhergehenden Meta-Evaluierungen als signifikante Zusammenhänge erwiesen hatten:

- Ein signifikanter positiver Zusammenhang besteht zwischen dem Umfang der Methoden und der methodischen Qualität. Besonders auffallend ist hierbei der Qualitätssprung zwischen PEV, die bis zu zwei Methoden einsetzen, und PEV, in denen drei oder mehr Methoden angewandt werden. Wie in den vergangenen Jahren ist allerdings der Methodenmix recht einseitig. Die PEV-Prüfteams arbeiten vornehmlich mit Interviews und Dokumentenauswertungen. Darüber hinaus erfolgt zum Teil eine eigene Auswertung von Monitoringdaten durch die PEV-Teams, und in einigen Fällen wird mit Fokusgruppen gearbeitet.

- Die zweite Stellschraube für die methodische Genauigkeit ist das Mengengerüst für die Gutachter.

Hier weisen die Daten darauf hin, dass die methodische Qualität mit steigender Anzahl an Gutachtertagen im Mittel zunächst ansteigt, der Zusammenhang dann jedoch abflacht. Es lässt sich vor allem ein positiver Zusammenhang zwischen der Anzahl der Gutachtertage für die Vorbereitung und Durchführung der PEV und deren methodischer Qualität ausmachen. Allerdings ist hierbei zu berücksichtigen, dass in den PEV-Berichten in der Regel nicht differenziert wird zwischen Anzahl der Gutachtertage für die Evaluierung und ggf. Anzahl der Tage für die Prüfung eines Folgevorhabens.

- Weiterhin lässt sich ein positiver Zusammenhang ausmachen zwischen einer klaren Darstellung des Evaluierungsgegenstands sowie der Auseinandersetzung mit dessen Wirkungslogik und der methodischen Qualität der Evaluierung. Darüber hinaus zeichnen sich PEV, die eine klare Trennung zwischen dem Dreischnitt Beschreibung, Analyse und Bewertung vornehmen, durch eine vergleichsweise höhere Genauigkeit aus. Dieser Sachverhalt entspricht dem allgemeinen Evaluierungsverständnis, dass intendierte Ziele und Wege zur Zielerreichung nachvollziehbar darstellt sein müssen, um eine differenzierte Auseinandersetzung mit den Ergebnissen zu gewährleisten.

Mit Blick auf die **Trends in der Qualität der dezentralen PEV** über vier Jahre hinweg lässt sich eine Verbesserung der methodischen Qualität in fast allen untersuchten Bewertungskriterien ausmachen. Von den 13 Bewertungskriterien, die in allen Meta-Evaluierungen erfasst wurden, ist für sieben Kriterien zwischen 2015 und 2018 eine durchschnittliche Steigerung von zehn Prozentpunkten und mehr auszumachen. Die stärkste Verbesserung ist dabei für das Bewertungskriterium zu validen und reliablen Informationen, sowie für das Bewertungskriterium zur Beschreibung des Evaluierungsgestandes auszumachen.

Auf Ebene der Indikatoren sind diese Entwicklungen vor allem auf Verbesserungen hinsichtlich der Datentriangulation und der Darstellung relevanter Wirkungshypothesen zurückzuführen. Eine rückläufige Tendenz zeigt sich hingegen für die Durchschnittswerte hinsichtlich des Bewertungskriteriums zur systematischen Fehlerprüfung. Nur minimale Veränderungen gab es hinsichtlich der Kriterien zu Beschreibung von Zweck und Vorgehen und zur Bearbeitung der Nachhaltigkeit.

In der Gesamtschau zeigt sich jährlich eine leichte, aber über die Jahre hinweg stetig positive Entwicklung der Qualität der PEVs, ohne dass sich die formalen Vorgaben an die dezentralen Evaluierungen im Untersuchungszeitraum verändert hätten. Aus Sicht des Meta-Evaluierungsteams sind mögliche Erklärungsfaktoren hierfür die Rückmeldungen der Stabsstelle Evaluierung an die PEV-Prüfteams z.B. im Rahmen von Qualitätschecks[39], aber auch eine allgemeine Sensibilisierung innerhalb der GIZ für Qualitätsaspekte wie bspw. die Auseinandersetzung mit dem Wirkungsmodell oder die Berücksichtigung von SMART-Qualitätskriterien für Indikatoren. Trotz der insgesamt positiven Entwicklungen wurden jedoch auch in den Jahren 2017 und 2018 zum Teil grundlegende Evaluierungsstandards in den PEV nicht eingehalten. Es ist daher zu begrüßen, dass mit der Umstellung von dezentralen auf zentralen Evaluierungen im GIZ Evaluierungssystem höhere methodische Anforderungen in den Vorgaben verankert und in der Abnahme der Berichte nachgehalten werden.

[39] In der Meta-Evaluierung 2017 / 2018 ließen sich keine signifikanten Qualitätsunterschiede feststellen zwischen PEVs, für die ein PEV-Check durchgeführt wurde, und solche, für die kein PEV-Check durchgeführt wurde. Dies schließt jedoch nicht aus, dass Evaluatoren, die in der Vergangenheit Rückmeldungen im Rahmen von PEV-Checks bekommen haben und später an der Umsetzung weiterer PEVs beteiligt waren, Feedback von der Stabsstelle berücksichtigt haben.

Literaturverzeichnis

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

DeGEval (2008). Standards für Evaluation. 4. unveränderte Auflage. Mainz.

Freimann, I., Krämer, M. (2016). Querschnittsauswertung (QSA) von Projektevaluierungen (PEV) 2015 – Meta-Evaluierung. Bonn und Eschborn. Herausgeber: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). Online unter https://www.giz.de/de/downloads/giz2016-de-Finaler_Bericht_Metaevaluierung_der_PEV.pdf .

GIZ (2016). Unternehmensorganigramm. Online unter https://www.giz.de/de/downloads/giz2016-unternehmensorganigramm-de.pdf (zuletzt abgerufen am 15.12.2016).

GIZ (2015 a). Handreichung zum methodischen Vorgehen bei der Projektevaluierung und Prüfung von TZ-Folgemaßnahmen im Auftrag des BMZ. Anlage 2 zum Leitfaden zur Vorbereitung von TZ-Maßnahmen im Auftrag des BMZ. Bonn und Eschborn.

GIZ (2015 b). Fahrplan Projektevaluierung.

GIZ (2015 c). Anleitung für die Nutzung der Checkliste für die Bewertung der OECD/DAC-Kriterien.

GIZ (2015 d). Annotierte Berichtsgliederung / Projektevaluierungsbericht.

GIZ (2015 e). Annotierte Berichtsgliederung / Projektevaluierung Kurzbericht.

GIZ (2015 f). Annotierte Berichtsgliederung / Managementanlage zum Projektevaluierungsbericht (intern).

GIZ (2015 g). Kommentierungsblatt für den Qualitätscheck von Projektevaluierungen.

GIZ (2015 h). Methoden-Toolbox. Bonn und Eschborn.

GIZ (2015 i). Das Wirkungsmodell der GIZ. Eine Arbeitshilfe. Bonn und Eschborn.

GIZ (2014). Indikatoren. Eine Arbeitshilfe. Bonn und Eschborn.

GIZ (2013). Policy für Monitoring und Evaluierung der GIZ. Bonn und Eschborn.

GIZ (2018). Terms of Reference für Projektevaluierung und Prüfung einer Folgemaßnahme.

GIZ (o. D. b). Selbsteinschätzung nach Capacity WORKS.

GIZ (o. D. c). Berichtsgliederung / Projektevaluierungsbericht.

GIZ (o. D. d). Berichtsgliederung / Projektevaluierung Kurzbericht.

GIZ (o. D. e). Berichtsgliederung / Managementanlage zum Projektevaluierungsbericht (intern).

JCSEE (2006). Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation”. 3. erweiterte und aktualisierte Auflage. VS Verlag für Sozialwissenschaften.

JCSEE (2000). Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation”. 2. Auflage. Opladen: Leske + Budrich.

OECD-DAC (2010). DAC-Reihe Leitlinien und Grundsatztexte. Qualitätsstandards für die Entwicklungsevaluierung.

OECD-DAC (1991). Principles for the Evaluation of Development Assistance.

Schäferhoff, M., Schrade, C., Corhs, T. (SEEK Development), Raetzell, L., Krämer, M. (Rambøll Management Consulting) (2013). GIZ Review Gesundheit: Hauptbericht. Querschnittauswertung mit Metaevaluierung, Effizienzanalyse und Evaluierungssynthese.

Stern, T., Scheller, O., Freimann, I. (2015). Externe Qualitätskontrolle der GIZ. Ergebnisbericht 2015.

Widmer, T. (1996). Meta-Evaluation. Kriterien zur Bewertung von Evaluationen. Bern, Stuttgart, Wien: Verlag Paul Haupt.

2.9 Anlagen

- Anlage 1: Überblick über die Ergebnisse für den Evaluierungsstandard Genauigkeit für die Jahre 2017 / 2018

- Anlage 2: Analyseraster 2018

- Anlage 3: Überblick über die Ergebnisse für den Evaluierungsstandard Genauigkeit für die Jahre 2015 – 2018 (digital)

Anlage 1: Überblick über die Ergebnisse im Evaluierungsstandard Genauigkeit für die Jahre 2017 / 2018[40]

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 1 Beschreibung des Evaluationsgegenstandes	Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert werden kann.	**G 1 Beschreibung des Evaluationsgegenstandes**						75%	**G 1 Beschreibung des Evaluationsgegenstandes**						81%	**G 1 Beschreibung des Evaluationsgegenstandes**						77%
G 1.1 Der Evaluierungsgegenstand wird genau beschrieben.	Der Indikator ist dann erfüllt, wenn die Evaluierung auf einer genauen Definition und Abgrenzung des Evaluierungsgegenstands basiert. Dies ist dann gegeben, wenn mindestens vier der folgenden sechs Aspekte beschrieben sind: zeitliche Abgrenzung (Phase), aufgewendete Mittel, regionale Abgrenzung, inhaltliche Abgrenzung (Mehrebenenansatz, Systemgrenze, CD Ebenen). **(Textanalyse der PEV-Berichte)**	36	29%	89	71%	125	0	71%	8	16%	43	84%	51	0	84%	44	25%	132	75%	176	0	75%
Checklist:	zeitliche Abgrenzung (Phase)	19	15%	106	85%	125	0		5	10%	46	90%	51	0		24	14%	152	86%	176	0	
	aufgewendete Mittel	66	53%	59	47%	125	0		27	53%	24	47%	51	0		93	53%	83	47%	176	0	
	regionale Abgrenzung	21	17%	104	83%	125	0		7	14%	44	86%	51	0		28	16%	148	84%	176	0	
	inhaltliche Abgrenzung (Mehrebenenansatz)	46	37%	79	63%	125	0		15	29%	36	71%	51	0		61	35%	115	65%	176	0	
	inhaltliche Abgrenzung (Systemgrenze)	33	26%	92	74%	125	0		16	31%	35	69%	51	0		49	28%	127	72%	176	0	
	inhaltliche Abgrenzung (CD Ebenen)	55	44%	70	56%	125	0		13	25%	38	75%	51	0		68	39%	108	61%	176	0	
G 1.2 Das Wirkungsmodell wird dargestellt.	Der Indikator ist dann erfüllt, wenn die Wirkungslogik der Maßnahme des aktuellen Moduls graphisch dargestellt oder textlich eindeutig beschrieben wird, so dass die Systemgrenzen des Vorhabens deutlich werden. Falls für die evaluierte Maßnahme kein Wirkungsmodell vorliegt, muss dies nachträglich (graphisch oder textlich) dargestellt werden. **(Textanalyse der PEV-Berichte)**	20	16%	105	84%	125	0	84%	5	10%	46	90%	51	0	90%	25	14%	151	86%	176	0	86%
G 1.3 Die relevanten Wirkungshypothesen werden dargestellt.	Der Indikator ist dann erfüllt, wenn die Wirkungslogik der Maßnahme anhand eines Narrativs dargestellt wird. Die Wirkungshypothesen müssen begründete Vermutungen über Zusammenhänge innerhalb des Wirkungsmodells darstellen. Es muss deutlich werden, was unter welchen Umständen und warum erwartet werden kann. Jede Hypothese muss zudem empirisch überprüfbar und falsifizierbar sein. Als Mindeststandard gilt, dass zumindest die Einordnung des Modulziels auf der Outcome Ebene und des Programmziels (falls vorhanden) auf der Impact Ebene und die entsprechenden Wirkungsannahmen (kausale Annahmen bei Outcome und zumindest plausible Annahmen bei Impact) auf diese Zielsetzungen hin deutlich werden sollen. **(Textanalyse der PEV-Berichte)**	37	30%	88	70%	125	0	70%	16	31%	35	69%	51	0	69%	53	30%	123	70%	176	0	70%

[40] Die sechs Indikatoren unter dem Buchstaben Z (blau markiert) sind neu und wurden separat ausgewertet. Sie sind in die aggregierte Analyse der bestehenden Bewertungskriterien somit nicht eingeflossen, um die Vergleichbarkeit mit den vorherigen Zeiträumen zu gewährleisten.

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 2 Rahmenbedingungen	Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden.	**G 2 Rahmenbedingungen**						**82%**	**G 2 Rahmenbedingungen**						**84%**	**G 2 Rahmenbedingungen**						**83%**
G 2.1 Der Evaluierungsgegenstand wird im Politikkontext des Partnerlandes verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht den Politikkontext beschreibt: Beschreibung von Strategien, Leitlinien oder Zielen des Partnerlandes. Kenntnisse des Politikkontexts sind erforderlich, um insbesondere die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**	9	7%	116	93%	125	0	93%	2	4%	49	96%	51	0	96.08%	11	6%	165	94%	176	0	94%
G 2.2 Der Evaluierungsgegenstand wird im Entwicklungskontext des Sektors im Partnerland verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht den Entwicklungskontext im Sektor bescheibt: sozioökonomische, politische oder kulturelle Rahmenbedingungen im Sektor. Kenntnisse des Entwicklungskontexts sind erforderlich, um insbesondere die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**	6	6%	94	94%	100	25	94%	1	3%	38	97%	39	12	97.44%	7	5%	132	95%	139	37	95%
G 2.3 Der Evaluierungsgegenstand wird in der Träger- und Partnerstruktur verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht die Träger- und der Partnerstruktur beschreibt: Kenntnisse der Träger- und Partnerstruktur gehören zum institutionellen Kontexts einer Maßnahme und sind erforderlich, um die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**	38	30%	87	70%	125	0	70%	16	32%	34	68%	50	1	68.00%	54	31%	121	69%	175	1	69%
G 2.4 In der Bewertung und in den Schlussfolgerungen wird ein Rückbezug auf die Kontextanalyse vorgenommen.	Der Indikator ist dann erfüllt, wenn deutlich wird, wie der untersuchte Kontext die Bewertung der Evaluierung beeinflusst. Der Einfluss der Kontextfaktoren auf die Ergebnisse der Evaluierung wird transparent gemacht. Dies hilft auch, eine mögliche Übertragbarkeit der Evaluierungsergebnisse auf andere Maßnahmen einschätzen zu können. **(Textanalyse der PEV-Berichte)**	33	26%	92	74%	125	0	74%	11	22%	40	78%	51	0	78.43%	44	25%	132	75%	176	0	75%
G 3 Beschreibung von Zwecken und Vorgehen	Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschrieben werden, so dass sie identifiziert und eingeschätzt werden können.	**G 3 Beschreibung von Zwecken und Vorgehen**						**64%**	**G 3 Beschreibung von Zwecken und Vorgehen**						**61%**	**G 3 Beschreibung von Zwecken und Vorgehen**						**63%**
G 3.1 Anlass, Zweck und beabsichtigte Verwendung der Evaluierung werden transparent beschrieben.	Der Indikator ist dann erfüllt, wenn deutlich wird, was der Anlass und Zweck (Zwischen- oder Schlussevaluierung) der Evaluierung sind und wer bzw. wie die Ergebnisse verwendet soll/en (Nutzer/ Nutzung). **(Textanalyse der PEV-Berichte)**	76	61%	49	39%	125	0	39%	31	61%	20	39%	51	0	39%	107	61%	69	39%	176	0	39%
G 3.2 Die spezifischen Ziele der Evaluierung werden deutlich.	Der Indikator ist dann erfüllt, wenn transparent dargestellt ist, was mit der Evaluierung erreicht werden soll (bspw. Überprüfung von Relevanz, Zielerreichung, Nachhaltigkeit, Handlungsempfehlungen für Folgemodul, Übertragbarkeit auf andere Kontexte, etc.). Die spezifischen Evaluierungsziele müssen aus den TOR abgeleitet sein. **(Textanalyse der PEV-Berichte)**	73	58%	52	42%	125	0	42%	38	75%	13	25%	51	0	25%	111	63%	65	37%	176	0	37%
G 3.3 Die TOR einschließlich der Evaluierungsfragen der Prüfmission befinden sich im Anhang des Prüfberichts.	Der Indikator ist dann erfüllt, wenn die TOR mit den Evaluierungsfragen ein Anhang des Evaluierungsberichts sind. So kann beurteilt werden, inwieweit sich die Prüfmission ausreichend mit den Evaluierungsfragen auseinander gesetzt hat. **(Textanalyse der PEV-Berichte)**	17	14%	108	86%	125	0	86%	6	12%	45	88%	51	0	88%	23	13%	153	87%	176	0	87%
G 3.4 Anhand der TOR lassen sich Aussagen zum Zweck der Evaluierung ableiten.	Der Indikator ist dann erfüllt, wenn die TOR den Anlass und Zweck der Evaluierung klar definieren. **(Textanalyse der PEV-Berichte)**	13	10%	112	90%	125	0	90%	4	8%	47	92%	51	0	92%	18	10%	159	90%	176	0	90%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 4 Angabe von Informationsquellen	Die im Rahmen einer Evaluation genutzten Informationsquellen sollen hinreichend genau dokumentiert werden, damit die Verlässlichkeit und Angemessenheit der Informationen eingeschätzt werden kann.	**G 4 Angabe von Informationsquellen**						**50%**	**G 4 Angabe von Informationsquellen**						**54%**	**G 4 Angabe von Informationsquellen**						**51%**
G 4.1 Der Bericht umfasst eine vollständige Liste der Gesprächspartner/innen und sonstigen Informationsquellen.	Der Indikator ist dann erfüllt, wenn die verwendeten Informationsquellen beschrieben werden, so dass die Aussagen im Bericht evidenzbasiert unterlegt sind. Falls Anonymität und das Recht auf Vertraulichkeit der einzelnen Gesprächspartner/innen die Darstellung der Informationsquellen einschränken, wird entsprechend darauf hingewiesen. **(Textanalyse der PEV-Berichte)**	41	33%	84	67%	125	0	67%	12	24%	39	76%	51	0	76%	53	30%	123	70%	176	0	70%
G 4.2 Der Bericht stellt transparent dar, welcher Systematik der Dokumentenauswertung zugrunde lag.	Der Indikator ist dann erfüllt, wenn die Auswertung der verwendeten Dokumente beschrieben werden. Hierdurch wird deutlich, welche Dokumente warum ausgewertet wurden und welches System der Dokumentenauswertung zugrunde lag. **(Textanalyse der PEV-Berichte)**	107	86%	18	14%	125	0	14%	39	76%	12	24%	51	0	24%	146	83%	30	17%	176	0	17%
Checklist:	Bezug Programmvorschlag	48	38%	77	62%	125	0		22	43%	29	57%	51	0		70	40%	106	60%	176	0	
	Bezug Operationsplan	47	38%	78	62%	125	0		16	31%	35	69%	51	0		63	36%	113	64%	176	0	
	Bezug Steuerungsstruktur	79	63%	46	37%	125	0		32	63%	19	37%	51	0		111	63%	65	37%	176	0	
	Bezug CD-Strategie	52	42%	73	58%	125	0		29	57%	22	43%	51	0		81	46%	95	54%	176	0	
	Bezug Akteursanalyse	40	32%	85	68%	125	0		18	35%	33	65%	51	0		58	33%	118	67%	176	0	
	Bezug Wirkungsmodell	44	35%	81	65%	125	0		16	31%	35	69%	51	0		60	34%	116	66%	176	0	
G 4.3 Die Auswahl von Gesprächspartner bzw. die Ziehung von Stichproben geschieht systematisch.	Der Indikator ist dann erfüllt, wenn die Auswahl von Gesprächspartner/-innen (bzw. Stichproben bei quantitativen Analysen) nach klar definierten und transparenten Kriterien geschieht (z.B. Reflektion der Erwartungen an die Aussagefähigkeit der TN, Zusammensetzung der TN, Rolle der Prüfmission bei der Auswahl der TN), und die Auswirkungen der Auswahl auf die Repräsentativität der Ergebnisse erläutert wird. Da im Rahmen einer Prüfmission nicht jeder befragt werden kann, der zur Grundgesamtheit gehört, muss eine begründete Auswahl aus der Grundgesamtheit vorgenommen werden. **(Textanalyse der PEV-Berichte)**	47	38%	78	62%	125	0	62%	22	43%	29	57%	51	0	57%	69	39%	107	61%	176	0	61%
G 4.4 Aufbauend auf die Bewertung der Verlässlichkeit der Daten des WOM, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Daten des WOM erfolgt und die Daten des WOM dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	19	15%	106	85%	125	0	85%	9	18%	42	82%	51	0	82%	28	16%	148	84%	176	0	84%
G 4.5 Aufbauend auf die Bewertung der Verlässlichkeit der Baseline Daten, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Baseline Daten erfolgt und die Baseline Daten dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	76	61%	49	39%	125	0	39%	28	55%	23	45%	51	0	45%	104	59%	72	41%	176	0	41%
G 4.6 Aufbauend auf die Bewertung der Verlässlichkeit der Partnerdaten, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Partnerdaten erfolgt und die Partnerdaten dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	74	74%	26	26%	100	25	26%	25	64%	14	36%	39	12	36%	99	71%	40	29%	139	37	29%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 5 Valide und reliable Informationen	Die Verfahren zur Gewinnung von Daten sollen so gewählt oder entwickelt und dann eingesetzt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien quantitativer und qualitativer Sozialforschung orientieren.	**G 5 Valide und reliable Informationen**						**75%**	**G 5 Valide und reliable Informationen**						**76%**	**G 5 Valide und reliable Informationen**						**75%**
G 5.1 Es wird eine Datentriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn Daten / Informationen zu einem gleichen Sachverhalt durch die Einbeziehung verschiedener Akteure erhoben werden. **(Textanalyse der PEV-Berichte)**	12	10%	113	90%	125	0	90%	5	10%	46	90%	51	0	90%	17	10%	159	90%	176	0	90%
G 5.2 Es wird eine Methodentriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn Informationen zu einem gleichen Sachverhalt durch verschiedene Methoden der Datensammlung erhoben werden. **(Textanalyse der PEV-Berichte)**	51	41%	74	59%	125	0	59%	19	37%	32	63%	51	0	63%	70	40%	106	60%	176	0	60%
G 6 Systematische Fehlerprüfung	Die in einer Evaluation gesammelten, aufbereiteten, analysierten und präsentierten Informationen sollen systematisch auf Fehler geprüft werden.	**G 6 Systematische Fehlerprüfung**						**68%**	**G 6 Systematische Fehlerprüfung**						**59%**	**G 6 Systematische Fehlerprüfung**						**65%**
G 6.1 Es wird eine Forschertriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn das PEV-Team seine Erfahrungen und Interpretationen / Perspektiven trianguliert. Hinweise auf die systematische Überprüfung der gesammelten Informationen sind z.B. Synthesetreffen, die dokumentiert wurden, oder die Darstellung auf Unstimmigkeiten der Mitglieder des PEV-Teams, die entsprechend diskutiert wurden. **(Textanalyse der PEV-Berichte)**	54	43%	71	57%	125	0	57%	29	57%	22	43%	51	0	43%	83	47%	93	53%	176	0	53%
G 6.2 Es wird eine Triangulation der Ergebnisse mit den Beteiligten und Betroffenen durchgeführt.	Der Indikator ist dann erfüllt, wenn eine Triangulation der Ergebnisse mit den Beteiligten und Betroffenen erfolgt. So soll eine Fehlerüberprüfung der Analyseergebnisse erfolgen. Hinweise auf die systematische Überprüfung der gesammelten Informationen sind z.B. gemeinsame Syntheseworkshops oder die Kommentierung des Berichts durch Beteiligte und Betroffene. **(Textanalyse der PEV-Berichte)**	26	21%	99	79%	125	0	79%	13	25%	38	75%	51	0	75%	39	22%	137	78%	176	0	78%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 7 (A) Analyse qualitativer und quantitativer Informationen	Qualitative und quantitative Informationen einer Evaluation sollen nach fachlichen Maßstäben angemessen und systematisch analysiert werden, damit die Fragestellungen der Evaluation effektiv beantwortet werden können.	**G 7 (A) Analyse qualitativer und quantitativer Informationen**						**63%**	**G 7 (A) Analyse qualitativer und quantitativer Informationen**						**63%**	**G 7 (A) Analyse qualitativer und quantitativer Informationen**						**63%**
G 7.1 (A) Die methodische Vorgehensweise beantwortet die Frage nach der Zuordnungs- und/ oder Beitragsanalyse.	Der Indikator ist dann erfüllt, wenn die methodische Vorgehensweise mit Hinweis auf die Beantwortung von Evaluierungsfragen, Bedingungen der Untersuchung und/ oder Besonderheiten des Evaluierungsgegenstandes begründet dargestellt wird. So kann bspw. eine Attributionsanalyse mit dem Hinweis auf den Evaluierungsgegenstand begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	96	77%	29	23%	125	0	23%	37	73%	14	27%	51	0	27%	133	76%	43	24%	176	0	24%
Checklist:	keine begründete Vorgehensweise	29	23%	96	77%	125	0		14	27%	37	73%	51	0		43	24%	133	76%	176	0	
	experimentelles Design	125	100%	0	0%	125	0		51	100%	0	0%	51	0		176	100%	0	0%	176	0	
	quasi-experimentelles Design	125	100%	0	0%	125	0		51	100%	0	0%	51	0		176	100%	0	0%	176	0	
	Ansatz einer Kontributionsanalyse	100	80%	25	20%	125	0		40	78%	11	22%	51	0		140	80%	36	20%	176	0	
	theoriebasierter Ansatz	121	97%	4	3%	125	0		48	94%	3	6%	51	0		169	96%	7	4%	176	0	
G 7.2 (A) Die Evaluierungsmethoden werden dargestellt.	Der Indikator ist dann erfüllt, wenn die verwendeten Methoden beschrieben werden, so dass die Aussagen evidenzbasiert unterlegt sind. **(Textanalyse der PEV-Berichte)**	6	5%	119	95%	125	0	95%	1	2%	50	98%	51	0	98%	7	4%	169	96%	176	0	96%
Checklist:	Allgemeine Beschreibung der Methoden	7	6%	118	94%	125	0		1	2%	50	98%	51	0		8	5%	168	95%	176	0	
	Es wird deutlich, wie die Methoden im Prozess der Datenerhebung eingesetzt wurden - Analyseraster, Datenerhebungsplan o.ä. analyseleitendes System	97	78%	28	22%	125	0		38	75%	13	25%	51	0		135	77%	41	23%	176	0	
	Es wird deutlich, wie die erhobenen Daten dokumentiert wurden - Ergebnisberichte, Matrizen, Transkription von Interviews, Interviewprotokolle, Fallbeispiele, etc.	121	97%	4	3%	125	0		51	100%	0	0%	51	0		172	98%	4	2%	176	0	
	Es wird deutlich, wie die erhobenen qualitativen Daten ausgewertet wurden - Codierung, Auswertungsraster, Inhaltsanalyse o.ä.	124	99%	1	1%	125	0		49	96%	2	4%	51	0		173	98%	3	2%	176	0	
	Es wird deutlich, wie die erhobenen quantitativen Daten ausgewertet wurden - uni-, bi- oder multivariate Auswertungen	125	100%	0	0%	125	0		51	100%	0	0%	51	0		176	100%	0	0%	176	0	
G 7.3 (A) Es gibt eine Methodenvielfalt.	Der Indikator ist dann erfüllt, wenn mehr als eine Methode verwendet wird. **(Textanalyse der PEV-Berichte)**	0	0%	125	100%	125	0	100%	0	0%	51	100%	51	0	100%	0	0%	176	100%	176	0	100%
Checklist:	Dokumentenauswertung	1	1%	124	99%	125	0		1	2%	50	98%	51	0		2	1%	174	99%	176	0	
	Fokusgruppen	98	78%	27	22%	125	0		41	80%	10	20%	51	0		139	79%	37	21%	176	0	
	Interviews	0	0%	125	100%	125	0		0	0%	51	100%	51	0		0	0%	176	100%	176	0	
	Befragung	120	96%	5	4%	125	0		43	84%	8	16%	51	0		163	93%	13	7%	176	0	
	eigene Auswertung der Monitoringdaten durch die Prüfmission	72	58%	53	42%	125	0		28	55%	23	45%	51	0		100	57%	76	43%	176	0	
G 7.4 (A) Die Vor- und Nachteile der gewählten Methoden werden dargestellt.	Der Indikator ist dann erfüllt, wenn die Methodenauswahl nachvollziehbar ist. Diese Nachvollziehbarkeit spiegelt sich in einer Klärung darüber, welche Vorteile die ausgesuchte Methode hat und welche möglichen Nachteile vorhanden sind und wie damit umgegangen werden soll, wider. **(Textanalyse der PEV-Berichte)**	83	66%	42	34%	125	0	34%	37	73%	14	27%	51	0	27%	120	68%	56	32%	176	0	32%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz	Die methodische Bearbeitung der Relevanzbewertung ist angemessen.	**G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz**						**72%**	**G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz**						**73%**	**G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz**						**72%**
G 7.5 (B) Die Nachvollziehbarkeit der zugrunde gelegten Rahmenbedingungen und Kernprobleme der Maßnahme ist gegeben.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Relevanz die Rahmenbedingungen und Kernbedarfe, auf die die Maßnahme beruht, überprüft wird. **(Textanalyse der PEV-Berichte)**	12	10%	113	90%	125	0	90%	3	6%	48	94%	51	0	94%	15	9%	161	91%	176	0	91%
G 7.6 (B) Die Mehrdimensionalität der Relevanz wird analysiert.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Relevanz entlang der Kernbedarfe von Zielgruppe und Partner sowie der Ziele des Auftraggebers analysiert wird. **(Textanalyse der PEV-Berichte)**	8	6%	117	94%	125	0	94%	5	10%	46	90%	51	0	90%	13	7%	163	93%	176	0	93%
G 7.7 (B) Die strategische Ausrichtung der Maßnahme an veränderten Rahmenbedingungen werden analysiert.	Der Indikator ist dann erfüllt, wenn die strategische Ausrichtung einer Maßnahme an Veränderungen in den Rahmenbedingungen analysiert werden. **(Textanalyse der PEV-Berichte)**	84	67%	41	33%	125	0	33%	34	67%	17	33%	51	0	33%	118	67%	58	33%	176	0	33%
Z 7 (B) Die Bewertung des OECD/DAC Kriteriums "Relevanz" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**	7	6%	118	94%	125	0	94%	2	4%	49	96%	51	0	96%	9	5%	167	95%	176	0	95%
G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität	Die methodische Bearbeitung der Effektivitätsbewertung ist angemessen.	**G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität**						**74%**	**G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität**						**69%**	**G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität**						**72%**
G 7.8 (C) Die Kausalität zwischen Maßnahmen und Wirkungen wird differenziert analysiert und eingeschätzt.	Der Indikator ist dann erfüllt, wenn beobachtete Zusammenhänge zwischen Maßnahmen und Wirkungen nicht ohne Weiteres als ursächliche Zusammenhänge interpretiert werden. Der Einfluss von Drittvariablen muss ausgeschlossen werden. Korrelationen dürfen nicht mit Kausalitäten gleichgesetzt werden. **(Textanalyse der PEV-Berichte)**	63	50%	62	50%	125	0	50%	29	57%	22	43%	51	0	43%	92	52%	84	48%	176	0	48%
G 7.9 (C) Die Zielerreichung wird anhand von Modulindikatoren bewertet.	Der Indikator ist dann erfüllt, wenn deutlich wird, welche Indikatoren zur Bewertung der Zielerreichung als Grundlage genommen wurden. Wurden während der Projektplanung für einzelne Aspekte von Effektivität keine Indikatoren definiert, müssen diese Indikatoren nachträglich gebildet werden. **(Textanalyse der PEV-Berichte)**	1	1%	124	99%	125	0	99%	1	2%	50	98%	51	0	98%	2	1%	174	99%	176	0	99%
G 7.10 (C) Die verwendeten Indikatoren zur Messung und Beurteilung der Zielerreichung sind SMART (spezifisch, messbar, erreichbar, relevant, zeitgebunden).	Der Indikator ist dann erfüllt, wenn ersichtlich ist, dass sich die Prüfer mit der Qualität der Indikatoren auseinandergesetzt haben (zum Beispiel durch die Nutzung von SMART-Kriterien). Wird eine geringe Qualität festgestellt, müssen neue bzw. veränderte Indikatoren für die Evaluation genutzt werden. **(Textanalyse der PEV-Berichte)**	34	27%	91	73%	125	0	73%	18	35%	33	65%	51	0	65%	52	30%	124	70%	176	0	70%
Z 7 (C) Die Bewertung des OECD/DAC Kriteriums "Effektivität" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**	14	11%	111	89%	125	0	89%	6	12%	45	88%	51	0	88%	20	11%	156	89%	176	0	89%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz	Die methodische Bearbeitung der Effizienzbewertung ist angemessen.	**G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz**						**41%**	**G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz**						**39%**	**G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz**						**40%**
G 7.11 (D) Die verschiedenen Ebenen der Effizienz einer Maßnahme werden analysiert.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht dargestellt ist, welche Ebenen der Effizienzmessung analysiert werden und die Auswahl begründet wird. Mögliche Ebenen sind: die Implementierunseffizienz (Analyse der Strukturen und Prozesse der Vorhabensumsetzung), der Produktionseffizienz (Verhältnis Input zu Output) und der Allokationseffizienz (Verhältnis von Input zu Outcome). **(Textanalyse der PEV-Berichte)**	40	32%	85	68%	125	0	68%	22	43%	29	57%	51	0	57%	62	35%	114	65%	176	0	65%
Checklist:	Implementierungseffizienz	50	40%	75	60%	125	0		25	49%	26	51%	51	0		75	43%	101	57%	176	0	
	Produktionseffizienz,	108	86%	17	14%	125	0		43	84%	8	16%	51	0		151	86%	25	14%	176	0	
	Allokationseffizienz	119	95%	6	5%	125	0		49	96%	2	4%	51	0		169	96%	7	4%	176	0	
G 7.12 (D) Die Auswahl von Methoden und Verfahren der Effizienzmessung wird begründet.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht die Auswahl der Methoden und Verfahren zur Messung der Effizient begründet wird. **(Textanalyse der PEV-Berichte)**	109	87%	16	13%	125	0	13%	46	90%	5	10%	51	0	10%	155	88%	21	12%	176	0	12%
Checklist:	Level 2 methods	125	100%	0	0%	125	0		51	100%	0	0%	51	0		176	100%	0	0%	176	0	
	Level 1 methods	120	96%	5	4%	125	0		48	94%	3	6%	51	0		168	95%	8	5%	176	0	
	Descriptive methods	111	89%	14	11%	125	0		46	90%	5	10%	51	0		157	89%	19	11%	176	0	
G 7.13 (D) Die Bearbeitung der Effizienz ermöglicht die Identifikation von Potenzialen zur Effizienzsteigerung.	Der Indikator ist dann erfüllt, wenn die verwendeten Methoden und Verfahren implizit oder explizit Komponenten/Teile der Maßnahme mit Alternativen vergleichen, um Verbesserungspotenziale aufzuzeigen. **(Textanalyse der PEV-Berichte)**	72	58%	53	42%	125	0	42%	26	51%	25	49%	51	0	49%	98	56%	78	44%	176	0	44%
Z 7 (D) Die Bewertung des OECD/DAC Kriteriums "Effizienz" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**	13	10%	112	90%	125	0	90%	5	10%	46	90%	51	0	90%	18	10%	158	90%	176	0	90%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact	Die methodisch Bearbeitung der Impactbewertung ist angemessen.	**G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact**						**61%**	**G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact**						**59%**	**G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact**						**60%**
G 7.14 (E) In der Analyse und Beurteilung der übergeordneten Wirkungen (Impacts) wird die Attributionslücke analysiert.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht bei der Analyse und Beurteilung der übergeordneten Wirkungen die Zuordnungs- bzw. Attributionslücke reflektiert wird. Es wird kein direkter und monokausaler Zusammenhang zwischen Maßnahme und übergeordneten Zielen hergestellt. **(Textanalyse der PEV-Berichte)**	40	32%	85	68%	125	0	68%	17	33%	34	67%	51	0	67%	57	32%	119	68%	176	0	68%
G 7.15 (E) Die Plausibilität der Hypothesen zu den intendierten langfristigen Wirkungen (Impacts) wird bewertet.	Der Indikator ist dann erfüllt, wenn sich die Evaluation mit der Plausibilität der Hypothesen zu übergeordneten langfristigen Wirkungen auseinandersetzt. **(Textanalyse der PEV-Berichte)**	55	44%	70	56%	125	0	56%	26	51%	25	49%	51	0	49%	81	46%	95	54%	176	0	54%
G 7.16 (E) Bewertungsmaßstäbe zur Analyse und Beurteilung des Beitrages der Maßnahme zu übergeordneten Wirkungen (Impacts) sind transparent dargestellt.	Der Indikator ist dann erfüllt, wenn beschreiben ist, anhand welcher Bewertungsmaßstäbe der Beitrag des Vorhabens zu übergeordneten entwicklungspolitischen Wirkungen bewertet wird. Diese Bewertungsmaßstäbe können entweder selbst vom Prüfteam entwickelt (zum Beispiel (Proxi-)Indikatoren) oder vorhandene Maßstäbe (zum Beispiel aus Teil A des PV) herangezogen werden. **(Textanalyse der PEV-Berichte)**	52	42%	73	58%	125	0	58%	19	37%	32	63%	51	0	63%	71	40%	105	60%	176	0	60%
Z 7 (E) Die Bewertung des OECD/DAC Kriteriums "Impact" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**	11	9%	114	91%	125	0	91%	3	6%	48	94%	51	0	94%	14	8%	162	92%	176	0	92%
G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit	Die methodische Bearbeitung der Nachhaltigkeitsbewertung ist angemessen.	**G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit**						**63%**	**G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit**						**63%**	**G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit**						**63%**
G 7.17 (F) Die Grenzen der Messung der Nachhaltigkeit werden beschrieben.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Nachhaltigkeit beschrieben wird, inwiefern zum Zeitpunkt der Evaluierung eine ergebnisorientierte Analyse und Beurteilung der Nachhaltigkeit von Effekten überhaupt möglich ist. **(Textanalyse der PEV-Berichte)**	84	67%	41	33%	125	0	33%	32	63%	19	37%	51	0	37%	116	66%	60	34%	176	0	34%
G 7.18 (F) Die angelegten Ansätze zur Schaffung von Nachhaltigkeit werden analysiert.	Der Indikator ist dann erfüllt, wenn in der Analyse und Beurteilung der Nachhaltigkeit die angelegten Ansätze zur Schaffung von Nachhaltigkeit berücksichtigt werden. Zum Beispiel wird evaluiert, inwiefern die Maßnahme eine Exit- oder Nachhaltigkeitsstrategie besitzt, Ansätze für die Anschlussfähigkeit der Maßnahme betreibt oder Aktivitäten zur Verantwortungsübergabe (Ownership) an den Partner betreibt. Die Grenzen der Messung werden dann bei der Analyse und Beurteilung eingehalten. **(Textanalyse der PEV-Berichte)**	31	25%	94	75%	125	0	75%	13	25%	38	75%	51	0	75%	44	25%	132	75%	176	0	75%
G 7.19 (F) Es werden mindestens zwei Ebenen der Nachhaltigkeit im Rahmen einer Prognose analysiert.	Der Indikator ist dann erfüllt, wenn der Mehrdimensionalität von Nachhaltigkeit im Rahmen einer Prognose Rechnung getragen wird und mindestens Nachhaltigkeit auf zwei Ebenen betrachtet wird (finanzielle, institutionelle, personelle, technologische, soziale, ökologische Nachhaltigkeit) überprüft. **(Textanalyse der PEV-Berichte)**	25	20%	100	80%	125	0	80%	12	24%	39	76%	51	0	76%	37	21%	139	79%	176	0	79%
Z 7 (F) Die Bewertung des OECD/DAC Kriteriums "Nachhaltigkeit" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**	16	13%	109	87%	125	0	87%	4	8%	47	92%	51	0	92%	20	11%	156	89%	176	0	89%

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Indikatoren / Deskriptoren 2017						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2018						Erfüllung im Durchschnitt	Indikatoren / Deskriptoren 2017/2018						Erfüllung im Durchschnitt
		Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme			Nein		Ja		Gesamtsumme		
		Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M	Anzahl	%	Anzahl	%	Anzahl	Fehlend	M
G 8 Begründete Analyse und begründete Schlussfolgerung	Die in einer Evaluation gezogenen Folgerungen sollen ausdrücklich begründet werden, damit die Adressaten und Adressatinnen diese einschätzen können.	**G 8 Begründete Analyse und begründete Schlussfolgerung**						**55%**	**G 8 Begründete Analyse und begründete Schlussfolgerung**						**61%**	**G 8 Begründete Analyse und begründete Schlussfolgerung**						**57%**
G 8.1 Es wird zwischen Beschreibung, Analyse und Bewertung unterschieden.	Der Indikator ist dann erfüllt, wenn der Dreischritt Beschreibung – Analyse – Bewertung aus evaluatorischer Einschätzung zumeist eingehalten wird. Das heißt, es wird wird nicht direkt von der Beschreibung zur Bewertung gesprungen bzw. die Analyse und Bewertung werden nicht vermischt. Der Dreischritt soll mindestens bei zwei der folgenden drei Aspekte eingehalten werden: Wirkungsmodell, Bewertung der OECD/DAC-Kriterien, Schlussfolgerungen. **(Textanalyse der PEV-Berichte)**	65	52%	60	48%	125	0	48%	23	45%	28	55%	51	0	55%	88	50%	88	50%	176	0	50%
Checklist:	Wirkungsmodell	49	39%	76	61%	125	0		20	39%	31	61%	51	0		69	39%	107	61%	176	0	
	Bewertung der OECD/ DAC-Kriterien	51	41%	74	59%	125	0		19	37%	32	63%	51	0		70	40%	106	60%	176	0	
	Schlussfolgerungen	104	83%	21	17%	125	0		45	88%	6	12%	51	0		149	85%	27	15%	176	0	
G 8.2 Beschreibungen und Analysen werden belegt.	Der Indikator ist dann erfüllt, wenn Beschreibung und Analyse im Bericht mit relevanten Zahlen, Daten und Fakten untermauert werden. Für alle Zahlen, Daten und Fakten, werden Quellen angegeben. Demnach sind bei der Beschreibung und Analyse von Ergebnissen der Datenerhebung die jeweiligen Informationsquellen ersichtlich. Hierbei werden die Daten, die von der Evaluierungsmission erhoben wurden, als solche gekennzeichnet. **(Textanalyse der PEV-Berichte)**	69	55%	56	45%	125	0	45%	28	55%	23	45%	51	0	45%	97	55%	79	45%	176	0	45%
G 8.3 Empfehlungen werden aus der Analyse abgeleitet und sind spezifisch, realistisch und termingebunden. Sie richten sich an die wesentlichen Nutzer und ihre Umsetzung ist messbar.	Der Indikator ist dann erfüllt, wenn für alle Handlungsempfehlungen ein Ziel bzw. ein Zweck definiert worden ist. Die Empfehlungen müssen realistisch sein und sich aus den Schlussfolgerungen ableiten. Die Realisierbarkeit aller Handlungsempfehlungen muss sichergestellt sein. Z.B indem Handlungsempfehlungen in Zusammenarbeit mit denjenigen entwickelt werden, die sie umsetzen müssen. Alle Handlungsempfehlungen müssen einen Zeitraum bzw. einen Zeitpunkt für die Umsetzung benennen und die Umsetzung muss empirisch erhoben werden können. **(Textanalyse der PEV-Berichte)**	72	58%	53	42%	125	0	42%	23	45%	28	55%	51	0	55%	95	54%	81	46%	176	0	46%
G 8.4 Die positiven und/ oder negativen nicht-intendierten Wirkungen der Maßnahme werden beschrieben.	Der Indikator ist dann erfüllt, wenn sich die Evaluierung mit möglichen positiven und/ oder negativen nicht-intendierten Wirkungen befasst. Hiermit soll gewährleistet werden, dass sich die Evaluierung möglicher Hinweise auf nicht-intendierte Wirkung bewusst ist. **(Textanalyse der PEV-Berichte)**	20	16%	105	84%	125	0	84%	6	12%	45	88%	51	0	88%	26	15%	150	85%	176	0	85%
Z 8 In den Schlussfolgerungen werden die untersuchten Zusammenhänge und Beiträge des Vorhabens differenziert bewertet.	Eine Kontributionsanalyse überprüft die Zusammenhänge aus der Theory-of-Change und ermöglicht dadurch Schlussfolgerungen darüber, welche Zusammenhänge mit Evidenz untermauert sind, welche weiteren Zusammenhänge identifiziert worden sind und welche Zusammenhänge revidiert, bzw. nicht belegt werden konnten. Der Indikator ist dann erfüllt, wenn in den Schlussfolgerungen die analysierten Zusammenhänge differenziert bewertet werden (zum Beispiel nach dem Grad der logischen Plausibilität, der vorgefundenen Evidenz, der vorgefundenen externen Erkärungsfaktoren, des wissenschaftlichen Kenntnisstandes oder der Aussagen von befragten Akteuren) **(Textanalyse der PEV-Berichte)**	53	42%	72	58%	125	0	58%	17	33%	34	67%	51	0	67%	70	40%	106	60%	176	0	60%
Genauigkeit insgesamt								**66%**							**67%**							**66%**

Anlage 2: Analyseraster[41]

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 1 Beschreibung des Evaluationsgegenstandes	Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert werden kann.								
G 1.1 Der Evaluierungsgegenstand wird genau beschrieben.	Der Indikator ist dann erfüllt, wenn die Evaluierung auf einer genauen Definition und Abgrenzung des Evaluierungsgegenstands basiert. Dies ist dann gegeben, wenn mindestens vier der folgenden sechs Aspekte beschrieben sind: zeitliche Abgrenzung (Phase), aufgewendete Mittel, regionale Abgrenzung, inhaltliche Abgrenzung (Mehrebenenansatz, Systemgrenze, CD Ebenen). **(Textanalyse der PEV-Berichte)**	**X**		**X**		**X**	**X**		
Checklist:	zeitliche Abgrenzung (Phase)	X		X		X	X		
	aufgewendete Mittel								
	regionale Abgrenzung	X		X		X	X		
	inhaltliche Abgrenzung (Mehrebenenansatz)	X				X			
	inhaltliche Abgrenzung (Systemgrenze)	X		X		X			
	inhaltliche Abgrenzung (CD Ebenen)	X		X		X			
G 1.2 Das Wirkungsmodell wird dargestellt.	Der Indikator ist dann erfüllt, wenn die Wirkungslogik der Maßnahme des aktuellen Moduls graphisch dargstellt oder textlich eindeutig beschrieben wird, so dass die Systemgrenzen des Vorhabens deutlich werden. Falls für die evaluierte Maßnahme kein Wirkungsmodell vorliegt, muss dies nachträglich (graphisch oder textlich) dargestellt werden. **(Textanalyse der PEV-Berichte)**	**X**		**X**		**X**	**X**		
G 1.3 Die relevanten Wirkungshypothesen werden dargestellt.	Der Indikator ist dann erfüllt, wenn die Wirkungslogik der Maßnahme anhand eines Narrativs dargestellt wird. Die Wirkungshypothesen müssen begründete Vermutungen über Zusammenhänge innerhalb des Wirkungsmodells darstellen. Es muss deutlich werden, was unter welchen Umständen und warum erwartet werden kann. Jede Hypothese muss zudem empirisch überprüfbar und falsifizierbar sein. Als Mindeststandard gilt, dass zumindest die Einordnung des Modulziels auf der Outcome Ebene und des Programmziels (falls vorhanden) auf der Impact Ebene und die entsprechenden Wirkungsannahmen (kausale Annahmen bei Outcome und zumindest plausible Annahmen bei Impact) auf diese Zielsetzungen hin deutlich werden sollen. **(Textanalyse der PEV-Berichte)**	**X**		**X**		**X**	**X**		

[41] Die sechs Indikatoren unter dem Buchstaben Z (blau markiert) sind neu und wurden separat ausgewertet. Sie sind in die aggregierte Analyse der bestehenden Bewertungskriterien somit nicht eingeflossen, um die Vergleichbarkeit mit den vorherigen Zeiträumen zu gewährleisten.

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 2 Rahmenbedingungen	Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden.								
G 2.1 Der Evaluierungsgegenstand wird im Politikkontext des Partnerlandes verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht den Politikkontext beschreibt: Beschreibung von Strategien, Leitlinien oder Zielen des Partnerlandes. Kenntnisse des Politikkontexts sind erforderlich, um insbesondere die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**	**X**							
G 2.2 Der Evaluierungsgegenstand wird im Entwicklungskontext des Sektors im Partnerland verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht den Entwicklungskontext im Sektor bescheibt: sozioökonomische, politische oder kulturelle Rahmenbedingungen im Sektor. Kenntnisse des Entwicklungskontexts sind erforderlich, um insbesondere die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**	**X**							
G 2.3 Der Evaluierungsgegenstand wird in der Träger- und Partnerstruktur verortet.	Der Indikator ist dann erfüllt, wenn der Evaluierungsbericht die Träger- und der Partnerstruktur beschreibt: Kenntnisse der Träger- und Partnerstruktur gehören zum institutionellen Kontexts einer Maßnahme und sind erforderlich, um die Gelingensbedingungen und die Wechselwirkungen der Maßnahme realistisch einzuschätzen. **(Textanalyse der PEV-Berichte)**								
G 2.4 In der Bewertung und in den Schlussfolgerungen wird ein Rückbezug auf die Kontextanalyse vorgenommen.	Der Indikator ist dann erfüllt, wenn deutlich wird, wie der untersuchte Kontext die Bewertung der Evaluierung beeinflusst. Der Einfluss der Kontextfaktoren auf die Ergebnisse der Evaluierung wird transparent gemacht. Dies hilft auch, eine mögliche Übertragbarkeit der Evaluierungsergebnisse auf andere Maßnahmen einschätzen zu können. **(Textanalyse der PEV-Berichte)**								
G 3 Beschreibung von Zwecken und Vorgehen	Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschrieben werden, so dass sie identifiziert und eingeschätzt werden können.								
G 3.1 Anlass, Zweck und beabsichtigte Verwendung der Evaluierung werden transparent beschrieben.	Der Indikator ist dann erfüllt, wenn deutlich wird, was der Anlass und Zweck (Zwischen- oder Schlussevaluierung) der Evaluierung sind und wer bzw. wie die Ergebnisse verwendet soll/en (Nutzer/ Nutzung). **(Textanalyse der PEV-Berichte)**							**X**	**X**
G 3.2 Die spezifischen Ziele der Evaluierung werden deutlich.	Der Indikator ist dann erfüllt, wenn transparent dargestellt ist, was mit der Evaluierung erreicht werden soll (bspw. Überprüfung von Relevanz, Zielerreichung, Nachhaltigkeit, Handlungsempfehlungen für Folgemodul, Übertragbarkeit auf andere Kontexte, etc.). Die spezifischen Evaluierungsziele müssen aus den TOR abgeleitet sein. **(Textanalyse der PEV-Berichte)**								
G 3.3 Die TOR einschließlich der Evaluierungsfragen der Prüfmission befinden sich im Anhang des Prüfberichts.	Der Indikator ist dann erfüllt, wenn die TOR mit den Evaluierungsfragen ein Anhang des Evaluierungsberichts sind. So kann beurteilt werden, inwieweit sich die Prüfmission ausreichend mit den Evaluierungsfragen auseinander gesetzt hat. **(Textanalyse der PEV-Berichte)**	**X**							
G 3.4 Anhand der TOR lassen sich Aussagen zum Zweck der Evaluierung ableiten.	Der Indikator ist dann erfüllt, wenn die TOR den Anlass und Zweck der Evaluierung klar definieren. **(Textanalyse der PEV-Berichte)**				**X**		**X**	**X**	**X**

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 4 Angabe von Informationsquellen	Die im Rahmen einer Evaluation genutzten Informationsquellen sollen hinreichend genau dokumentiert werden, damit die Verlässlichkeit und Angemessenheit der Informationen eingeschätzt werden kann.								
G 4.1 Der Bericht umfasst eine vollständige Liste der Gesprächspartner/innen und sonstigen Informationsquellen.	Der Indikator ist dann erfüllt, wenn die verwendeten Informationsquellen beschrieben werden, so dass die Aussagen im Bericht evidenzbasiert unterlegt sind. Falls Anonymität und das Recht auf Vertraulichkeit der einzelnen Gesprächspartner/innen die Darstellung der Informationsquellen einschränken, wird entsprechend darauf hingewiesen. **(Textanalyse der PEV-Berichte)**	**X**							
G 4.2 Der Bericht stellt transparent dar, welcher Systematik der Dokumentenauswertung zugrunde lag.	Der Indikator ist dann erfüllt, wenn die Auswertung der verwendeten Dokumente beschrieben werden. Hierdurch wird deutlich, welche Dokumente warum ausgewertet wurden und welches System der Dokumentenauswertung zugrunde lag. **(Textanalyse der PEV-Berichte)**								
Checklist:	Bezug Programmvorschlag								
	Bezug Operationsplan								
	Bezug Steuerungsstruktur								
	Bezug CD-Strategie								
	Bezug Akteursanalyse								
	Bezug Wirkungsmodell								
G 4.3 Die Auswahl von Gesprächspartner bzw. die Ziehung von Stichproben geschieht systematisch.	Der Indikator ist dann erfüllt, wenn die Auswahl von Gesprächspartner/-innen (bzw. Stichproben bei quantitativen Analysen) nach klar definierten und transparenten Kriterien geschieht (z.B. Reflektion der Erwartungen an die Aussagefähigkeit der TN, Zusammensetzung der TN, Rolle der Prüfmission bei der Auswahl der TN), und die Auswirkungen der Auswahl auf die Repräsentativität der Ergebnisse erläutert wird. Da im Rahmen einer Prüfmission nicht jeder befragt werden kann, der zur Grundgesamtheit gehört, muss eine begründete Auswahl aus der Grundgesamtheit vorgenommen werden.								
G 4.4 Aufbauend auf die Bewertung der Verlässlichkeit der Daten des WOM, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Daten des WOM erfolgt und die Daten des WOM dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	**X**							
G 4.5 Aufbauend auf die Bewertung der Verlässlichkeit der Baseline Daten, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Baseline Daten erfolgt und die Baseline Daten dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**	**X**				**X**			
G 4.6 Aufbauend auf die Bewertung der Verlässlichkeit der Partnerdaten, werden diese entweder verwendet oder ausgeschlossen.	Der Indikator ist dann erfüllt, wenn zunächst eine kritische Bewertung der Verlässlichkeit der Partnerdaten erfolgt und die Partnerdaten dann, ausgehend von der Bewertung, entweder verwendet oder, falls nicht oder nur eingeschränkt auf die Daten zurück gegriffen wird, begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**							**X**	**X**

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 5 Valide und reliable Informationen	Die Verfahren zur Gewinnung von Daten sollen so gewählt oder entwickelt und dann eingesetzt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien quantitativer und qualitativer Sozialforschung orientieren.								
G 5.1 Es wird eine Datentriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn Daten / Informationen zu einem gleichen Sachverhalt durch die Einbeziehung verschiedener Akteure erhoben werden. **(Textanalyse der PEV-Berichte)**			**X**		**X**	**X**	**X**	**X**
G 5.2 Es wird eine Methodentriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn Informationen zu einem gleichen Sachverhalt durch verschiedene Methoden der Datensammlung erhoben werden. **(Textanalyse der PEV-Berichte)**			**X**		**X**	**X**	**X**	**X**
G 6 Systematische Fehlerprüfung	Die in einer Evaluation gesammelten, aufbereiteten, analysierten und präsentierten Informationen sollen systematisch auf Fehler geprüft werden.								
G 6.1 Es wird eine Forschertriangulation durchgeführt.	Der Indikator ist dann erfüllt, wenn das PEV-Team seine Erfahrungen und Interpretationen / Perspektiven trianguliert. Hinweise auf die systematische Überprüfung der gesammelten Informationen sind z.B. Synthesetreffen, die dokumentiert wurden, oder die Darstellung auf Unstimmigkeiten der Mitglieder des PEV-Teams, die entsprechend diskutiert wurden. **(Textanalyse der PEV-Berichte)**			**X**					
G 6.2 Es wird eine Triangulation der Ergebnisse mit den Beteiligten und Betroffenen durchgeführt.	Der Indikator ist dann erfüllt, wenn eine Triangulation der Ergebnisse mit den Beteiligten und Betroffenen erfolgt. So soll eine Fehlerüberprüfung der Analyseergebnisse erfolgen. Hinweise auf die systematische Überprüfung der gesammelten Informationen sind z.B. gemeinsame Syntheseworkshops oder die Kommentierung des Berichts durch Beteiligte und Betroffene. **(Textanalyse der PEV-Berichte)**			**X**					

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 7 (A) Analyse qualitativer und quantitativer Informationen	Qualitative und quantitative Informationen einer Evaluation sollen nach fachlichen Maßstäben angemessen und systematisch analysiert werden, damit die Fragestellungen der Evaluation effektiv beantwortet werden können.								
G 7.1 (A) Die methodische Vorgehensweise beantwortet die Frage nach der Zuordnungs- und/ oder Beitragsanalyse.	Der Indikator ist dann erfüllt, wenn die methodische Vorgehensweise mit Hinweis auf die Beantwortung von Evaluierungsfragen, Bedingungen der Untersuchung und/ oder Besonderheiten des Evaluierungsgegenstandes begründet dargestellt wird. So kann bspw. eine Attributionsanalyse mit dem Hinweis auf den Evaluierungsgegenstand begründet ausgeschlossen werden. **(Textanalyse der PEV-Berichte)**						X		
Checklist:	keine begründete Vorgehensweise						X		
	experimentelles Design						X		
	quasi-experimentelles Design						X		
	Ansatz einer Kontributionsanalyse						X		
	theoriebasierter Ansatz						X		
G 7.2 (A) Die Evaluierungsmethoden werden dargestellt.	Der Indikator ist dann erfüllt, wenn die verwendeten Methoden beschrieben werden, so dass die Aussagen evidenzbasiert unterlegt sind. Dies ist dann gegeben, wenn mindestens eine der folgenden Anforderungen erfüllt ist. **(Textanalyse der PEV-Berichte)**	X		X		X	X		
Checklist:	Allgemeine Beschreibung der Methoden	X		X		X	X		
	Es wird deutlich, wie die Methoden im Prozess der Datenerhebung eingesetzt wurden - Analyseraster, Datenerhebungsplan o.ä. analyseleitendes System								
	Es wird deutlich, wie die erhobenen Daten dokumentiert wurden - Ergebnisberichte, Matrizen, Transkription von Interviews, Interviewprotokolle,								
	Es wird deutlich, wie die erhobenen qualitativen Daten ausgewertet wurden - Codierung, Auswertungsraster, Inhaltsanalyse o.ä.								
	Es wird deutlich, wie die erhobenen quantitativen Daten ausgewertet wurden - uni-, bi- oder multivariate Auswertungen								
G 7.3 (A) Es gibt eine Methodenvielfalt.	Der Indikator ist dann erfüllt, wenn mehr als eine Methode verwendet wird. **(Textanalyse der PEV-Berichte)**			X		X	X	X	X
Checklist:	Dokumentenauswertung						X		
	Fokusgruppen						X		
	Interviews						X		
	Befragung						X		
	eigene Auswertung der Monitoringdaten durch die Prüfmission								
G 7.4 (A) Die Vor- und Nachteile der gewählten Methoden werden dargestellt.	Der Indikator ist dann erfüllt, wenn die Methodenauswahl nachvollziehbar ist. Diese Nachvollziehbarkeit spiegelt sich in einer Klärung darüber, welche Vorteile die ausgesuchte Methode hat und welche möglichen Nachteile vorhanden sind und wie damit umgegangen werden soll, wider. **(Textanalyse der PEV-Berichte)**	X				X	X		

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 7 (B) Analyse qualitativer und quantitativer Informationen: Relevanz	Die methodische Bearbeitung der Relevanzbewertung ist angemessen.								
G 7.5 (B) Die Nachvollziehbarkeit der zugrunde gelegten Rahmenbedingungen und Kernprobleme der Maßnahme ist gegeben.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Relevanz die Rahmenbedingungen und Kernbedarfe, auf die die Maßnahme beruht, überprüft wird. **(Textanalyse der PEV-Berichte)**								
G 7.6 (B) Die Mehrdimensionalität der Relevanz wird analysiert.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Relevanz entlang der Kernbedarfe von Zielgruppe und Partner sowie der Ziele des Auftraggebers analysiert wird. **(Textanalyse der PEV-Berichte)**	X	X						
G 7.7 (B) Die strategische Ausrichtung der Maßnahme an veränderten Rahmenbedingungen werden analysiert.	Der Indikator ist dann erfüllt, wenn die strategische Ausrichtung einer Maßnahme an Veränderungen in den Rahmenbedingungen analysiert werden. **(Textanalyse der PEV-Berichte)**		X						
Z 7 (B) Die Bewertung des OECD/DAC Kriteriums "Relevanz" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**								X
G 7 (C) Analyse qualitativer und quantitativer Informationen: Effektivität	Die methodische Bearbeitung der Effektivitätsbewertung ist angemessen.								
G 7.8 (C) Die Kausalität zwischen Maßnahmen und Wirkungen wird differenziert analysiert und eingeschätzt.	Der Indikator ist dann erfüllt, wenn beobachtete Zusammenhänge zwischen Maßnahmen und Wirkungen nicht ohne Weiteres als ursächliche Zusammenhänge interpretiert werden. Der Einfluss von Drittvariablen muss ausgeschlossen werden. Korrelationen dürfen nicht mit Kausalitäten gleichgesetzt werden. **(Textanalyse der PEV-Berichte)**			X		X	X		
G 7.9 (C) Die Zielerreichung wird anhand von Modulindikatoren bewertet.	Der Indikator ist dann erfüllt, wenn deutlich wird, welche Indikatoren zur Bewertung der Zielerreichung als Grundlage genommen wurden. Wurden während der Projektplanung für einzelne Aspekte von Effektivität keine Indikatoren definiert, müssen diese Indikatoren nachträglich gebildet werden. **(Textanalyse der PEV-Berichte)**	X	X			X	X		
G 7.10 (C) Die verwendeten Indikatoren zur Messung und Beurteilung der Zielerreichung sind SMART (spezifisch, messbar, erreichbar, relevant, zeitgebunden).	Der Indikator ist dann erfüllt, wenn ersichtlich ist, dass sich die Prüfer mit der Qualität der Indikatoren auseinandergesetzt haben (zum Beispiel durch die Nutzung von SMART-Kriterien). Wird eine geringe Qualität festgestellt, müssen neue bzw. veränderte Indikatoren für die Evaluation genutzt werden. **(Textanalyse der PEV-Berichte)**	X		X		X	X		
Z 7 (C) Die Bewertung des OECD/DAC Kriteriums "Effektivität" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**								X

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 7 (D) Analyse qualitativer und quantitativer Informationen: Effizienz	Die methodische Bearbeitung der Effizienzbewertung ist angemessen.								
G 7.11 (D) Die verschiedenen Ebenen der Effizienz einer Maßnahme werden analysiert.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht dargestellt ist, welche Ebenen der Effizienzmessung analysiert werden und die Auswahl begründet wird. Mögliche Ebenen sind: die Implementierunseffizienz (Analyse der Strukturen und Prozesse der Vorhabensumsetzung), der Produktionseffizienz (Verhältnis Input zu Output) und der Allokationseffizienz (Verhältnis von Input zu Outcome). **(Textanalyse der PEV-Berichte)**	**X**	**X**						
Checklist:	Implementierungseffizienz	X	X						
	Produktionseffizienz,	X	X						
	Allokationseffizienz	X	X						
G 7.12 (D) Die Auswahl von Methoden und Verfahren der Effizienzmessung wird begründet.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht die Auswahl der Methoden und Verfahren zur Messung der Effizient begründet wird. **(Textanalyse der PEV-Berichte)**								
Checklist:	Level 2 methods								
	Level 1 methods								
	Descriptive methods								
G 7.13 (D) Die Bearbeitung der Effizienz ermöglicht die Identifikation von Potenzialen zur Effizienzsteigerung.	Der Indikator ist dann erfüllt, wenn die verwendeten Methoden und Verfahren implizit oder explizit Komponenten/Teile der Maßnahme mit Alternativen vergleichen, um Verbesserungspotenziale aufzuzeigen. **(Textanalyse der PEV-Berichte)**								
Z 7 (D) Die Bewertung des OECD/DAC Kriteriums "Effizienz" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**								**X**

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 7 (E) Analyse qualitativer und quantitativer Informationen: Impact	Die methodisch Bearbeitung der Impactbewertung ist angemessen.								
G 7.14 (E) In der Analyse und Beurteilung der übergeordneten Wirkungen (Impacts) wird die Attributionslücke analysiert.	Der Indikator ist dann erfüllt, wenn im Evaluierungsbericht bei der Analyse und Beurteilung der übergeordneten Wirkungen die Zuordnungs- bzw. Attributionslücke reflektiert wird. Es wird kein direkter und monokausaler Zusammenhang zwischen Maßnahme und übergeordneten Zielen hergestellt. **(Textanalyse der PEV-Berichte)**	**X**							
G 7.15 (E) Die Plausibilität der Hypothesen zu den intendierten langfristigen Wirkungen (Impacts) wird bewertet.	Der Indikator ist dann erfüllt, wenn sich die Evaluation mit der Plausibilität der Hypothesen zu übergeordneten langfristigen Wirkungen auseinandersetzt. **(Textanalyse der PEV-Berichte)**		**X**						
G 7.16 (E) Bewertungsmaßstäbe zur Analyse und Beurteilung des Beitrages der Maßnahme zu übergeordneten Wirkungen (Impacts) sind transparent dargestellt.	Der Indikator ist dann erfüllt, wenn beschreiben ist, anhand welcher Bewertungsmaßstäbe der Beitrag des Vorhabens zu übergeordneten entwicklungspolitischen Wirkungen bewertet wird. Diese Bewertungsmaßstäbe können entweder selbst vom Prüfteam entwickelt (zum Beispiel (Proxi-)Indikatoren) oder vorhandene Maßstäbe (zum Beispiel aus Teil A des PV) herangezogen werden. **(Textanalyse der PEV-Berichte)**	**X**	**X**	**X**		**X**	**X**		
Z 7 (E) Die Bewertung des OECD/DAC Kriteriums "Impact" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**								**X**
G 7 (F) Analyse qualitativer und quantitativer Informationen: Nachhaltigkeit	Die methodisch Bearbeitung der Nachhaltigkeitsbewertung ist angemessen.								
G 7.17 (F) Die Grenzen der Messung der Nachhaltigkeit werden beschrieben.	Der Indikator ist dann erfüllt, wenn bei der Bearbeitung der Nachhaltigkeit beschrieben wird, inwiefern zum Zeitpunkt der Evaluierung eine ergebnisorientierte Analyse und Beurteilung der Nachhaltigkeit von Effekten überhaupt möglich ist. **(Textanalyse der PEV-Berichte)**								
G 7.18 (F) Die angelegten Ansätze zur Schaffung von Nachhaltigkeit werden analysiert.	Der Indikator ist dann erfüllt, wenn in der Analyse und Beurteilung der Nachhaltigkeit die angelegten Ansätze zur Schaffung von Nachhaltigkeit berücksichtigt werden. Zum Beispiel wird evaluiert, inwiefern die Maßnahme eine Exit- oder Nachhaltigkeitsstrategie besitzt, Ansätze für die Anschlussfähigkeit der Maßnahme betreibt oder Aktivitäten zur Verantwortungsübergabe (Ownership) an den Partner betreibt. Die Grenzen der Messung werden dann bei der Analyse und Beurteilung eingehalten. **(Textanalyse der PEV-Berichte)**		**X**						
G 7.19 (F) Es werden mindestens zwei Ebenen der Nachhaltigkeit im Rahmen einer Prognose analysiert.	Der Indikator ist dann erfüllt, wenn der Mehrdimensionalität von Nachhaltigkeit im Rahmen einer Prognose Rechnung getragen wird und mindestens Nachhaltigkeit auf zwei Ebenen betrachtet wird (finanzielle, institutionelle, personelle, technologische, soziale, ökologische Nachhaltigkeit) überprüft. **(Textanalyse der PEV-Berichte)**								
Z 7 (F) Die Bewertung des OECD/DAC Kriteriums "Nachhaltigkeit" ist nachvollziehbar	Die Bewertung des Kriteriums wird anhand der Analyseergebnisse begründet. Die Punktevergabe und die Bewertung passen zueinander. **(Textanalyse der PEV-Berichte)**								**X**

Bewertungskriterien / Indikatoren / Deskriptoren	Definitionen	Quellen zur Ableitung des Indikators							
		Annotierte Gliederung	Checklist OECD	Method. Vorgehen PEV	Muster TOR	Checklist QS	Methoden Toolbox	M&E Policy	Abstimmung GIZ Stabsstelle
G 8 Begründete Analyse und begründete Schlussfolgerung	Die in einer Evaluation gezogenen Folgerungen sollen ausdrücklich begründet werden, damit die Adressaten und Adressatinnen diese einschätzen können.								
G 8.1 Es wird zwischen Beschreibung, Analyse und Bewertung unterschieden.	Der Indikator ist dann erfüllt, wenn der Dreischritt Beschreibung – Analyse – Bewertung aus evaluatorischer Einschätzung zumeist eingehalten wird. Das heißt, es wird wird nicht direkt von der Beschreibung zur Bewertung gesprungen bzw. die Analyse und Bewertung werden nicht vermischt. Der Dreischritt soll mindestens bei zwei der folgenden drei Aspekte eingehalten werden: Wirkungsmodell, Bewertung der OECD/DAC-Kriterien, Schlussfolgerungen. **(Textanalyse der PEV-Berichte)**					X	X		
Checklist:	Wirkungsmodell	X				X			
	Bewertung der OECD/ DAC-Kriterien					X			
	Schlussfolgerungen					X			
G 8.2 Beschreibungen und Analysen werden belegt.	Der Indikator ist dann erfüllt, wenn Beschreibung und Analyse im Bericht mit relevanten Zahlen, Daten und Fakten untermauert werden. Für alle Zahlen, Daten und Fakten, werden Quellen angegeben. Demnach sind bei der Beschreibung und Analyse von Ergebnissen der Datenerhebung die jeweiligen Informationsquellen ersichtlich. Hierbei werden die Daten, die von der Evaluierungsmission erhoben wurden, als solche gekennzeichnet. **(Textanalyse der PEV-Berichte)**	X				X	X		
G 8.3 Empfehlungen werden aus der Analyse abgeleitet und sind spezifisch, realistisch und termingebunden. Sie richten sich an die wesentlichen Nutzer und ihre Umsetzung ist messbar.	Der Indikator ist dann erfüllt, wenn für alle Handlungsempfehlungen ein Ziel bzw. ein Zweck definiert worden ist. Die Empfehlungen müssen realistisch sein und sich aus den Schlussfolgerungen ableiten. Die Realisierbarkeit aller Handlungsempfehlungen muss sichergestellt sein. Z.B indem Handlungsempfehlungen in Zusammenarbeit mit denjenigen entwickelt werden, die sie umsetzen müssen. Alle Handlungsempfehlungen müssen einen Zeitraum bzw. einen Zeitpunkt für die Umsetzung benennen und die Umsetzung muss empirisch erhoben werden können. **(Textanalyse der PEV-Berichte)**					X	X		
G 8.4 Die positiven und/ oder negativen nicht-intendierten Wirkungen der Maßnahme werden beschrieben.	Der Indikator ist dann erfüllt, wenn sich die Evaluierung mit möglichen positiven und/ oder negativen nicht-intendierten Wirkungen befasst. Hiermit soll gewährleistet werden, dass sich die Evaluierung möglicher Hinweise auf nicht-intendierte Wirkung bewusst ist. **(Textanalyse der PEV-Berichte)**	X	X						
Z 8 In den Schlussfolgerungen werden die untersuchten Zusammenhänge und Beiträge des Vorhabens differenziert bewertet.	Eine Kontributionsanalyse überprüft die Zusammenhänge aus der Theory-of-Change und ermöglicht dadurch Schlussfolgerungen darüber, welche Zusammenhänge mit Evidenz untermauert sind, welche weiteren Zusammenhänge identifiziert worden sind und welche Zusammenhänge revidiert, bzw. nicht belegt werden konnten. Der Indikator ist dann erfüllt, wenn in den Schlussfolgerungen die analysierten Zusammenhänge differenziert bewertet werden (zum Beispiel nach dem Grad der logischen Plausibilität, der vorgefundenen Evidenz, der vorgefundenen externen Erkärungsfaktoren, des wissenschaftlichen Kenntnisstandes oder der Aussagen von befragten Akteuren) **(Textanalyse der PEV-Berichte)**								X

Fotonachweise und Quellen

Fotonachweise/Quellen:
© GIZ / Thomas L. Kelly, Markus Kirchgessner, Florian Kopp, Dirk Ostermeier

URL-Verweise:
In dieser Publikation befinden sich ggf. Verweise zu externen Internetseiten. Für die Inhalte der aufgeführten externen Seiten ist stets der jeweilige Anbieter verantwortlich. Die GIZ hat beim erstmaligen Verweis den fremden Inhalt daraufhin überprüft, ob durch ihn eine mögliche zivilrechtliche oder strafrechtliche Verantwortlichkeit ausgelöst wird. Eine permanente inhaltliche Kontrolle der Verweise auf externe Seiten ist jedoch ohne konkrete Anhaltspunkte einer Rechtsverletzung nicht zumutbar. Wenn die GIZ feststellt oder von anderen darauf hingewiesen wird, dass ein externes Angebot, auf das sie verwiesen hat, eine zivil- oder strafrechtliche Verantwortlichkeit auslöst, wird sie den Verweis auf dieses Angebot unverzüglich aufheben. Die GIZ distanziert sich ausdrücklich von derartigen Inhalten.

Kartenmaterial:
Kartografischen Darstellungen dienen nur dem informativen Zweck und beinhalten keine völkerrechtliche Anerkennung von Grenzen und Gebieten. Die GIZ übernimmt keinerlei Gewähr für die Aktualität, Korrektheit oder Vollständigkeit des bereitgestellten Kartenmaterials. Jegliche Haftung für Schäden, die direkt oder indirekt aus der Benutzung entstehen, wird ausgeschlossen.

Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Sitz der Gesellschaft
Bonn und Eschborn

Friedrich-Ebert-Allee 36 + 40
53113 Bonn, Deutschland
T +49 228 44 60-0
F +49 228 44 60-17 66

Dag-Hammarskjöld-Weg 1-5
65760 Eschborn, Deutschland
T +49 61 96 79-0
F +49 61 96 79-11 15

E info@giz.de
I www.giz.de