Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen Entwicklungszusammenarbeit

Guffler, Kerstin; Kunert, Laura; Wittenberg, Marian; Herforth, Nico

Veröffentlichungsversion / Published Version
Monographie / monograph

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

2022

DEval
DEUTSCHES EVALUIERUNGSINSTITUT DER ENTWICKLUNGS-ZUSAMMENARBEIT

In der organisationsübergreifenden Meta-Evaluierung werden in Deutschland (mit-)verantwortete und BMZ-(mit-)geförderte Projektevaluierungen von elf staatlichen und nichtstaatlichen Organisationen hinsichtlich ihres Qualitätsverständnisses in Evaluierungen und ihrer Anwendung von international geltenden Qualitätsstandards, insbesondere der OECD-DAC- und der DeGEval-Standards, untersucht. Weiterhin werden Faktoren analysiert, die mit der Anwendung der Qualitätsstandards zusammenhängen. Insgesamt zeigt sich, dass die Anwendung der untersuchten Qualitätsstandards bei den beteiligten Organisationen weitgehend in ihrer Evaluierungspraxis verankert ist. Darüber hinaus werden sie aber noch nicht durchgehend beziehungsweise systematisch in den Organisationsdokumenten und -prozessen sowie auf Ebene der einzelnen Evaluierung verschriftlicht oder nachvollziehbar dokumentiert. Faktoren, die die Anwendung mehrerer Qualitätsstandards bedingen, konnten nicht identifiziert werden. Neben den Empfehlungen an die beteiligten Organisationen, eine systematische Verankerung der Qualitätsstandards sicherzustellen und zukünftig gemeinsames Lernen über einen systematisierten Erfahrungsaustausch zu gewährleisten, wird dem BMZ empfohlen, auf Basis der in Kraft getretenen BMZ-Leitlinien Evaluierung ein Analyseraster für die Anwendung der Qualitätsstandards zu entwickeln und für die Organisationen bereitzustellen.

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

2022

IMPRESSUM

Verfasst von
Dr. Kerstin Guffler
Laura Kunert
Marian Wittenberg
Dr. Nico Herforth

Verantwortlich
Amélie Gräfin zu Eulenburg

Gestaltung Umschlag und Grafiken
Katharina Mayer, DEval

Lektorat
Marcus Klein, PhD

Bildnachweis
Titelseite: VektorMine, Shutterstock

Bibliografische Angabe
Guffler, K., L. Kunert, M. Wittenberg und N. Herforth (2022), *Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen Entwicklungszusammenarbeit*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval), Bonn.

Druck
Bonifatius, Paderborn

ISBN 978-3-96126-171-0 (gebundene Ausgabe)
ISBN 978-3-96126-172-7 (PDF)

Herausgeber

Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval)
Fritz-Schäffer-Straße 26
53113 Bonn, Germany

Tel: +49 (0)228 33 69 07-0
E-Mail: *info@DEval.org*
www.DEval.org

Das Deutsche Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) ist vom Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) mandatiert, Maßnahmen der deutschen Entwicklungszusammenarbeit unabhängig und nachvollziehbar zu analysieren und zu bewerten.

Mit seinen Evaluierungen trägt das Institut dazu bei, die Entscheidungsgrundlage für eine wirksame Gestaltung des Politikfeldes zu verbessern und die Transparenz zu den Ergebnissen zu erhöhen.

Der vorliegende Bericht ist auch auf der DEval-Website als PDF-Download verfügbar unter:
https://www.deval.org/de/publikationen

Anfragen nach einer gebundenen Ausgabe richten Sie bitte an: *info@DEval.org*

Eine Stellungnahme des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) findet sich unter:
https://www.bmz.de/de/ministe-rium/evaluierung/bmz-stellungnahmen-19404

DANKSAGUNG

Das Evaluierungsteam hat im Verlauf der Arbeit an der Meta-Evaluierung zur Qualität von (Projekt-) Evaluierungen in der deutschen EZ von verschiedenen Stakeholder*innen Unterstützung erhalten, für die wir uns an dieser Stelle bedanken möchten.

Ein ausdrücklicher Dank gilt den *Mitgliedern der Referenzgruppe*, die aus dem BMZ sowie Vertreter*innen der Evaluierungseinheiten/-stellen der beteiligten Organisationen – konkret der Bundesanstalt für Geowissenschaften und Rohstoffe, CARE Deutschland e. V., der Deutschen Gesellschaft für Internationale Zusammenarbeit, des Deutschen Roten Kreuzes, des Deutschen Volkshochschul-Verbands International, des Evangelischen Werks für Diakonie und Entwicklung e. V., der Heinrich-Böll-Stiftung, der KfW Entwicklungsbank, der Konrad-Adenauer-Stiftung, MISEREOR und der Physikalisch-Technischen Bundesanstalt – sowie Vertreter*innen von VENRO bestand. Wir danken dafür, dass sie uns im Rahmen ihrer Rolle als „Sounding Board" bei der korrekten Darstellung des Evaluierungsgegenstands unterstützt, uns Daten und Dokumente bereitgestellt und unsere Arbeit schriftlich und mündlich kommentiert haben. Daneben bedanken wir uns auch für den positiven, konstruktiven und wertschätzenden Austausch, den wir über die Zeit hinweg erfahren durften.

Wir bedanken uns des Weiteren bei Prof. Dr. Wolfang Beywl und Prof. Dr. Thomas Widmer, die uns im Rahmen ihrer Tätigkeiten als *externer Peer-Reviewer* und *Gutachter für die Qualität von Evaluierungen* über den gesamten Evaluierungsprozess hinweg begleitet und unterstützt haben. Dazu gehörten schriftliche Kommentierungen als auch konstruktive Rückmeldungen in Workshops zu verschiedenen Zwischenprodukten. DEval-intern bedanken wir uns bei Dr. Martin Noltze, der bereits die vorangegangene organisationsübergreifende Meta-Evaluierung durchführte und als *interner Peer-Reviewer* nicht nur einen wichtigen Beitrag zur Qualitätssicherung geleistet hat, sondern auch sein wertvolles Fachwissen und frühere Erfahrungen mit uns teilte.

Weiterhin bedanken wir uns für die punktuelle Unterstützung ausgewählter *Expert*innen* für Meta-Evaluierungen, OECD-DAC- beziehungsweise DeGEval-Standards, Stichprobenziehungen und die Nutzung der Software MAXQDA, durch die wir wichtige Hinweise und Erkenntnisse für die Ausgestaltung der vorliegenden Arbeit erhalten konnten.

Nicht zuletzt gilt unser Dank unseren Kolleg*innen Dr. Thomas Wencker, Jens Eger und des gesamten Teams der *Community of Practice* am DEval/KZM.

Besten Dank!

ZUSAMMENFASSUNG

Einleitung

Meta-Evaluierungen, die als Evaluierung von Evaluierungen bezeichnet werden können, gewinnen in der EZ zunehmend an Bedeutung. Laut Caracelli und Cooksy (2009, S. 2 f., eigene Übersetzung) werden Meta-Evaluierungen „durchgeführt, um den Evaluierungsprozess zu verbessern, die Stärken und Schwächen einer Evaluierung systematisch zu reflektieren, die zukünftige Evaluierungsarbeit zu verbessern oder Informationen zur Glaubwürdigkeit der Ergebnisse für die Nutzenden zur Verfügung zu stellen“. Diesem Verständnis folgt auch die vorliegende Meta-Evaluierung. In der internationalen Entwicklungszusammenarbeit (EZ) besteht inzwischen eine größere Anzahl an Meta-Evaluierungen, aber auch in der deutschen EZ werden organisationsinterne und -übergreifende Meta-Evaluierungen durchgeführt.

Die vorliegende Meta-Evaluierung, die zeitweise parallel zur Erstellung der Leitlinien Evaluierung des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) lief, schließt inhaltlich unter anderem an frühere Systemprüfungen des deutschen EZ-Systems, das Umsetzungsmonitoring einer Systemprüfung und die Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) an. Im Jahr 1999 wurde im Auftrag des BMZ eine erste systematische Untersuchung der Evaluierungspraxis in der deutschen EZ vorgenommen (Borrmann et al., 1999) und 2009 eine zweite Systemprüfung abgeschlossen (Borrmann und Stockmann, 2009). Auf Letzterer aufbauend führte das DEval im Jahr 2015 ein Umsetzungsmonitoring zu den Ergebnissen und Empfehlungen durch (Lücking et al., 2015). Im Jahr 2018 wurde eine organisationsübergreifende Meta-Evaluierung[1] zur Analyse der Qualität von Projektevaluierungen der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) und der KfW Entwicklungsbank (KfW) vom DEval veröffentlicht (Noltze et al., 2018).

Vor Abschluss der Meta-Evaluierung wurden als bisher letzter Meilenstein die BMZ-Leitlinien zur Evaluierung in der EZ (BMZ, 2021) veröffentlicht, unter anderem um das Qualitätsverständnis in der deutschen EZ weiter zu festigen. Darin sind insbesondere die Standards des Entwicklungsausschusses (Development Assistance Committee, DAC) der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (Organisation for Economic Co-operation and Development, OECD), aber auch der Gesellschaft für Evaluation (DeGEval) als verbindlich für die Durchführungsorganisationen beziehungsweise als Orientierung für die nichtstaatlichen Organisationen verschriftlicht worden. Da der Untersuchungszeitraum der Meta-Evaluierung vor Verabschiedung der Leitlinien lag, konnten diese nicht in die Analyse einbezogen werden. Gleichwohl wurden wichtige Erkenntnisse für die zukünftige Ausgestaltung der Leitlinien generiert.

Eine Besonderheit der vorliegenden Meta-Evaluierung liegt darin, dass die Anwendung der Qualitätsstandards in Evaluierungen bei einer Vielzahl staatlicher und nichtstaatlicher Organisationen untersucht wurde. Die nachfolgenden Organisationen waren an der Meta-Evaluierung beteiligt: Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), CARE Deutschland e. V. (CARE), Deutscher Volkshochschul-Verband International (DVV), Deutsches Rotes Kreuz (DRK), Evangelisches Werk für Diakonie und Entwicklung e. V. (EWDE), GIZ, Heinrich-Böll-Stiftung (hbs), Konrad-Adenauer-Stiftung (KAS), KfW, MISEREOR und Physikalisch-Technische Bundesanstalt (PTB). Untersucht wurden in Deutschland (mit-)verantwortete und zwischen Oktober 2016 und Dezember 2020 umgesetzte Projektevaluierungen. Dabei wurden staatliche Organisationen vollumfänglich und nichtstaatliche entlang von Kriterien bezüglich ihrer strukturellen Heterogenität ausgewählt, sodass die Ergebnisse ein möglichst breites Spektrum an Erfahrungen mit der Anwendung von Qualitätsstandards abbilden konnten.

[1] In den BMZ-Leitlinien zur Evaluierung in der EZ werden die DEval-Meta-Evaluierungen als ein Teil der Qualitätssicherung des Evaluierungssystems benannt (BMZ, 2021).

Die Anwendung der Qualitätskriterien wurde entlang der Verpflichtungsgrundlage der Organisationen zu den Standarddokumenten der OECD-DAC- und/oder der DeGEval- und organisationsspezifischer Qualitätskriterien untersucht sowie für GIZ und KfW entlang der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit.

Mit der organisationsübergreifenden Meta-Evaluierung sollen Erkenntnisse zum Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen sowie zu den Stärken und Schwächen in der Anwendung der Qualitätsstandards geliefert werden. Darüber hinaus wurden Faktoren identifiziert und untersucht, die mit der Anwendung der Qualitätsstandards zusammenhängen. Im Sinne des zukünftigen Lernens wurde in dieser Meta-Evaluierung auch das Ziel verfolgt, Erklärungen für die Nichtanwendung von Qualitätsstandards aufzuzeigen. Die nachfolgenden Evaluierungsfragen wurden dabei untersucht:

Evaluierungsfragen

1. Wie ist das Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen in der deutschen EZ?
2. Inwieweit werden Qualitätsstandards bei Evaluierungen der beteiligten Organisationen in der deutschen EZ angewandt?

a) Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

b) Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der organisationsspezifischen Qualitätsstandards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

c) Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW?

3. Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen?

Theoretische und empirische Herleitung

Qualitätsverständnis

In der vorliegenden Meta-Evaluierung wird die Evaluierungsqualität mit der Anwendung der einschlägigen beziehungsweise der für die Organisationen verpflichtenden Qualitätsstandards gleichgesetzt und dementsprechend untersucht. Der Qualitätsbegriff schließt vor allem – aber nicht ausschließlich – die Inhalte der OECD-DAC- und der DeGEval-Standarddokumente mit ein. Die Grundlage für die Identifizierung von Belegen guter Evaluierungen bilden die Qualitätsstandards des OECD DAC und der DeGEval aufgrund ihres international geltenden Charakters, ihres Bezugs zur EZ und ihrer Relevanz für deutsche EZ-Organisationen. Der Begriff „Anwendung der Qualitätsstandards“ wurde im Konsens mit der Referenzgruppe[2] gewählt und beschreibt, inwieweit der Nachweis erbracht werden kann (das heißt schriftlich dokumentiert war oder schriftlich auf Nachfrage zurückgemeldet wurde), ob und wie die Qualitätskriterien in den untersuchten Evaluierungen berücksichtigt wurden. Bei den OECD-DAC- und den DeGEval-Standards handelt es sich um Maximalstandards. Das bedeutet, dass die beteiligten Organisationen nicht alle Qualitätsstandards in allen Evaluierungen anwenden müssen. Das Qualitätsverständnis wurden im Weiteren um organisationsspezifische Qualitätsstandards und die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ergänzt.

2 Die Referenzgruppe bestand aus Vertreter*innen der beteiligten Organisationen und VENRO sowie aus Referent*innen des BMZ-Referats GS 22 „Evaluierung und Ressortforschung, DEval, IDOS“. Ihre Mitglieder begleiteten den Prozess der Evaluierung in allen Evaluierungsphasen (zum Beispiel über virtuelle Treffen oder Kommentierungen von Evaluierungsdokumenten; DEval, 2021a).

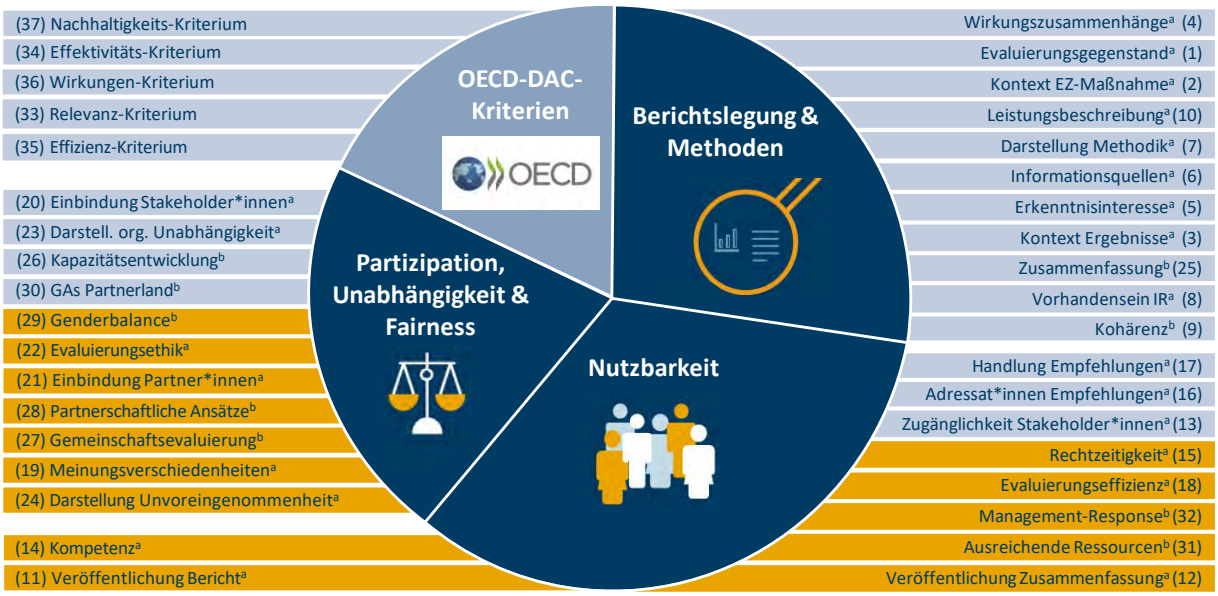
Analyseraster

Das Analyseraster beinhaltet Qualitätskriterien, die aus den OECD-DAC- und den DeGEval-Standarddokumenten abgeleitet wurden, organisationsspezifische Qualitätskriterien und Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018). Die OECD-DAC- und die DeGEval-Qualitätskriterien lassen sich drei Bereichen zuordnen: 1) der Überschneidung zwischen den OECD-DAC- und den DeGEval-Standarddokumenten, 2) dem OECD-DAC-Standarddokument ohne Überschneidung mit dem DeGEval-Standarddokument (OECD DAC only) und 3) den OECD-DAC-Kriterien. Da alle Organisationen im Untersuchungszeitraum den „OECD-DAC-Kriterien" (BMZ, 2006) verpflichtet waren, wird dies als separater Bereich aufgeführt, obwohl ihre Anwendung Teil der OECD-DAC-Standards ist (Qualitätsstandard 2.8; OECD DAC, 2010). Für DRK, EWDE, GIZ und hbs umfasste das Analyseraster darüber hinaus weitere organisationsspezifische Qualitätskriterien[3] und für GIZ und KfW Qualitätskriterien, die bereits in der vorangegangenen Meta-Evaluierung Nachhaltigkeit erhoben worden waren.

Standardcluster

Mit Ausnahme der fünf OECD-DAC-Kriterien wurden die Qualitätskriterien drei inhaltlichen Standardclustern – „Berichtslegung und Methoden", „Partizipation, Unabhängigkeit und Fairness" und „Nutzbarkeit" – zugeordnet.[4] Das Standardcluster „Berichtslegung und Methoden" umfasst vor allem Qualitätskriterien, die sich auf die Darstellung von Informationen zur Methodik der Evaluierung beziehen oder auf das Vorliegen beziehungsweise den Informationsgehalt ausgewählter Evaluierungsdokumente (Abbildung 1). Dem Standardcluster „Partizipation, Unabhängigkeit und Fairness" sind vor allem Qualitätskriterien zugeordnet, die sich mit der Berücksichtigung unterschiedlicher Personengruppen in der Evaluierung auseinandersetzen. Das Standardcluster „Nutzbarkeit" fokussiert insbesondere auf die Nützlichkeit der Evaluierungen, während die aktive Nutzung nur eine kleine Rolle spielt und der Nutzen nicht untersucht wird.[5]

Abbildung 1 Zuordnung der 37 Qualitätskriterien zu den Standardclustern und OECD-DAC-Kriterien



Quelle: DEval, eigene Darstellung

Anmerkung: blauer Balken = Qualitätskriterium wurden je Evaluierung untersucht; gelber Balken = Qualitätskriterium wurde auf Organisationsebene über alle Evaluierungen hinweg untersucht; GAs = Gutachtende; Darstell. = Darstellung; org. = organisationale [a] Qualitätskriterium kommt aus der Überschneidung von OECD-DAC- und DeGEval-Standards; [b] Qualitätskriterium kommt aus dem Bereich „OECD DAC only".

3 Organisationsspezifische Qualitätskriterien wurden definiert als Anforderungen, die unabhängig von den OECD-DAC- oder den DeGEval-Standards eine große Bedeutung für eine Organisation hinsichtlich der Qualität ihrer Evaluierungen haben.

4 Die Benennungen der drei gebildeten Standardcluster zeigen Ähnlichkeiten zu den Benennungen der DeGEval-Standardgruppen (1. Nützlichkeit, 2. Durchführbarkeit, 3. Fairness und 4. Genauigkeit). Da die identifizierten Qualitätskriterien allerdings zum Teil die Überschneidung zwischen den OECD-DAC- und den DeGEval-Standards darstellen, war eine identische Benennung inhaltlich nicht zutreffend. Die OECD-DAC-Standards sind überwiegend entlang von Evaluierungsphasen strukturiert, sodass diese Benennungen nicht berücksichtigt wurden.

5 Weitere Details zu den Begriffsdefinitionen finden sich im Haupttext des Evaluierungsberichts in Abschnitt 2.1.

Faktoren für die Anwendung der Qualitätsstandards

Die nachfolgend beschriebenen Faktoren wurden in den Analysen untersucht, wenn sie drei Merkmale erfüllten, nämlich 1) eine eindeutige organisationsübergreifende Definition aufwiesen, 2) klare Wirkungszusammenhänge mit ausgewählten Qualitätsstandards beschrieben werden konnten und 3) Daten bei den Organisationen oder in Sekundärdatenbanken verfügbar waren. Um die Faktoren zu identifizieren, wurden drei Fokusgruppendiskussionen mit den Verantwortlichen der beteiligten Organisationen durchgeführt sowie wissenschaftliche und empirische Literatur gesichtet. Die Faktoren wurden anschließend entlang der 1) länderkontext-, 2) evaluierungs- und 3) organisationsspezifischen Einflussdimensionen systematisiert.

Methodisches Vorgehen

Datengrundlage und -analyse

Nachdem die Organisationen in die Meta-Evaluierung aufgenommen waren, wurde eine geschichtete Stichprobe von insgesamt 296 Evaluierungen für die Untersuchung der Anwendung der Qualitätsstandards gezogen. Die Organisationen wurden anhand von vier Kriterien ausgewählt, um eine möglichst große strukturelle Heterogenität abzudecken und je Organisation ausreichend Evaluierungen untersuchen zu können. Insgesamt haben die Evaluierungseinheiten/-stellen der Organisationen im Untersuchungszeitraum von Oktober 2016 bis Dezember 2020 839 Evaluierungen in Deutschland (mit-)verantwortet. In die Grundgesamtheit wurden daraufhin die 576 Evaluierungen aufgenommen, die das BMZ entweder (mit-)gefördert hatte oder in denen eine vom BMZ (mit-)geförderte EZ-Maßnahme untersucht wurde. Die daraus gezogene Stichprobe umfasste 296 Evaluierungen.

Qualitätsverständnis (Evaluierungsfrage 1): Für die Erarbeitung des Qualitätsverständnisses beziehungsweise der Verpflichtungsgrundlage der beteiligten Organisationen wurden Organisationsdokumente sowie BMZ-Vorgaben mithilfe einer qualitativen Inhaltsanalyse ausgewertet.

OECD-DAC- und/oder DeGEval-Standards (Evaluierungsfrage 2a): Um die Anwendung der Qualitätskriterien in den 296 Evaluierungen zu untersuchen, wurden Evaluierungsdokumente (Evaluierungsberichte und -anhänge, Leistungsbeschreibungen, Inception Reports) und Organisationsdokumente (zum Beispiel Evaluierungskonzepte, Leitfäden und Handreichungen zur Durchführung von Evaluierungen) herangezogen. Für 14 der 37 OECD-DAC- und DeGEval-Qualitätskriterien konnten in der Inter-Kodierenden-Phase keine oder nur sehr wenig Informationen in den von den Organisationen bereitgestellten Evaluierungsdokumenten kodiert werden. Um keine fehlerhaften Rückschlüsse auf eine Nichtanwendung zu ziehen, wurden die Verantwortlichen der Evaluierungseinheiten/-stellen in einem weiteren Schritt zur Anwendung dieser Qualitätskriterien in ihrer Organisation befragt.[6] Die im Analyseraster festgelegten Qualitätskriterien wurden entlang von ordinalen oder binären Bewertungsstufen kodiert. Den von den Verantwortlichen der Evaluierungseinheiten/-stellen in der Onlinebefragung zurückgemeldeten durchschnittlichen Häufigkeiten der Anwendung der Qualitätskriterien wurden ebenfalls Werte zugeordnet. Für jedes Qualitätskriterium wurden je Organisation verschiedene Kennzahlen (zum Beispiel Mittelwerte) berechnet. Die Kennzahlen wurden nachfolgend in Prozentwerte umgewandelt und den vorab festgelegten Schwellenwerten des Anspruchsniveaus zugeordnet. Anschließend wurde ein Mittelwert über alle Organisationen hinweg ermittelt und mit Verfahren der deskriptiven Statistik ausgewertet.

Organisationsspezifische Qualitätsstandards (Evaluierungsfrage 2b): Für die Untersuchung der Anwendung der elf organisationsspezifischen Qualitätskriterien wurden diese in den Evaluierungsdokumenten der vier betroffenen Organisationen kodiert und anschließend wie die OECD/DAC- und die DeGEval-Qualitätskriterien berechnet und ausgewertet.

[6] Da die Qualitätskriterien der Onlinebefragung durchschnittlich rund 6 Prozent weniger angewandt wurden als die Qualitätskriterien der Dokumentenanalyse, gab es keinen Anlass anzunehmen, dass sich die Organisationen systematisch besser bewerteten, als sie durch die objektive Kodierung bewertet worden wären.

Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit (Evaluierungsfrage 2c): Für die Analyse der Anwendung beziehungsweise der erneuten Anwendung der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit wurden 15 Qualitätskriterien aus der vorangegangenen Meta-Evaluierung von Noltze et al. (2018) herangezogen. Acht Qualitätskriterien waren bereits als Qualitätskriterien im OECD-DAC- und im DeGEval-Analyseraster berücksichtigt und wurden lediglich transformiert, die übrigen sieben wurden in den aktuellen Evaluierungen erneut kodiert. Für die Auswertung wurden deskriptive Statistiken und für die Differenz der Ergebnisse zwischen der Meta-Evaluierung Nachhaltigkeit und der vorliegenden Meta-Evaluierung Strukturgleichungsmodelle berechnet (Weiber und Mühlhaus, 2010).

Um die Zusammenhänge zwischen ausgewählten Faktoren und der Anwendung der Qualitätskriterien zu untersuchen, wurden multivariate Regressionsanalysen geschätzt (Evaluierungsfrage 3). Regressionsanalysen erlauben die Identifikation von statistischen Zusammenhängen zwischen den Faktoren (unabhängige Variablen) und den Qualitätskriterien sowie dem Standardcluster „Berichtslegung und Methoden“ (abhängige Variablen; Backhaus et al., 2011). Konkret wurden Faktoren in 1) der Länderkontext-, 2) der Evaluierungs- und 3) der Organisationsdimension untersucht. Die Informationen für die einzelnen Faktoren wurden mithilfe von Daten gewonnen, die von den Organisationen übermittelt worden waren, sowie über Sekundärdatenbanken.

Bewertung der Anwendung der Qualitätsstandards

Die Anwendung der Qualitätskriterien wurde bei Organisationen mit Verpflichtungsgrundlage zur Anwendung der Qualitätskriterien sowohl untersucht als auch bewertet (Gruppe 1), bei Organisationen ohne Verpflichtungsgrundlage ausschließlich untersucht (Gruppe 2). Die Schwellenwerte des Anspruchsniveaus für die Anwendung der Qualitätskriterien wurden im Austausch mit der Referenzgruppe festgelegt und dienten als Basis für die Bewertung. Das Anspruchsniveau stellte die ex ante festgelegte Einschätzung dar, ab wann ein Qualitätskriterium in einer Evaluierung als kaum, teilweise, größtenteils und vollständig angewandt gilt. Bei der Festlegung der Schwellenwerte in 25-Prozent-Schritten (0 ≤ 25 Prozent = „kaum angewandt“, > 25 ≤ 50 Prozent = „teilweise angewandt“, > 50 ≤ 75 Prozent = „größtenteils angewandt“, > 75 ≤ 100 Prozent = „vollständig angewandt“) wurde berücksichtigt, dass die Qualitätsstandards als Maximalstandards zu verstehen sind. Die Bewertung wurde entlang der Schwellenwerte des Anspruchsniveaus mit Hinzunahme der Extremwerte 0 (verfehlt) und 100 (übertroffen) für Gruppe 1 vorgenommen. Da eine erneute Untersuchung der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit stattfand, wurde an dieser Stelle der Anspruch gestellt, dass sich die Anwendung der Qualitätskriterien seit der Meta-Evaluierung Nachhaltigkeit verbessert hatte. Bei den Qualitätskriterien der OECD-DAC- und der DeGEval-Standards wurden beide Gruppen hinsichtlich ihres Grads der Anwendung analysiert, Gruppe 1 wurde zusätzlich bewertet. Bei den OECD-DAC-Kriterien (BMZ, 2006), den organisationsspezifischen Qualitätskriterien und den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit wurden ausschließlich Organisationen der Gruppe 1 untersucht und bewertet.

Stärken und Herausforderungen des methodischen Vorgehens

Aufgrund der Auswahl der Organisationen entlang ihrer strukturellen Heterogenität wurde eine große Bandbreite in der Anwendung einzelner Qualitätskriterien untersucht und dargestellt. Die Erkenntnisse der Meta-Evaluierung gelten für die beteiligten Organisationen. Die staatlichen Organisationen wurden vollständig abgebildet. Die Auswahl der beteiligten nichtstaatlichen Organisationen ist nicht repräsentativ für alle Organisationen der nichtstaatlichen Entwicklungszusammenarbeit. Entsprechend der Auswahl der Organisationen entlang ihrer strukturellen Heterogenität, können sich nicht beteiligte nichtstaatliche Organisationen innerhalb dieser Bandbreite verorten und Erkenntnisse der Meta-Evaluierung für sich nutzen. Eine Übertragbarkeit der Erkenntnisse auf die Grundgesamtheit der Evaluierungen je Organisation war im Rahmen der gewählten statistischen Kennwerte für die Stichprobenziehung gewährleistet, eine Übertragbarkeit auf andere Evaluierungstypen einer Organisation nicht.

Aufgrund der systematischen Ableitung des Analyserasters entlang der OECD-DAC- und der DeGEval-Standards kann es auch von anderen Organisationen genutzt werden. Das Analyseraster der Meta-Evaluierung kann somit zukünftig als Grundlage für die Erstellung eines Analyserasters auf der Basis der BMZ-Leitlinien Evaluierung herangezogen werden.

Bei der Analyse der Qualitätskriterien aus der Onlinebefragung bestanden Einschränkungen hinsichtlich der Methoden-Triangulation. Die Hinzunahme einer weiteren Datenerhebungsmethode hätte allerdings angesichts des Aufwand-Nutzen-Verhältnisses in keiner angemessenen Relation gestanden. Neben der Befragung der Verantwortlichen der Evaluierungseinheiten/-stellen hätte zur Triangulation der Daten zum Beispiel auch die Bewertung der Gutachtenden der jeweiligen Evaluierungen herangezogen werden können. Da in der vorliegenden Meta-Evaluierung aber eine große Anzahl an Evaluierungen untersucht wurde, lag es außerhalb des Umfangs dieser Untersuchung, anstatt oder in Ergänzung zu den Evaluierungsverantwortlichen ehemalige Gutachtende der Evaluierungen zu befragen. Darüber hinaus hätten aufgrund von Personalfluktuation die Gutachtenden zum Teil nicht mehr ausfindig gemacht und befragt werden können. Die Stichprobengröße wäre somit gemindert gewesen.

Generell bestanden Grenzen in der Messung einiger Qualitätskriterien. Bei bestimmten Qualitätskriterien bedürfte es eines hohen Aufwands, um eine „gute" Anwendung zu untersuchen. Es gibt Qualitätskriterien, die nur mit viel Aufwand in der Tiefe untersucht werden können. Zum Beispiel ist beim Qualitätskriterium „Einbindung der Stakeholder*innen" und „Zugänglichkeit für Stakeholder*innen" sowohl die angemessene Anzahl der Stakeholder*innen, die eingebunden werden können, als auch die Intensität der Einbindung in den verschiedenen Evaluierungsphasen schwer ermittelbar.

Die Operationalisierungen der Qualitätskriterien wurden über alle Organisationen hinweg entwickelt. Einige dieser Operationalisierungen trafen nicht auf die Evaluierungspraxis aller Organisationen zu, sodass die Bewertung der Anwendung für diese Organisationen niedriger war, als sie entlang alternativer Operationalisierungen ausgefallen wäre. Es gibt Qualitätskriterien, bei denen mehr Spielraum für die Operationalisierung bestand (beispielsweise bei den „Qualitätssicherungsprozessen"). Hier besteht entsprechend ein Zielkonflikt zwischen dem Erkenntnisinteresse der Anwendung ausgewählter Qualitätskriterien über Organisationen hinweg und der Heterogenität der Anwendung der Qualitätskriterien.

Aufgrund eines uneinheitlichen organisationsübergreifenden Verständnisses der Messung der „Kosten der Evaluierung" konnten Analysen zu Erklärungen der (Nicht-)Anwendung von Qualitätsstandards zum Teil nur eingeschränkt durchgeführt werden.

Die wiederholte Untersuchung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ermöglichte eine Untersuchung der Differenz in der Anwendung der Qualitätskriterien von GIZ und KfW über die Zeit. Sie liefert somit Hinweise, inwieweit organisationsinterne Reformen und die Unterstützung externer Akteure (BMZ und DEval) die Anwendung der Qualitätskriterien verbessern konnten. Sie zeigte darüber hinaus Herausforderungen auf, die mit einer längsschnittlichen Untersuchung einhergehen (zum Beispiel die Anhebung der Schwellenwerte des Anspruchsniveaus und die gegebenenfalls angemessenen Anpassungen von Qualitätskriterien über die Zeit).

Ergebnisse

Evaluierungsfrage 1: Wie ist das Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen in der deutschen EZ?

Das Qualitätsverständnis der beteiligten Organisationen beruhte überwiegend auf den OECD-DAC- und/ oder den DeGEval- und gegebenenfalls organisationsspezifischen Qualitätsstandards. Mit diesen Qualitätsstandards hatten sich die beteiligten Organisationen zu Beginn der Meta-Evaluierung zum Teil wenig systematisch auseinandergesetzt. Darüber hinaus variierten die BMZ-Vorgaben zur Anwendung der Qualitätsstandards im Untersuchungszeitraum in den ausgewählten Haushaltstiteln – zum Teil wurden die OECD-DAC-Standards als verpflichtend gekennzeichnet, zum Teil bestanden keine Vorgaben.

Evaluierungsfrage 2a: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

Insgesamt zeichnete sich ein positives Bild in Bezug auf die Anwendung der OECD-DAC- und der DeGEval-Qualitätsstandards ab. Die beteiligten Organisationen der deutschen EZ wandten die Qualitätsstandards bei circa zwei Drittel ihrer Evaluierungen an. Dies traf auch – zu einem etwas niedrigeren Grad – auf Organisationen ohne Verpflichtungsgrundlage zu. Die Anwendung der Qualitätsstandards zwischen den Organisationen wich dabei zum Teil deutlich voneinander ab. Dies war aufgrund der gewählten Auswahlkriterien für die Aufnahme der beteiligten Organisationen in die Stichprobe zu erwarten und ermöglichte somit ein heterogenes Bild über die unterschiedlichen Anwendungsgrade hinweg.

Jedoch zeigte sich, dass die Qualitätsstandards in den Organisationsdokumenten überwiegend noch nicht vollumfänglich von den Organisationen identifiziert worden waren und ihre (Nicht-)Anwendung nicht systematisch verschriftlicht worden ist. Dies trifft ebenfalls auf die Nachvollziehbarkeit der Anwendung und Nichtanwendung bei einigen ausgewählten Qualitätsstandards auf Ebene der einzelnen Evaluierung zu. Es ist anzumerken, dass die Anwendung einiger Qualitätsstandards aus verschiedenen Gründen nicht auf Evaluierungsebene, sondern auf Ebene der Organisation erfasst wurden. Dies könnte an der gewählten Operationalisierung einiger Qualitätskriterien in der Meta-Evaluierung liegen, durch eine fehlende Dokumentation der (Nicht-)Anwendung bedingt sein oder auch dadurch, dass die Dokumentation der Anwendung ausschließlich auf Organisationsebene und nicht auf Ebene der einzelnen Evaluierung erfolgte. Hier besteht deutlicher Verbesserungsbedarf, da eine externe Untersuchung ohne Informationen zu einer (begründeten) (Nicht-)Anwendung der Qualitätsstandards auf Ebene der Evaluierung nur eingeschränkt möglich war. Es war somit nicht nachvollziehbar, ob ein Qualitätsstandard (mit oder ohne Begründung) nicht angewandt oder angewandt, aber nicht dokumentiert wurde.

Die OECD-DAC-Kriterien wurden zu einem sehr großen Teil von den beteiligten Organisationen in der deutschen EZ erfüllt. Allerdings ist explizit darauf hinzuweisen, dass die Operationalisierung entlang der OECD-DAC-Standards und nicht der OECD-DAC-Kriterien erfolgte und eine Erfüllung somit leicht möglich war. Bei der Anwendung der OECD-DAC-Kriterien bestehen darüber hinaus erste Dokumentationen einer Nichtanwendung auf Organisations- und Evaluierungsebene. Hierin unterscheidet sich die Anwendung der OECD-DAC-Kriterien bereits – wenn auch nur in kleinerem Umfang – von der Anwendung der meisten anderen Qualitätskriterien. Es ist anzunehmen, dass in zukünftigen Evaluierungen die Dokumentation der Nichtanwendung der OECD-DAC-Kriterien weiter ansteigen wird, da ab 2020 (BMZ, 2020) die aktualisierte BMZ-Orientierungslinie zu den Evaluierungskriterien eine begründete und transparente Schwerpunktsetzung vorsieht.

Im Berichtsanhang im Abschnitt 7.1 werden die Erkenntnisse für die vier staatlichen Organisationen BGR, GIZ, KfW und PTB zusätzlich auf Ebene der einzelnen Organisation dargestellt und eingeordnet.

Evaluierungsfrage 2b: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der organisationsspezifischen Qualitätsstandards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

Es zeigt sich ebenfalls ein positives Bild für die Anwendung der organisationsspezifischen Qualitätsstandards durch DRK, EWDE, GIZ und hbs. Sie wurden durchschnittlich „größtenteils angewandt“. Verbesserungspotenzial zeigen sich erneut in der Nachvollziehbarkeit der Nichtanwendung auf Ebene der Evaluierung.

Evaluierungsfrage 2c: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW?

Hinsichtlich der Anwendung der Qualitätskriterien zeichnet sich – mit wenigen Ausnahmen – ein positives Bild ab. Konkret wurden die Qualitätskriterien durchschnittlich zu circa 75 Prozent erfüllt. Dies entspricht einer etwas höheren Ausprägung der Anwendung als bei den OECD-DAC- und den DeGEval-Standards. Herausforderungen bestanden allerdings weiterhin in der Anwendung der Qualitätskriterien „Auswahlverfahren der Gesprächspartner beschrieben“ und „Kontroll-/Vergleichsgruppe einbezogen“. Darüber hinaus ist anzumerken, dass sich die Anwendung der Qualitätskriterien seit der Meta-Evaluierung Nachhaltigkeit in allen Qualitätskriterien – zum Teil bedeutsam – verbesserte. Insgesamt zeigt sich ein

durchschnittlicher Unterschied von 36 Prozent. Die Veränderungen deuten darauf hin, dass die nach der Meta-Evaluierung Nachhaltigkeit durchgeführten Maßnahmen zur Verbesserung der Evaluierungspraxis von GIZ und KfW einen Einfluss auf die Anwendung gehabt haben könnten. Dies ist vor dem Hintergrund der damit einhergehenden umfassenden Anstrengungen einer Vielzahl an Akteuren ein sehr positives Ergebnis. Anzumerken ist aber, dass Alternativerklärungen nicht ausgeschlossen werden können (beispielsweise relativ leicht zu erfüllende Operationalisierungen der Qualitätskriterien oder veränderte Dokumentationsweisen).

Evaluierungsfrage 3: Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen?

Insgesamt geben die Ergebnisse hinsichtlich der in der Literatur und aus den Fokusgruppendiskussionen identifizierten Faktoren sowie der Anwendung der Qualitätskriterien wenig Hinweise auf signifikante Zusammenhänge. Dies sind insbesondere Faktoren aus der evaluierungsspezifischen Dimension „Anzahl der internen und externen Gutachtenden“ und die Umsetzung verschiedener „Qualitätssicherungsprozesse“.

Schlussfolgerungen und Empfehlungen

Die Empfehlungen der Meta-Evaluierung leiten sich vor allem aus den Evaluierungsfragen zur Anwendung ausgewählter Qualitätsstandards und -kriterien (OECD-DAC-, DeGEval- und/oder organisationsspezifischer Standards als auch Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit, Evaluierungsfragen 2a bis 2c) ab. Die Empfehlungen sind allgemein formuliert, sodass alle beteiligten Organisationen sich entlang ihrer organisationsspezifischen Ergebnisse selbst innerhalb der Empfehlungen verorten müssen, da der organisationsübergreifende Mittelwert der Ergebnisse für die Bewertung der Anwendung der Qualitätsstandards einzelner Organisationen nicht ausreichend aussagekräftig ist. Da in der kriterienbasierten Auswahl der nichtstaatlichen Organisationen ein Fokus auf ihre strukturelle Heterogenität gelegt und somit die Bandbreite möglicher Anwendungsgrade und -formen für unterschiedliche Organisationen abgebildet wurde, können sich auch nicht beteiligte nichtstaatliche Organisationen in den Ergebnissen verorten und damit an den Schlussfolgerungen und Empfehlung orientieren. Die an das BMZ gerichteten Empfehlungen beziehen sich auf das BMZ-Referat für Evaluierung.

Identifikation (nicht) relevanter Qualitätsstandards und Verschriftlichung dieser in Organisationsdokumenten

Empfehlung 1

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten im Rahmen einer Revision ihrer Evaluierungspraxis – wenn noch nicht vorhanden – die für ihre Organisation verpflichtenden Qualitätsstandards identifizieren, in Organisationsdokumenten explizit benennen und ihre Anwendung in Evaluierungsprozessen festlegen. Die Identifikation und die systematische Verschriftlichung der Qualitätsstandards sollten in regelmäßigen Abständen überprüft werden. Dabei sollten die Organisationen ihren Anspruch an den Grad der Anwendung der einzelnen Qualitätsstandards konkret bestimmen. (Ergebnis: 2, 5.1, 9.1)
b) Das BMZ sollte im Rahmen anstehender Aktualisierungen von Förderrichtlinien oder Nebenbestimmungen für einzelne Haushaltstitel einen Beitrag dazu leisten, die BMZ-Leitlinien Evaluierung als Referenzdokument für Evaluierungen zu stärken. Im Rahmen der Aktualisierungen sollte das BMZ gemeinsam mit den betroffenen nichtstaatlichen Organisationen organisationale Besonderheiten (beispielsweise wie bei den Förderrichtlinien der politischen Stiftungen) festhalten und verschriftlichen. Das Maximalstandardprinzip sollte dabei erhalten bleiben. (Ergebnis: 3)
c) Das BMZ sollte basierend auf den BMZ-Leitlinien Evaluierung und im Austausch mit staatlichen und nichtstaatlichen Organisationen sowie unter Berücksichtigung des Analyserasters der vorliegenden Meta-Evaluierung ein Analyseraster für die Anwendung der Qualitätsstandards erarbeiten und den staatlichen und nichtstaatlichen Organisationen bereitstellen.

Empfehlung 2

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten in Organisationsdokumenten die generelle Nichtanwendung einzelner für sie verpflichtender Qualitätsstandards begründen und verschriftlichen. (Ergebnis: 5.1, 6.2)

b) Das BMZ sollte sich mit den staatlichen Organisationen bezüglich einer Anwendung und (begründeten) Nichtanwendung der in den BMZ-Leitlinien Evaluierung beschriebenen Qualitätsstandards verständigen, um eine Nichtanwendung auf Organisationsebene gemeinsam festzulegen oder Unstimmigkeiten zu dokumentieren.

Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung relevanter Qualitätsstandards auf Ebene der Evaluierung

Empfehlung 3

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten – wenn noch nicht vorhanden – die Anwendung der auf Organisationsebene festgelegten Qualitätsstandards (Empfehlung 1) in den einzelnen Evaluierungen weiter verbessern, insbesondere die kaum oder teilweise angewandten Qualitätsstandards. Darüber hinaus sollte eine Anwendung oder (begründete) Nichtanwendung aller Qualitätsstandards auf Ebene jeder Evaluierung nachvollziehbar sein und von den Organisationen regelmäßig untersucht werden. (Ergebnis: 4.1, 4.2, 5.2, 5.3, 6.1, 7.1, 8.1, 9.2, 9.3)
b) Das BMZ sollte die staatlichen Organisationen zur Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung der relevanten Qualitätsstandards auf Ebene der Evaluierung anhalten.

Gemeinsames Lernen

Empfehlung 4

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sowie Vertreter*innen von VENRO sollten ihre unterschiedlichen Erfahrungen in der Identifikation, Verschriftlichung, Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung aller Qualitätsstandards regelmäßig untereinander austauschen. Der Austausch sollte auch nicht beteiligte Organisationen integrieren und weitere Evaluierungstypen beinhalten – zum Beispiel dezentrale Evaluierungen –, um die Anwendung der Qualitätsstandards weiter zu verbessern. (Ergebnis: 4.2, 4.3)
b) Das BMZ sollte den Austausch zur Identifikation, Verschriftlichung, Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätsstandards mit und zwischen den Organisationen finanziell unterstützen.

Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit

Empfehlung 5

a) Das BMZ sollte im Zuge der Erarbeitung des Analyserasters für die in den BMZ-Leitlinien Evaluierung beschriebenen Qualitätsstandards (Empfehlung 1) die Übernahme der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit überprüfen und diese gegebenenfalls in das Analyseraster aufnehmen.

b) GIZ und KfW sollten (angelehnt an Empfehlung 5a) die Anwendung und (Nicht-)Anwendung der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit, die in ein BMZ-Analyseraster übernommen wurden, sicherstellen beziehungsweise verbessern und die Nachvollziehbarkeit der (begründeten) (Nicht-)Anwendung je Evaluierung gewährleisten. (Ergebnis: 10.1, 10.2, 11.1)

INHALT

Abbildungen

Tabellen

Kästen

ABKÜRZUNGEN UND AKRONYME

BGR	Bundesanstalt für Geowissenschaften und Rohstoffe
BMZ	Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
CARE	CARE Deutschland e. V.
DAC	Development Assistance Committee (Entwicklungsausschuss)
DeGEval	Gesellschaft für Evaluation e. V.
DO	Durchführungsorganisation
DRK	Deutsches Rotes Kreuz
DVV	Deutscher Volkshochschul-Verband International
EWDE	Evangelisches Werk für Diakonie und Entwicklung e. V.
EZ	Entwicklungszusammenarbeit
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit
hbs	Heinrich-Böll-Stiftung
IR	Inception Report
KAS	Konrad-Adenauer-Stiftung
KfW	KfW Entwicklungsbank
OECD	Organisation for Economic Co-operation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
PTB	Physikalisch-Technische Bundesanstalt
ToR	Terms of References (Leistungsbeschreibung)

1. EINLEITUNG

Im einleitenden Kapitel wird auf den Hintergrund, den Evaluierungsgegenstand, die beteiligten Organisationen sowie die Ziele und Evaluierungsfragen der Meta-Evaluierung eingegangen.

1.1 Hintergrund

Evaluierungen sind ein anerkanntes Mittel, um zu untersuchen, wie erfolgreich eine EZ-Maßnahme war, neue Erkenntnisse für Entscheidungstragende zu schaffen, Lernen aus Schlussfolgerungen und Empfehlungen zu fördern und die Rechenschaftslegung über eine unabhängige Bewertung zu stärken. Um dies zu erreichen, werden in Evaluierungen „möglichst objektive und empirisch fundierte Analysen und Bewertungen des Ausmaßes des erzielten Erfolgs von Entwicklungsmaßnahmen“ (BMZ, 2021, S. 4) durchgeführt. Zielgruppen von Evaluierungen sind häufig die Verantwortlichen und Partner*innen der Maßnahme der Entwicklungszusammenarbeit (EZ), politische Entscheidungstragende, aber beispielsweise auch die Evaluierungscommunity und die Zielgruppen der evaluierten EZ-Maßnahme selbst. Damit hochwertige Evaluierungen entstehen, sollen bei deren Umsetzung Qualitätsstandards eingehalten werden.[7] Die Anwendung dieser Qualitätsstandards gewährleistet, dass wesentliche Aspekte bei der Konzeption, Durchführung und Nutzung einer Evaluierung berücksichtigt werden (beispielweise eine möglichst belastbare Analyse der EZ-Maßnahme, die Partizipation verschiedener Stakeholder*innen sowie die Nützlichkeit der Empfehlungen). Um diese Anwendung organisationsübergreifend und unabhängig bei ausgewählten Organisationen in der deutschen EZ zu untersuchen, wurde die „Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen EZ“ in das mehrjährige Evaluierungsprogramm des DEval aufgenommen (DEval, 2021b).

Meta-Evaluierungen, die als Evaluierung von Evaluierungen bezeichnet werden können, gewinnen in der EZ zunehmend an Bedeutung. Laut Caracelli und Cooksy (2009, S. 2 f., eigene Übersetzung) werden Meta-Evaluierungen „durchgeführt, um den Evaluierungsprozess zu verbessern, die Stärken und Schwächen einer Evaluierung systematisch zu reflektieren, die zukünftige Evaluierungsarbeit zu verbessern oder Informationen zur Glaubwürdigkeit der Ergebnisse für die Nutzenden zur Verfügung zu stellen“.[8] Diesem Verständnis wird ebenfalls in der vorliegenden Meta-Evaluierung gefolgt. Meta-Evaluierungen grenzen sich von Meta-Analysen und Evaluierungssynthesen ab; in den Ersteren werden verschiedene Ergebnisse quantitativ zusammengefasst und ausgewertet, während in den Letzteren eine Auswertung von Evaluierungen mit inhaltlichem Fokus (Caspari, 2012) vorgenommen wird. In der internationalen EZ gibt es inzwischen eine größere Anzahl an Meta-Evaluierungen (zum Beispiel der Agentur der Österreichischen EZ oder des finnischen Außenministeriums). Aber auch in der deutschen EZ werden organisationsinterne (zum Beispiel von der Deutschen Gesellschaft für Internationale Zusammenarbeit [GIZ], der Welthungerhilfe, der Friedrich-Ebert-Stiftung, MISEREOR und von World Vision) und erste organisationsübergreifende Meta-Evaluierungen (zum Beispiel der Projektevaluierungen von GIZ und KfW Entwicklungsbank [KfW] durch das DEval) durchgeführt.[9]

Die vorliegende Meta-Evaluierung, die zeitweise parallel zur Erstellung der Leitlinien Evaluierung des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) lief, schließt inhaltlich unter anderem an frühere Systemprüfungen des deutschen EZ-Systems, das Umsetzungsmonitoring einer Systemprüfung und die Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) an. Im Jahr 1999 wurde im Auftrag des BMZ eine erste systematische Untersuchung der Evaluierungspraxis in der deutschen EZ vorgenommen (Borrmann et al., 1999), zehn Jahre später eine zweite Systemprüfung abgeschlossen. Letztere

7 Auch die OECD (2022) empfiehlt aktuell die Qualität von Evaluierungen im Bereich der öffentlichen Politik zu fördern, beispielsweise durch die Festlegung von Standards.

8 Eine weitere Definition von Meta-Evaluierung des OECD DAC (2002, S. 27, eigene Übersetzung) lautet: „Der Begriff […] Meta-Evaluierung bezeichnet die Evaluierung einer Evaluierung, um deren Qualität zu beurteilen“.

9 Beispiele für weitere Meta-Evaluierungen, in denen die Qualität entlang verschiedener international geltender oder eigens entwickelter Qualitätsstandards untersucht wird, sind Caspari (2010, 2011), FES (2015), Freiman et al. (2016, 2017), Hageboeck et al. (2013), HTSPE Limited (2011), Koy et al. (2016), Krämer et al. (2019), Mauthofer und Silvestrini (2018), Noltze et al.(2018), Queiroz de Souza (2017), Silvestrini und Bäthge (2019), Silvestrini et al. (2018), UNFPA (2020) und Väth et al. (2022). Eine Liste durchgeführter Meta-Evaluierungen findet sich im Onlineanhang Kapitel 1.

beinhaltet systemweite Handlungsempfehlungen zur Weiterentwicklung der Evaluierungspraxis (Borrmann und Stockmann, 2009). Im Rahmen dieser wurde ein Leitfaden erarbeitet, der unter anderem eine vorläufige Version der Standards des Entwicklungsausschusses (Development Assistance Committee, DAC) der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (Organisation for Economic Co-operation and Development, OECD), die Erstfassung der Standards der Gesellschaft für Evaluation (DeGEval), die OECD-DAC-Kriterien und die OECD-DAC-Prinzipien berücksichtigte. Darauf aufbauend führte das DEval im Jahr 2015 ein Umsetzungsmonitoring zu den Ergebnissen und Empfehlungen der Systemprüfung von 2009 durch (Lücking et al., 2015).[10] Im Jahr 2018 wurde schließlich eine organisationsübergreifende Meta-Evaluierung[11] zur Analyse der Qualität von GIZ- und KfW-Projektevaluierungen vom DEval veröffentlicht (Noltze et al., 2018). In dieser sowie der vorliegenden Meta-Evaluierung wird die Qualität von Evaluierungen und nicht der Evaluierungspraxis als System untersucht.

Im zuvor dargestellten Zeitraum gab es regelmäßigen Austausch zwischen verschiedenen Evaluierungseinheiten/-stellen deutscher staatlicher und nichtstaatlicher Organisationen sowie dem BMZ zur Qualität der Evaluierungspraxis. Vor Abschluss der vorliegenden Meta-Evaluierung wurden als aktuellster Meilenstein die BMZ-Leitlinien Evaluierung (BMZ, 2021) veröffentlicht, unter anderem auch um das Qualitätsverständnis von Evaluierungen in der deutschen EZ weiter zu festigen. Darin wurden insbesondere die OECD-DAC-, aber gleichfalls die DeGEval-Standards als verbindlich für die Durchführungsorganisationen beziehungsweise als Orientierung für die nichtstaatlichen Organisationen verschriftlicht. Da der Untersuchungszeitraum der Meta-Evaluierung vor der Verabschiedung der Leitlinien lag, konnten diese nicht in die Untersuchung einbezogen werden. Gleichwohl wurden wichtige Erkenntnisse für die zukünftige Ausgestaltung der Leitlinien generiert.

In die Meta-Evaluierung wurden Evaluierungen von elf deutschen staatlichen und nichtstaatlichen EZ-Organisationen einbezogen. Das erlaubt einen organisationsübergreifenden Blick auf die Anwendung von Qualitätsstandards in der deutschen EZ. Dabei wurden grundsätzlich nur Projektevaluierungen untersucht, die in Deutschland (mit-)verantwortet, zwischen Oktober 2016 und Dezember 2020[12] umgesetzt und vom BMZ (mit-)gefördert wurden. Eine BMZ-Förderung lag vor, wenn entweder die Projektevaluierung oder die EZ-Maßnahme der Evaluierung (mit-)gefördert wurde. Es wurden Projektevaluierungen[13] einbezogen, strategische (zum Beispiel Meta- oder organisationsstrategische Evaluierungen) sowie dezentrale Evaluierungen hingegen nicht.[14] Eine Besonderheit der vorliegenden Meta-Evaluierung liegt darin, dass die Anwendung der Qualitätsstandards in Evaluierungen bei einer Vielzahl staatlicher (Bundesanstalt für Geowissenschaften und Rohstoffe [BGR], GIZ, KfW, Physikalisch-Technische Bundesanstalt [PTB]) und nichtstaatlicher Organisationen (CARE Deutschland e. V. [CARE], Deutsches Rotes Kreuz [DRK], Deutscher Volkshochschul-Verband International [DVV], Evangelisches Werk für Diakonie und Entwicklung e. V. [EWDE], Heinrich-Böll-Stiftung [hbs], Konrad-Adenauer-Stiftung [KAS] und MISEREOR) untersucht wurde. Obwohl das DEval-Mandat zur Evaluierung der EZ-Akteure zum Zeitpunkt der Meta-Evaluierung nur für die vier staatlichen Durchführungsorganisationen über die BMZ-Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit eindeutig festgelegt war, nahmen auch die anderen Organisationen das Angebot an, sich im Rahmen einer unabhängigen, externen Prüfung untersuchen zu lassen.

10 Da die Systemprüfung und das Monitoring der Systemprüfung auch einen Bezug zu den OECD-DAC- und den DeGEval-Standards herstellten, ergeben sich stellenweise für die vorliegende Meta-Evaluierung Anknüpfungsmöglichkeiten, die im Onlineanhang in Abschnitt 4.1.4 weiter ausgeführt werden, aber nicht im Fokus dieser Meta-Evaluierung stehen.

11 In den BMZ-Leitlinien Evaluierung werden DEval-Meta-Evaluierungen als ein Teil der Qualitätssicherung des Evaluierungssystems benannt (BMZ, 2021).

12 Der Zeitraum wurde festgelegt, um bei der Untersuchung mit der Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) anschlussfähig zu sein. Von der Heinrich-Böll-Stiftung liegen Evaluierungen im Zeitraum von Januar 2016 bis Oktober 2020, bei CARE und GIZ von Januar 2018 bis Dezember 2020 vor.

13 Im Folgenden wird „Projekt“ als Präfix nur noch selten angeführt, da nicht alle Organisationen eine einzelne EZ-Maßnahme, sondern zum Teil auch Länderbüros evaluieren.

14 Darüber hinaus gibt es sogenannte Bündelevaluierungen, die als eine Mischform zwischen Projekt- und strategischen Evaluierungen angesehen werden. Diese wurden im Einzelfall zur Grundgesamtheit dazugezählt (beispielsweise bei einer eher technischen Bündelung) oder weggelassen (beispielsweise bei einer eher strategischen Fragestellung). Die beschriebenen Begrifflichkeiten sind unscharf, da auch Projektevaluierungen strategisch genutzt werden können. Sie decken sich aber mit der Evaluierungspraxis einer Vielzahl von Organisationen und werden daher beibehalten.

Bei der Auswahl der Organisationen wurde auf ihre strukturelle Heterogenität geachtet, sodass die Ergebnisse ein möglichst breites Spektrum an Erfahrungen mit der Anwendung von Qualitätsstandards abbilden konnten. Die Heterogenität der Organisationen zeigte sich unter anderem in der Größe der Evaluierungseinheiten/-stellen: Die durchschnittliche Anzahl an Vollzeitäquivalenten in den Evaluierungseinheiten/-stellen im Untersuchungszeitraum lag zwischen 0,5 und 20 Vollzeitäquivalenten je Jahr. Auch die Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Organisationsgröße (Anzahl der Vollzeitäquivalente je Jahr für die Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl Vollzeitäquivalente der Organisation) unterschied sich zwischen den Organisationen und lag zwischen 0,2 und 5,2 Prozent. Die ungefähre durchschnittliche BMZ-Förderung der EZ-Maßnahmen beziehungsweise die Höhe der Aufträge je Jahr variierte zwischen 1,3 Millionen und 1,5 Milliarden Euro, die Anzahl an (mit-)verantworteten Evaluierungen zwischen zwei und 57 je Jahr. Der Fokus auf die Heterogenität der Organisationen ermöglichte es, unterschiedliche strukturelle Herausforderungen bei der Anwendung der Qualitätsstandards zu untersuchen sowie in den Schlussfolgerungen und Empfehlungen zu berücksichtigen, sodass diese auch für nicht beteiligte Organisationen Anknüpfungsmöglichkeiten bieten. Heterogenität besteht zudem nicht nur zwischen den Organisationen, sondern ebenso zwischen den Evaluierungen. So wurden diese zum Beispiel in verschiedenen Regionen (Europa, Afrika, Asien und Amerika) und in unterschiedlichen Sektoren (wie der sozialen Infrastruktur, Produktionssektoren, Verschuldung, der Humanitären Hilfe, Klima und Gender) umgesetzt.

Die Anwendung der Qualitätsstandards wurde entlang der Verpflichtungsgrundlage der Organisationen zu den Standarddokumenten[15] der OECD-DAC- und/oder der DeGEval-Standards und organisationsspezifischer Qualitätsstandards untersucht. **Darüber hinaus besteht eine Einordnung der Erkenntnisse der staatlichen Organisationen, da diese die BMZ-Leitlinien mittelbar umsetzen müssen, während sie nichtstaatlichen als Orientierung dienen.** Die Anwendung von Qualitätsstandards wurde als verpflichtend eingeordnet, wenn es im Untersuchungszeitraum schriftliche Vorgaben zu ihrer Anwendung in Organisationsdokumenten, im Rahmen der BMZ-Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit (BMZ, 2007) beziehungsweise anderen verbindlichen Dokumenten für die beteiligten Organisationen gab (beispielsweise Förderrichtlinien der „privaten Träger“ oder „Sozialstrukturträger“) oder eine Mitgliedschaft in der DeGEval vorlag. Entsprechend wurden die Organisationen für die Untersuchung in zwei Gruppen eingeteilt: Gruppe 1 bestand aus beteiligten Organisationen mit und Gruppe 2 aus beteiligten Organisationen ohne Verpflichtungsgrundlage zur Anwendung der Qualitätsstandards. Daneben besteht eine historisch gewachsene Einteilung in staatliche und nichtstaatliche Organisationen. Als nichtstaatlich werden Organisationen mit zivilgesellschaftlichem, kommunalem und wirtschaftlichem Engagement bezeichnet. Basierend auf dem Primat der instrumentellen Differenzierung der bilateralen staatlichen EZ müssen staatliche Organisationen die BMZ-Leitlinien mittelbar umsetzen, für nichtstaatliche Organisationen gelten diese als Orientierung. Entsprechend gibt es für die staatlichen Organisationen eine gesonderte Darstellung der Erkenntnisse im Berichtsanhang (Abschnitt 7.1) und in den auf das BMZ zugeschnittenen Empfehlungen wurden staatliche und nichtstaatliche Organisationen zum Teil unterschiedlich adressiert.

1.2 Ziele der Evaluierung und Evaluierungsfragen

Mit der organisationsübergreifenden Meta-Evaluierung sollen Erkenntnisse zum Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen sowie zu den Stärken und Schwächen in der Anwendung von Qualitätsstandards geliefert werden. Darüber hinaus werden Faktoren identifiziert und untersucht, die den Grad der Anwendung der Qualitätsstandards erklären können. Ein erstes Ziel bestand darin, das Qualitätsverständnis von Evaluierungen bei unterschiedlichen Organisationen der deutschen EZ herauszuarbeiten (Lernen). Hierbei ging es vor allem darum zu identifizieren, ob und, wenn ja, inwieweit sich die beteiligten Organisationen an den beiden international geltenden Qualitätsstandards für Evaluierungen – OECD-

15 Die „Qualitätsstandards für die Entwicklungsevaluierung“ (OECD DAC, 2010) und die „Standards für Evaluation“ (DeGEval, 2016) werden im Folgenden als Standarddokumente bezeichnet.

DAC- und DeGEval-Standards[16] – und/oder an organisationsspezifischen Qualitätsstandards orientierten. Angelehnt an das Qualitätsverständnis der beteiligten Organisationen sollte darüber hinaus die Anwendung der Qualitätsstandards untersucht werden (Lernen und Rechenschaftslegung). Zudem wurde analysiert, inwieweit GIZ und KfW im Anschluss an die Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) die Anwendung der Qualitätskriterien im Rahmen ihrer Evaluierungspraxis verbessert haben. Im Sinne des zukünftigen Lernens ging es schließlich ebenfalls darum, Erklärungen für die Nichtanwendung von Qualitätsstandards aufzuzeigen (Lernen).

Evaluierungsfrage 1: Wie ist das Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen in der deutschen EZ?

Bei den beteiligten Organisationen bestand im Untersuchungszeitraum ein heterogenes Qualitätsverständnis für Evaluierungen. In einem ersten Schritt war es daher notwendig, dieses zu erfassen, zu systematisieren und darzustellen. Dabei konnte zwischen international geltenden und organisationsspezifischen Qualitätsstandards unterschieden werden.

Evaluierungsfrage 2: Inwieweit werden Qualitätsstandards bei Evaluierungen der beteiligten Organisationen in der deutschen EZ angewandt?

Um diese Frage zu beantworten, wurden die Qualitätsstandards in drei Bereiche eingeteilt und untersucht: 1) international geltende Qualitätsstandards, konkret OECD-DAC- und DeGEval-Standards, 2) organisationsspezifische Qualitätsstandards und 3) Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit. Für jeden Bereich wurde eine separate Evaluierungsfrage formuliert.

Evaluierungsfrage 2a: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

International geltende Qualitätsstandards beschreiben Merkmale für „qualitativ gute Evaluierungen".[17] Um die beteiligten Organisationen entlang eines einheitlichen Analyserasters untersuchen zu können, wurde aus den zwei relevanten international geltenden Qualitätsstandards – OECD-DAC-Standards als Qualitätsstandards für Evaluierungen in der EZ im Spezifischen und DeGEval-Standards als Qualitätsstandards für Evaluierungen im Allgemeinen – die Schnittmenge abgeleitet. Einzelne nicht darin aufgenommene OECD-DAC-Standards wurden zusätzlich analysiert (normativer Bezugsrahmen).

16 Weitere internationale Standarddokumente sind beispielsweise die „African Evaluation Guidelines" der African Evaluation Association (AfrEA, 2020), die „Evaluation Standards for Latin America and the Caribbean" (Rodríguez Bilella et al., 2016) oder die „Normen für Standards und Evaluierung" der United Nations Evaluation Group (UNEG, 2016). Die BMZ-Leitlinien Evaluierung waren zum Zeitpunkt der Analyse noch nicht in Kraft gesetzt und konnten entsprechend nicht berücksichtigt werden.

17 Der OECD DAC (2010, S. 5) beschreibt diesen Anspruch in seinem Standarddokument wie folgt: „In den DAC-Qualitätsstandards für die Entwicklungsevaluierung werden die zentralen Voraussetzungen für Evaluierungsprozesse und -produkte von hoher Qualität identifiziert". Auch die DeGEval hebt die Bedeutung ihrer Qualitätsstandards hervor. Die Qualitätsstandards „sollen Evaluierenden ebenso wie Auftraggebenden […] Orientierung geben, wie gute Evaluationen zu gestalten sind" (DeGEval, 2016, S. 5). In den aktuellen BMZ-Leitlinien Evaluierung (2021) steht, dass insbesondere die OECD-DAC-, aber auch die DeGEval-Standards als relevant angesehen werden.

Evaluierungsfrage 2b: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der organisationsspezifischen Qualitätsstandards in den Evaluierungen der beteiligten deutschen EZ-Organisationen?

Darüber hinaus gibt es organisationsspezifische Qualitätsstandards, die über die Qualitätsstandards des OECD DAC und der DeGEval hinausgehen, für die Organisationen eine besondere Bedeutung darstellen und sich entsprechend in den jeweiligen Evaluierungen widerspiegeln. Diese wurden ebenfalls untersucht.

Evaluierungsfrage 2c: Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW?

Außerdem wurde – angelehnt an die Untersuchung und die Ergebnisse der Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) – für GIZ und KfW die Anwendung der zusätzlichen Qualitätskriterien erneut untersucht. Diese Ergebnisse wurden im Anschluss den vorangegangenen Ergebnissen gegenübergestellt.

Evaluierungsfrage 3: Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen?

Um den Verantwortlichen der beteiligten Organisationen Lernmöglichkeiten hinsichtlich der Anwendung von Qualitätsstandards aufzuzeigen, wurden Faktoren identifiziert und empirisch analysiert, die mit der Anwendung der Qualitätsstandards zusammenhängen. Um eine systematische Untersuchung möglicher Faktoren zu gewährleisten, wurden diese entlang der 1) länderkontext-, 2) evaluierungs- und 3) organisationsspezifischen Einflussdimensionen systematisiert.

2. THEORETISCHE UND EMPIRISCHE HERLEITUNGEN

Im nachfolgenden Kapitel wird im ersten Abschnitt das Qualitätsverständnis der vorliegenden Meta-Evaluierung dargelegt, das Analyseraster mit den herangezogenen Qualitätskriterien hergeleitet und die Zuordnung der Qualitätskriterien zu den drei Standardclustern „Berichtslegung und Methoden", „Partizipation, Unabhängigkeit und Fairness" und „Nutzbarkeit" beschrieben. Im zweiten Abschnitt erfolgt die Herleitung relevanter Faktoren, die mit der Anwendung der Qualitätsstandards positiv oder negativ zusammenhängen können.

2.1 Qualitätsverständnis, Analyseraster und Standardcluster

Um die Anwendung der Qualitätsstandards bei den beteiligten Organisationen zu untersuchen, ist ein abgestimmtes Qualitätsverständnis sowie ein für alle beteiligten Organisationen relevantes Analyseraster notwendig. Daher wurde das Analyseraster systematisch aus den OECD-DAC- und den DeGEval-Standarddokumenten, den Organisationsdokumenten sowie den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit abgeleitet. Da die BMZ-Leitlinien Evaluierung zeitgleich mit dem Analyseraster der Meta-Evaluierung verfasst wurden, konnten sie nicht vollständig berücksichtigt werden.

Qualitätsverständnis

In der vorliegenden Meta-Evaluierung wird die Evaluierungsqualität mit der Anwendung der einschlägigen beziehungsweise der für die Organisationen verpflichtenden Qualitätsstandards gleichgesetzt und dementsprechend untersucht. Der im Konsens mit der Referenzgruppe[18] gewählte Begriff „Anwendung der Qualitätsstandards" beschreibt, inwieweit Belege erkennbar (das heißt schriftlich dokumentiert oder schriftlich auf Nachfrage zurückgemeldet) sind, dass die Qualitätsstandards in den untersuchten Evaluierungen bearbeitet wurden beziehungsweise die Anwendung sichergestellt wurde. Grundlegend für die Qualität von Evaluierungen sind die Standards des OECD DAC und der DeGEval, und zwar aufgrund ihres international geltenden Charakters[19], ihres Bezugs zur EZ und ihrer Relevanz für deutsche EZ-Organisationen[20]. Die international geltenden Qualitätsstandards wurden um organisationsspezifische Qualitätsstandards und die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ergänzt. Da andere Qualitätsverständnisse bestehen, bedeutet ein hoher Grad der „Anwendung der Qualitätsstandards" in Evaluierungen nicht, dass diese in einem alternativen Qualitätsverständnis ebenfalls hoch bewertet werden würden.[21]

Bei den OECD-DAC- und den DeGEval-Standards handelt es sich um Maximalstandards. Das bedeutet, dass die beteiligten Organisationen nicht alle Qualitätsstandards in allen Evaluierungen anwenden müssen. Die beteiligten Organisationen können einzelne Qualitätsstandards aus verschiedenen Gründen nicht anwenden, beispielsweise weil sie nicht mit der (strategischen) Ausrichtung des Evaluierungskonzeptes einer Organisation übereinstimmen (zum Beispiel „Berücksichtigung von Gemeinschaftsevaluierungen" bei nichtstaatlichen Organisationen), sie in einzelnen Evaluierungen nicht anwendbar sind (zum Beispiel die Gewährleistung der „Veröffentlichung des Evaluierungsberichts" aufgrund vertraulicher Informationen) oder sie sich gegenseitig ausschließen (gegebenenfalls „Darstellung der Angemessenheit des methodisches Vorgehens" und „Recht-

18 Die Referenzgruppe bestand aus Vertreter*innen der beteiligten Organisationen und VENRO sowie aus Referent*innen des BMZ-Referats GS 22 „Evaluierung und Ressortforschung, DEval, IDOS". Die Mitglieder begleiteten den Prozess der Evaluierung in allen Evaluierungsphasen (zum Beispiel über virtuelle Treffen oder Kommentierungen von Evaluierungsdokumenten; DEval, 2021a).

19 Ein erster Entwurf der OECD-DAC-Standards wurde zwischen 2006 und 2009 pilotiert. Diese Pilotierung wurde entlang der Kommentare der Mitglieder überarbeitet und im Jahr 2010 eingeführt (OECD, 2013). Im OECD-DAC-Standarddokument heißt es dazu: „Diese Standards, die im internationalen Konsens erarbeitet wurden, sollen als Anreiz und Anregung für die Verbesserung der Evaluierungspraxis dienen" (OECD DAC, 2010, S. 1).

20 Die OECD-DAC- und die DeGEval-Standards haben über die BMZ-Leitlinien Evaluierung einen bindenden Charakter für die staatlichen Durchführungsorganisationen und bieten Orientierung für deutsche zivilgesellschaftliche Organisationen (BMZ, 2021).

21 Beispielsweise kann im Rahmen dieser Meta-Evaluierung eine unvollständige, nicht zufällige Erhebung innerhalb einer Evaluierung als „vollständig angewandt" eingestuft werden, wenn das methodische Vorgehen angemessen begründet und die Limitationen beschrieben werden (zum Beispiel in bestimmten Situationen in fragilen Kontexten). Wenn das methodische Vorgehen aber aus Sicht einer rigorosen Wirkungsmessung betrachtet werden würde, könnte es als verfehlt bewertet werden.

zeitigkeit der Erkenntnisse“). Darüber hinaus kann die Anwendung einzelner Qualitätsstandards stark zwischen Organisationen variieren, da diese in unterschiedlichen Ländern, Sektoren und Kontexten arbeiten beziehungsweise verschiedene Zielsetzungen und Werte vertreten, die ihre Anwendung beeinflussen können. In der vorliegenden Meta-Evaluierung wurde eine Maximalbetrachtung aller Qualitätsstandards angesetzt, um so ein erstes, umfassendes Bild über die aktuelle Anwendungspraxis der beteiligten Organisationen erhalten zu können. Bei der Bewertung dieser Betrachtung wurde das Prinzip der Maximalstandards jedoch berücksichtigt.

Analyseraster

Das Analyseraster beinhaltet 37 Qualitätskriterien, die aus den OECD-DAC- und den DeGEval-Standarddokumenten abgeleitet wurden, elf organisationsspezifische Qualitätskriterien und weitere acht Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit. [22] Der Kodierleitfaden für die OECD-DAC- und die DeGEval-Qualitätskriterien findet sich im Onlineanhang in Kapitel 6.

Das Analyseraster beinhaltet aus den OECD-DAC- und den DeGEval-Standarddokumenten abgeleitete Qualitätskriterien. Diese lassen sich drei Bereichen zuordnen: 1) der Überschneidung zwischen den OECD-DAC- und den DeGEval-Standarddokumenten, 2) dem OECD-DAC-Standarddokument ohne Überschneidung mit dem DeGEval-Standarddokument („OECD DAC only“) und 3) den OECD-DAC-Kriterien. Laut DeGEval (2016, S. 25) sollen die Standards „als Kriterien für die Meta-Evaluation, also die Evaluation von Evaluationen, dienen, indem sie definieren, welche Merkmale gute Evaluationen aufweisen sollten“. Die beiden international geltenden Standarddokumente vereinen zum Teil ähnliche (zum Beispiel „Beschreibung des Evaluierungsgegenstands“) und zum Teil unterschiedliche Aspekte (beispielsweise „Berücksichtigung partnerschaftlicher Ansätze“) in ihrem Qualitätsverständnis. Im Verlauf der Identifikation der Schnittmenge beider Standarddokumente zeigte sich, dass Textteile aus allen DeGEval-Standards in Textteilen der OECD-DAC-Standards aufgehen. Die Textteile, die die Überschneidung abbilden, wurden in einem nächsten Schritt benannt und mit einem Qualitätskriterium hinterlegt. Darüber hinaus wurden in das Analyseraster Qualitätsstandards der OECD-DAC-Standards einbezogen, die keine Überschneidung mit den DeGEval-Standards aufweisen („OECD DAC only“) und mit Qualitätskriterien hinterlegt (Tabelle 1; eine Übersicht über die Herleitung aller Qualitätskriterien der OECD DAC und DeGEval findet sich im Onlineanhang in Abschnitt 2.1).[23] Für die Untersuchung von sieben Qualitätsstandards wurde mehr als ein Qualitätskriterium zur Analyse verwendet. Einen weiteren Bereich bilden die OECD-DAC-Kriterien (Abbildung 2; DEval, 2020). Da alle beteiligten Organisationen den OECD-DAC-Kriterien verpflichtet waren, werden sie als separater Bereich aufgeführt, obwohl ihre Anwendung Teil der OECD-DAC-Standards ist (Qualitätsstandard 2.8; OECD DAC, 2010).

Insgesamt wurden 26 Qualitätsstandards in das Analyseraster aufgenommen, die mit 37 Qualitätskriterien hinterlegt sind. Die Qualitätskriterien wurden – wenn möglich – entlang der Kriterien der Meta-Evaluierung Nachhaltigkeit (Noltze et al., 2018) oder des Monitorings der Systemprüfung (Lücking et al., 2015) operationalisiert. Sie stellen sowohl die Operationalisierung der Qualitätsstandards als auch den Referenzmaßstab für die Bewertung dar.

22 „Qualitätsstandards“ bezeichnen den Originaltext aus den Standarddokumenten des OECD DAC und der DeGEval, „Qualitätskriterien“ stellen daraus abgeleitete Operationalisierungen für die Untersuchung dar. Manche Qualitätsstandards beinhalten mehrere Aspekte die entsprechend durch mehrere Qualitätskriterien untersucht wurden.

23 Bei der Herleitung der Operationalisierungen wurden 1) „Teilaspekte“ der originalen Qualitätsstandards nicht untersucht, da sie nicht wesentlicher Teil der Überschneidung waren (zum Beispiel spezifiziert OECD-DAC-Standard 3.15, dass Meinungsverschiedenheiten im Evaluierungsbericht in Fußnoten oder Anhängen wiedergegeben werden sollen. Diese Spezifikation stellte keine wesentliche Überschneidung zu dem DeGEval-Standard N5 dar, der auf die transparente Dokumentation von unterschiedlichen Perspektiven abzielt und nicht beschreibt, ob dies in den Fußnoten oder dem Anhang stattfinden soll) und 2) teilweise inhaltsschwere Begrifflichkeiten aus den Standarddokumenten (zum Beispiel „Sicherheit“) entweder für die Operationalisierung ebenfalls nicht weiter ausdifferenziert oder eine durch das Evaluierungsteam gewählte Spezifikation herangezogen. In beiden Fällen ist ein inhaltlich übereinstimmendes Verständnis mit den Qualitätsstandards der Standarddokumente nicht sicher gegeben. In den Standarddokumenten sind darüber hinaus die OECD-DAC-Standards in der Regel eher ausführlich und die DeGEval-Standards eher allgemein gehalten, sodass die Zuordnung der Textstellen zueinander nicht frei von Interpretation der Textstellen ist.

Tabelle 1 Herleitung des Qualitätskriteriums „Beschreibung des Evaluierungsgegenstands“

Name	Standard-dokument	Überschneidung	Qualitätskriterium
Beschreibung des Evaluierungs-gegenstands	OECD DAC 2.3	„Die zu evaluierende Entwicklungsmaßnahme (der Gegenstand der Evaluierung) wird klar definiert“ (OECD DAC, 2010, S. 8).	Das Qualitätskriterium ist erfüllt, wenn 1. Ziel(e), 2. Zielgruppe(n) sowie 3. relevante Akteure (politische Partner und/oder Träger) der EZ-Maßnahme genannt sind.
	DeGEval G1	„Konzept des Evaluationsgegenstands […] genau und umfassend beschrieben und dokumentiert“ (DeGEval, 2016, S. 20)	

Quelle: DEval, eigene Darstellung

Anmerkung: Das Qualitätskriterium wurde anhand einer vierstufigen Bewertungsskala kodiert: Wenn keiner der drei im Qualitätskriterium beschriebenen Punkte in der Evaluierung beschrieben wurde, wurde es als „nicht erfüllt“ (1) bewertet, bei einem als „eher nicht erfüllt“ (2), bei zwei als „eher erfüllt“ (3) und bei drei als „vollständig erfüllt“ (4). In Kapitel 6 des Onlineanhangs ist der Kodierleitfaden dargestellt und damit sind alle Bewertungsstufen für jedes Qualitätskriterium abgebildet.

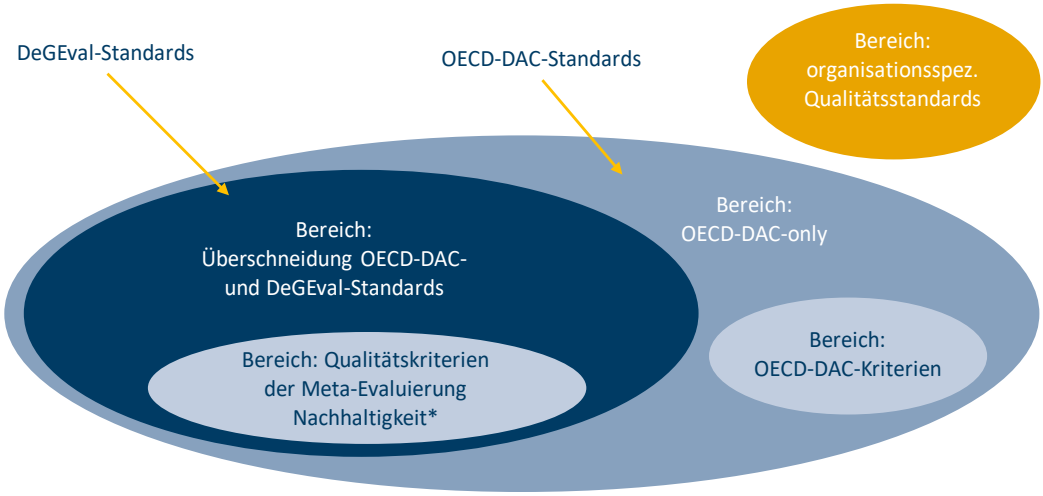
Für DRK, EWDE, GIZ und hbs umfasste das Analyseraster weitere organisationsspezifische Qualitätskriterien. Organisationsspezifische Qualitätskriterien stellen einen weiteren Bereich dar. Sie wurden definiert als Anforderungen, die unabhängig von den OECD-DAC- oder den DeGEval-Standards eine große Bedeutung für eine Organisation hinsichtlich der Qualität ihrer Evaluierungen haben. Beim OECD DAC (2010, S. 5) heißt es dazu: „Die Standards […] schließen auch nicht die Nutzung anderer Qualitätsstandards oder ähnlicher Dokumente für die Evaluierung aus, wie z. B. solchen, die von einzelnen Entwicklungsorganisationen, Evaluierungsgesellschaften oder Netzwerken ausgearbeitet wurden“. Elf organisationsspezifische Qualitätskriterien wurden zunächst über Organisationsdokumente und dann im Austausch mit der Organisation identifiziert und operationalisiert. Sie umfassen neben inhaltlichen Themen auch verschiedene Methoden sowie die Rolle der*s Partners*in der EZ-Maßnahme. Inhaltliche Qualitätskriterien von zwei Organisationen beziehen sich vor allem auf die Verschriftlichung von Genderthemen in verschiedenen Evaluierungsdokumenten wie der Leistungsbeschreibung und bei den Ergebnissen im Evaluierungsbericht. Für die GIZ wurden im Bereich der Methoden die Kontributions- und die Effizienzanalyse als organisationsspezifische Qualitätskriterien identifiziert. Eine andere Organisation legte den Fokus hingegen auf die Rolle der*s Partner*in zum Beispiel bei der Adressat*innenorientierung der Empfehlungen.

Für KfW und GIZ wurden darüber hinaus bereits in der vorangegangenen Meta-Evaluierung Nachhaltigkeit erhobene Qualitätskriterien ins Analyseraster aufgenommen und erneut untersucht. Die Qualitätskriterien fokussieren inhaltlich auf den Bereich des methodischen Vorgehens (zum Beispiel die Durchführung eines „Vorher-Nachher-Vergleichs“). Sieben der insgesamt 15 Qualitätskriterien in diesem Bereich wurden in der vorliegenden Meta-Evaluierung erneut kodiert, wobei teilweise Anpassungen in den Kodierregeln vorgenommen werden mussten, da sich die Struktur und die Inhalte der Evaluierungsberichte seit der Meta-Evaluierung Nachhaltigkeit verändert hatten.[24] Die anderen acht Qualitätskriterien konnten jeweils einem Qualitätskriterium aus dem Analyseraster der international geltenden Qualitätsstandards zugeordnet werden. Die

24 Die Qualitätskriterien lassen sich nicht vollständig aus den Überschneidungen der beiden Standarddokumente ableiten, können ihnen aber inhaltlich zugeordnet werden. Dies betrifft die Qualitätskriterien „Indikatoren verwendet“, „Auswahlverfahren beschrieben“, „Vorher-Nachher-Vergleich“, „Kontroll-/Vergleichsgruppen“, „Kausalität über Plausibilitäten“, „Methoden-Triangulation“ und „Datengrundlage ausreichend“.

Vergleichbarkeit dieser acht (transformierten) Qualitätskriterien ist zum Teil eingeschränkter als die der wiederholt kodierten.[25] Die zusätzlichen Qualitätskriterien wurden nur für GIZ und KfW und nicht die anderen neun beteiligten Organisationen untersucht, da nur Evaluierungen dieser beiden Organisationen in der Meta-Evaluierung Nachhaltigkeit untersucht worden waren. Weitere Details zu den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit finden sich im Onlineanhang in Abschnitt 2.1.

Abbildung 2 Überblick über die fünf Analysebereiche



Quelle: DEval, eigene Darstellung

*Anmerkung: * Die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit lassen sich nicht vollständig aus der Überschneidung der beiden Standarddokumente ableiten, können ihnen aber inhaltlich zugeordnet werden.*

Standardcluster

Mit Ausnahme der fünf OECD-DAC-Kriterien wurden die Qualitätskriterien drei inhaltlichen Standardclustern – „Berichtslegung und Methoden", „Partizipation, Unabhängigkeit und Fairness" und „Nutzbarkeit"[26] – zugeordnet.[27] Da ein Qualitätskriterium inhaltlich mehreren Standardclustern zugeordnet werden könnte (zum Beispiel kann die „Einbindung der Stakeholder*innen" sowohl nützlich sein, da die Stakeholder*innen gegebenenfalls die Evaluierungsergebnisse eher berücksichtigen, als auch fair, da sie sich in die Evaluierung einbringen können), findet die Zuordnung entlang der eindeutigsten inhaltlichen Ausrichtung statt (Tabelle 2). Die Benennungen der drei gebildeten Standardcluster zeigen Ähnlichkeiten zu den Benennungen der DeGEval-Standardgruppen (1. Nützlichkeit, 2. Durchführbarkeit, 3. Fairness und 4. Genauigkeit). Da die identifizierten Qualitätskriterien allerdings zum Teil die Überschneidung zwischen den OECD-DAC- und den DeGEval-Standards darstellen, war eine identische Benennung inhaltlich nicht zutreffend.[28]

25 Insbesondere trifft dies auf die Qualitätskriterien „Methodisches Vorgehen beschrieben" und „Befragte Stakeholder*innen identifiziert" sowie „Schlussfolgerungen referenziert" und „Schlussfolgerungen plausibel" zu, da diese jeweils mit demselben Qualitätskriterium der vorliegenden Meta-Evaluierung verglichen wurden. Das Qualitätskriterium „Daten-Triangulation" wurde nicht erneut erhoben, da es über das Qualitätskriterium „Methoden-Triangulation" inhaltlich abgedeckt wurde. Weitere Details finden sich im Onlineanhang, Abschnitt 2.1. Darüber hinaus fand eine Untersuchung von zwei der in der Meta-Evaluierung Nachhaltigkeit nicht bewerteten Datenerhebungsmethoden statt (Onlineanhang Abschnitt 4.1.2).

26 Der Name „Nutzbarkeit" wurde gewählt, um eine Verwechslung mit der Standardgruppe „Nützlichkeit" der DeGEval-Standards zu vermeiden. Das Cluster „Nutzbarkeit" weicht zum Teil inhaltlich ab beziehungsweise geht darüber hinaus.

27 Die Qualitätsstandards der DeGEval-Standardgruppe „Durchführbarkeit" wurden nicht ausgelassen, sondern inhaltlich von diversen Qualitätskriterien in den Standardclustern „Partizipation, Unabhängigkeit und Fairness" als auch „Nutzbarkeit" abgedeckt.

28 Die OECD-DAC-Standards sind überwiegend entlang von Evaluierungsphasen strukturiert. Die Evaluierungsphasen der OECD-DAC-Standards lauten: 1. Allgemeine Betrachtungen, 2. Zweck, Planung und Konzeption, 3. Durchführung und Berichterstattung, 4. Follow-up, Nutzung der Evaluierung, Lernprozess. Im Vorwort der DeGEval-Standards wird ausgeführt, dass Standards mehreren Evaluierungsphasen zugeordnet werden können und daher auf eine chronologische Einteilung verzichtet wird: „Auch wurde bewusst keine Neugliederung nach typischen Prozessphasen der Evaluation vorgenommen[,] da eine solche nicht einheitlich für alle Evaluationsfälle anwendbar wäre" (DeGEval, 2016, S. 14). Dennoch befindet sich in Kapitel 6 der DeGEval-Standards eine Zuordnung der Standards zu sechs Phasen einer Evaluierung, um eine allgemeine Orientierung zu ermöglichen.

Tabelle 2 Anzahl der Qualitätskriterien je Bereich und Standardcluster

Bereich	untersuchte Organisationen	Berichts-legung und Methoden	Partizipation, Unabhängigkeit und Fairness	Nutzbarkeit	eigenes Cluster	Summe
Überschneidung der OECD-DAC- und der DeGEval-Standards	alle	10	6	8	/	37
OECD DAC *only*	alle	1	5	2	/	
OECD-DAC-Kriterien	alle	/	/	/	5	
Summe	***alle***	***11***	***11***	***10***	***5***	***37***
organisations-spezifische QKs	DRK, EWDE, GIZ, hbs	3	5	3	/	11
QKs Meta-Evaluierung Nachhaltigkeit	GIZ, KfW	15	/	/	/	15

Quelle: DEval, eigene Darstellung

Anmerkung: QK = Qualitätskriterium. 19 Qualitätsstandards (QSs) wurden aus einem Qualitätskriterium berechnet, fünf QSs aus zwei QKs, ein QS aus drei QKs und ein weiterer aus fünf; entsprechend wurden 37 Qualitätskriterien von 26 Qualitätsstandards abgebildet.

Berichtslegung und Methoden: Das Standardcluster umfasst vor allem Qualitätskriterien, die sich auf die Darstellung von Informationen zur Methodik der Evaluierung beziehen (beispielsweise die „Darstellung der Wirkungszusammenhänge“, die „Nachvollziehbarkeit der Informationsquellen“ oder die „Darstellung der Angemessenheit des methodischen Vorgehens“) oder das Vorhandensein beziehungsweise den Informationsgehalt ausgewählter Evaluierungsdokumente (zum Beispiel den „Informationsgehalt der Leistungsbeschreibung“, die „Qualitätssicherung mit Inception Report [IR]“ oder den „Informationsgehalt der Zusammenfassung“). Abbildung 3 stellt die 32 Qualitätskriterien aus der Überschneidung von OECD-DAC- und DeGEval-Standards und dem Bereich „OECD DAC only“ entsprechend ihrer Zuordnung zu den drei Standardclustern dar.

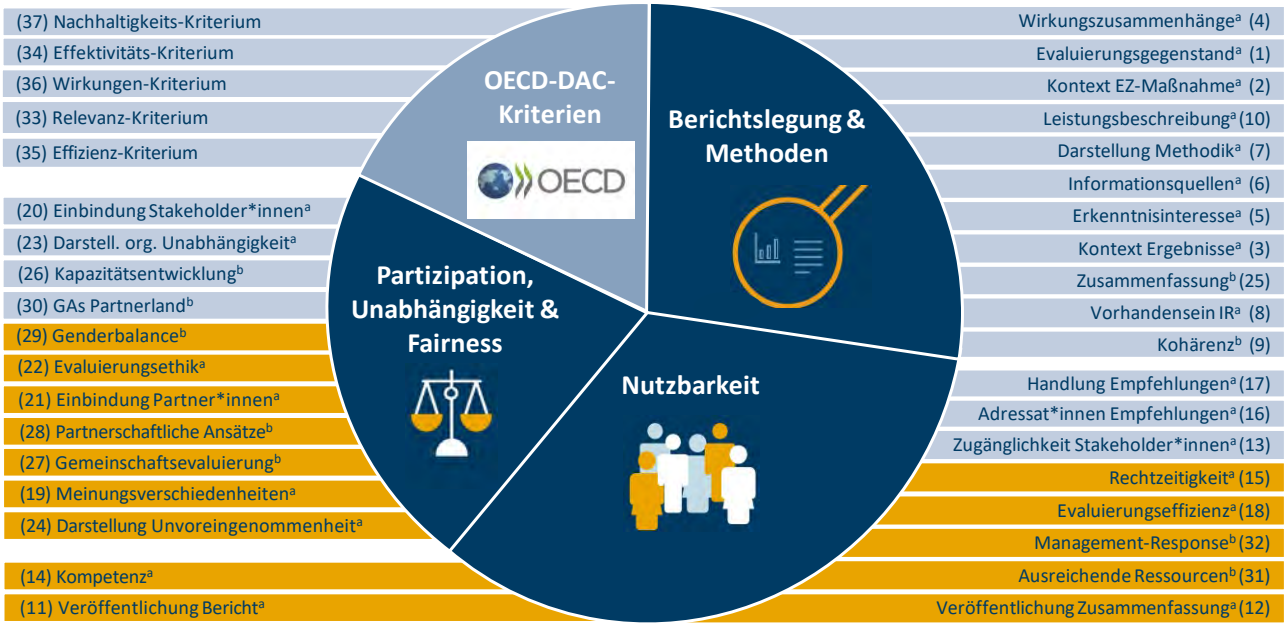
Partizipation, Unabhängigkeit und Fairness: Dem Standardcluster werden vor allem Qualitätskriterien zugeordnet, die sich mit der Berücksichtigung unterschiedlicher Personengruppen in der Evaluierung auseinandersetzen (zum Beispiel der „Einbindung der internen und externen Stakeholder*innen“, dem „Einbezug von Gutachtenden aus dem Partnerland“ sowie der „Genderbalance im Team“). Darüber hinaus sind in dem Standardcluster Unabhängigkeitsaspekte (zum Beispiel die „Darstellung Unvoreingenommenheit der Gutachtenden“) und ethische Aspekte (zum Beispiel die „Evaluierungsethik“) abgedeckt.

Nutzbarkeit: In diesem Standardcluster ist für die Meta-Evaluierung vor allem die Nützlichkeit (theoretisch) relevant, während die Nutzung (praktisch) nur eine kleine Rolle spielt und der Nutzen alleinstehend nicht untersucht wird. Nützlichkeit wird dabei definiert als das Potenzial zur Nutzung (und durch die Nutzung auch das Potenzial zur Entstehung von Nutzen) der Evaluierung, das vor, während und nach ihrer Durchführung beeinflusst werden kann.[29] Nutzung wird als die direkte Reaktion auf die Inhalte der Evaluierung definiert, wobei diese eine textliche oder sprachliche Reaktion auf unterschiedliche Aspekte der Evaluierung darstellen kann. Nutzen beschreibt den tatsächlichen Vorteil, der durch eine Evaluierung entsteht (zum Beispiel die

29 Somit können theoretisch nahezu alle Qualitätskriterien aus allen drei Standardclustern nützlich sein. Das Standardcluster „Nutzbarkeit“ beinhaltet die Qualitätskriterien, die direkt und immer mit der Nützlichkeit einer Evaluierung verbunden sind.

Steigerung der Effizienz auf Basis der Empfehlungen einer Evaluierung). Qualitätskriterien in diesem Standardcluster sind zum Beispiel die „Veröffentlichung des Evaluierungsberichts“, die „Rechtzeitigkeit der Erkenntnisse“ und die „Handlungsorientierung der Empfehlungen“.

Abbildung 3 Zuordnung der 37 Qualitätskriterien zu den Standardclustern und OECD-DAC-Kriterien



Quelle: DEval, eigene Darstellung

Anmerkung: blauer Balken = Qualitätskriterium wurden je Evaluierung untersucht; gelber Balken = Qualitätskriterium wurde auf Organisationsebene über alle Evaluierungen hinweg untersucht; GAs = Gutachtende; Darstell. = Darstellung; org. = organisationale [a] Qualitätskriterium kommt aus der Überschneidung von OECD-DAC- und DeGEval-Standards; [b] Qualitätskriterium kommt aus dem Bereich „OECD DAC only“. Die organisationsspezifischen Qualitätskriterien und die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit werden nicht abgebildet, da sie nicht für alle Organisationen untersucht wurden.

2.2 Faktoren für die Anwendung der Qualitätskriterien

Im folgenden Kapitel soll ein Überblick darüber gegeben werden, welche Faktoren (hergeleitet aus Literatur und Fokusgruppendiskussionen) mit der Anwendung von Qualitätskriterien zusammenhängen beziehungsweise diese erklären können. Die identifizierten Faktoren sind die Grundlage für die in Abschnitt 4.3 geschätzten Regressionsanalysen (Evaluierungsfrage 3). Details zu den Definitionen, den Wirkungszusammenhängen, den Hypothesen und den Operationalisierungen der Faktoren finden sich im Onlineanhang in Abschnitt 2.2.[30]

Die nachfolgend beschriebenen Faktoren wurden in den Analysen untersucht, wenn sie drei Merkmale erfüllten, nämlich 1) eine eindeutige organisationsübergreifende Definition aufwiesen, 2) klare Wirkungszusammenhänge mit ausgewählten Qualitätskriterien beschrieben werden konnten und 3) Daten bei den Organisationen

[30] Obwohl in den folgenden Abschnitten Zusammenhänge zwischen den Faktoren und den Qualitätskriterien auf Basis der zugrunde liegenden Informationen zum Teil kausal formuliert werden, werden die in Abschnitt 4.3 geschätzten Regressionsanalysen ausschließlich korrelativ interpretiert, da mit dem gewählten Design keine Messung von rigorosen Wirkungszusammenhängen möglich war.

oder in der Sekundärdatenbanken verfügbar waren. Um die Faktoren zu identifizieren, wurden drei Fokusgruppendiskussionen[31] mit den Verantwortlichen der beteiligten Organisationen durchgeführt und wissenschaftliche und empirische Literatur gesichtet. Die Faktoren wurden anschließend entlang der 1) länderkontext-, 2) evaluierungs-[32] und 3) organisationsspezifischen Einflussdimensionen systematisiert (Tabelle 3).

Länderkontextdimension[33]

Empirische Untersuchungen und theoretische Ausführungen zeigen auf, dass „Konflikt/Fragilität" keinen eindeutigen Zusammenhang mit der Anwendung des Standardclusters „Berichtslegung und Methoden" hat; „Pandemien" und der „kulturelle Kontext" könnten alle drei Standardcluster beeinflussen. In der erfahrungsbasierten Literatur gibt es Hinweise darauf, dass „Konflikte" sowie „Pandemien" die Datenerhebung in Evaluierungen und die Zusammenarbeit mit lokalen Organisationen aufgrund von Zugangsbeschränkungen, eines erhöhten Sicherheitsrisikos und/oder einer mangelnden Infrastruktur erschweren (Church und Shouldice, 2002; OECD, 2012). In einer Evaluierungssynthese des DEval zur Arbeit der EZ in fragilen Kontexten wird jedoch gezeigt, dass „Fragilität" statistisch nicht mit der methodischen Qualität in Evaluierungsberichten zusammenhängt (Wencker und Verspohl, 2019). Bisher nicht untersucht wurde, ob „Fragilität/Konflikt" mit Qualitätskriterien aus den Standardclustern „Partizipation, Unabhängigkeit und Fairness", „Nutzbarkeit" oder einzelnen Qualitätskriterien aus dem in der vorliegenden Meta-Evaluierung untersuchten Standardcluster „Berichtslegung und Methoden" in Verbindung steht. Ein weiterer in den Fokusgruppendiskussionen identifizierter Faktor ist der „kulturelle Kontext". Dieser könnte bei einer starken Berücksichtigung in der Konzeptionsphase der Evaluierung sowohl „Berichtslegung und Methoden" als auch „Partizipation, Unabhängigkeit und Fairness" der Evaluierung beeinflussen.

Evaluierungsdimension

In der erfahrungsbasierten Literatur finden sich Hinweise, dass sich eine Remote-Datenerhebung positiv auf die Anwendung der Qualitätskriterien „Einbezug von Gutachtenden aus Partnerland", „Berücksichtigung Kapazitätsentwicklung" und negativ auf die „Darstellung der Angemessenheit des methodischen Vorgehens" auswirkt. Seit Beginn der Covid-19-Pandemie hat sich die Durchführung von Remote-Evaluierungen unter anderem auf zwei Aspekte ausgewirkt: 1) die Zusammenarbeit in (internationalen) Teams beziehungsweise die Kooperation mit Gutachtenden aus dem Partnerland und 2) die Methoden für die Datenerhebung. So ist die Zusammenarbeit mit Gutachtenden aus dem Partnerland bei der Durchführung von Remote-Evaluierungen von großer Bedeutung, da sie beispielswiese die richtigen Stakeholder*innen identifizieren können, gut vernetzt sind sowie den Projekt- beziehungsweise Länderkontext gut kennen (von Gumppenberg et al., 2022; Mäder, 2020; World Bank, 2020a, 2020b). Durch ihren Einbezug kann auch die „Kapazitätsentwicklung" in den Partnerländern gefördert werden (von Gumppenberg et al., 2022). Darüber hinaus gehen Remote- beziehungsweise Semi-remote-Evaluierungen häufig mit Reisebeschränkungen einher, sodass Daten für eine Evaluierung nicht mehr vor Ort erhoben werden können. Dies kann gegebenenfalls Einschränkungen in Bezug auf die „Darstellung der Angemessenheit des methodischen Vorgehens" mit sich bringen (World Bank, 2020a; Hundt und Bräuer, 2021; Lange et al., 2020).

31 Fokusgruppendiskussionen gelten als ein erprobtes Instrument, um Erfahrungen und Einschätzungen zu einem bestimmten Thema mit oft heterogen strukturierten Organisationsgruppen in einem relativ kurzen Zeitraum zu bearbeiten (Morgan, 1999). Vier weitere Faktoren aus der Organisationsdimension wurden darüber hinaus im Rahmen einer Diskussion in der Referenzgruppensitzung „Erste Ergebnisse und Schlussfolgerungen" identifiziert.

32 Einige Faktoren können inhaltlich sowohl der Evaluierungs- als auch der Organisationsdimension zugeordnet werden, zum Beispiel der Faktor „Qualitätssicherung mit Inception Report". Da diese Faktoren aber auf der Evaluierungsebene nicht immer gleich angewandt beziehungsweise umgesetzt werden, wurden sie der Evaluierungs- und nicht der Organisationsdimension zugeordnet.

33 Die Faktoren innerhalb dieser Dimension sind von den Organisationen schwer beeinflussbar, allerdings können sie gegebenenfalls dennoch erklären, warum Qualitätskriterien mehr oder weniger angewandt werden.

Sowohl in Meta-Evaluierungen als auch in Rückmeldungen der Verantwortlichen der Evaluierungseinheiten/-stellen der beteiligten Organisationen aus den Fokusgruppendiskussionen hat sich gezeigt, dass der „Informationsgehalt der Leistungsbeschreibung", die „Qualitätssicherung mit Inception Report" und die „Einbindung der Stakeholder*innen" positiv im Zusammenhang mit der Anwendung der Qualitätskriterien aus den Standardclustern „Berichtslegung und Methoden" und „Nutzbarkeit" stehen. In mehreren Meta-Evaluierungen werden die beschriebenen positiven Zusammenhänge gezeigt, zum Beispiel, dass durch den „Informationsgehalt der Leistungsbeschreibung" eine stärkere Auseinandersetzung mit dem „Evaluierungsgegenstand", den „Partner*innen" sowie mit „Berichtslegung und Methoden" stattfindet (Caspari, 2010; FES, 2015; Queiroz de Souza, 2017; Silvestrini und Bäthge, 2019; Väth et al., 2022).

In anderen Meta-Evaluierungen, gleich wie in den Fokusgruppendiskussionen, zeigt sich, dass die „Anzahl der Gutachtenden", die „Anzahl der Arbeitstage der Gutachtenden"[34], die „Kompetenz der Gutachtenden", die „Kosten der Evaluierung" sowie das „Jahr der Evaluierung" die Anwendung der Qualitätskriterien im Standardcluster „Berichtslegung und Methoden" sowie weitere Aspekte der Qualität beeinflussen. In bereits durchgeführten Meta-Evaluierungen wird empirisch belegt, dass eine höhere „Anzahl der Gutachtenden" die „methodische Qualität" steigert (Freimann et al., 2016; Krämer und Almqvist, 2019), jedoch nicht die „Nützlichkeit" (Freimann et al., 2017). Die „Berichtslegung und Methoden" erhöht sich ebenfalls mit der „Anzahl der Arbeitstage der Gutachtenden". Allerdings trifft dies nur bis zu einer bestimmten „Anzahl der Arbeitstage der Gutachtenden" zu, danach ist der Zusammenhang nicht mehr statistisch signifikant beziehungsweise sogar rückläufig (Freimann et al., 2016; Krämer und Almqvist, 2019). In bestehenden Meta-Evaluierungen (Koy et al., 2016; Queiroz de Souza, 2017) und unter den beteiligten Organisationen wird zudem thematisiert, dass die „Kosten der Evaluierung" theoretisch die „methodische Qualität" beeinflussen können, dies konnte allerdings bisher nicht empirisch untersucht werden. Vor allem basiert dies auf der Annahme, dass mit mehr finanziellen Ressourcen – also höheren Kosten – mehr Mittel und somit beispielsweise eine höhere „Anzahl der Arbeitstage der Gutachtenden" zur Verfügung stehen. Hageboeck et al. (2013) führen ebenfalls an, dass die Kosten der Evaluierung als Einflussfaktor untersucht werden sollten, dies aber aufgrund geringer Datenverfügbarkeit empirisch nicht umgesetzt werden konnte. Somit besteht aktuell keine klare empirische Evidenz bezüglich des Zusammenhangs zwischen den „Kosten der Evaluierung" und der Anwendung der Qualitätskriterien. Darüber hinaus wurde in den Fokusgruppendiskussionen angemerkt, dass erfahrungsbasiert das „Jahr der Evaluierung" positiv mit der Anwendung der Qualitätskriterien zusammenhängt, vor allem mit „Berichtslegung und Methoden", da sich die Evaluierungssysteme der Organisationen meist über die Zeit weiterentwickeln beziehungsweise verbessern.

[34] Dabei ist zu berücksichtigen, dass die Faktoren „Anzahl der Gutachtenden" und „Anzahl der Arbeitstage der Gutachtenden" miteinander zusammenhängen beziehungsweise sich gegenseitig bedingen können.

Kasten 1 Zusammenhänge zwischen den Qualitätskriterien

Die einzelnen Qualitätskriterien werden in den beiden Standarddokumenten – von OECD DAC und DeGEval – gleichwertig nebeneinandergestellt. Dennoch ist es wahrscheinlich, dass einige Qualitätskriterien andere beeinflussen (beziehungsweise ihnen vorausgehen). So werden in der Literatur wie auch in den Fokusgruppendiskussionen einzelne Qualitätskriterien als erklärende Faktoren für die Anwendung anderer Qualitätskriterien thematisiert (zum Beispiel, dass die „Kompetenz der Gutachtenden" einen Einfluss auf die „Darstellung der Angemessenheit des methodischen Vorgehens" hat) (Silvestrini und Bäthge, 2019). Vermuten lässt sich dies, da im OECD-DAC-Standarddokument die Qualitätskriterien Evaluierungsprozessphasen zugeordnet werden und somit zumindest zum Teil zeitlich aufeinanderfolgen. Um diesem Verständnis und den bisherigen Erkenntnissen Rechnung zu tragen, wurden ausgewählte Qualitätskriterien als Faktoren wie auch als beeinflusste Qualitätskriterien in den Analysen untersucht.

Organisationsdimension[35]

Über einen Erfahrungsaustausch mit den beteiligten Organisationen wurden die „Größe der Evaluierungseinheiten/-stellen", die „Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen ", die „Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Größe der Organisation" sowie die „Evaluierungstätigkeit" als potenzielle Faktoren ermittelt, die mit der Anwendung der Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" zusammenhängen. Aus Sicht der Verantwortlichen der Evaluierungseinheiten/-stellen der beteiligten Organisationen hängen Faktoren mit Bezug zu den verfügbaren Ressourcen für die Verantwortung von Evaluierungen – konkret die Faktoren „Größe der Evaluierungseinheiten/-stellen", „Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen" und „Größe der Evaluierungseinheiten/-stellen im Verhältnis zur Größe der Organisation" – positiv mit der Anwendung der Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" zusammen. Darüber hinaus steht eine höhere „Evaluierungstätigkeit" positiv mit der Anwendung der Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" in Verbindung.[36] Tabelle 3 gibt einen Überblick über die im Rahmen der Meta-Evaluierung untersuchten Faktoren.

35 Innerhalb dieser Einflussdimension zeigen sich inhaltliche Schnittmengen mit der Systemprüfung beziehungsweise deren Monitoring (Lücking et al., 2015), da in dieser auch Aspekte des Organisationskontexts untersucht wurden. Im Rahmen der Meta-Evaluierung wird allerdings der Zusammenhang zwischen organisationsspezifischen Faktoren und der Anwendung von Qualitätsstandards in Evaluierungen untersucht und nicht, inwieweit sich organisationsspezifische Aspekte zwischen den Organisationen unterscheiden.

36 In den Fokusgruppendiskussionen wurden darüber hinaus „Strukturierter Planungsprozess", „Evaluierungseinheit vorhanden" und „Evaluierungs- und Lernkultur" als Faktoren in der Organisationsdimension genannt. Diese konnten nicht untersucht werden, da sie die drei Merkmale (organisationsübergreifende Definition, Wirkungszusammenhänge, Datenverfügbarkeit) nicht erfüllten. Für einige Faktoren, die eines oder mehrere dieser Merkmale nicht erfüllen konnten, konnten jedoch Proxys ermittelt werden. Ein Proxy ist eine Stellvertretervariable, die der relevanten Variable inhaltlich möglichst ähnlich ist (zum Beispiel konnten für „Kompetenz Gutachtende" die „Anzahl der Gutachtenden" [die unterschiedliche Expertisen besitzen] und der durchschnittliche „Tagessatz externe Gutachtende" als Proxys verwendet werden).

Tabelle 3 Überblick über die untersuchten Faktoren

Dimension	Faktor	M1 – Definition	M2 – Wirkungszusammenhänge	M3 – Datenverfügbarkeit	Proxy des Faktors*	Skalenniveau
Länder-Kontext-dimension	Fragilität	Ja	Ja	Nein	Konflikt	binär
	Kultureller Kontext	Nein	Nein	Nein	Sozialkapital-Index	metrisch
	Pandemie	Ja	Ja	Nein	Jahr 2020	binär
Evaluierungsdimension	Kompetenz Gutachtende	Nein	Ja	Nein	Anzahl interne und externe Gutachtende	metrisch
					Tagessatz externe Gutachtende	metrisch
	Qualitätssicherungs-prozesse	Nein	Ja	Nein	Informationsgehalt LB	ordinal
					Qualitätssicherung mit Inception Report	binär
					Einbindung der Stakeholder*innen	ordinal
	Remote-Datenerhebung	Ja	Ja	Ja	/	ordinal
	Kosten der Evaluierung	Nein	Ja	Nein	Gutachtenden-Tage im Verhältnis zu Kosten der EZ-Maßnahme	metrisch
	Jahr der Evaluierung	Ja	Ja	Ja	/	metrisch
	Datenverfügbarkeit bei Stakeholder*innen**	Nein	Nein	Nein	/	

Dimension	Faktor	M1 – Definition	M2 – Wirkungszusammenhänge	M3 – Datenverfügbarkeit	Proxy des Faktors*	Skalenniveau
Organisations-dimension	Strukturierter Planungsprozess	Nein	Ja	Nein	Organisation	nominal
	Evaluierungs- und Lernkultur	Nein	Ja	Nein		
	Evaluierungseinheit vorhanden	Nein	Ja	Nein		
	Größe der Evaluierungs-einheiten/-stellen	Ja	Ja	Ja	/	metrisch
	Größe der Evaluierungs-einheiten/-stellen im Verhältnis zur Anzahl der Evaluierungen	Ja	Ja	Ja	/	metrisch
	Größe der Evaluierungs-einheiten/-stellen im Verhältnis zur Größe der Organisation	Ja	Ja	Ja	/	metrisch
	Evaluierungstätigkeit	Ja	Ja	Ja	/	metrisch

Quelle: DEval, eigene Darstellung

*Anmerkung: M = Merkmal; M1 – M3 beziehen sich auf die zu Beginn des Kapitels dargestellten Merkmale, die für die empirische Untersuchung der Faktoren erfüllt werden müssen: M1 = eine eindeutige organisationsübergreifende Definition des Faktors; M2 = klare Wirkungszusammenhänge des Faktors mit ausgewählten Qualitätsstandards; M3 = Verfügbarkeit der Daten zu den Faktoren bei den Organisationen; LB = Leistungsbeschreibung. * Wenn der Faktor nicht direkt untersucht werden konnte, weil zum Beispiel keine genaue beziehungsweise keine organisationsübergreifende Definition gefunden werden konnte, wurde stattdessen ein Proxy bestimmt, der große inhaltliche Überschneidungen mit dem originalen Faktor hat. „/" bedeutet, dass keine Bestimmung eines Proxys notwendig war und der Faktor direkt untersucht werden konnte (also alle Merkmale erfüllt waren). ** Der Faktor „Datenverfügbarkeit" wird inhaltlich über die Faktoren „Remote-Datenerhebung" und „Konflikt" abgedeckt.*

3. METHODISCHES VORGEHEN

Im Folgenden werden sowohl die Datengrundlage als auch die durchgeführten Analysen beschrieben. Detailliertere Informationen zu beiden Bereichen finden sich im Onlineanhang in Kapitel 3. Darüber hinaus wird dargestellt, wie die Anwendung der Qualitätskriterien bewertet wurde und welche Stärken und Herausforderungen im methodischen Vorgehen der Meta-Evaluierung bestanden.

3.1 Datengrundlage und -analyse

Auswahl der beteiligten Organisationen und Evaluierungen

Für die Identifikation der Evaluierungen wurde ein zweistufiges Verfahren gewählt. Zuerst wurden die Organisationen und im Anschluss die Evaluierungen der ausgewählten Organisationen bestimmt.

Im ersten Schritt wurden die vier staatlichen – BGR, GIZ, KfW und PTB – und sieben nichtstaatlichen Organisationen – CARE, DRK, DVV, EWDE, hbs, KAS und MISEREOR – in die Meta-Evaluierung aufgenommen. Die Organisationen wurden anhand von vier Kriterien ausgewählt, um eine möglichst große strukturelle Heterogenität der Organisationen abzudecken (Kriterien 1 bis 3) und je Organisation ausreichend Evaluierungen untersuchen zu können (Kriterium 4). Dabei bezog sich Kriterium 1 auf die Höhe der BMZ-Zuwendungen (je zwei Organisationen mit den durchschnittlich höchsten und niedrigsten absoluten BMZ-Zuwendungen pro Jahr), Kriterium 2 auf die relative Evaluierungstätigkeit (je eine Organisation mit den durchschnittlich niedrigsten und höchsten BMZ-Zuwendungen pro Evaluierung pro Jahr) und Kriterium 3 auf den Haushaltstitel der Organisationen (mindestens eine Organisation je Haushaltstitel[37]). Die Auswahl der Organisationen (Fälle) entlang von Kriterien, die auf eine möglichst große Heterogenität abzielen (*diverse case method*), ermöglichte zwar keine Rückschlüsse auf die Verteilung der Anwendung der Qualitätsstandards über alle nichtstaatlichen Organisationen der deutschen EZ hinweg; sie erlaubte allerdings Aussagen zur Anwendung der Qualitätsstandards für eine große Bandbreite an unterschiedlichen Organisationen. Organisationen, die nicht untersucht wurden, könnten die Erkenntnisse somit als Anknüpfpunkte für ihre eigene Evaluierungspraxis heranziehen. Damit kommt die *diverse case method* einer repräsentativen Untersuchung näher als andere Fallauswahlen bei kleinen Stichproben (beispielsweise einer Auswahl entlang von gleichen Fällen oder Extremfällen; Seawright und Gerring, 2008).[38] Kriterium 4 adressierte die Evaluierungshäufigkeit aller Organisationen, um eine ausreichende Anzahl an Evaluierungen je Organisation zu gewährleisten (circa zwei Evaluierungen pro Jahr).[39] Weitere Details finden sich im Onlineanhang in Abschnitt 3.1.

Im zweiten Schritt wurden auf Basis einer nach Organisation und Jahr geschichteten, zufälligen Stichprobe 296 von 576 Evaluierungen gezogen. Insgesamt haben die Evaluierungseinheiten/-stellen der Organisationen im Untersuchungszeitraum von Oktober 2016 bis Dezember 2020[40] 849 Evaluierungen in Deutschland (mit-)verantwortet. In die Grundgesamtheit wurden die 576 Evaluierungen aufgenommen, die das BMZ entweder (mit-)gefördert hatte oder in denen eine vom BMZ (mit-)geförderte EZ-Maßnahme untersucht wurde (durchschnittlicher Deckungsgrad = 62 Prozent). Die gezogene Stichprobe umfasste 296 Evaluierungen

37 Es wurden Organisationen aufgenommen, die in Kapitel „Zivilgesellschaftliches, kommunales und wirtschaftliches Engagement“ des Einzelplans 23 verortet sind, konkret in den BMZ-Haushaltstiteln „Förderung entwicklungswichtiger Vorhaben der Sozialstruktur“, „Förderung entwicklungswichtiger Vorhaben der politischen Stiftungen“, „Förderung entwicklungswichtiger Vorhaben der Kirchen“ und „Förderung entwicklungswichtiger Vorhaben privater deutscher Träger“ (BMF, 2020). Weitere Informationen finden sich im Onlineanhang, Abschnitt 3.1.

38 Da angemessene Kriterien für eine repräsentative Auswahl der nichtstaatlichen Organisationen im Rahmen der Meta-Evaluierung nicht bereitstanden oder im zeitlichen Rahmen hätten ermittelt werden können, wurde auf die *diverse case method* zurückgegriffen.

39 Bei den Kriterien 1 und 2 konnte eine Organisation nur einmal ausgewählt werden (wenn eine Organisation zum Beispiel sowohl die niedrigsten BMZ-Zuwendungen erhielt als auch die höchste Evaluierungstätigkeit aufwies, wurde sie nur über ein Kriterium bestimmt, beim zweiten Kriterium wurde dann die Organisation auf dem nachfolgenden Platz ausgewählt). Bei Kriterium 1 wurden je zwei und bei Kriterium 2 je eine Organisation je Ausprägung ausgewählt, da die Höhe der BMZ-Zuwendungen und der damit verfügbaren Ressourcen je Evaluierung höher gewichtet wurde als die Evaluierungstätigkeit der Organisation. Kriterium 3 wurde ergänzend berücksichtigt und Kriterium 4 für alle identifizierten Organisationen herangezogen.

40 Von der hbs liegen Evaluierungen im Zeitraum von Januar 2016 bis Oktober 2020, bei CARE und GIZ von Januar 2018 bis Dezember 2020 vor.

(durchschnittlich 75,4 Prozent der Grundgesamtheit; Tabelle 4). Aufgrund der gewählten statistischen Auswahlparameter[41] führten die Unterschiede in der Anzahl an Evaluierungen je Organisation dazu, dass Organisationen mit einer geringeren Anzahl an Evaluierungen stärker in der untersuchten Stichprobe vertreten waren. Um die Mittelwerte für die Anwendung der Qualitätskriterien berechnen zu können, wurde bei jeder Organisation das Verhältnis der Evaluierungen aus der Stichprobe im Vergleich zur Grundgesamtheit durch Gewichtung berücksichtigt (disproportionale Stichprobenziehung).

41 Je Organisation wurde die Anzahl an Evaluierungen in der Stichprobe so gewählt, dass bei einem Konfidenzniveau von 95 Prozent eine Fehlerspanne von 10 Prozent nicht überschritten wurde. Die Fehlerspanne steht dabei für eine mögliche Abweichung der gefundenen Ergebnisse von ungefähr 10 Prozent vom realen Wert. Da ordinale Variablen quasimetrisch verwendet werden, kann eine Fehlerspanne definiert und eine Verteilungsannahme getroffen werden. Das Konfidenzlevel besagt, wie sicher die gefundenen Ergebnisse sind. Zusammen mit der Fehlerspanne lässt sich somit zum Beispiel die Aussage treffen, dass bei einem identifizierten Prozentanteil von 40 Prozent in der Stichprobe für ein Qualitätskriterium mit 95 Prozent Wahrscheinlichkeit der Wert in der Grundgesamtheit zwischen 30 und 50 Prozent liegt. Bei einer Organisation mit weniger als zehn Evaluierungen wurden alle Evaluierungen untersucht.

Tabelle 4 Beteiligte Organisationen und Anzahl der Evaluierungen

Nr.	Organisation	Haushaltstitel	GZ der (mit-) verantworteten Evaluierung[a]	GG[b]	SP[c]	Deckungsgrad[d]	Anteil der GG je Organisation an GG für alle Organisationen	Anteil der SP je Organisation an SP für alle Organisationen	Anteil der SP an GG je Organisation
1	BGR	FZ/TZ	31	21	18	67,7 %	3,6 %	6,1 %	85,7 %
2	CARE[e, f]	PT	6	6	6	100,0 %	1,0 %	2,0 %	100,0 %
3	DRK	SST	64	20	17	31,3 %	3,5 %	5,7 %	85,0 %
4	DVV	SST	56	20	17	35,7 %	3,5 %	5,7 %	85,0 %
5	EWDE	Kirche	63	14	13	22,2 %	2,4 %	4,4 %	92,9 %
6	GIZ[e]	FZ/TZ	109	62	38	56,9 %	10,8 %	12,8 %	61,3 %
7	hbs[g]	POS	27	22	19	81,5 %	3,8 %	6,4 %	86,4 %
8	KfW	FZ/TZ	239	230	68	96,2 %	39,9 %	23,0 %	29,6 %
9	KAS	POS	39	20	17	51,3 %	3,5 %	5,7 %	85,0 %
10	MISEREOR	Kirche	158	123	55	77,8 %	21,4 %	18,6 %	44,7 %
11	PTB	FZ/TZ	57	38	28	66,7 %	6,6 %	9,6 %	73,7 %
		Summe	**849**	**576**	**296**	**Ø: 62,5 %[h]**	**100,0 %**	**100,0 %**	**Ø: 75,4 %[i]**

Quelle: DEval, eigene Darstellung

Anmerkung: FZ/TZ = bilaterale staatliche finanzielle und technische Zusammenarbeit; GG = Grundgesamtheit; GZ = Gesamtzahl; POS = politische Stiftung; PT = privater Träger; SP = Stichprobe; SST = Sozialstrukturträger. [a] Gesamtzahl aller Evaluierungen, für deren Berichtsabnahme die (zentralen) Evaluierungseinheiten/-stellen (mit-)verantwortlich waren; [b] Anzahl aller Evaluierungen, für deren Berichtsabnahme die (zentralen) Evaluierungseinheiten/-stellen (mit-)verantwortlich waren und die vom BMZ in einer Form (mit-)gefördert worden waren; [c] Anzahl der untersuchten Evaluierungen; [d] Prozentanteil der Grundgesamtheit an der Gesamtzahl der (mit-)verantworteten Evaluierungen; [e] Zahlen beziehen sich auf die Jahre 2018 bis 2020; [f] Vollerhebung, da weniger als zehn Evaluierungen vorlagen; [g] Zahlen beziehen sich auf den Zeitraum Januar 2016 bis Oktober 2020; [h] Deckungsgrad liegt bei 67,8 Prozent, wenn der Anteil der Grundgesamtheit aller Organisationen an der Gesamtzahl der (mit-)verantworteten Evaluierungen aller Organisationen berechnet wird; [i] Anteil der Stichprobe an der Grundgesamtheit über die Evaluierungen aller Organisationen hinweg liegt bei 51,4 Prozent.

Datenerhebung und -analyse

Im Folgenden wird die Datenerhebung und -analyse für die Evaluierungsfragen 1 (Qualitätsverständnis), 2 (Anwendung der Qualitätsstandards) und 3 (Zusammenhänge zwischen ausgewählten Faktoren und den Qualitätsstandards) dargestellt.

Qualitätsverständnis

Für die Erarbeitung des Qualitätsverständnisses und der Verpflichtungsgrundlage der beteiligten Organisationen wurden Organisationsdokumente sowie BMZ-Vorgaben sowie Interviews mit den Evaluierungsverantwortlichen mithilfe einer qualitativen Inhaltsanalyse ausgewertet.

Anwendung der Qualitätskriterien

Um die Anwendung der Qualitätskriterien in den 296 Evaluierungen zu untersuchen, wurden Evaluierungs- und Organisationsdokumente herangezogen. Zusätzlich wurden die Verantwortlichen der Evaluierungseinheiten/-stellen online befragt, damit eine falsch negative Bewertung der Anwendung ausgeschlossen werden konnte. Insgesamt wurden zur Beantwortung der Evaluierungsfrage 2a (Anwendung der OECD-DAC- und der DeGEval-Standards) ungefähr 1.000 Evaluierungsdokumente (Evaluierungsberichte und -anhänge, Leistungsbeschreibungen, Inception Reports) und weitere Dokumente auf Organisationsebene ausgewertet (zum Beispiel Evaluierungskonzepte, Leitfäden und Handreichungen zur Durchführung von Evaluierungen, standardisierte Vorlagen für Evaluierungsberichte). Für 14 der 37 OECD-DAC- und DeGEval-Qualitätskriterien konnten in der Inter-Kodierenden-Phase[42] keine oder nur sehr wenig Informationen in den von den Organisationen bereitgestellten Evaluierungsdokumenten kodiert werden. Um keine fehlerhaften Rückschlüsse auf eine Nichtanwendung zu ziehen, wurden die Verantwortlichen der Evaluierungseinheiten/-stellen deshalb in einem weiteren Schritt zur Anwendung dieser Qualitätskriterien in ihrer Organisation befragt. Dies geschah online – somit zeitlich flexibel – und standardisiert.[43] Ein Nachteil dieses Vorgehens war, dass die Onlinebefragung auf der Ebene der Organisation erfolgte und somit Erkenntnisse dieser Qualitätskriterien nicht auf Ebene der einzelnen Evaluierung erfasst wurden. Kritisch angemerkt werden muss auch, dass es sich bei den Antworten um Selbstangaben der Verantwortlichen der Evaluierungseinheiten/-stellen handelt, die nicht für die einzelnen Evaluierungen nachvollzogen werden konnten.[44] Die Ergebnisse der Dokumentenanalyse und der Onlinebefragung werden entsprechend nebeneinander – mit unterschiedlichen Farben – in den Ergebnisgrafiken abgebildet, sodass die unterschiedliche Datengrundlage erkennbar ist.

42 Die Inter-Kodierenden-Phase stellte die erste Phase der Kodierung dar, in der jedes Qualitätskriterium in 10 Prozent der Evaluierungen (N = 30 Evaluierungen) von allen Kodierenden kodiert wurde (Döring und Bortz, 2016). In dieser Phase wurde die Anwendung eines Qualitätskriteriums mit „−99“ kodiert, wenn keine Informationen in der Evaluierung identifiziert wurden. Im Anschluss an diese Phase wurde überprüft, ob die Informationen bei mehr als 24 Evaluierungen und mindestens neun Organisationen durchgehend nicht kodiert werden konnten. Wenn dies der Fall war, wurden die Qualitätskriterien in die Onlinebefragung integriert; wenn dies nicht der Fall war, wurde die −99 in eine 1 transformiert und somit das Nichtvorhandensein der Information in den Evaluierungsdokumenten als „kaum angewandt“ eingestuft. Diese Transformation führte dazu, dass einige Qualitätskriterien eher niedrig bewertet wurden (dies gilt insbesondere für die Qualitätskriterien „Zugänglichkeit Stakeholder*innen“, „Einbindung Stakeholder*innen [intern/extern]“, „Darstellung organisationale Unabhängigkeit GAs“ und „Kapazitätsentwicklung“). Weitere Details zur Onlinebefragung finden sich im Onlineanhang, Abschnitt 3.4.

43 Im Gegensatz zur Dokumentenanalyse bestand in der Onlinebefragung für die Organisationen die Möglichkeit, keine Angaben zur Anwendung einzelner Qualitätskriterien zu machen, ohne dass dies als „kaum angewandt“ gezählt wurde. Den Verantwortlichen der Evaluierungseinheiten/-stellen wurde dadurch ermöglicht, keine unangemessene Einschätzung zur Anwendung eines Qualitätskriteriums über alle Evaluierungen hinweg vornehmen zu müssen. In acht Fällen wurden einzelne Rückmeldungen der Onlinebefragung nachträglich begründet verändert (in sechs Fällen wurden die Bewertungen herauf- und in zwei herabgesetzt).

44 Da die Qualitätskriterien der Onlinebefragung durchschnittlich rund 6 Prozent weniger angewandt wurden als die Qualitätskriterien der Dokumentenanalyse, gab es keinen Anlass anzunehmen, dass sich die Organisationen systematisch besser bewerteten, als sie durch die objektive Kodierung bewertet worden wären.

Mögliche Gründe für die fehlende Information zur Anwendung der 14 Qualitätskriterien auf Ebene der einzelnen Evaluierung können 1) in der Komplexität der Qualitätsstandards liegen, 2) der fehlenden Dokumentation einer (begründeten) Nichtanwendung, 3) der Verschriftlichung der Anwendung in nicht berücksichtigten Evaluierungsdokumenten oder 4) der Dokumentation der Anwendung auf Organisationsebene.

1) Einzelne Qualitätskriterien bilden komplexe Inhalte ab (zum Beispiel „Evaluierungsethik", „Evaluierungseffizienz", „Ausreichende Ressourcen vorhanden"), sodass eine über alle Organisationen hinweg sinnvolle Operationalisierung für diese Qualitätskriterien kaum möglich war, da bis zu elf unterschiedliche Operationalisierungen bestehen können (für organisationsinterne Meta-Evaluierungen kann auf eine organisationsspezifische Operationalisierung zurückgegriffen werden). Die Einschätzung der Verantwortlichen der Evaluierungseinheiten/-stellen in der Onlinebefragung bildete an diesen Stellen die Anwendung verschiedener Operationalisierungen ab.
2) Die Auseinandersetzung mit der (Nicht-)Anwendung von Qualitätskriterien hat aktuell noch keinen Eingang in die Evaluierungspraxis gefunden. Somit bestehen auf Ebene der einzelnen Evaluierung keine Belege für die Anwendung einiger Qualitätskriterien (beispielsweise „Transparenz von Meinungsverschiedenheiten", „Berücksichtigung von Gemeinschaftsevaluierungen" und „Berücksichtigung partnerschaftlicher Ansätze").
3) Die Anwendung der Qualitätskriterien kann in Dokumenten verschriftlicht worden sein, die dem Evaluierungsteam aus verschiedenen Gründen nicht zur Verfügung gestellt wurden (zum Beispiel „Kompetenz der Gutachtenden" in den Bewerbungsunterlagen der Gutachtenden); oder sie finden sich auf Kanälen, die nicht für alle Evaluierungen untersucht wurden (beispielsweise „Veröffentlichung des Berichts" und „Veröffentlichung der Zusammenfassung" auf den Webseiten der Organisationen).
4) Die Organisationen hatten die Anwendung der Qualitätskriterien auf Ebene der Organisation und nicht auf Ebene der einzelnen Evaluierung dokumentiert (zum Beispiel „Vorhandensein einer Management-Response", „Evaluierungseffizienz").

Für die Analyse der Anwendung der aus den OECD-DAC- und/oder den DeGEval-Standards abgeleiteten Qualitätskriterien wurden quantitative Inhaltsanalysen sowie Verfahren der deskriptiven Statistik durchgeführt. Für die Analyse wurden die im Analyseraster (Abschnitt 2.1) beschriebenen Qualitätskriterien entlang von ordinalen (1 = „nicht angewandt", 2 = „eher nicht angewandt", 3 = „eher angewandt", 4 = „vollständig angewandt") oder binären Bewertungsstufen (1 = „nicht angewandt", 4 = „vollständig angewandt") kodiert. Ein Vorteil der quantitativen Inhaltsanalyse liegt darin, dass die Anwendung der Qualitätskriterien möglichst intersubjektiv nachvollziehbar und auf Ebene der einzelnen Evaluierung erfasst und ausgewertet werden konnte (Döring und Bortz, 2016).[45] Auch den von den Verantwortlichen der Evaluierungseinheiten/-stellen in der Onlinebefragung zurückgemeldeten durchschnittlichen Häufigkeiten der Anwendung der Qualitätskriterien wurden Werte zugeordnet (1 = „nie", 2 = „selten", 3 = „teilweise", 4 = „überwiegend/häufig", 5 = „immer"). Für jedes Qualitätskriterium wurden je Organisation verschiedene Kennzahlen (zum Beispiel Mittelwerte, Mediane, Maximal- und Minimalwerte) berechnet. Die Kennzahlen wurden nachfolgend in Prozentwerte[46] umgewandelt und den vorab festgelegten Schwellenwerten des Anspruchsniveaus zugeordnet (Abschnitt 3.2). Anschließend wurde für beide Gruppen gemäß der Verpflichtungsgrundlage (Abschnitt 3.2)

45 Die durchschnittliche Inter-Kodierenden-Übereinstimmung lag bei einem guten Wert von 0,77, mit einem niedrigsten Wert von 0,63 und einem höchsten von 0,97. Der Koeffizient von Krippendorffs Alpha kann zwischen 0,00 (fehlende Reliabilität) und 1,00 (perfekte Reliabilität) liegen (Werte von $0{,}80 \leq \alpha \leq 1{,}00$ sind als gut und Werte von $0{,}67 < \alpha \leq 0{,}80$ als akzeptabel anzusehen; Krippendorff, 2012, S. 241). Detaillierte Informationen zum Vorgehen und zur Berechnung der Inter-Kodierenden-Übereinstimmung finden sich im Onlineanhang, Abschnitt 3.3.

46 Für die Analyse wurden die Kennzahlen unter Berücksichtigung der disproportionalen Stichprobenziehung (mittels Designgewichten) berechnet und mithilfe einer Normalisierung in Prozentangaben umgewandelt. Da Bewertungsstufen binär oder ordinal sind, bilden die Prozentangaben Werte ab, die in den Bewertungsstufen nicht definiert waren. Insgesamt bestanden 19 Qualitätsstandards aus einem Qualitätskriterium, fünf Qualitätsstandards aus zwei Qualitätskriterien, ein Qualitätsstandard aus drei Qualitätskriterien und ein weiterer aus fünf. Für die sieben Qualitätsstandards, die aus mehreren Qualitätskriterien bestanden, wurde der Wert separat berechnet. Da eine exakte Schätzung der Mittelwerte über die Evaluierungen hinweg für die Evaluierungseinheiten/-stellen schwierig war, gingen die Qualitätskriterien der Dokumentenanalyse zu 100 Prozent und die der Onlinebefragung zu 50 Prozent in den Wert des Qualitätsstandards ein.

jeweils aus den einzelnen Mittelwerten der Organisationen ein organisationsübergreifender Mittelwert berechnet und für Gruppe 1 bewertet (Abbildung 4). Da der durchschnittliche organisationsübergreifende Mittelwert je Qualitätskriterium auf Basis der Mittelwerte je Organisation berechnet wurde, ist dieser unabhängig von der Anzahl an Evaluierungen der einzelnen Organisation. Für die Berechnung des Standardclusters „Berichtslegung und Methoden" wurden zudem ein Summenindex (Wert über alle Qualitätskriterien hinweg) und Faktorwerte ermittelt.[47]

Abbildung 4 Prozessschritte zur Ermittlung der Bewertung der Anwendung der Qualitätskriterien

Quelle: DEval, eigene Darstellung

Anmerkung: QK = Qualitätskriterium; QS = Qualitätsstandard. Schritte 1 und 2 werden in Abschnitt 2.1 (Qualitätsverständnis, Analyseraster und Standardcluster), Schritte 3 bis 8 im vorherigen Abschnitt beschrieben.

Die Gründe für die (Nicht-)Anwendung der Qualitätskriterien, die auf Ebene der Organisationen untersucht wurden, wurden mit einer qualitativen, ihre Dokumentation mithilfe einer quantitativen Inhaltsanalyse ausgewertet. Die inhaltlich strukturierende qualitative Inhaltsanalyse (Kuckartz, 2014) ermöglichte eine systematische Auswertung schriftlicher Informationen aus der Onlinebefragung und somit ein umfassendes Bild zur formalen und gelebten Evaluierungspraxis. Darüber hinaus wurde mit einer quantitativen Inhaltsanalyse der Organisationsdokumente untersucht, ob und inwieweit die (Nicht-)Anwendung dieser Qualitätskriterien verschriftlicht wurde. In der vorliegenden Meta-Evaluierung wurde eine verschriftlichte begründete Nichtanwendung in den Organisationsdokumenten als „vollständig erfüllt" bewertet. Dies trat für zwei Qualitätskriterien bei fünf Organisationen auf (bei 14 untersuchten Qualitätskriterien in der Onlinebefragung für je elf Organisationen entspricht dies circa 2,6 Prozent der Fälle). Die begründete Nichtanwendung auf Ebene der Organisation wurde positiv im Sinne der Anwendung bewertet, da es sich um einen transparenten und nachvollziehbaren Umgang mit den Qualitätskriterien handelte und dies auf alle Evaluierungen der Organisation Auswirkungen hatte.[48]

47 In den Summenindex flossen die Werte aller Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" zu gleichen Teilen – gleich gewichtet – ein. Für die Standardcluster „Partizipation, Unabhängigkeit und Fairness" sowie „Nutzbarkeit" wurden aufgrund der geringen Anzahl an Informationen der Qualitätskriterien auf Ebene der einzelnen Evaluierung keine Summenindices berechnet („Partizipation, Unabhängigkeit und Fairness": vier von elf Qualitätskriterien; „Nutzbarkeit": drei von zehn Qualitätskriterien). Darüber hinaus wurden Faktorwerte für das Standardcluster „Berichtslegung und Methoden" ermittelt, die eine gewichtete Zusammenfassung der Qualitätskriterien darstellen und mittels einer explorativen Faktorenanalyse bestimmt wurden. Die explorative Faktorenanalyse ist eine statistische Methode, mit der untersucht wurde, ob die Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" auch empirisch zusammengefasst werden können (Backhaus et al., 2015; Brown, 2006).

48 Im Rahmen dieses Vorgehens wurde nicht die Güte der Begründung bewertet. Alternativ könnten (begründete) Nichtanwendungen auch separat zur Anwendung der Qualitätskriterien dargestellt werden.

Bei vier Organisationen wurden insgesamt elf organisationsspezifische Qualitätskriterien entlang binärer oder ordinaler Bewertungsstufen kodiert. Zur Beantwortung der Evaluierungsfrage 2b (Anwendung der organisationsspezifischen Qualitätskriterien) wurden organisationsspezifische Qualitätskriterien (bis zu drei je Organisation, dargestellt in Abschnitt 2.1) in den Evaluierungsdokumenten der Organisationen kodiert und im Anschluss wie die OECD-DAC- und die DeGEval-Qualitätskriterien berechnet und ausgewertet.[49]

Die Evaluierungen von GIZ und KfW wurden zusätzlich entlang der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit untersucht. Für die Beantwortung der Evaluierungsfrage 2c (Anwendung beziehungsweise erneute Anwendung der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit) wurden 15 Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit von Noltze et al. (2018) herangezogen. Acht Qualitätskriterien waren bereits als Qualitätskriterien im OECD-DAC- und im DeGEval-Analyseraster berücksichtigt und wurden lediglich transformiert (zum Beispiel, indem vier Bewertungsstufen in zwei umgewandelt wurden), die übrigen sieben wurden in den aktuellen Evaluierungen erneut kodiert (1 = „nicht angewandt“, 4 = „vollständig angewandt“).[50] Anschließend wurde die Differenz der Ergebnisse zwischen der Meta-Evaluierung Nachhaltigkeit (t0) und der vorliegenden Studie (t1) mithilfe eines Strukturgleichungsmodells untersucht (Weiber und Mühlhaus, 2010). Dies ermöglichte die zeitgleiche Berechnung des Faktors „Zeit“ (t0 versus t1; unabhängige Variable) mit mehreren Qualitätskriterien (abhängigen Variablen). Außerdem wurden dadurch die statistischen Zusammenhänge zwischen den Qualitätskriterien berücksichtigt.

Zusammenhänge zwischen ausgewählten Faktoren und den Qualitätsstandards

Um die Zusammenhänge zwischen ausgewählten Faktoren und der Anwendung der Qualitätskriterien zu untersuchen, wurden multivariate Regressionsanalysen geschätzt. Regressionsanalysen erlauben die Identifikation von statistischen Zusammenhängen[51] zwischen den Faktoren (unabhängige Variablen) und den Qualitätskriterien sowie dem Standardcluster „Berichtslegung und Methoden“ (abhängige Variablen; Backhaus et al., 2011). Konkret wurden Faktoren in 1) der Länderkontext-, 2) der Evaluierungs- und 3) der Organisationsdimension untersucht (Tabelle 5). Die Informationen für die einzelnen Faktoren wurden mithilfe von Daten gewonnen, die von den Organisationen übermittelt worden waren, sowie über Sekundärdatenbanken.

3.2 Bewertung der Anwendung der Qualitätskriterien

Die Anwendung der Qualitätskriterien wurde bei Organisationen mit Verpflichtungsgrundlage zur Anwendung der Qualitätskriterien sowohl untersucht als auch bewertet (Gruppe 1), bei Organisationen ohne Verpflichtungsgrundlage ausschließlich untersucht (Gruppe 2). Bei den Qualitätskriterien der OECD-DAC- und der DeGEval-Standards wurden beide Gruppen hinsichtlich ihres Grads der Anwendung analysiert, Gruppe 1 wurde zusätzlich bewertet. Bei den OECD-DAC-Kriterien (BMZ, 2006)[52], den organisationsspezifischen und

49 Die Intra-Kodierer*innen-Übereinstimmung (erneute Selektion und Klassifikation bei mindestens 10 Prozent der Evaluierungen je Organisation zu zwei unterschiedlichen Zeitpunkten) lag durchschnittlich bei einem guten Wert von 0,91, mit einem niedrigsten bei 0,68 und dem höchsten bei 1,00. Eine Intra-Kodierer*innen-Übereinstimmung wurde berechnet, wenn die Evaluierungen überwiegend von einer Person kodiert wurden. Detaillierte Informationen zur Intra-Kodierer*innen-Übereinstimmung finden sich im Onlineanhang, Abschnitt 3.3.

50 Um die Reliabilität der Qualitätskriterien zwischen den Meta-Evaluierungen sicherzustellen, wurde ein dreistufiges Verfahren gewählt: 1) Die zusätzlichen Qualitätskriterien wurden mit Bezug zu Selektion und Klassifikation in einem diskursiven Verfahren zwischen einem Kodierenden der Meta-Evaluierung Nachhaltigkeit und den Kodierenden der vorliegenden Meta-Evaluierung besprochen. 2) Eine reduzierte Stichprobe von elf Evaluierungen (ungefähr 10 Prozent von 106 Evaluierungen) wurde randomisiert gezogen und von den Kodierenden der vorliegenden Meta-Evaluierung kodiert. Die Inter-Kodierer*innen-Übereinstimmung lag bei 1,00. 3) Da die 106 Evaluierungen darüber hinaus fast ausschließlich von einer Person kodiert wurden, wurde zudem die Intra-Kodierer*innen-Übereinstimmung berechnet. Diese lag ebenfalls durchgehend bei einem Wert von 1,00. Details finden sich im Onlineanhang, Abschnitt 3.3.

51 Da keine rigorose Wirkungsanalyse durchgeführt wurde, wurden die Ergebnisse der multivariaten Regressionsanalysen korrelativ interpretiert. In den Regressionsanalysen wurden die einzelnen Organisationen aufgenommen, um für Unterschiede zwischen diesen kontrollieren zu können. Die Ergebnisse sind über alle Organisationen hinweg gültig, und das unabhängig von der Anzahl an Evaluierungen, die von ihnen beigesteuert wurde.

52 Alle Organisationen wurden im Bereich „OECD-DAC-Kriterien“ Gruppe 1 zugeordnet, da sie sich in ihren Organisationsdokumenten zur Anwendung der OECD-DAC-Kriterien verpflichtet hatten.

den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit wurden ausschließlich Organisationen der jeweiligen Gruppe 1 untersucht und bewertet (Tabelle 5).

Tabelle 5 Zuordnung der Organisationen zu Gruppe 1 und 2 je Bereich

Bereich	BGR	CARE	DRK	DVV	EWDE	GIZ	hbs	KAS	KfW	MISEREOR	PTB
Überschneidung der OECD-DAC- und der DeGEval-Standards	G1	G2	G2	G1	G1	G1	G1	G1	G1	G1	G1
OECD DAC *only*	G1	G2	G2	G2	G2	G1	G1	G1	G1	G2	G1
OECD-DAC-Kriterien	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1
organisations-spezifische QKs	/	/	G1	/	G1	G1	G1	/	/	/	/
QKs der Meta-Evaluierung Nachhaltigkeit	/	/	/	/	/	G1	/	/	G1	/	/

Quelle: DEval, eigene Darstellung

Anmerkung: QK = Qualitätskriterium; G1 = Organisationen mit Verpflichtungsgrundlage (Anwendung der Qualitätskriterien wurde untersucht und bewertet); G2 = Organisationen ohne Verpflichtungsgrundlage (Anwendung der Qualitätskriterien wurde untersucht, aber nicht bewertet); / = keine Untersuchung

Die im Austausch mit der Referenzgruppe festgelegten Schwellenwerte des Anspruchsniveaus für die Anwendung der Qualitätskriterien dienten als Basis für die Bewertung. Das Anspruchsniveau stellte die ex ante festgelegte Einschätzung dar, ab wann ein Qualitätskriterium in einer Evaluierung als kaum, teilweise, größtenteils und vollständig angewandt galt. Bei der Festlegung der Schwellenwerte in 25-Prozent-Schritten (0 ≤ 25 Prozent = „kaum angewandt", > 25 ≤ 50 Prozent = „teilweise angewandt", > 50 ≤ 75 Prozent = „größtenteils angewandt", > 75 ≤ 100 Prozent = „vollständig angewandt") wurde berücksichtigt, dass die Qualitätsstandards als Maximalstandards zu verstehen sind.[53] Es ist somit nachvollziehbar, „dass nicht alle Standards in allen Konstellationen vollumfänglich zu realisieren sind" (DeGEval, 2016, S. 28). Da eine erneute Untersuchung der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit stattfand, wurde an dieser Stelle der Anspruch gestellt, dass sich die Anwendung der Qualitätskriterien seit der Meta-Evaluierung Nachhaltigkeit verbessert hatte.

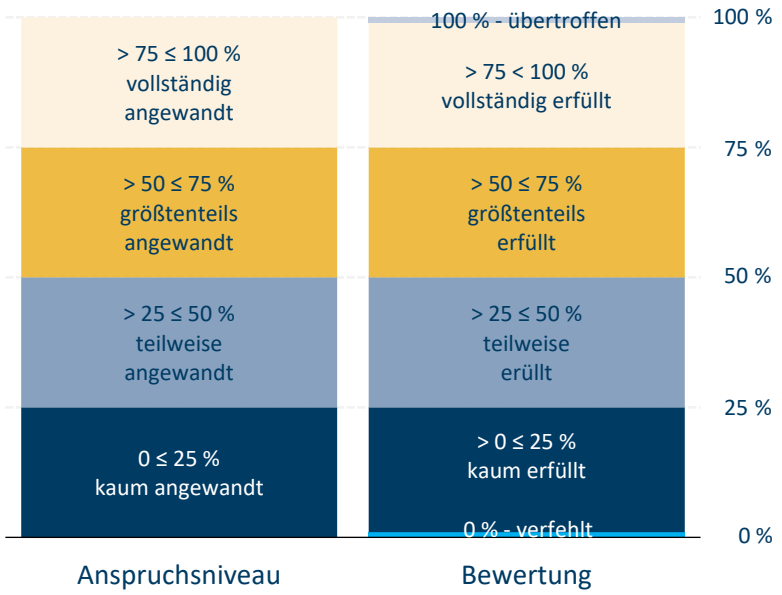
Die Bewertung wurde entlang der Schwellenwerte des Anspruchsniveaus und durch die Hinzunahme der Extremwerte 0 (verfehlt) und 100 (übertroffen) für Gruppe 1 vorgenommen. Zur Vereinheitlichung der Bewertung in DEval-Evaluierungen wurden die DEval-Bewertungsmaßstäbe (2020)[54] erstellt. Daran angelehnt wurden sechs Bewertungskategorien entlang der Schwellenwerte des Anspruchsniveau festgelegt: 0 = „verfehlt", > 0 ≤ 25 Prozent = „kaum erfüllt", > 25 ≤ 50 Prozent = „teilweise erfüllt", > 50 ≤ 75 Prozent = „größtenteils erfüllt", > 75 < 100 Prozent = „vollständig erfüllt", 100 Prozent = „übertroffen" (Abbildung 5 und Abschnitt 7.3 im Berichtsanhang). Um die Bewertung der Anwendung der Qualitätskriterien

53 Da keine wissenschaftlichen oder empirischen Vorlagen für die Einteilung der Schwellenwerte des Anspruchsniveaus in 25-Prozent-Schritten bestanden, beruht die Einteilung auf der nachvollziehbaren und transparenten Herleitung, der Machbarkeit der Ermittlung der Werte und der Akzeptanz durch die beteiligten Organisationen, VENRO und das BMZ.

54 Das DEval verwendet weitgehend einheitliche Bewertungsmaßstäbe, um an den Evaluierungen beteiligten Organisationen die Einordnung zu erleichtern.

angemessen vornehmen zu können, ist auch die Kenntnis über deren Nichtanwendung erforderlich.[55] Eine fehlende Dokumentation der Nichtanwendung wurde als verfehlt bewertet. Dabei ist anzumerken, dass eine fehlende Dokumentation nicht die tatsächliche Nichtanwendung des Qualitätskriteriums in der Evaluierung darstellen muss; aufgrund der fehlenden Nachvollziehbarkeit wurde dies allerdings als fehlend bewertet. Die Organisationen wurden durch dieses Vorgehen in der Anwendung bestimmter Qualitätskriterien eher niedrig bewertet, dies wird bei der Ergebnisbeschreibung der entsprechenden Qualitätskriterien transparent gemacht.

Abbildung 5 Beziehung zwischen Anspruchsniveau und Bewertung



Quelle: DEval, eigene Darstellung

3.3 Stärken und Herausforderungen im methodischen Vorgehen

Aufgrund der Auswahl der Organisationen entlang ihrer strukturellen Heterogenität wurde eine große Bandbreite in der Anwendung einzelner Qualitätskriterien untersucht und dargestellt. Entsprechend können sich nicht beteiligte nichtstaatliche Organisationen innerhalb dieser Bandbreite verorten und Erkenntnisse der Meta-Evaluierung für sich nutzen. Die Erkenntnisse der Meta-Evaluierung gelten für die beteiligten Organisationen. Die staatlichen Organisationen werden vollständig abgebildet. Die Auswahl der beteiligten nichtstaatlichen Organisationen ist nicht repräsentativ. Dies bedeutet aber nicht, dass einzelne Befunde nicht auch für andere nichtstaatliche Organisationen zutreffen und somit nützlich sein könnten.[56] Aufgrund der Auswahl der Organisationen entlang ihrer strukturellen Heterogenität ist zwar die Verteilung der Anwendung für alle Organisationen der nichtstaatlichen EZ nicht bekannt, die Bandbreite möglicher Anwendungen beziehungsweise Auseinandersetzungen mit den Qualitätsstandards allerdings dargestellt.

55 Dies wird nur in den DeGEval-Standards explizit ausgeführt, kann aber auch auf die OECD-DAC-Standards übertragen werden, da ohne eine Dokumentation einer Nichtanwendung weder eine angemessene Überprüfung der Anwendung noch Querschnittsanalysen möglich sind. In den DeGEval-Standards (2016, S. 29) heißt es entsprechend: „Die teilweise oder vollständige Nichterfüllung von Einzelstandards sollte immer offen und nachvollziehbar etwa im Rahmen der Berichterstattung dokumentiert und begründet werden“.

56 Merkmale, die eine repräsentative Stichprobenziehung aller nichtstaatlichen Organisationen ermöglichten, waren zu Beginn der Meta-Evaluierung nicht bekannt und konnten in ihrem zeitlichen Rahmen nicht identifiziert werden.

Eine Übertragbarkeit der Erkenntnisse auf die Grundgesamtheit der Evaluierungen je Organisation ist im Rahmen der gewählten statistischen Kennwerte für die Stichprobenziehung gewährleistet, eine Übertragbarkeit auf andere Evaluierungstypen einer Organisation nicht. Es gibt keine systematische Analyse, ob beziehungsweise inwieweit die Erkenntnisse zur Anwendung der Qualitätskriterien für andere Evaluierungstypen gelten, die nicht untersucht wurden. Möglich ist, dass Evaluierungen, die anhand derselben Prozesse in einer Organisation umgesetzt, aber beispielsweise mittels anderer Finanzierungsquellen gefördert wurden, ähnliche Ergebnisse abbilden. Hingegen ist anzunehmen, dass mittels anderer Prozesse durchgeführte Evaluierungen (beispielsweise dezentrale Evaluierungen) eher andere Ergebnisse hervorbringen (BMZ, 2021; Koy et al., 2016).

Aufgrund der systematischen Ableitung des Analyserasters entlang der OECD-DAC- und der DeGEval-Standards kann es auch von anderen Organisationen genutzt werden. Da sich das Analyseraster auf die OECD-DAC- und die DeGEval-Standarddokumente zur Evaluierung bezieht (die auch in den BMZ-Leitlinien Evaluierung herangezogen werden; BMZ, 2021), erhöht sich die Nützlichkeit der Meta-Evaluierung. Das Analyseraster der Meta-Evaluierung kann somit zukünftig als Grundlage für die Erstellung eines Analyserasters auf der Basis der BMZ-Leitlinien Evaluierung dienen.

Bei der Analyse der Qualitätskriterien aus der Onlinebefragung bestanden Einschränkungen hinsichtlich der Methoden-Triangulation. Die Hinzunahme einer weiteren Datenerhebungsmethode hätte allerdings angesichts des Aufwand-Nutzen-Verhältnisses in keiner angemessenen Relation gestanden. Bei den Qualitätskriterien, die mittels der Befragung der Verantwortlichen der Evaluierungseinheiten/-stellen erhoben wurden (Selbstangaben über alle Evaluierungen hinweg auf Ebene der Organisation), hätten zur Triangulation der Daten die Einschätzungen der ehemaligen Gutachtenden der Evaluierung herangezogen werden können. Da in der vorliegenden Meta-Evaluierung aber eine große Anzahl an Evaluierungen unterschiedlicher Organisationen über einen Zeitraum von über vier Jahren behandelt wurde, lag es außerhalb des Umfangs dieser Untersuchung, anstatt oder in Ergänzung zu den Evaluierungsverantwortlichen ehemalige Gutachtende der Evaluierungen zu befragen. Insbesondere aufgrund von Personalfluktuation hätten die Gutachtenden zum Teil nicht mehr ausfindig gemacht und befragt werden können. Die Stichprobengröße wäre somit gemindert gewesen. Darüber hinaus hätte eine Befragung ebenfalls Risiken mit sich gebracht und damit in keinem ausreichenden Verhältnis zum Nutzen gestanden. Zur transparenten Darstellung werden im vorliegenden Bericht die Erkenntnisse aus der Dokumentenanalyse und der Onlinebefragung voneinander getrennt dargestellt (siehe Abschnitt 4.2.1).

Generell bestehen Grenzen in der Messung einiger Qualitätskriterien. Bei bestimmten Qualitätskriterien bedürfte es eines hohen Aufwands, um eine „gute" Anwendung zu untersuchen. Es gibt Qualitätskriterien, die nur mit viel Aufwand in der Tiefe untersucht werden können. Zum Beispiel ist beim Qualitätskriterium „Einbindung der Stakeholder*innen" und „Zugänglichkeit für Stakeholder*innen" sowohl die angemessene Anzahl der Stakeholder*innen, die eingebunden werden können, als auch die Intensität der Einbindung in den verschiedenen Evaluierungsphasen schwer ermittelbar. Dies betraf ebenso das Qualitätskriterium „Darstellung der Angemessenheit des methodischen Vorgehens". Hierfür wurde beispielsweise folgende Operationalisierung herangezogen: „Das Qualitätskriterium ist erfüllt, wenn a) nachvollziehbar begründet wird, warum die angewandten Methoden angemessen sind und b) Limitationen des methodischen Vorgehens diskutiert werden." Im Rahmen der Kodierung wurde folglich nicht überprüft, ob die dargestellten Methoden tatsächlich angemessen waren; es ging darum, ob die Begründungen nachvollziehbar waren und die Limitationen diskutiert wurden. Somit erfassen die gewählten Operationalisierungen zwar relevante Aspekte der Qualitätskriterien, bilden aber nicht unbedingt die Tiefe des zugrunde liegenden Qualitätsanspruchs ab. Auch die Qualitätskriterien „Darstellung Unvoreingenommenheit der Gutachtenden" und „Darstellung organisationale Unabhängigkeit der Gutachtenden" würden eine umfassende qualitative Analyse der Vergabe- und Einstellungsprozesse der Organisationen sowie der Überzeugungen der Gutachtenden erfordern, um die Unvoreingenommenheit beziehungsweise Unabhängigkeit gemäß dem Qualitätsverständnis einschätzen zu können. Eine Möglichkeit, die Einschätzungen zukünftig zu verbessern, wäre es, die notwendigen Informationen über Rückmeldungen der Gutachtenden oder der Verantwortlichen der Evaluierungseinheiten/-stellen zum Zeitpunkt der Evaluierung festzuhalten. Entsprechend könnten dann auch die Operationalisierungen dieser Qualitätskriterien weiter präzisiert werden.

Für die organisationsübergreifende Meta-Evaluierung wurden die Operationalisierungen der Qualitätskriterien über alle Organisationen hinweg entwickelt. Einige dieser Operationalisierungen trafen nicht auf die Evaluierungspraxis aller Organisationen zu, sodass die Bewertung der Anwendung für diese Organisationen niedriger war, als sie entlang alternativer Operationalisierungen ausgefallen wäre. Es gibt Qualitätskriterien, für die die Operationalisierung relativ eindeutig aus den Standarddokumenten abgeleitet werden konnte (zum Beispiel „Informationsgehalt der Leistungsbeschreibung" entlang der Darstellung verschiedener Aspekte); bei anderen Qualitätskriterien bestand hingegen mehr Spielraum (beispielsweise bei den „Qualitätssicherungsprozessen"). Da es im Interesse der vorliegenden Meta-Evaluierung lag, organisationsübergreifende Erkenntnisse zu generieren, wurde in den meisten Fällen eine Operationalisierung festgelegt. Hier besteht entsprechend ein Zielkonflikt zwischen dem Erkenntnisinteresse der Anwendung ausgewählter Qualitätskriterien über Organisationen hinweg und der Heterogenität der Anwendung der Qualitätskriterien. In den Ergebnissen (Abschnitt 4.2.1) wird das Vorhandensein mehrerer Formen der Anwendung bestimmter Qualitätskriterien transparent gemacht.

Aufgrund eines uneinheitlichen organisationsübergreifenden Verständnisses der Messung der „Kosten der Evaluierung" konnten Analysen zu Erklärungen der (Nicht-)Anwendung von Qualitätsstandards zum Teil nur eingeschränkt durchgeführt werden. Er war nicht möglich, die Kosten für Evaluierungen zu ermitteln beziehungsweise diesen Faktor in der Analyse zu untersuchen. Entsprechend musste ein Proxy für die Analyse der Zusammenhänge zwischen den Kosten einer Evaluierung und der Anwendung ausgewählter Qualitätsstandards herangezogen werden. Ohne ein einheitliches Verständnis beziehungswiese eine klare Definition der Evaluierungskosten können diesbezüglich keine Erkenntnisse gewonnen werden.

Die wiederholte Untersuchung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit ermöglichte es, die Differenz in der Anwendung der Qualitätskriterien von GIZ und KfW über die Zeit zu betrachten. Sie lieferte somit Hinweise, inwieweit gegebenenfalls organisationsinterne Anstrengungen und Maßnahmen die Anwendung der Qualitätskriterien verbessern konnten. Die Differenz der Ergebnisse über die Zeit gibt Hinweise, inwieweit die Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit zum Beispiel mittels organisationsinterner Reformen und der Unterstützung externer Akteure (BMZ und DEval) verbessert werden konnte. Sie zeigt darüber hinaus Herausforderungen auf, die mit einer längsschnittlichen Untersuchung einhergehen (zum Beispiel die Anhebung der Schwellenwerte des Anspruchsniveaus und die gegebenenfalls angemessenen Anpassungen von Qualitätskriterien über die Zeit). Da die wiederholte Umsetzung von Meta-Evaluierungen in den meisten Organisationen noch nicht Eingang in die regelmäßige Evaluierungspraxis gefunden hat, bilden diese Erkenntnisse auch Lernmöglichkeiten für andere Organisationen ab.

4. ERGEBNISSE

Das Ergebniskapitel ist in drei Abschnitte eingeteilt. In jedem Abschnitt wird eine der drei Evaluierungsfragen thematisiert. Zuerst wird dargestellt, welche Qualitätsverständnisse bei den beteiligten Organisationen bestanden (Evaluierungsfrage 1). Im Anschluss wird beschrieben, inwieweit die beteiligten Organisationen die OECD-DAC- und die DeGEval-Qualitätskriterien (2a), die organisationsspezifischen Qualitätskriterien (2b) und Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit (2c) anwandten. Abschließend werden die Zusammenhänge zwischen den identifizierten erklärenden Faktoren und der Anwendung der Qualitätskriterien dargestellt (3). Im Berichtsanhang werden die Erkenntnisse für die vier staatlichen Organisationen BGR, GIZ, KfW und PTB zusätzlich auf Ebene der einzelnen Organisation dargestellt und eingeordnet (Abschnitt 7.1).

4.1 Qualitätsverständnis bei den beteiligten Organisationen

Im ersten Abschnitt wird das Qualitätsverständnis der beteiligten Organisationen während des Untersuchungszeitraums beschrieben. Im Kasten 2 werden das Fazit und die Hauptergebnisse dargestellt.

Kasten 2 Fazit zum Qualitätsverständnis

Wie ist das Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen in der deutschen EZ? (Evaluierungsfrage 1)

Das Qualitätsverständnis der beteiligten Organisationen beruhte überwiegend auf den OECD-DAC- und/oder den DeGEval- und gegebenenfalls organisationsspezifischen Qualitätsstandards. Mit diesen Qualitätsstandards hatten sich die beteiligten Organisationen zu Beginn der Meta-Evaluierung zum Teil wenig systematisch auseinandergesetzt.

- Bei sechs von elf Organisationen lag eine Verpflichtungsgrundlage zur Anwendung der OECD-DAC- und der DeGEval-Standards vor (BGR, GIZ, hbs, KAS, KfW und PTB), bei drei bestand ausschließlich eine Verpflichtung zu den DeGEval-Standards (DVV, EWDE, Misereor) und bei zwei Organisation gab es keine Verpflichtungsgrundlage (CARE und DRK). Alle elf Organisationen verpflichteten sich zu den OECD-DAC-Kriterien und vier Organisationen berücksichtigten darüber hinaus organisationsspezifische Qualitätsstandards (DRK, EWDE, GIZ und hbs). (Ergebnis: 1)
- Die OECD-DAC- und/oder die DeGEval- und gegebenenfalls organisationsspezifische Qualitätsstandards waren zu Beginn der Meta-Evaluierung bei den beteiligten Organisationen zum Teil nicht vollumfänglich bekannt, systematisch in Organisationsdokumenten verschriftlicht und operationalisiert. (Ergebnis: 2)
- Die BMZ-Vorgaben hinsichtlich der Anwendung der Qualitätsstandards variierten im Untersuchungszeitraum in den relevanten Haushaltstiteln – zum Teil wurden die OECD-DAC-Standards als verpflichtend gekennzeichnet, zum Teil bestanden keine Vorgaben.57 (Ergebnis: 3)

Bei den beteiligten Organisationen lagen unterschiedliche Verpflichtungsgrundlagen zur Anwendung der OECD-DAC-, DeGEval- und/oder organisationsspezifischen Qualitätsstandards vor, zum größten Teil mussten diese mit den Organisationen zu Beginn der Meta-Evaluierung erst geklärt werden. Bis auf CARE und DRK verpflichteten sich alle Organisationen im Rahmen einer Mitgliedschaft[58] und/oder der Beschreibung in ihren Organisationsdokumenten zur Anwendung der DeGEval-Standards, GIZ und KAS[59] zusätzlich zu den

57 In den 2021 veröffentlichten BMZ-Leitlinien Evaluierung sind insbesondere die OECD-DAC-, aber auch die DeGEval-Standards als verbindlich für die Durchführungsorganisationen beziehungsweise als Orientierung für die nichtstaatlichen Organisationen verschriftlicht worden.

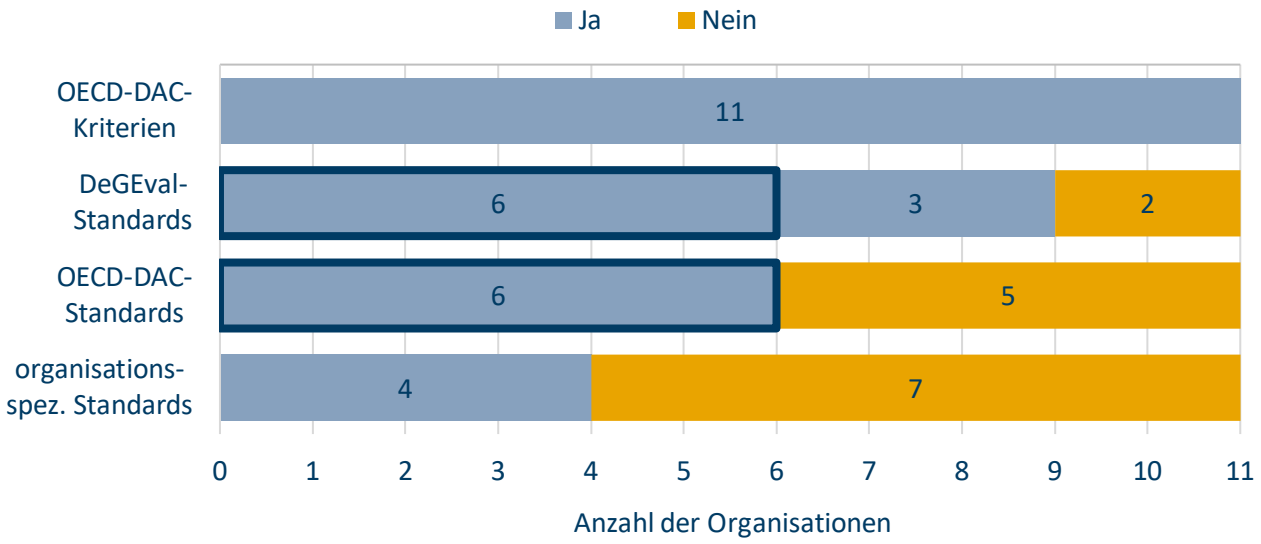
58 Seit 2009 verpflichtet sich eine Organisation mit der Unterschrift der Beitrittserklärung zur Orientierung an den DeGEval-Standards. Zur aktuellen revidierten Fassung der DeGEval-Standards verpflichten sich die Organisationen, die seit deren Verabschiedung am 21. September 2016 die (aktuelle) Beitrittserklärung unterzeichnet haben (DeGEval, 2021). Für beteiligte Organisationen, die vor 2016 beziehungsweise 2009 der DeGEval beigetreten sind, wurde eine mögliche Selbstverpflichtung zu den DeGEval-Standards über die Organisationsdokumente und/oder im Austausch mit den Verantwortlichen der Evaluierungsstellen/-einheiten ermittelt.

59 Die KAS verpflichtete sich den OECD-DAC-Standards von 2016 bis 2019.

OECD-DAC-Standards. Die vier Durchführungsorganisationen und zwei politischen Stiftungen waren außerdem über die „Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit“ (BMZ, 2007) beziehungsweise die „Förderrichtlinien für politische Stiftungen“ (BMZ, 2016) zur Anwendung der OECD-DAC-Standards verpflichtet. Für die Organisationen der Haushaltstitel „Private Träger“, „Sozialstrukturträger“ und „Kirchen“ bestanden seitens des BMZ keine Vorgaben. Insgesamt waren im Untersuchungszeitraum sechs Organisationen beiden Standarddokumenten verpflichtet. In den BMZ-Vorgaben lag der Fokus auf den OECD-DAC-Standards, während in den Organisationsdokumenten häufiger eine Verpflichtung zu den DeGEval-Standards beschrieben wurde. Alle elf beteiligten Organisationen waren (und sind) über ihre Organisationsdokumente zur Anwendung der auch in den OECD-DAC-Standards integrierten OECD-DAC-Kriterien verpflichtet (Abbildung 6).[60] Die zum Zeitpunkt des Untersuchungszeitraums der Meta-Evaluierung noch nicht in Kraft gesetzten BMZ-Leitlinien Evaluierung (2021) haben das BMZ-Qualitätsverständnis sowie die Verpflichtung zur Anwendung der Qualitätsstandards weiter formalisiert und konkretisiert, dabei bezieht sich das BMZ insbesondere auf die Anwendung der OECD-DAC- und darüber hinaus auf die DeGEval-Standards.

Die elf identifizierten organisationsspezifischen Qualitätskriterien von DRK, EWDE, GIZ und hbs fokussierten auf inhaltliche, methodische und partnerschaftliche Aspekte von Evaluierungen. Organisationsspezifische Qualitätsstandards ergänzten die OECD-DAC- und die DeGEval-Standards und hatten für die Organisationen eine besondere Bedeutung. Zwei weitere Organisationen verwiesen auf organisationsspezifische Qualitätsstandards, die sich mit Qualitätsstandards des bestehenden Analyserasters überschnitten; diese stellten somit eine Fokussierung auf ausgewählte international geltende und keine zusätzlichen Qualitätsstandards dar. Zu den organisationsspezifischen Qualitätsstandards, die nicht über die Standarddokumente abgedeckt waren, gehörten beispielsweise die „Berücksichtigung des*r Partners*in bei den Empfehlungen“ oder die „Durchführung einer Kontributionsanalyse“. Bei der Sichtung der Organisationsdokumente stellte sich heraus, dass organisationsspezifische Qualitätsstandards – sofern vorhanden – nur zum Teil explizit ausgewiesen waren und somit deren Vorliegen einem gewissen Interpretationsspielraum unterlag.[61]

Abbildung 6 Verpflichtung der Organisationen zu ausgewählten Qualitätsstandards



Quelle: DEval, eigene Darstellung

Anmerkung: Die dunkelblaue Umrandung weist darauf hin, dass bei sechs Organisationen (Durchführungsorganisationen und politische Stiftungen) sowohl eine Verpflichtung zu den OECD-DAC- als auch zu den DeGEval-Standards vorlag. Die Anwendung der OECD-DAC-Kriterien ist ein Qualitätsstandard der OECD-DAC-Standards.

60 Die Verpflichtungsgrundlage einer Organisation war auch ausschlaggebend für die Zuordnung der Organisation zu Gruppe 1 oder 2 bei der Darstellung der Ergebnisse (Tabelle 5, Abschnitt 3.2).

61 Für detailliertere Informationen zu den organisationsspezifischen Qualitätskriterien siehe im Onlineanhang, Abschnitt 3.3.

4.2 Bewertung der Anwendung der Qualitätskriterien

Im zweiten Abschnitt wird die Anwendung der Qualitätskriterien in vier Unterabschnitten beschrieben: 1) Überschneidung der OECD-DAC-, der DeGEval- sowie der OECD-DAC-only-Standards, 2) die OECD-DAC-Kriterien, 3) organisationsspezifische Qualitätskriterien und 4) Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit (inklusive Differenz der Anwendung über die Zeit). Im Kasten 3 wird ein Gesamtfazit gegeben.

Kasten 3 Gesamtfazit

Es zeichnete sich ein positives Bild dahingehend ab, dass die beteiligten Organisationen die OECD-DAC-, die DeGEval- sowie organisationsspezifische Qualitätsstandards und Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit durchschnittlich größtenteils anwandten. Dies zeigt, dass die Anwendung der Qualitätsstandards in der Evaluierungspraxis der Organisationen gelebt wurde. Die Ergebnisse der Organisationen waren dabei sehr heterogen. Darüber hinaus zeigte sich, dass die Qualitätsstandards in den Organisationsdokumenten überwiegend noch nicht vollumfänglich von den Organisationen identifiziert worden waren und ihre (Nicht-)Anwendung nicht systematisch verschriftlicht worden ist. Dies traf ebenfalls auf die Nachvollziehbarkeit der Anwendung und Nichtanwendung bei einigen ausgewählten Qualitätsstandards auf Ebene der einzelnen Evaluierung zu.

4.2.1 OECD-DAC- und DeGEval-Qualitätskriterien

Im ersten Unterabschnitt werden die Stärken und Schwächen bei der Anwendung der Überschneidung von OECD-DAC- und DeGEval-Standards sowie der OECD-DAC-only-Standards gezeigt. Im Kasten 4 werden das Fazit und die Hauptergebnisse dargestellt.

Kasten 4 Fazit zur Anwendung der OECD-DAC- und der DeGEval-Qualitätsstandards

Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards in den Evaluierungen der beteiligten deutschen EZ-Organisationen? (Evaluierungsfrage 2a)

Insgesamt zeichnete sich ein positives Bild in Bezug auf die Anwendung der OECD-DAC- und der DeGEval-Qualitätsstandards ab. Die beteiligten Organisationen der deutschen EZ wandten die Qualitätsstandards bei circa zwei Drittel ihrer Evaluierungen an. Dies traf auch – zu einem etwas niedrigeren Grad – auf Organisationen ohne Verpflichtungsgrundlage zu. Die Unterschiede bei der Anwendung der Qualitätsstandards zwischen den Organisationen wichen dabei zum Teil deutlich voneinander ab. Dies war aufgrund der gewählten Auswahlkriterien für die Aufnahme der beteiligten Organisationen in die Stichprobe zu erwarten und ermöglichte somit ein heterogenes Bild über die unterschiedlichen Anwendungsgrade hinweg.

Einschränkend ist anzumerken, dass die Anwendung einiger Qualitätsstandards aus verschiedenen Gründen nicht auf Evaluierungsebene, sondern auf Ebene der Organisation erfasst wurden. Dies könnte an der gewählten Operationalisierung einiger Qualitätskriterien in der Meta-Evaluierung liegen, durch eine fehlende Dokumentation der (Nicht-)Anwendung bedingt sein oder auch dadurch, dass die Dokumentation der Anwendung ausschließlich auf Organisationsebene und nicht auf Ebene der einzelnen Evaluierung erfolgte. Hier besteht deutlicher Verbesserungsbedarf, da eine externe Untersuchung ohne Informationen zu einer (begründeten) (Nicht-)Anwendung der Qualitätsstandards auf Ebene der Evaluierung nur eingeschränkt möglich war. Es war somit nicht nachvollziehbar, ob ein Qualitätsstandard (mit oder ohne Begründung) nicht angewandt oder angewandt, aber nicht dokumentiert wurde.

a) Stärken zeigten sich in der Anwendung der Qualitätsstandards.[62] (Ergebnis: 4)

- Die 37 Qualitätskriterien wurden im Durchschnitt zu 68 Prozent angewandt. Ungefähr drei Viertel der Qualitätskriterien (29 von 37) wurden durchschnittlich größtenteils/vollständig und ein Viertel teilweise oder kaum angewandt. Die drei am häufigsten angewandten Qualitätskriterien waren „Beschreibung des Evaluierungsgegenstands (1)“, „Evaluierungsethik (22)“ und die durchschnittliche „Anwendung der OECD-DAC-Kriterien (33–37)“. (Ergebnis: 4.1)
- Organisationen unterschieden sich zum Teil stark in der Anwendung der Qualitätskriterien. (Ergebnis: 4.2)
- Auch Organisationen ohne Verpflichtungsgrundlage zur Anwendung der Qualitätskriterien wandten diese zu durchschnittlich 61 Prozent in ihrer Evaluierungspraxis an. (Ergebnis: 4.3)

b) Schwächen zeigten sich in der Nachvollziehbarkeit der (Nicht-)Anwendung einiger Qualitätsstandards auf Evaluierungsebene und der Identifikation und systematischen Verschriftlichung relevanter Qualitätsstandards in Organisationsdokumenten. (Ergebnis: 5)

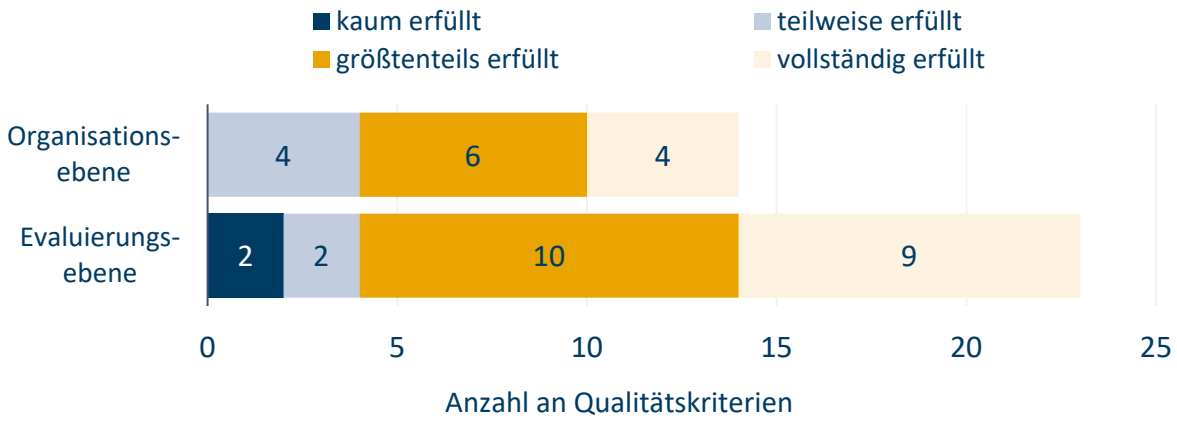
- Die für die Organisationen verpflichtenden Qualitätskriterien wurden zum Teil nicht identifiziert und in den Organisationsdokumenten nur teilweise systematisch verschriftlicht. (Ergebnis: 5.1)
- Circa ein Viertel der Qualitätskriterien wurde teilweise oder kaum angewandt. Die vier am wenigsten angewandten Qualitätskriterien waren „Einbezug von Gutachtenden aus Partnerland (30)“, „Berücksichtigung Kapazitätsentwicklung (26)“, „Darstellung Unvoreingenommenheit der Gutachtenden (24)“ und „Berücksichtigung partnerschaftlicher Ansätze (28)“. (Ergebnis: 5.2)
- Die Anwendung von circa 38 Prozent der Qualitätskriterien (14 von 37) wurde nicht auf Ebene der einzelnen Evaluierung, sondern über alle Evaluierungen hinweg für eine Organisation mittels einer Onlinebefragung erfasst. Ein Grund lag darin, dass eine Nichtanwendung der Qualitätskriterien fast nie in den einzelnen Evaluierungen verschriftlicht wurde. (Ergebnis: 5.3)
- Gründe für die Nichtanwendung von ausgewählten Qualitätskriterien wurden mit organisationalen oder evaluierungsspezifischen Aspekten erklärt oder mit Anwendungsformen, die nicht den Operationalisierungen der vorliegenden Meta-Evaluierung entsprachen. (Ergebnis: 5.4)

Übergeordnete Ergebnisse

Die 37 Qualitätskriterien wurden im Durchschnitt zu 68 Prozent angewandt und somit als „größtenteils erfüllt“ bewertet. Die Anwendung variierte deutlich zwischen den Organisationen. Auch Organisationen ohne Verpflichtungsgrundlage wandten die Qualitätskriterien an. Ungefähr drei Viertel der Qualitätskriterien wurden durchschnittlich größtenteils oder vollständig erfüllt (N = 29), ein Viertel teilweise oder kaum (N = 8; Abbildung 7). Die drei am höchsten erfüllten Qualitätskriterien waren die durchschnittliche „Anwendung der OECD-DAC-Kriterien (33–37)“, „Beschreibung des Evaluierungsgegenstands (1)“ und „Evaluierungsethik (22)“, die vier am niedrigsten „Einbezug von Gutachtenden aus Partnerland (30)“, „Berücksichtigung Kapazitätsentwicklung (26)“, „Darstellung Unvoreingenommenheit der Gutachtenden (24)“ und „Berücksichtigung partnerschaftlicher Ansätze (28)“. Darüber hinaus wichen bei rund 57 Prozent der Qualitätskriterien (N = 21) der minimale und der maximale Wert der Organisationen der Gruppe 1 (Organisationen mit Verpflichtungsgrundlage) um mehr als 50 Prozent voneinander ab. Die absoluten Werte der Gruppe 1 lagen bei sieben Qualitätskriterien um mindestens 20 Prozent höher, bei vier niedriger als bei Gruppe 2 (Organisationen ohne Verpflichtungsgrundlage).

62 In diesem Abschnitt werden die Ergebnisse größtenteils auf der Ebene der Qualitätskriterien dargestellt. Ausnahmen sind übergeordnete Erkenntnisse sowie die Darstellung der Ergebnisse von Qualitätsstandards, die aus zwei oder mehr Qualitätskriterien bestehen.

Abbildung 7 Anzahl der OECD-DAC- und DeGEval-Qualitätskriterien je Grad der Erfüllung



Quelle: DEval, eigene Darstellung

Anmerkung: N = 37 Qualitätskriterien. Organisationsebene = Datenerfassung mittels Onlinebefragung über alle Evaluierungen hinweg; Evaluierungsebene = Datenerfassung mittels Dokumentenanalyse für jede einzelne Evaluierung. Kein Qualitätskriterium wurde durchschnittlich mit 0 Prozent (verfehlt) oder 100 Prozent (übertroffen) bewertet.

Das durchschnittliche Ergebnis über die beteiligten Organisationen hinweg zeigte, dass sich die meisten vollständig und größtenteils erfüllten Qualitätskriterien in den Standardclustern „Nutzbarkeit" und „Berichtslegung und Methoden" befanden. Im Standardcluster „Partizipation, Unabhängigkeit und Fairness" schnitten vor allem Qualitätskriterien mit Bezug auf Partner*innen schlechter ab. Im Standardcluster „Partizipation, Unabhängigkeit und Fairness" schnitten die Qualitätskriterien „Berücksichtigung partnerschaftliche Ansätze (28)", „Berücksichtigung Kapazitätsentwicklung (26)" und „Einbezug von Gutachtenden aus Partnerland (30)" mit am niedrigsten ab, sodass sich insgesamt eine eher geringere Berücksichtigung der Partner*innen erkennen ließ. Organisationen aus Gruppe 2 wiesen bei diesen Qualitätskriterien durchschnittlich höhere Ergebnisse auf. Das Qualitätskriterium „Einbindung der internen und externen Stakeholder*innen (20)" wurde von Gruppe 1 durchschnittlich „größtenteils erfüllt" und stellte einen positiven Aspekt der Partizipation dar.

Die Anwendung der Qualitätskriterien wurde bei rund 38 Prozent der Qualitätskriterien (14 von 37) nicht auf Ebene der einzelnen Evaluierung, sondern über alle Evaluierungen der jeweiligen Organisationen hinweg erfasst. Die fehlende Erfassung auf Evaluierungsebene kann unter anderem daran liegen, dass 1) eine (Nicht-)Anwendung dieser Qualitätskriterien von den Organisationen nicht transparent dargelegt wurde, 2) die Anwendung in dem Evaluierungsteam nicht zugänglich gemachten Organisationsdokumenten verschriftlicht worden ist, 3) die Anwendung nicht auf Ebene der Evaluierung, sondern ausschließlich in den Organisationsdokumenten verschriftlicht wurde oder 4) es keine angemessene organisationübergreifende Operationalisierung gab. Die Anwendung aller Qualitätskriterien im Standardcluster „Berichtslegung und Methoden" und im Bereich „OECD-DAC-Kriterien" konnte auf Ebene der einzelnen Evaluierung ermittelt werden, die Anwendung der Qualitätskriterien in den Standardclustern „Partizipation, Unabhängigkeit und Fairness" und „Nutzbarkeit" wurde überwiegend auf Ebene der Organisation – also über alle Evaluierungen hinweg – erfasst.

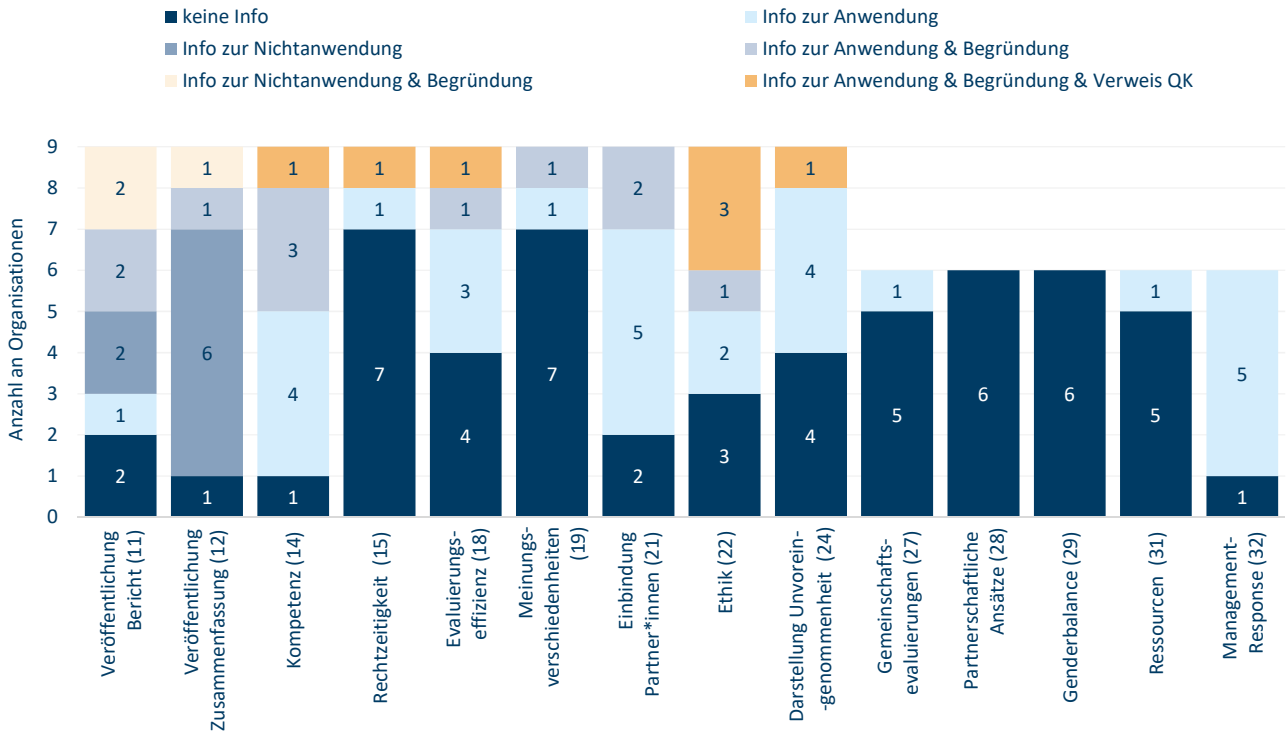
Eine (begründete) Nichtanwendung eines Qualitätskriteriums wurde auf Organisationsebene kaum dokumentiert – Beispiele sind die „Veröffentlichung des Evaluierungsberichts" beziehungsweise die „Veröffentlichung der Zusammenfassung". Auf Evaluierungsebene bestanden wenige begründete Nichtanwendungen bei der „Anwendung der OECD-DAC-Kriterien". Eine begründete Nichtanwendung wurde bei zwei der 14 untersuchten Qualitätskriterien verschriftlicht und somit insgesamt als „kaum erfüllt" bewertet[63]

63 Im DeGEval-Standarddokument (DeGEval, 2016, S. 29) heißt es dazu: „Die teilweise oder vollständige Nichterfüllung von Einzelstandards sollte immer offen und nachvollziehbar etwa im Rahmen der Berichterstattung dokumentiert und begründet werden. Damit wird nicht zuletzt eine Beurteilung der Evaluationsqualität möglich." Da ohne eine Verschriftlichung auch die Beurteilung der Evaluierungsqualität der OECD-DAC-Standards nur eingeschränkt möglich ist, wird eine Dokumentation auch für sie vorausgesetzt.

(N = 2 Organisationen, im Qualitätskriterium „Veröffentlichung des Evaluierungsberichts [11]“; N = 1 Organisation, im Qualitätskriterium „Veröffentlichung der Zusammenfassung [12]“). Bei diesen zwei Qualitätskriterien wurden darüber hinaus auch Nichtanwendungen beschrieben, ohne diese zu begründen (N = 2 Organisationen, im Qualitätskriterium „Veröffentlichung des Evaluierungsberichts [11]“; N = 6 Organisationen, im Qualitätskriterium „Veröffentlichung der Zusammenfassung [12]“; Abbildung 8).

In den untersuchten Organisationsdokumenten wurde kaum ein Bezug zu den einzelnen Qualitätsstandards hergestellt. Konkret wurden bei fünf von 14 Qualitätskriterien von mindestens einer Organisationen ein Verweis in den Organisationsdokumenten auf die Anwendung von Qualitätsstandards verschriftlicht (N = 1 Organisation, im Qualitätskriterium „Kompetenz Gutachtende [14]“; N = 1 Organisation, im Qualitätskriterium „Rechtzeitigkeit Erkenntnisse [15]“; N = 1 Organisation, im Qualitätskriterium „Evaluierungseffizienz [18]“; N = 3 Organisationen, im Qualitätskriterium „Evaluierungsethik [22]“; N = 1 Organisation, im Qualitätskriterium „Darstellung Unvoreingenommenheit“). Bei „Partnerschaftliche Ansätze [28]“ und „Genderbalance [29]“ lagen bei keiner Organisation Informationen in den Organisationsdokumenten vor (Abbildung 8).

Abbildung 8 Dokumentation der (Nicht-)Anwendung ausgewählter Qualitätskriterien in den Organisationsdokumenten der Gruppe 1



Quelle: DEval, eigene Darstellung

Anmerkung: QK = Qualitätskriterium. Die Ergebnisse der Gruppe 2 finden sich im Onlineanhang, Abschnitt 4.1.1. „Info zur Nichtanwendung & Begründung & Verweis QK“ wurde ebenfalls untersucht, konnte aber für kein Qualitätskriterium ermittelt werden.

Auf aktive Nachfrage in der Onlinebefragung wurden für untersuchten Qualitätskriterien vielfältige Gründe für eine (teilweise) Nichtanwendung zurückgemeldet. Bei einem Teil der Qualitätskriterien ähnelten sich die Begründungen der Organisationen, bei anderen unterschieden sie sich. Die zurückgemeldeten Motive für die Nichtanwendung lagen darüber hinaus auf unterschiedlichen Ebenen: 1) Organisationsebene (zum Beispiel lag die Verantwortung für die Anwendung außerhalb der Evaluierungseinheiten/-stellen oder das Qualitätskriterium hatte für die Organisation keine Relevanz), 2) Evaluierungsebene (zum Beispiel hatte

das Qualitätskriterium für die vorliegende Evaluierung keine Relevanz) und/oder 3) Anwendung der Qualitätskriterien anhand einer anderen Form als es in der Meta-Evaluierung untersucht wurde (zum Beispiel wurde die Qualitätssicherung nicht über einen Inception Report „Qualitätssicherung IR [8]“ gewährleistet). Details zu den Gründen für eine (teilweise) Nichtanwendung finden sich im Onlineanhang in Abschnitt 4.1.1.

Die nachfolgenden Abbildungen und Textabschnitte sind einheitlich aufgebaut.

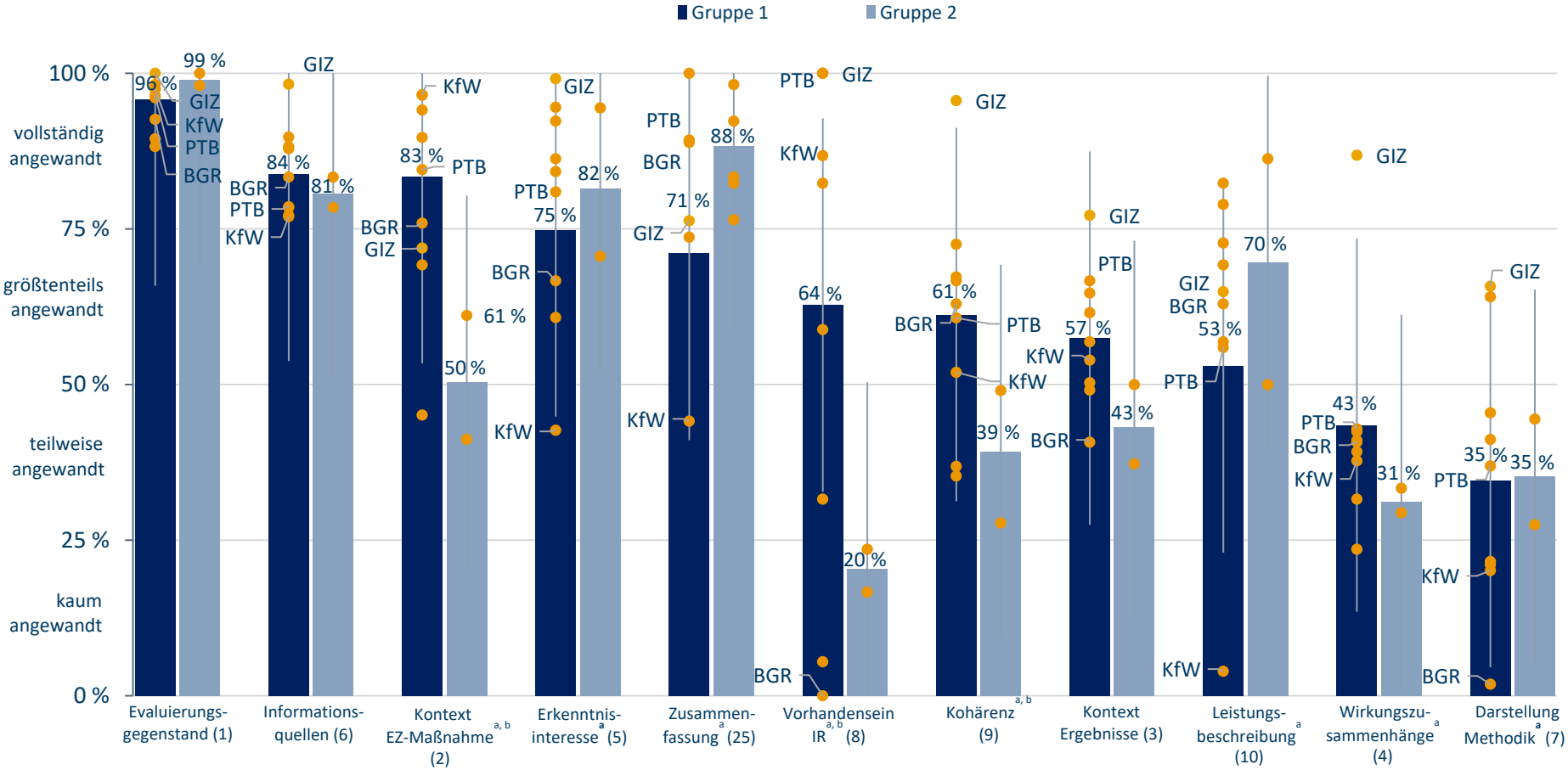
In den Abbildungen zu den durchschnittlichen Ergebnissen der beiden Gruppen und Organisationen (Abbildung 9, Abbildung 12, Abbildung 14 und Abbildung 16) sind die Qualitätskriterien eines Standardclusters von links nach rechts in absteigender Reihenfolge gemäß dem durchschnittlichen Gruppenergebnis der Gruppe 1 (mit Verpflichtungsgrundlage) angeordnet. Der dunkelblaue/-gelbe Balken in jedem Qualitätskriterium stellt das durchschnittliche Ergebnis als Prozentwert über die Organisationen der Gruppe 1 dar, der hellblaue/-gelbe für Gruppe 2 (ohne Verpflichtungsgrundlage). Die individuellen durchschnittlichen Organisationsergebnisse sind auf den Längsachsen der Gruppenbalken als gelbe (Abbildung 9 und Abbildung 16) oder blaue (Abbildung 12 und Abbildung 14) Punkte dargestellt. Der Prozentwert wird in die Intervalle der y-Achse eingeordnet und ermöglicht, die Bewertung abzulesen. Aufgrund des zu Beginn der Meta-Evaluierung nicht eindeutig geklärten Mandats des DEval bezüglich der Durchführung von Meta-Evaluierungen bei einigen nichtstaatlichen Organisationen wurden zur Wahrung der Anonymität nur die Ergebnisse der vier staatlichen Durchführungsorganisationen in den Abbildungen namentlich kenntlich gemacht. Die dünne schwarze Linie auf der Längsachse aller Balken dient der visuellen Unterstützung, damit eindeutig erkennbar ist, welchem Gruppenbalken die gelben Punkte zugeordnet sind. Bei den staatlichen Organisationen sind die durchschnittlichen Ergebnisse beschriftet, bei den nichtstaatlichen werden diese anonymisiert abgebildet.

In den Häufigkeitsgrafiken (Abbildung 10, Abbildung 13, Abbildung 15 und Abbildung 16) werden bei den Qualitätskriterien je Standardcluster die prozentualen Anteile der zwei/vier Bewertungsstufen aus den 296 Evaluierungen dargestellt. Somit lässt sich also beispielsweise ablesen, wie viele Evaluierungen in einer Gruppe die höchste Bewertungsstufe erreicht haben. Bei den Standardclustern „Partizipation, Unabhängigkeit und Fairness“ und „Nutzbarkeit“ unterscheidet sich die Reihenfolge zwischen den Abbildungen zu den durchschnittlichen Ergebnissen und den Häufigkeiten (also zwischen Abbildung 12, Abbildung 13, Abbildung 14 und Abbildung 15). Dies liegt daran, dass in diesen Standardclustern auch Qualitätskriterien aus der Onlinebefragung eingeflossen sind, deren Skalen von denen auf der Evaluierungsebene abweichen. Daher werden in den Häufigkeitsgrafiken Qualitätskriterien auf Organisations- und Evaluierungsebene getrennt dargestellt. Die Logik folgt jedoch auf beiden Ebenen erneut der absteigenden Anordnung gemäß dem Ergebnis von Gruppe 1.

In den Textabschnitten werden die Ergebnisse je Qualitätskriterium (Qualitätsstandard) beschrieben und es wird besonders hervorgehoben, wenn die Organisationen der Gruppe 1 innerhalb eines Qualitätskriteriums eine große Differenz zwischen dem minimalen und dem maximalen Organisationsergebnis aufweisen (> 50 Prozent) oder sich die durchschnittlichen Ergebnisse der beiden Gruppen in absoluten Werten mehr als 20 Prozent voneinander unterscheiden. Alle Qualitätskriterien sind im Abschnitt 4.2.1 und 4.2.2 einheitlich nummeriert. Eine Übersicht zur Nummerierung findet sich im Berichtsanhang in Abschnitt 7.1. Die im Text beschriebenen Ergebnisse beziehen sich in der Regel auf Gruppe 1, Ausnahmen werden explizit hervorgehoben. Wenn die „Anwendung“ eines Qualitätskriteriums beschrieben wird, bezieht sich die Aussage auf einen Befund, bei „Erfüllung“ auf eine Bewertung. Der Text, der den Abbildungen folgt, ist immer gemäß derselben Anordnung der Qualitätskriterien der Grafik gegliedert. Manche Textabschnitte können jedoch von der Reihenfolge der Abbildung abweichen, wenn es sich um mehrere Qualitätskriterien handelt, die demselben Qualitätsstandard zugeordnet worden sind. In diesen Fällen werden in einem Abschnitt direkt alle zugehörigen Qualitätskriterien behandelt, sobald das in der Abbildung erste Qualitätskriterium beschrieben wird. Bei einigen Qualitätskriterien werden darüber hinaus Good-Practice-Beispiele vorgestellt.

Standardcluster „Berichtslegung und Methoden“

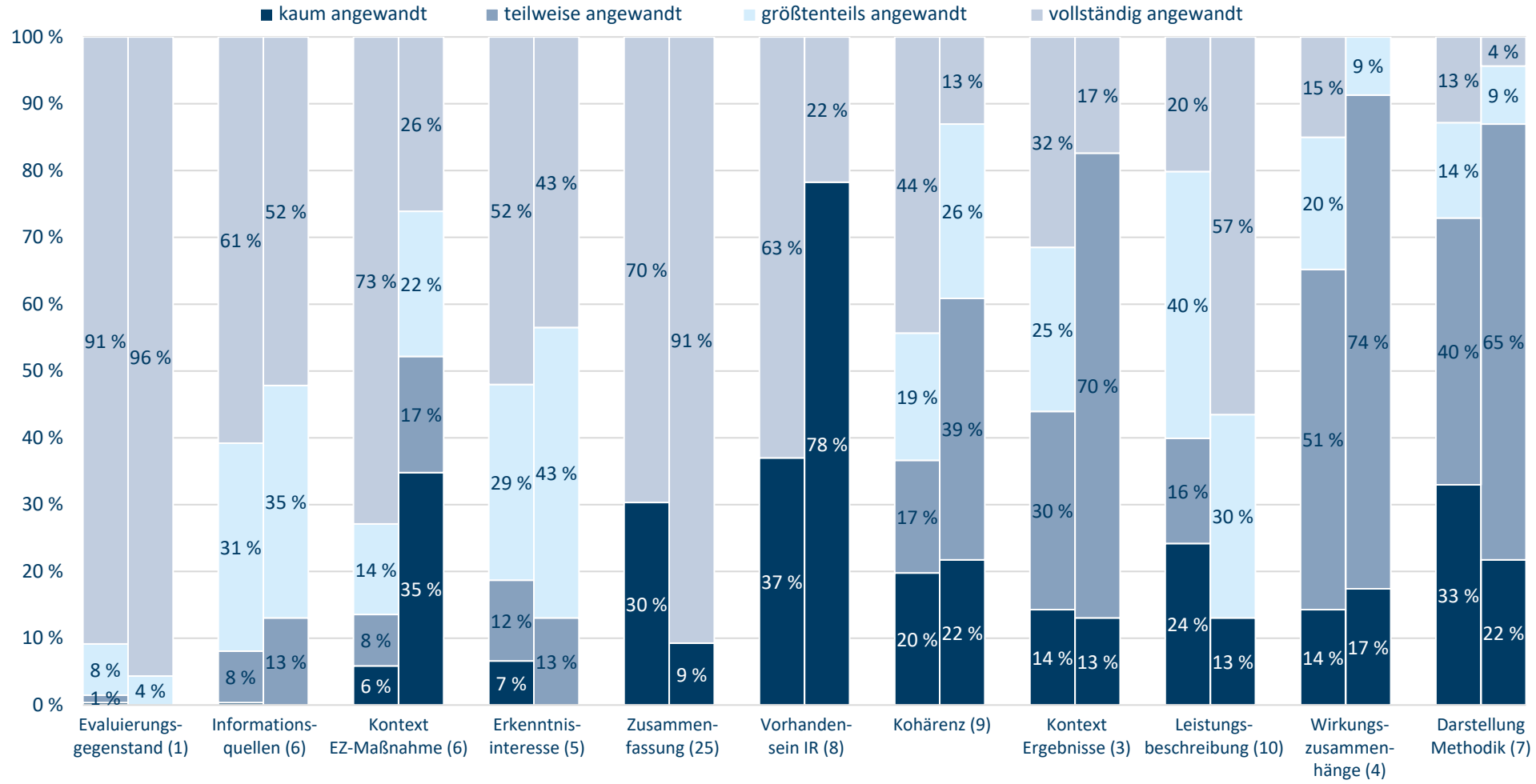
Abbildung 9 Anwendung der Qualitätskriterien im Standardcluster „Berichtslegung und Methoden“



Quelle: DEval, eigene Darstellung

Anmerkung: Einige Organisationen wandten als Qualitätssicherungsprozess andere Formen an als die „Qualitätssicherung mit IR (8)“ (beispielsweise „Kommentierung der Evaluierungsberichte“). Da diese Formen in der vorliegenden Meta-Evaluierung nicht untersucht wurden, werden diese Organisationen in ihrer Anwendung unterschätzt. [a] Qualitätskriterien zeigen in Gruppe 1 eine Differenz zwischen dem minimalen und dem maximalen Wert der Organisationen von > 50 Prozent; [b] Qualitätskriterien unterscheiden sich mit Gruppe 1 > Gruppe 2 (20 Prozent).

Abbildung 10 Häufigkeiten der Bewertungsstufen der Qualitätskriterien im Standardcluster „Berichtslegung und Methoden“



Quelle: DEval, eigene Darstellung

Anmerkung: Der jeweils linke Balken der zwei nah beieinander abgebildeten Balken = Gruppe 1; rechter Balken = Gruppe 2. Bei der Addition der Prozente je Qualitätskriterium kann aufgrund von Rundungen eine Abweichung von +/− 1 Prozent zu 100 Prozent auftreten. Die Nummerierung der Qualitätskriterien dient dem Abgleich mit dem Text. Eine Übersicht zur Nummerierung findet sich im Berichtsanhang, Abschnitt 7.1.

***Beschreibung des Evaluierungsgegenstands (1)*: Das Qualitätskriterium wurde durchschnittlich „vollständig erfüllt". Damit gewährleisteten alle Organisationen fast durchgehend in ihren Evaluierungsberichten, dass nachvollziehbar beschrieben wurde, welche EZ-Maßnahme evaluiert wurde.** Die Abweichung von einer Bewertung mit 100 Prozent bedeutet, dass in einigen wenigen Evaluierungen (9 Prozent) Ziele, Zielgruppen und/oder relevante Akteure für die evaluierte EZ-Maßnahme nicht beschrieben wurden (Abbildung 10). Insbesondere die Nennung der Zielgruppe oder der relevanten Akteure wurde in diesen Fällen ausgelassen. Die GIZ übertraf den Qualitätsstandard, das heißt, in jeder Evaluierung wurden alle relevanten Aspekte des Evaluierungsgegenstands beschrieben; ihre Evaluierungsberichte können diesbezüglich als Good Practice angesehen werden (ein Good-Practice-Beispiel zur ausführlichen Beschreibung der Ziele, Zielgruppen und der relevanten Akteure einer Evaluierung findet sich im Onlineanhang, Abschnitt 4.1.1).

***Nachvollziehbarkeit der Informationsquellen (6)*: Durchschnittlich wurde das Qualitätskriterium „vollständig erfüllt". In ungefähr 92 Prozent der Evaluierungsberichte wurden die Informationsquellen mindestens größtenteils dargestellt.** Dies bedeutet, dass in den Evaluierungsdokumenten nachvollziehbar beschrieben wurde, welche spezifischen Dokumente oder Befragungen für welche Analyse als Informationsquellen herangezogen wurden. Bei den restlichen 8 Prozent wurden die Informationsquellen rudimentärer dargestellt (zum Beispiel wurde ausschließlich aufgeführt, dass Interviews geführt und Sekundärdaten verwendet wurden). Innerhalb einer Evaluierung wurden die verschiedenen Informationsquellen unterschiedlich detailliert dargestellt. In der Analyse möglicher Faktoren auf die Anwendung dieses Qualitätskriteriums zeigte sich, dass unter anderem die Hinzunahme unterschiedlicher Qualitätssicherungsprozesse (zum Beispiel „Qualitätssicherung mit IR [8]" und „Einbindung Stakeholder*innen [20]") mit einer besseren Nachvollziehbarkeit der Informationsquellen einherging.

***Einbindung des Kontexts (2 + 3)*: Die zwei Qualitätskriterien des Qualitätsstandards wurden durchschnittlich „vollständig erfüllt" und „größtenteils erfüllt". Insgesamt wurde die „Beschreibung des Kontexts der Entwicklungsmaßnahme (2)" häufiger erläutert, aber kaum systematisch bei den ermittelten Ergebnissen berücksichtigt.** Das Qualitätskriterium „Beschreibung des Kontexts der EZ-Maßnahme" untersuchte, wie viele Kontextelemente (zum Beispiel politische, wirtschaftliche) und wie umfänglich diese im Evaluierungsbericht und in den Evaluierungsanhängen beschrieben wurden. Die Historie der Organisation und ihre Tätigkeiten im Partnerland wurden dabei nicht als Kontextelemente berücksichtigt. Besonders die beteiligten politischen Stiftungen, die sich per se im politischen Kontext bewegen, wiesen in vielen ihrer Evaluierungsberichte eine detaillierte und umfangreiche kontextuelle Verortung der EZ-Maßnahme auf. Das Qualitätskriterium „Berücksichtigung des Kontexts bei den Ergebnissen (3)" stellte dar, wie systematisch Kontexte bei den Ergebnissen (Effektivität) bedacht wurden. Nur zwei Organisationen berücksichtigten den Kontext im Sinne einer Darstellung der hemmenden und fördernden Faktoren auf die Ergebnisse. Für beide Qualitätskriterien konnten unter anderem bei PTB, KfW und GIZ Good-Practice-Beispiele identifiziert werden. In diesen zeigte sich, dass das Qualitätskriterium besonders gut erfüllt wurde, wenn im Rahmen einer standardisierten Struktur des Evaluierungsberichts ein Abschnitt für die Beschreibung des Kontexts bei der Einführung als auch bei der Beschreibung der Ergebnisse vorgesehen war (Onlineanhang, Abschnitt 4.1.1).

***Beschreibung des Erkenntnisinteresses (5):* Das Qualitätskriterium wurde durchschnittlich „größtenteils erfüllt".**[64] **Konkret wurden in circa 52 Prozent der Evaluierungsberichte/-anhänge Zweck, Ziel und Evaluierungsfragen der Evaluierung klar dargestellt.** Der häufigste Aspekt, der in Evaluierungsberichten/-anhängen nicht deutlich erläutert wurde, ist der übergeordnete Zweck der Evaluierung (zum Beispiel die Überprüfung, ob eine EZ-Maßnahme verlängert werden soll). Evaluierungsfragen wurden in fast allen und die Ziele einer Evaluierung (zum Beispiel die Bewertung der Effektivität einer EZ-Maßnahme) häufig in Evaluierungsberichten beschrieben. Das Qualitätskriterium wurde unter anderem bei MISEREOR erfüllt, da durch die Verwendung einer Mustervorlage für Leistungsbeschreibungen (Terms of References, ToR) alle drei Aspekte des Erkenntnisinteresses standardisiert formuliert werden mussten (Onlineanhang, Abschnitt 4.1.1).

64 Die „Beschreibung Erkenntnisinteresse (5)" stellt das einzige Qualitätskriterium dar, in dem die Einstufung der Anwendung entlang der Berechnung der Mediane höher ausfällt (das heißt vollständig statt größtenteils) als bei der Berechnung der Mittelwerte.

Informationsgehalt der Zusammenfassung (25): **Das Qualitätskriterium wurde durchschnittlich „größtenteils erfüllt". In rund 70 Prozent der Zusammenfassungen wurden die Ergebnisse und Schlussfolgerungen oder Empfehlungen der Evaluierung beschrieben.** Der Informationsgehalt einer Zusammenfassung wurde als „vollständig erfüllt" bewertet, wenn die Ergebnisse und entweder Schlussfolgerungen oder Empfehlungen der Evaluierung genannt wurden. Der überwiegende Teil der Organisationen erfüllte das Qualitätskriterium vollständig, eine Organisation größtenteils und eine teilweise. Da eine Evaluierung nicht notwendigerweise Empfehlungen aussprechen muss, wurden gemäß den Standarddokumenten auch die Schlussfolgerungen als Grundlage der Bewertung herangezogen. In den Regressionsanalysen zur Erklärung der Anwendung des Qualitätskriteriums wurde ermittelt, dass der Informationsgehalt der Zusammenfassung stieg, je mehr Gutachtende an der Evaluierung beteiligt waren.

Qualitätssicherung mit Inception Report (8)**: Das Qualitätskriterium wurde durchschnittlich „größtenteils erfüllt". Das Qualitätskriterium fokussiert auf das Vorhandensein eines Inception Reports und somit auf eine Form der Qualitätssicherung.** Auf Basis einer Fokusgruppendiskussion mit den beteiligten Organisationen und bestehender Literatur (Queiroz de Souza, 2017) wurde das Anfertigen eines Inception Reports als wichtiges Merkmal für Qualitätssicherungsprozesse identifiziert. Ein Inception Report ist ein effektives Instrument, um ein gemeinsames Verständnis zum Vorgehen und zur Durchführung der Evaluierung zu schaffen und erste kritische Aspekte zu diskutieren und gegebenenfalls anzupassen. Bei rund 63 Prozent der Evaluierungen lag ein Inception Report vor. Bei PTB und GIZ wurden für alle Evaluierungen Inception Reports angefertigt, bei anderen Organisationen für keine Evaluierung, sodass eine große Streuung in der Anwendung des Qualitätskriteriums besteht. Durch die Onlinebefragung zeigte sich, dass neben dem Inception Report alternative Anwendungsformen eines Qualitätssicherungsprozesses von Organisationen umgesetzt wurden (unter anderem das Verwenden von Vorlagen der Leistungsbeschreibung oder annotierten Gliederungen für den Berichtsentwurf; siehe Onlineanhang, Abschnitt 4.1.1). Entsprechend kann davon ausgegangen werden, dass die tatsächliche Umsetzung der Qualitätssicherung von Evaluierungen nicht vollständig durch die Operationalisierung dieses Qualitätskriteriums abgebildet wurde.

Kohärenz von Daten-Ergebnissen-Schlussfolgerungen (9): **Das Qualitätskriterium wurde durchschnittlich „größtenteils erfüllt". In ungefähr 44 Prozent aller Evaluierungsberichte baute die Mehrheit der Schlussfolgerungen kohärent auf Datenanalyse und Ergebnissen auf.** In rund 37 Prozent der Evaluierungsberichte war bei der Mehrheit der Schlussfolgerungen kaum oder nur teilweise nachvollziehbar, auf welchen Ergebnissen und/oder auf welchen Datenanalysen die Befunde gründeten. Die GIZ wandte das Qualitätskriterium als einzige Organisation vollständig an. Als Good-Practice-Beispiel aus den Evaluierungen der PTB hat sich herausgestellt, dass das Qualitätskriterium gut erfüllt wurde, wenn bei der Verschriftlichung der Schlussfolgerungen durch Referenzen in Klammern (zum Beispiel durch Seitenzahlen und/oder Fußnoten) direkt Bezug auf die Ergebnisse genommen wurde. Eine weitere Möglichkeit stellt das Einrahmen der Schlussfolgerungen im gleichen Absatz dar (Good-Practice-Beispiele im Onlineanhang, Abschnitt 4.1.1).

Informationsgehalt der Leistungsbeschreibung (10): **Das Qualitätskriterium wurde durchschnittlich „größtenteils erfüllt". Bei circa 60 Prozent der Leistungsbeschreibungen wurden mindestens vier von acht Aspekten spezifiziert (Zweck, Nutzende, Ziele, Methoden, Zeitrahmen, verfügbare Mittel, Veröffentlichungsrechte und mitwirkende Personen der Evaluierung).** Dabei wurden die Aspekte Veröffentlichungsrechte sowie der Nennung der Nutzenden und der mitwirkenden Personen am wenigsten beschrieben. Bei einer Organisation wurde das Qualitätskriterium als „kaum erfüllt" bewertet. Ein Grund war, dass eine Leistungsbeschreibung bei dieser Organisation nicht notwendigerweise Bestandteil einer Evaluierung war.

Darstellung der Wirkungszusammenhänge: **Das Qualitätskriterium wurde durchschnittlich teilweise erfüllt. Obwohl fast alle Evaluierungen mit Elementen einer „Wirkungslogik" (*Input – Output – Outcome – Impact*) arbeiteten, wurden diese häufig lückenhaft oder nicht für alle Ziele einer EZ-Maßnahme abgebildet.** Eine vollständige Darstellung der Wirkungslogik erleichtert das Verständnis der Funktionsweise der EZ-Maßnahme und kann als Grundlage für wirkungsorientierte Evaluierungen genutzt werden (UNDAF, 2017). In 14 Prozent der Evaluierungen wurden keine, in 71 Prozent unvollständige Wirkungslogiken formuliert (in 51 Prozent wurden unvollständige Wirkungslogiken und in 20 Prozent mindestens für ein Ziel der EZ-Maßnahme, aber nicht für alle Ziele vollständige Wirkungslogiken dargestellt). In 15 Prozent wurden vollständige

Wirkungslogiken für alle Ziele der EZ-Maßnahme konzipiert. Diese stammten überwiegend von Evaluierungen der GIZ (Ausreißer nach oben). Eine niedrige Anwendung wurde von manchen Organisationen durch fehlende verfügbare Ressourcen oder eine fehlende Verpflichtung durch den Auftraggeber erklärt. Möglich ist auch eine Darstellung der Wirkungslogik in Dokumenten, die für die Meta-Evaluierung nicht zur Verfügung gestellt wurden. Als Good-Practice-Beispiele eignen sich Evaluierungen der GIZ, in denen in einem gesonderten Abschnitt die Wirkungslogik anhand einer Grafik und einer zugehörigen textlichen Erläuterung der Wirkzusammenhänge dargestellt wurde (Good-Practice-Beispiele im Onlineanhang, Abschnitt 4.1.1).

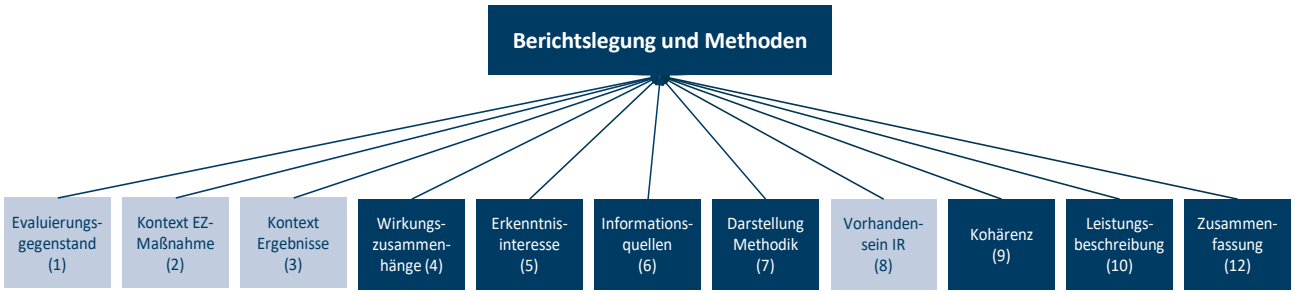
***Darstellung der Angemessenheit des methodischen Vorgehens (7)*: Das Qualitätskriterium wurde durchschnittlich „teilweise erfüllt". In der Mehrheit der Evaluierungsberichte wurden Limitationen des Vorgehens beschrieben, selten wurde begründet, warum das methodische Vorgehen gewählt worden war, und fast nie fand eine Diskussion zu alternativen Vorgehensweisen statt.** Das Qualitätskriterium untersuchte, ob beziehungsweise inwieweit nachvollziehbar begründet wurde, warum die angewandten Methoden angemessen waren und ob beziehungsweise inwieweit Limitationen des methodischen Vorgehens diskutiert wurden. Keine Organisation wandte das Qualitätskriterium vollständig an und vier begründeten ihr Vorgehen durchschnittlich kaum. Die Begründungen stellen kritische Informationen für die Nutzenden der Evaluierung dar, um die Reliabilität und Validität der Ergebnisse einschätzen zu können. Insgesamt wurde in 13 Prozent der Evaluierungsberichte umfassend begründet, warum die gewählten Methoden für die Evaluierung angemessen waren und welche Limitationen diese produzierten.[65] In 14 Prozent wurden beide Aspekte mit wenigen Sätzen, aber nicht umfassend beschrieben, und in 40 Prozent entweder Begründungen oder Limitationen aufgeführt. Bei 33 Prozent der Evaluierungen wurden weder Begründungen noch Limitationen diskutiert. Zusammengenommen erfüllten damit mehr als 73 Prozent der Evaluierungen das Qualitätskriterium kaum oder teilweise. Zwischen den Organisationsergebnissen zeigen sich große Unterschiede. Good-Practice-Beispiele konnten in den Evaluierungen des DVV identifiziert werden, in denen ausführliche Abschnitte ausschließlich den Vor- und Nachteilen sowie den Limitationen der verwendeten Methoden gewidmet wurden (Good-Practice-Beispiele im Onlineanhang, Abschnitt 4.1.1.).

Abschließend zeigten die Ergebnisse einer empirischen Untersuchung, dass die meisten Qualitätskriterien des Standardclusters „Berichtslegung und Methoden" zusammen entlang eines Faktors (dem Standardcluster) abgebildet werden konnten. Die Qualitätskriterien wurden inhaltlich Standardclustern zugeordnet. Ob diese Standardcluster nicht nur theoretisch, sondern auch empirisch gemeinsame Aspekte abbilden, konnte mittels einer explorativen Faktorenanalyse untersucht werden. Die Analyse hat gezeigt, dass sieben Qualitätskriterien empirisch durch das Standardcluster „Berichtslegung und Methoden" abgebildet werden konnten: „Darstellung der Wirkungszusammenhänge (4)", „Beschreibung des Erkenntnisinteresses (5)", „Nachvollziehbarkeit der Informationsquellen (6)", „Darstellung der Angemessenheit des methodischen Vorgehens (7)", „Kohärenz von Daten-Ergebnissen-Schlussfolgerungen (9)", „Informationsgehalt der Leistungsbeschreibung (10)" und „Informationsgehalt der Zusammenfassung (25)". Vier Qualitätskriterien konnten empirisch nicht abgebildet werden („Beschreibung des Evaluierungsgegenstands [1]", „Beschreibung des Kontexts der Entwicklungsmaßnahme [2]", „Berücksichtigung des Kontexts bei Ergebnissen [3]" und „Qualitätssicherung mit Inception Report [8]"; Abbildung 11).[66]

65 Mithilfe von Text Mining wurde zudem untersucht, ob in den untersuchten Evaluierungen Leitfadeninterviews oder Fokusgruppendiskussionen als Datenerhebungsmethoden verwendet wurden. Die Ergebnisse der Analyse zeigen, dass in 98 Prozent der Evaluierungen (N = 291) Leitfadeninterviews und in 47 Prozent der Evaluierungen (N = 138) Fokusgruppendiskussionen zum Einsatz kamen. Für mehr Informationen zur Methode Text Mining und den Ergebnissen siehe Onlineanhang, Abschnitt 4.1.2.

66 Die Qualitätsstandards „Beschreibung Evaluierungsgegenstand (1)" und „Einbindung Kontext" werden gegebenenfalls nicht durch das Standardcluster abgebildet, da sie inhaltlich Informationen zur EZ-Maßnahme wiedergeben und nicht zur Evaluierung. Sie bilden einen eigenen Faktor ab. Für detailliertere Informationen zur Faktorenanalyse siehe Onlineanhang, Abschnitt 3.4.

Abbildung 11 Explorative Faktorenanalyse des Standardclusters „Berichtslegung und Methoden“

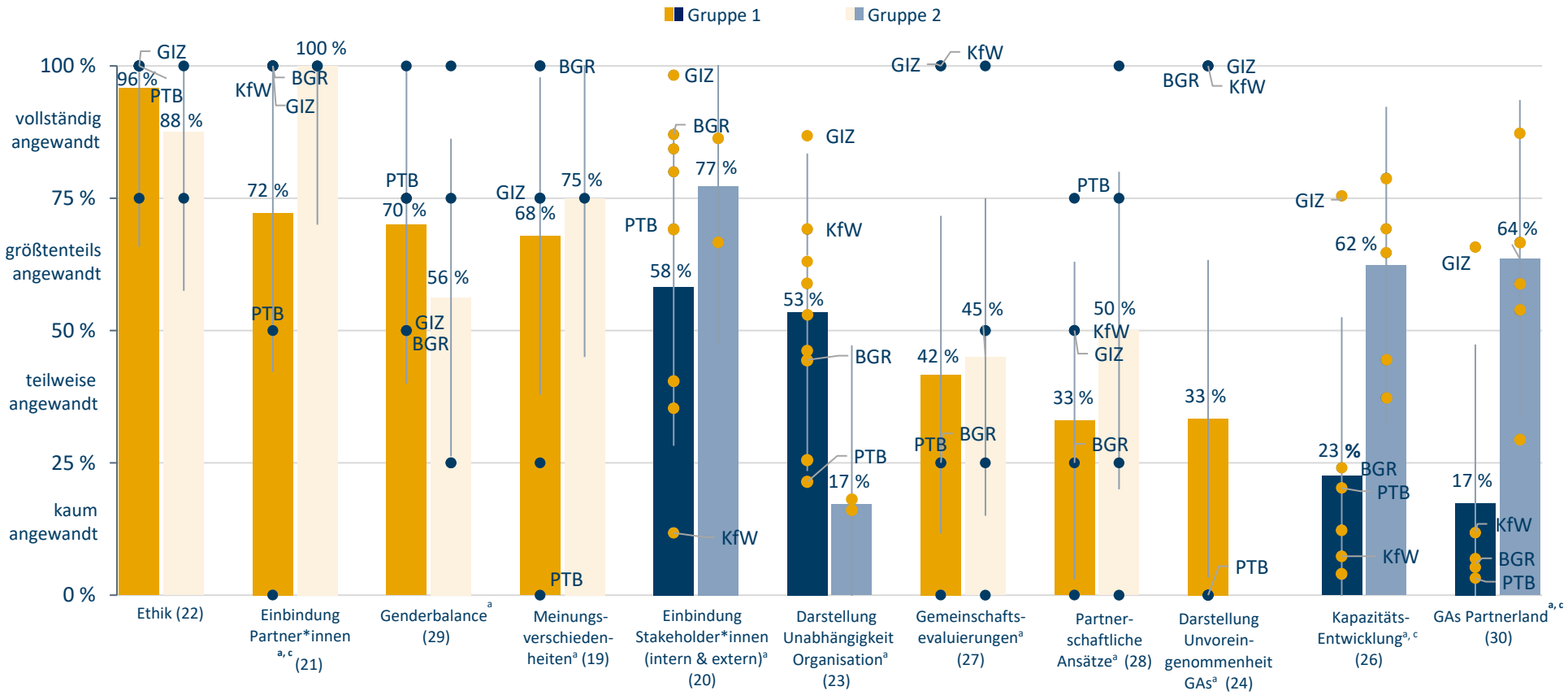


Quelle: DEval, eigene Darstellung

Anmerkung: Die Linien stellen Faktorladungen dar und geben an, dass überprüft wurde, ob das Qualitätskriterium mit dem Standardcluster zusammenhängt. Die grau unterlegten Qualitätskriterien sind keine Bestandteile des durch die explorative Faktorenanalyse identifizierten Konstrukts/Standardclusters.

Standardcluster „Partizipation, Unabhängigkeit und Fairness“

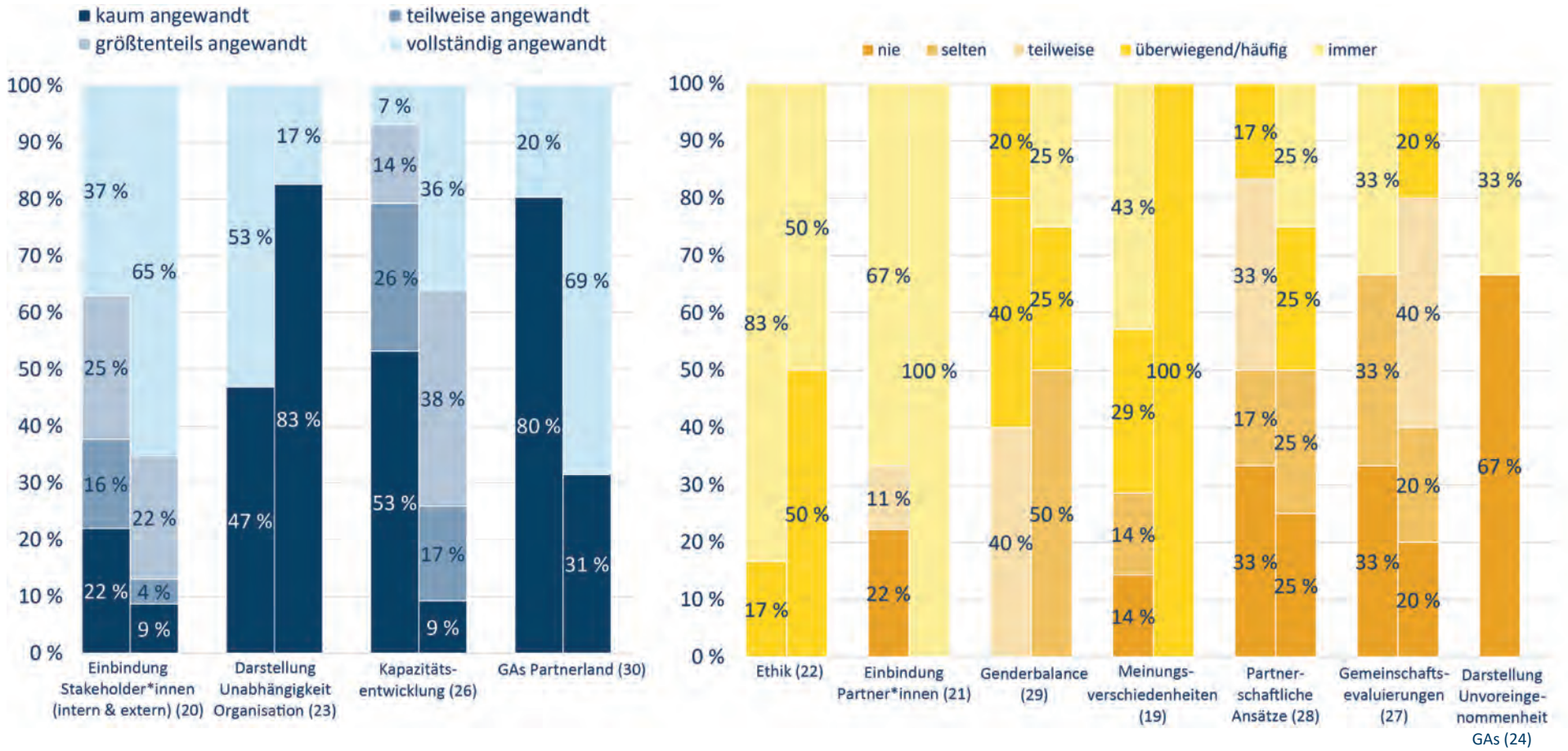
Abbildung 12 Anwendung der Qualitätskriterien im Standardcluster „Partizipation, Unabhängigkeit und Fairness“



Quelle: DEval, eigene Darstellung

Anmerkung: GAs= Gutachtende; blau = Qualitätskriterien (QKs), die auf Ebene einzelner Evaluierungen untersucht wurden; gelb = QKs, die auf Organisationsebene untersucht wurden. Beim Qualitätskriterium „Darstellung Unvoreingenommenheit GAs (24)“ machten Organisationen der Gruppe 2 in der Onlinebefragung keine Angaben, daher wurde der helle Balken ausgelassen. Einige Organisationen wandten dieses Qualitätskriterium in anderen Formen an (beispielsweise im Rahmen von Auswahlprozessen und nicht über eine unterschriebene Erklärung zur Unabhängigkeit). Da diese Formen in der vorliegenden Meta-Evaluierung nicht untersucht wurden, werden diese Organisationen in ihrer Anwendung unterschätzt. [a] Qualitätskriterien zeigen eine Differenz zwischen dem minimalen und dem maximalen Wert der Organisationen von > 50 Prozent; [b] Qualitätskriterien unterscheiden sich mit Gruppe 1 > Gruppe 2 (20 Prozent); [c] Qualitätskriterien unterscheiden sich mit Gruppe 1 < Gruppe 2 (20 Prozent). Eine Übersicht zur Nummerierung findet sich im Berichtsanhang im Abschnitt 7.1.

Abbildung 13 Häufigkeiten der Bewertungsstufen der Qualitätskriterien im Standardcluster „Partizipation, Unabhängigkeit und Fairness“



Quelle: DEval, eigene Darstellung

Anmerkung: der jeweils linke Balken der zwei nah beieinander abgebildeten Balken = Gruppe 1; rechter Balken = Gruppe 2; blau = Qualitätskriterien (QKs), die auf Ebene einzelner Evaluierungen untersucht wurden; gelb = QKs, die auf Organisationsebene erfasst wurden. Bei der Addition der Prozente je Qualitätskriterium kann aufgrund von Rundungen eine Abweichung von +/− 1 Prozent zu 100 Prozent auftreten. Für das Qualitätskriterium „Darstellung Unvoreingenommenheit GAs (24)“ machten Organisationen der Gruppe 2 keine Angaben, weshalb nur der Balken der Gruppe 1 abgebildet wird. Die Anordnung der Qualitätskriterien kann von der Anordnung in der vorhergehenden Abbildung abweichen (siehe auch Erklärung der Abbildungen in Unterabschnitt 4.2.1). Die Nummerierung der Qualitätskriterien dient der besseren Orientierung, eine Übersicht findet sich im Berichtsanhang im Abschnitt 7.1.

***Evaluierungsethik (22)*: Das Qualitätskriterium wurde durchschnittlich „vollständig erfüllt", da alle Organisationen online zurückmeldeten, dass sie Vorgaben zur Sicherheit und den Rechten der Evaluierungsbeteiligten immer oder zumindest häufig in den Evaluierungen machten.** Eine Eingrenzung des Qualitätskriteriums wurde nicht vorgenommen, sodass die Anwendung unterschiedliche Inhalte (beispielsweise Datenschutz, Menschenrechte oder Schutz von Personen oder Personengruppen, die in fragilen Kontexten befragt werden) und Umsetzungsvarianten einschließen konnte (zum Beispiel anhand eines Verhaltenskodex, der gesondert von Gutachtenden unterschrieben wurde oder durch Schulungen zu ethischem Handeln in Evaluierungen). Bei sechs von neun Organisationen wurden in Organisationsdokumenten (zum Beispiel Evaluierungsleitfäden oder Verträgen zwischen Gutachtenden und Organisationen) Vorgaben zur Anwendung des Qualitätskriteriums verschriftlicht, bei drei inklusive eines Bezugs zu den Qualitätsstandards.

***Einbindung der internen und externen Stakeholder*innen (20) und der Partner*innen (21):* Beide Qualitätskriterien wurden durchschnittlich „größtenteils erfüllt". Es zeigte sich, dass Stakeholder*innen und Partner*innen durchschnittlich mehr von den beteiligten Organisationen ohne Verpflichtungsgrundlage eingebunden wurden.** Das Qualitätskriterium beschreibt die Einbindung von mindestens einem*r Stakeholder*in in unterschiedliche Evaluierungsphasen. Als interne Stakeholder*innen werden beispielsweise die Auftraggebenden und als externe die Partnerorganisation im Partnerland angesehen.[67] Gruppe 2 erfüllte das Qualitätskriterium durchschnittlich vollständig und schnitt somit besser ab als Gruppe 1. Darüber hinaus zeigte sich, dass die Organisationen das Qualitätskriterium „Einbindung der Partner*innen (21)" in der Onlinebefragung als durchschnittlich „größtenteils angewandt" angaben.[68] Gründe für eine Nichteinbindung lagen zum Beispiel darin, dass Partner*innen aufgrund eines zu hohen Koordinationsaufwands oder aus politischen Sensibilitäten nicht miteinbezogen werden konnten. Letzteres trifft unter anderem auf die beteiligten politischen Stiftungen zu. Entsprechend wurde in einer Vereinbarung zwischen dem BMZ und den politischen Stiftungen festgehalten, dass Evaluierungsprinzipien wie die Partizipation von Partner*innen nur eingeschränkt berücksichtigt werden können (BMZ, 2016).[69]

Die Messung des Qualitätskriteriums hat allerdings Grenzen. Es ist schwer überprüfbar, in welcher Form (zum Beispiel freie Äußerung) und Intensität die Einbindung von Stakeholder*innen im Gesamtprozess oder in einzelnen Evaluierungsphasen sichergestellt wurde. Eine noch detailliertere Operationalisierung und eine eingehendere Analyse dieses Qualitätskriteriums wären notwendig, um tiefergehende Informationen zu erhalten. Kritisch anzumerken ist darüber hinaus, dass die Einbindung von nur einem*r Stakeholder*in in unterschiedliche Evaluierungsphasen auch problematisch sein kann (zum Beispiel, wenn diese*r eine einseitige Perspektive auf die Ergebnisse abbildet). Als Good-Practice-Beispiel hat sich in Evaluierungen der GIZ herausgestellt, dass das Qualitätskriterium vollständig erfüllt wurde, wenn sich im Anhang des Evaluierungsberichts ein Zeitplan befand, in dem für die Konzeptions-, Durchführungs- und Berichtslegungsphase in einer zusätzlichen Spalte erkenntlich wurde, ob und inwiefern Stakeholder*innen in die Evaluierung involviert waren. In der Konzeptionsphase wurde dies häufig erreicht, indem die Stakeholder*innen an der Entwicklung der Evaluierungsfragen beteiligt wurden oder in gemeinsamen Workshops erarbeitet wurde, mit welchen Methoden welche Informationen erhoben werden sollten. In der Berichtslegungsphase hatten Stakeholder*innen indes die Gelegenheit, den Berichtsentwurf zu kommentieren

67 Stakeholder*innen der Evaluierung werden dabei als Personen, Personengruppen oder Organisationen betrachtet, die etwas zu verlieren oder zu gewinnen haben. Es können Personen sein, die für die Planung und Umsetzung der Evaluierung zuständig sind: Auftraggebende; Geldgebende; Personen, die für das evaluierte Projekt zuständig sind; Personen, die die Evaluierung nutzen wollen/könnten (angelehnt an Beywl und Niestroj, 2009).

68 Da das Qualitätskriterium nicht zwischen internen und externen Stakeholder*innen unterscheidet und vorwiegend interne Stakeholder*innen in den Evaluierungsdokumenten kodiert wurden, die Formulierung des Qualitätsstandards aber Interpretationsspielraum lässt und auch ausschließlich externe Stakeholder*innen gemeint sein könnten, wurden zusätzliche Informationen zur „Einbindung der Partner*innen" in der Onlinebefragung erhoben. Beide Qualitätskriterien werden daher dargestellt und bewertet.

69 Da der Wortlaut in der Vereinbarung nicht eindeutig ist, wurden die politischen Stiftungen entsprechend ihrer Anwendung des Qualitätskriteriums abgebildet und die Dokumentation nicht als begründete Nichtanwendung gewertet.

oder die Ergebnisse und Empfehlungen zu besprechen, bevor sie veröffentlicht wurden (Good-Practice-Beispiele im Onlineanhang, Abschnitt 4.1.1).

Zusammensetzung der Gutachtenden (29 + 30): **Bei staatlichen Organisationen und den beteiligten politischen Stiftungen kam es durchschnittlich kaum zum „Einbezug von Gutachtenden aus [dem] Partnerland (30)", während nichtstaatliche Organisationen der Gruppe 2 diese durchschnittlich größtenteils in die Gutachtenden-Teams einbezogen. Eine ausgewogene „Genderbalance im Team (29)" wurde bei beiden Gruppen größtenteils berücksichtigt, in Organisationsdokumenten aber nicht formal festgehalten.**[70] Insgesamt wurden in rund 20 Prozent der Evaluierungen der Gruppe 1 Gutachtende aus Partnerländern einbezogen, bei Gruppe 2 in rund 70 Prozent. Die GIZ wandte das Qualitätskriterium größtenteils an und stellt damit bei den Organisationen mit Verpflichtungsgrundlage eine positive Ausnahme dar. Eine Nichteinbindung von Gutachtenden aus dem Partnerland wurde häufig damit begründet, dass nur eine Person und kein Gutachtenden-Team eingestellt wurde. Die PTB setzte darüber hinaus eine andere Form der Anwendung des Qualitätskriteriums um. Sie nahm Gutachtende aus Nachbarländern des Partnerlandes unter Vertrag, um eine potenzielle Befangenheit aufgrund persönlicher Beziehungen zu den Partnerorganisationen im nationalen Kontext zu vermeiden. In der Messung des Qualitätskriteriums bestanden Herausforderungen unter anderem darin, dass Gutachtende aus Partnerländern in den Evaluierungen von EZ-Maßnahmen in mehreren Ländern einbezogen werden können sowie die Zugehörigkeit der Gutachtenden zum Partnerland schwer identifizierbar war. Daher ist das Erfüllen dieses Qualitätskriteriums bei Evaluierungen von EZ-Maßnahmen die mehrere Länder betreffen schwer zu erreichen und das Ergebnis des Qualitätskriteriums konservativ einzuschätzen. In Bezug auf die Berücksichtigung einer „Genderbalance im Team (29)" gaben einige Organisationen an, nicht für die einzelne Evaluierung, sondern über Evaluierungen hinweg das Geschlechterverhältnis zu beachten. In den Organisationsdokumenten fanden sich bei zwei Organisationen Hinweise auf die Berücksichtigung eines ausgewogenen Männer-Frauen-Verhältnisses.

Transparenz von Meinungsverschiedenheiten (19)**: Das Qualitätskriterium wurde durchschnittlich „größtenteils" erfüllt. Die transparente Darstellung von Meinungsverschiedenheiten zwischen Mitgliedern des Gutachtenden-Teams im Evaluierungsbericht erschien nicht allen Organisationen erstrebenswert.** Einige Organisationen gaben in der Onlinebefragung als Grund für eine Nichtanwendung an, dass Konsens unter den Gutachtenden für sie eher das Ziel sei als die Darstellung von Differenzen. Insgesamt war das Qualitätskriterium schwer auf Ebene der einzelnen Evaluierung erfassbar, da zumeist unklar war, ob Meinungsunterschiede zwischen den Gutachtenden vorlagen.

Unabhängigkeit der Gutachtenden (23 + 24): **Das Qualitätskriterium berücksichtigte die Aspekte „Darstellung organisationale Unabhängigkeit der Gutachtenden (23)" (durchschnittlich „größtenteils erfüllt") und „Darstellung Unvoreingenommenheit der Gutachtenden (24)" (durchschnittlich „teilweise erfüllt").** In 53 Prozent der Evaluierungen waren die Gutachtenden weder politisch noch operativ noch beratend an der EZ-Maßnahme beteiligt und gehörten nicht zu ihrer Zielgruppe, in 47 Prozent waren sie organisational abhängig. Davon ließen jedoch 40 Prozent der Evaluierungen keine Rückschlüsse auf die Unabhängigkeit zu und wurden als nicht unabhängig eingestuft– da eine Unabhängigkeit nicht dokumentiert war, aber hätte beschrieben werden können. Zur persönlichen „Darstellung [der] Unvoreingenommenheit der Gutachtenden (24)" gegenüber dem Evaluierungsgegenstand gaben drei Organisationen (33 Prozent) an, dass sie dies durch eine unterschriebene „Erklärung zur Unabhängigkeit" formal bestätigen ließen. Die verbleibenden Organisationen erklärten, andere Anwendungsformen der Gewährleistung der Unabhängigkeit vorgenommen zu haben (beispielsweise dies in Vertragsanhängen erläutert zu haben). Insgesamt bestanden auch bei diesem Qualitätskriterium Grenzen in der Messung. Beispielsweise wäre es nur schwer feststellbar gewesen, ob Personen vor

[70] Die beiden Qualitätskriterien wurden aus dem OECD-DAC-Standard 3.1 „Evaluierungsteam" abgeleitet: „Bei der Zusammenstellung des Teams wird auf einen ausgewogenen Anteil von Männern und Frauen geachtet, und unter seinen Mitgliedern befinden sich Fachleute aus den betroffenen Partnerländern beziehungsweise -regionen".

ihrer Gutachtendentätigkeit für die evaluierte EZ-Maßnahme gearbeitet haben und sie somit gegenüber deren Arbeit voreingenommen hätten sein können. Da beide Qualitätskriterien eher konservativ[71] geschätzt wurden, wurde das Ergebnis des übergeordneten Qualitätsstandards ebenfalls konservativ interpretiert.

Berücksichtigung von Gemeinschaftsevaluierungen (27): **Die Prüfung der Möglichkeit einer Gemeinschaftsevaluierung mit Gebern und/oder der Partnerregierung in einem Partnerland wurde von den Organisationen „teilweise erfüllt".** Die Organisationen gaben als Gründe für eine Nichtanwendung unter anderem an, dass Gemeinschaftsevaluierungen aufgrund von komplexen Kooperationsverhältnissen mit unterschiedlichen Schwerpunkten nicht möglich gewesen seien. Der Aufwand für eine Gemeinschaftsevaluierung hätte auch nicht im Verhältnis zur Größe der einzelnen EZ-Maßnahmen und/oder der Evaluierung gestanden. Da nichtstaatliche Organisationen in ihren EZ-Maßnahmen selten mit Partnerregierungen kooperieren, wurde das Qualitätskriterium für sie tendenziell unterschätzt.

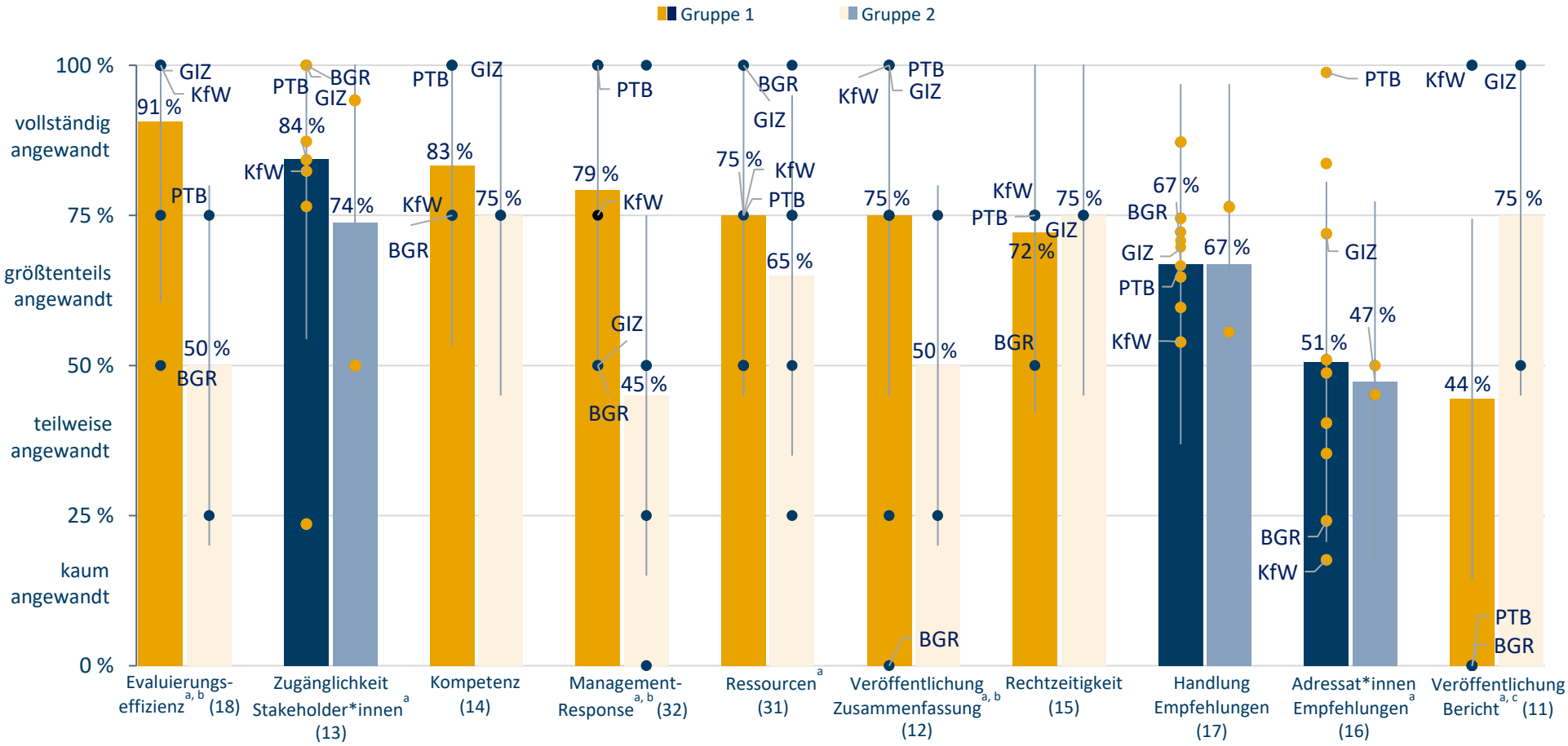
Berücksichtigung partnerschaftlicher Ansätze (28)**: Das Qualitätskriterium wurde durchschnittlich „teilweise erfüllt". Der überwiegende Teil der Organisationen gab an, dass „lokale oder nationale Evaluierungsaktivitäten und -pläne" nicht systematisch in den Evaluierungen berücksichtigt worden seien.** Als partnerschaftliche Ansätze werden dabei zum Beispiel nationale Evaluierungsvorgaben oder -standards verstanden. Die Ergebnisse weisen eine starke Streuung zwischen den Organisationen auf. In der Onlinebefragung gaben die Organisationen als Gründe für eine Nichtanwendung an, dass eine Berücksichtigung oft nicht relevant sei oder die eigenen EZ-Maßnahmen häufig zu spezifisch seien, um einen Bezug herzustellen. Auch in den Organisationsdokumenten fanden sich nahezu keine Hinweise auf die Berücksichtigung von partnerschaftlichen Ansätzen. Ebenso wie die Qualitätskriterien „Einbindung der Partner*innen (21)", „Berücksichtigung Kapazitätsentwicklung (26)" und „Einbezug von Gutachtenden aus Partnerland (30)" wandte Gruppe 2 den Qualitätsstandard in absoluten Werten häufiger an.

Berücksichtigung Kapazitätsentwicklung (26)**: Das Qualitätskriterium wurde bei den staatlichen Organisationen und den beteiligten politischen Stiftungen durchschnittlich „kaum erfüllt"; es stellt somit eines der am wenigsten angewandten Qualitätskriterien dar. Die verbleibenden nichtstaatlichen Organisationen wandten es im Gegensatz durchschnittlich größtenteils an.** Bei ungefähr 53 Prozent der Evaluierungen wurde Kapazitätsentwicklung nicht erfüllt. Dies zeigt, dass in Gruppe 1 kein Fokus auf diesen Qualitätsstandard gelegt wurde. In die gleiche Richtung geht auch eine Empfehlung des OECD DAC in seinem Peer-Review (2021, S. 7), in dem geraten wird, dass „Deutschland […] den Aufbau von Evaluierungskapazitäten in den Partnerländern weiter stärken und mehr investieren [sollte], um aus Evaluierungen von Sonderinitiativen und seiner Portfolios auf Länder-, Regional- und Programmebene zu lernen". Organisationen ohne Verpflichtungsgrundlage integrierten lokale/nationale Gutachtende und Partnerorganisationen (beispielsweise durch die gemeinsame Erarbeitung von Evaluierungsfragen) aktiv in die Konzeption und Durchführung der Evaluierung und diskutierten die Evaluierungsergebnisse häufiger mit ihnen als Organisationen mit Verpflichtungsgrundlage, mit Ausnahme der GIZ. Da nichtstaatliche Organisationen in ihren EZ-Maßnahmen selten mit Partnerregierungen kooperierten, wurde einer von vier Aspekten der Kapazitätsentwicklung von ihnen überwiegend nicht erfüllt und das Qualitätskriterium für sie tendenziell unterschätzt. Ebenso wie bei der Unabhängigkeit der Gutachtenden konnten bei vielen Evaluierungen keine Rückschlüsse auf die Anwendung des Qualitätskriteriums gezogen werden, sodass die Bewertung eher konservativ zu interpretieren ist. Als Good-Practice-Beispiel kann erneut der Zeitplan in Evaluierungen der GIZ angeführt werden, in dem erkenntlich wird, dass Partner*innen und relevante Ministerien in verschiedenen Phasen in Aktivitäten eingebunden wurden, die die Kapazitätsentwicklung förderten (Good-Practice-Beispiele im Onlineanhang, Abschnitt 4.1.1).

71 Konservativ bedeutet, dass das Ergebnis den unteren Wert darstellt und der wahre Wert wahrscheinlich höher liegt.

Standardcluster „Nutzbarkeit“

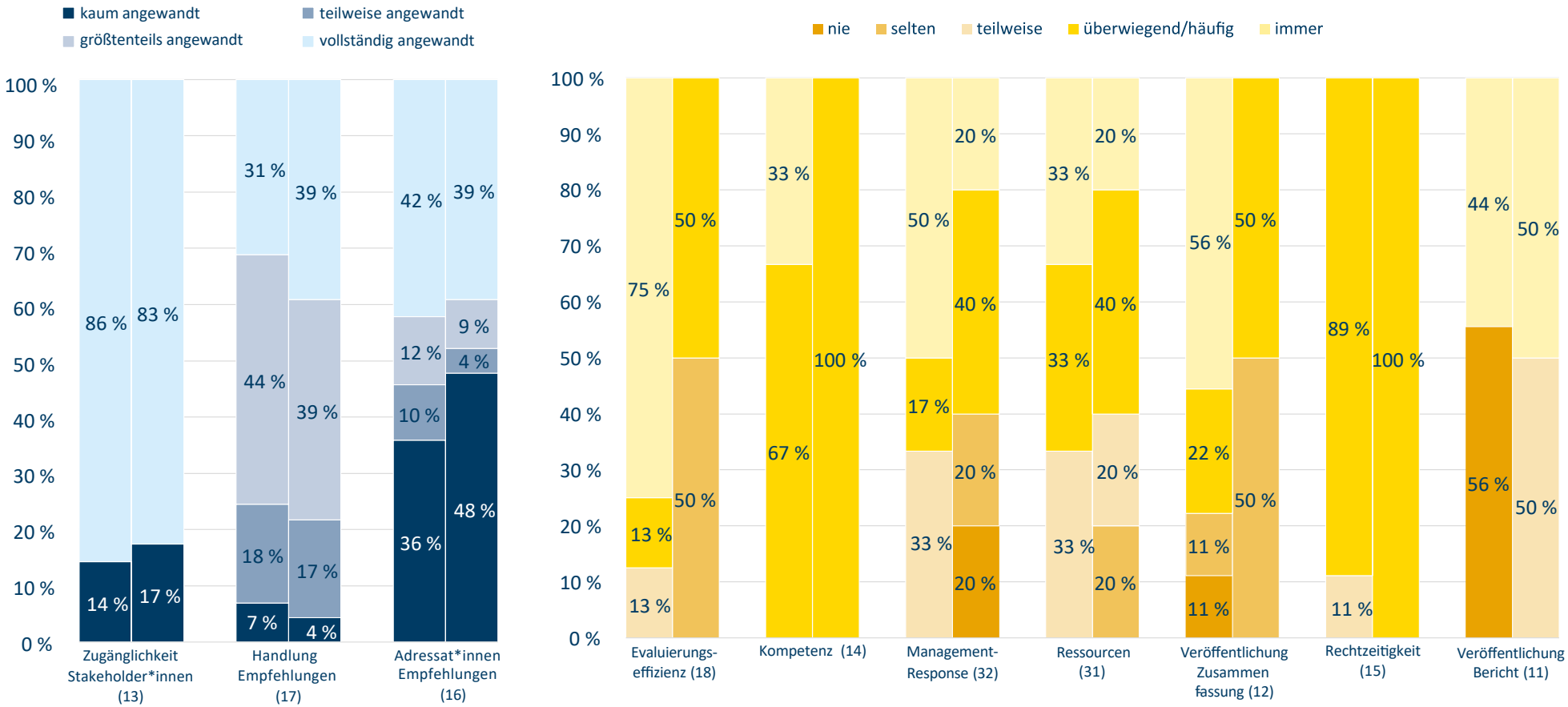
Abbildung 14 Anwendung der Qualitätskriterien Standardcluster „Nutzbarkeit“



Quelle: DEval, eigene Darstellung

Anmerkung: blau = Qualitätsstandards (QSs), die auf Evaluierungsebene untersucht wurden; gelb = QSs, die auf Organisationsebene untersucht wurden. [a] Qualitätskriterien zeigen eine Differenz zwischen dem minimalen und dem maximalen Wert der Organisationen von > 50 Prozent; [b] Qualitätskriterien unterscheiden sich mit Gruppe 1 > Gruppe 2 (20 Prozent); [c] Qualitätskriterien unterscheiden sich mit Gruppe 1 < Gruppe 2 (20 Prozent).

Abbildung 15 Häufigkeiten der Bewertungsstufen der Qualitätskriterien Standardcluster „Nutzbarkeit“



Quelle: DEval, eigene Darstellung

Anmerkung: der jeweils linke Balken der zwei nah beieinander abgebildeten Balken = Gruppe 1; rechter Balken = Gruppe 2; blau = Qualitätskriterien (QKs), die auf Ebene einzelner Evaluierungen untersucht wurden; gelb = QKs, die auf Organisationsebene erfasst wurden. Bei der Addition der Prozente je Qualitätskriterium kann aufgrund von Rundungen eine Abweichung von +/− 1 Prozent zu 100 Prozent auftreten. Die Anordnung der Qualitätskriterien kann von der Anordnung in der vorhergehenden Abbildung abweichen (siehe auch Erklärung der Ergebnisgrafiken in Unterabschnitt 4.2.1). Die Nummerierung der Qualitätskriterien dient der besseren Orientierung, eine Übersicht zur Nummerierung findet sich im Berichtsanhang, Abschnitt 7.1.

Evaluierungseffizienz (18): **Das Qualitätskriterium wurde durchschnittlich „vollständig erfüllt". Fast alle Organisationen gaben an, den Aufwand der Evaluierungen – immer oder häufig – im Verhältnis zu ihrem Nutzen reflektiert zu haben.** Bei GIZ, KfW und PTB wurde die Evaluierungseffizienz berücksichtigt, indem durch eine repräsentative Stichprobenziehung nicht alle EZ-Maßnahmen evaluiert wurden (bei der PTB seit 2020, zuvor Vollerhebung). Darüber hinaus wurde erwähnt, dass bei geringem erwartetem Nutzen eine Schreibtischstudie mit Telefoninterviews anstatt von Vor-Ort-Missionen durchgeführt würden. Andere Organisationen erklärten, dass Evaluierungen für das Projektmanagement und die Konzeption weiterer Projekte immer nützlich seien und damit immer im Verhältnis zu ihrem Aufwand stünden. Letzteres kann darauf hindeuten, dass die Evaluierungseffizienz vorwiegend ausgehend vom Nutzen reflektiert wird und weniger von den Kosten oder ihrem Verhältnis zueinander. Obwohl das Qualitätskriterium die Reflexion des Aufwand-Nutzen-Verhältnisses vor (ob eine Evaluierung durchgeführt werden soll) und während der Durchführung (wie stehen Entscheidungen in der Durchführung zum erwarteten Nutzen) umschließen könnte, bezogen sich die Rückmeldungen überwiegend auf die Erwägung vor Beginn der Evaluierung. Dies zeigt auf, dass zwischen den Organisationen kein einheitliches Verständnis von Evaluierungseffizienz besteht und gegebenenfalls selten eine Nachjustierung im Verlauf der Evaluierung stattfand.

Zugänglichkeit (11 + 12 + 13)**: Die Zugänglichkeit zu den Erkenntnissen der Evaluierung wurde für drei Anwendungsformen untersucht: 1) Zugänglichkeit der Evaluierungsergebnisse für Stakeholder*innen (zum Beispiel in Form einer gemeinsamen Diskussion), 2) Zugänglichkeit der Zusammenfassung für die Öffentlichkeit und 3) Zugänglichkeit des Evaluierungsberichts für die Öffentlichkeit. Der Qualitätsstandard insgesamt wurde durchschnittlich „größtenteils erfüllt" wobei dies in Einzelfällen durch eine begründete Nichtanwendung bedingt war. Auch zeigten sich deutliche Unterschiede in den drei Anwendungsformen.** Die Ergebnisse wurden den Stakeholder*innen durchschnittlich vollständig zugänglich gemacht, die Evaluierungsberichte wurden durchschnittlich teilweise und die Zusammenfassungen größtenteils veröffentlicht. Durch eine begründete Nichtanwendung in den Organisationsdokumenten von drei Organisationen wurde die Bewertung des Qualitätskriteriums „Veröffentlichung des Evaluierungsberichts (11)" als durchschnittlich „vollständig erfüllt" bewertet.[72] Dies gilt ebenso für eine Organisation im Qualitätskriterium „Veröffentlichung Zusammenfassung (12)". Zusammenfassend wurde in rund 86 Prozent der Evaluierungen die Evaluierung den Stakeholder*innen in Form des finalen Evaluierungsberichts, einer schriftlichen Zusammenfassung oder einer abschließenden Präsentation zugänglich gemacht. Die meisten Organisationen gaben an, zum Schutz der Evaluierungsbeteiligten und Partner*innen auf eine Veröffentlichung der vollständigen Evaluierungsberichte zu verzichten. Mittlerweile wird dies von den BMZ-Leitlinien Evaluierung gefordert.[73] Eine Organisation sah bei sensiblen Evaluierungen von einer Publizierung ab. Eine andere Organisation erstellte zwei separate Berichtsteile und veröffentlichte nur den einen Teil ohne sensible Informationen. Den politischen Stiftungen wurde darüber hinaus in einer BMZ-Vereinbarung ein zurückhaltender Umgang mit dem OECD-DAC-Evaluierungsprinzip Transparenz eingeräumt (BMZ, 2016), weil beispielsweise die politischen Verhältnisse mancher Länder eine besondere Schutzbedürftigkeit von Partnerstrukturen verlangen. Eine Zusammenfassung wurde von den Organisationen öfter über eine Webseite veröffentlicht als der gesamte Evaluierungsbericht. Bei der Analyse der Organisationsdokumente zeigte sich, dass Zugänglichkeit der einzige Qualitätsstandard war, bei dem eine Nichtanwendung begründet beschrieben wurde.

Kompetenz der Gutachtenden (14): **Bei allen Organisationen verfügten die Gutachtenden-Teams durchschnittlich vollständig über Evaluierungs-, Fach-/Sektor- und Regional-/Länderexpertise.** Dabei zeichnet sich ab, dass Organisationen unterschiedliche Kompetenzen bei der Auswahl des Gutachtenden-Teams priorisierten. So bevorzugten sie bei Bedarf das Fach-/Sektor- oder Evaluierungs- gegenüber dem Länderwissen und berücksichtigten zum Teil nicht die Länder-, sondern regionale Fachkenntnisse. Es wurde auch zurückge-

72 Eine begründete Nichtanwendung ist qualitativ nicht als gleichwertig mit einer tatsächlichen Anwendung anzusehen. Sie sollte nachvollziehbar und angemessen auf den jeweiligen Organisationskontext zugeschnitten sein.

73 In den verabschiedeten BMZ-Leitlinien Evaluierung (BMZ, 2021, S. 21) wird das Qualitätskriterium „Veröffentlichung Evaluierungsbericht (11)" wie folgt hervorgehoben: „Die Berichte werden im Sinne der Transparenz vorzugsweise vollständig veröffentlicht".

meldet, dass versucht wurde, Schwächen in der Evaluierungsexpertise durch eine enge Betreuung der Evaluierungsverantwortlichen auszugleichen. Kritisch angemerkt werden sollte, dass die Kompetenz von den Verantwortlichen der Evaluierungseinheiten für alle Evaluierungen der jeweiligen Organisation eingeschätzt wurde und dies nicht notwendigerweise die tatsächliche Kompetenz der Gutachtenden für jede einzelne Evaluierung widerspiegelt. Zudem ist die Untersuchung dieses Qualitätskriteriums über die zur Verfügungstellung von objektiven Daten schwer möglich, da es sich hierbei um vertrauliche Informationen der Gutachtenden handelt (zum Beispiel einen Lebenslauf).

Vorhandensein einer Management-Response (32): **Das Qualitätskriterium wurde von Gruppe 1 durchschnittlich „vollständig erfüllt" und von Gruppe 2 „teilweise erfüllt". Nicht jede Organisation verfügte über einen etablierten Prozess zum Verfassen einer Management-Response oder zur Umsetzungsplanung der Empfehlungen.** Da eine Management-Response im Anschluss an den Evaluierungsbericht verfasst wird, konnte das Vorhandensein nur über zusätzliche Dokumente beziehungsweise die Onlinebefragung überprüft werden. Drei der sechs Organisationen der Gruppe 1 erstellten immer eine Management-Response, eine häufig und zwei teilweise. Manche Organisationen beschrieben, dass sie andere Formen der Stellungnahme anwandten, zum Beispiel wurden Umsetzungshinweise direkt in die Folgekonzeption des evaluierten Projekts integriert oder mit den Partner*innen zum Abschluss der Evaluierung besprochen. Bei fünf der sechs Organisationen war das Qualitätskriterium in den Organisationsdokumenten verschriftlicht, allerdings ohne Bezug zu den Standarddokumenten (Abbildung 8). Die neu in Kraft getretenen BMZ-Leitlinien Evaluierung verlangen für die Empfehlungen zumindest eine Nachverfolgung der Umsetzung (BMZ, 2021).

Ausreichende Ressourcen vorhanden (31): **„Finanzielle, zeitliche und personelle Ressourcen" genügten laut Evaluierungsverantwortlichen durchschnittlich größtenteils, um die Ziele der Evaluierungen zu erreichen.** Die Durchführungsorganisationen in Gruppe 1 werden fast ausschließlich durch Steuergelder und nicht durch weitere Mittel finanziert. Diese Organisationen gaben an, dass die Ressourcen immer oder häufig ausgereicht hätten, um die Ziele zu realisieren. Gruppe 2 mit ausschließlich nichtstaatlichen Organisationen weist eine Streuung zwischen den Organisationsergebnissen auf und gab durchschnittlich an, dass ihre Ressourcen größtenteils ausgereicht hätten. Zum Teil wurde in Gruppe 2 die Finanzierung von Evaluierungen durch Spendenmittel ergänzt oder die Leistungskomponente der Gutachtenden reduziert, wenn interne Budgetverschiebungen nicht möglich waren.

Rechtzeitigkeit der Erkenntnisse (15): **Durchschnittlich gaben die Organisationen an, dass ihre Evaluierungen größtenteils rechtzeitig zum vereinbarten Zeitpunkt abgeschlossen worden seien.** Wenn dies nicht der Fall war, begründeten die Organisationen dies entweder mit externen (zum Beispiel Sicherheitslage im Partnerland oder Covid-19-Pandemie) und/oder internen Faktoren (zum Beispiel Abstimmungsschleifen zwischen den Gutachtenden oder Verfügbarkeit der Partner*innen). Die Rechtzeitigkeit des Abschlusses der Evaluierungen wurde lediglich in Gruppe 1 in zwei von neun Organisationdokumenten thematisiert (Abbildung 8).

Nützlichkeit der Empfehlungen (16 + 17)**: Der aus den beiden Qualitätskriterien 1) „Adressat*innen der Empfehlungen (16)" und 2) „Handlungsorientierung der Empfehlungen (17)" bestehende Qualitätsstandard wurde durchschnittlich „größtenteils erfüllt".** Rund 36 Prozent der Evaluierungen benannten in weniger als ein Viertel ihrer Empfehlungen Adressat*innen (beispielweise die Partner*innen der EZ-Maßnahme oder die Evaluierungseinheit der Organisationen selbst). Bei 31 Prozent war die Umsetzung bei mehr als der Hälfte der Empfehlungen nachvollziehbar. Sie galten beispielsweise als handlungsorientiert, wenn sie für jeweils konkrete Umsetzungshinweise beschrieben waren und somit klar erkennbar war, wann und wie eine Empfehlung umgesetzt werden sollte. Im Unterschied zum Qualitätskriterium „Handlungsorientierung der Empfehlungen (17)" wies „Adressat*innen der Empfehlungen (16)" eine große Streuung zwischen den Organisationsergebnissen auf. In KfW-Evaluierungen wurden auch Schlussfolgerungen anstatt Empfehlungen bewertet.

4.2.2 OECD-DAC-Kriterien

Im zweiten Unterabschnitt werden die Ergebnisse der Anwendung der fünf OECD-DAC-Kriterien beschrieben. Da alle elf Organisationen den OECD-DAC-Kriterien verpflichtet waren, wird keine Gruppe 2 abgebildet. Im Kasten 5 wird ein Überblick über das Fazit des Abschnitts geboten.

Kasten 5 Fazit zur Anwendung der OECD-DAC-Kriterien

Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards in den Evaluierungen der beteiligten deutschen EZ-Organisationen? (Evaluierungsfrage 2a)

Die OECD-DAC-Kriterien wurden zu einem sehr großen Teil von den beteiligten Organisationen in der deutschen EZ erfüllt. Allerdings ist explizit darauf hinzuweisen, dass die Operationalisierung entlang der Untersuchung einzelner Prüffragen aus den BMZ-Orientierungslinien (2006) und nicht der umfassenden Inhalte der OECD-DAC-Kriterien erfolgte und eine Erfüllung somit leicht möglich war. Bei der Anwendung der OECD-DAC-Kriterien bestanden darüber hinaus erste Dokumentationen einer Nichtanwendung auf Organisations- und Evaluierungsebene. Hierin unterschied sich die Anwendung der OECD-DAC-Kriterien bereits – wenn auch nur in kleinerem Umfang – von der Anwendung der meisten anderen Qualitätskriterien. Es ist anzunehmen, dass in zukünftigen Evaluierungen die Dokumentation der Nichtanwendung der OECD-DAC-Kriterien weiter ansteigen wird, da ab 2020 (BMZ, 2020) die aktualisierte BMZ-Orientierungslinie zu den Evaluierungskriterien eine begründete und transparente Schwerpunktsetzung vorsieht.

a) Stärken zeigten sich in der Anwendung der OECD-DAC-Kriterien. (Ergebnis: 6)

- Die OECD-DAC-Kriterien wurden durchschnittlich zu über 95 Prozent erfüllt. (Ergebnis: 6.1)
- Es gab erste Anfänge in der Dokumentation der begründeten Nichtanwendung; erwähnenswert ist hierbei vor allem die BMZ-Förderrichtlinie für die politischen Stiftungen, in der deren Besonderheiten im Umgang mit den OECD-DAC-Kriterien Berücksichtigung finden. (Ergebnis: 6.2)

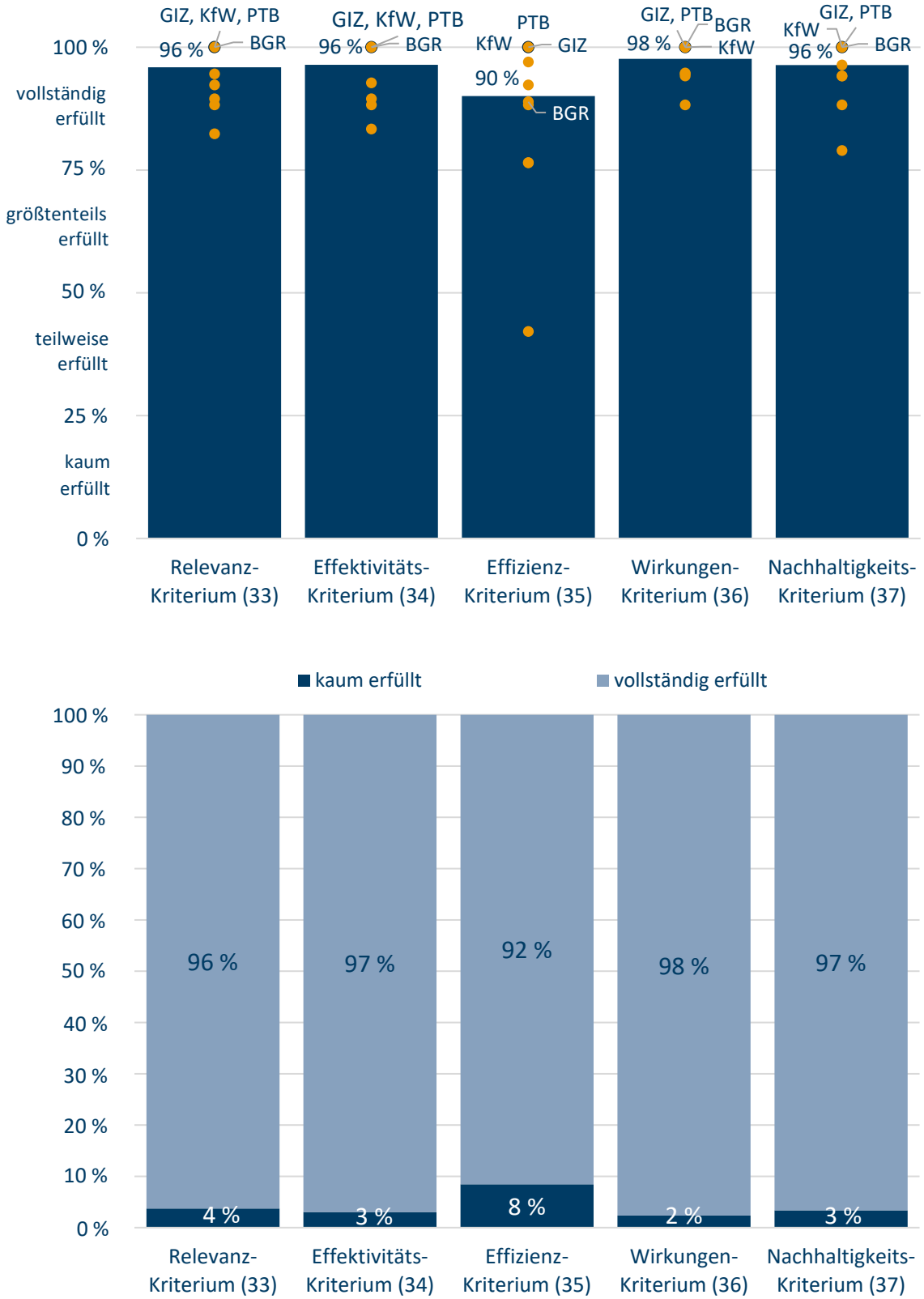
b) Schwächen zeigten sich in der Nachvollziehbarkeit der Nichtanwendung. (Ergebnis: 7)

- Eine (teilweise) Nichtanwendung der Qualitätskriterien war nicht immer nachvollziehbar dokumentiert und begründet. (Ergebnis: 7.1)

Anwendung der OECD-DAC-Kriterien: **Alle OECD-DAC-Kriterien wurden über alle elf Organisationen hinweg durchschnittlich „vollständig erfüllt".** Bei den OECD-DAC-Kriterien wurde überprüft, ob in den Evaluierungen mindestens eine Prüffrage der BMZ-Orientierungslinie (BMZ, 2006) in Bezug auf „Relevanz (33)", „Effektivität (34)", „Effizienz (35)", „Wirkungen (36)" und „Nachhaltigkeit (37)" angemessen bearbeitet wurde.[74] Die staatlichen Organisationen erfüllten alle Kriterien mindestens vollständig, in mehreren Fällen übertrafen sie die Kriterien (Abbildung 16). Eine Organisation wandte das OECD-DAC-Kriterium „Effizienz" teilweise an und verwies in Evaluierungen zum Teil auf eine Regelung in den BMZ-Richtlinien, die politischen Stiftungen einen eingeschränkteren Umgang mit allen Kriterien einräumt (BMZ, 2016). Eine Begründung für eine Nichtanwendung wurde bei allen Kriterien positiv bewertet. Die Ergebnisse stellten jedoch fast immer eine tatsächliche angemessene Erfüllung und selten eine Erfüllung durch eine begründete Nichtanwendung dar. Der geschätzte Anteil der Evaluierungen, in denen ein Kriterium durch eine begründete Nichtanwendung positiv bewertet wurde, liegt bei unter 5 Prozent.

74 Die Überprüfung der angemessenen Anwendung der OECD-DAC-Kriterien leitete sich aus den OECD-DAC-Standards und nicht aus der Publikation des OECD DAC zu den Kriterien ab. Die Überprüfung wurde auf Basis einer Prüffrage aus den BMZ-Orientierungslinien (2006) durchgeführt. Das Kriterium „Kohärenz", das seit 2020 für Organisationen verpflichtend ist (BMZ, 2020), wurde aufgrund des gewählten Untersuchungszeitraums der Meta-Evaluierung nicht aufgenommen.

Abbildung 16 Erfüllung und Häufigkeiten der Bewertungsstufen im Bereich „OECD-DAC-Kriterien“



Quelle: DEval, eigene Darstellung

Anmerkung: Die dunkelblauen Balken der oberen Grafik stellen die durchschnittlichen Ergebnisse über alle Organisationen hinweg dar. Die untere Grafik mit vorwiegend hellblauen Balken zeigt die Verteilung der Häufigkeiten über alle Berichte hinweg. Für die Ab-bildung existieren keine zwei Gruppen, da alle Organisationen zur Anwendung der OECD-DAC-Kriterien verpflichtet waren.

4.2.3 Organisationsspezifische Qualitätskriterien

Im dritten Unterabschnitt werden die Ergebnisse hinsichtlich der Anwendung der elf organisationsspezifischen Qualitätskriterien von vier Organisationen ausgeführt. Organisationsspezifische Qualitätskriterien fokussierten auf Inhalte, die ergänzend zu den OECD-DAC- und den DeGEval-Standards für eine Organisation wichtig waren und in jeder Evaluierung berücksichtigt werden sollten. Im Kasten 6 wird das Fazit und die Hauptergebnisse der Untersuchung dargestellt.

Kasten 6 Fazit zur Anwendung organisationsspezifischer Qualitätsstandards

Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der organisationsspezifischen Qualitätsstandards in den Evaluierungen der beteiligten deutschen EZ-Organisationen? (Evaluierungsfrage 2b)

Es zeigte sich ebenfalls ein positives Bild für die Anwendung der organisationsspezifischen Qualitätsstandards durch DRK, EWDE, GIZ und hbs. Sie wurden durchschnittlich „größtenteils erfüllt". Verbesserungspotenzial zeigte sich erneut in der Nachvollziehbarkeit der Nichtanwendung auf Ebene der Evaluierung.

a) Stärken zeigten sich in der Anwendung der Qualitätsstandards. (Ergebnis: 8)

- Alle organisationsspezifischen Qualitätskriterien wurden durchschnittlich „größtenteils erfüllt" (71 Prozent) und somit etwas weniger als die für diese vier Organisationen verpflichtenden OECD-DAC- und DeGEval-Qualitätskriterien (77 Prozent). (Ergebnis: 8.1)

b) Schwächen lagen in der Nachvollziehbarkeit der Nichtanwendung der Qualitätsstandards. (Ergebnis: 9)

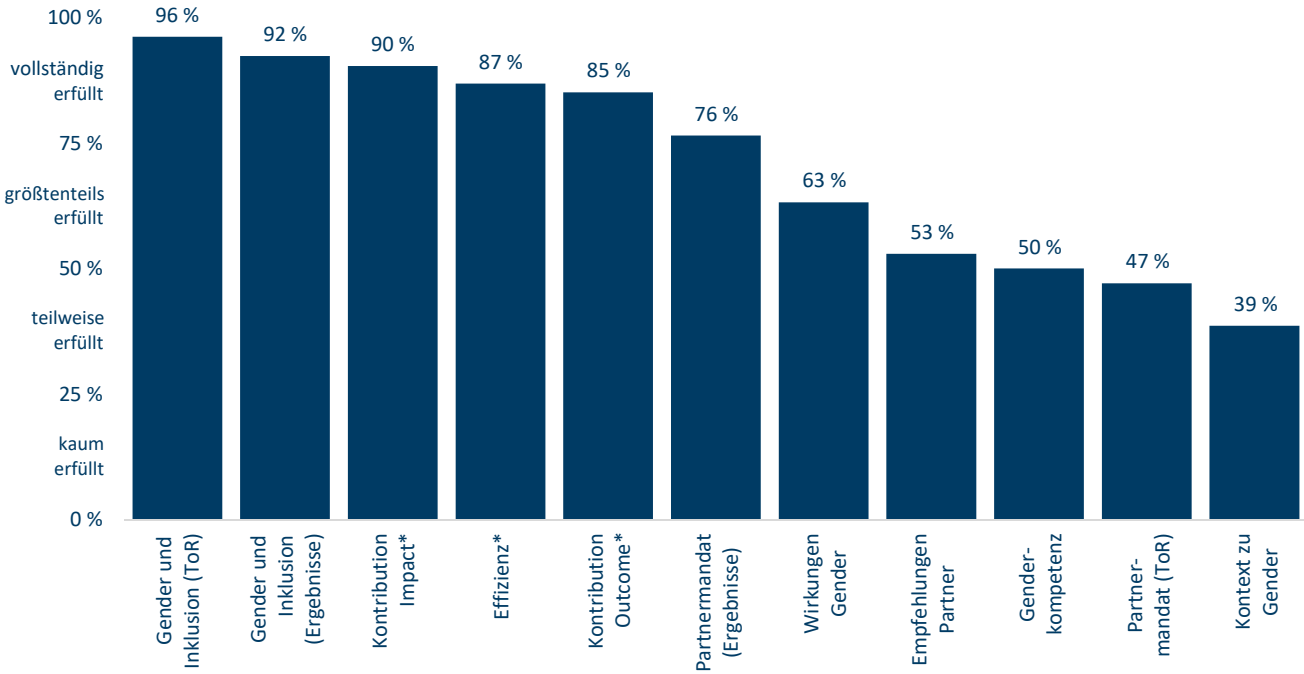
- Eine eindeutige Identifikation und Verschriftlichung der organisationsspezifischen Qualitätskriterien lag in den Organisationsdokumenten nicht bei allen Qualitätskriterien vor. (Ergebnis: 9.1)
- Ein Qualitätskriterium im Bereich „Gender" und eines im Bereich „Partnerrolle" wurde je „teilweise erfüllt". (Ergebnis: 9.2)
- Eine Nichtanwendung wurde in den Evaluierungen nicht festgehalten beziehungsweise begründet. (Ergebnis: 9.3)

Für die Untersuchung der Anwendung mussten die Operationalisierungen der Qualitätskriterien zuerst erarbeitet werden. Bis auf zwei wurden alle elf organisationsspezifischen Qualitätskriterien mindestens „größtenteils erfüllt". In der Anwendung lagen DRK, EWDE, GIZ und hbs durchschnittlich bei circa 71 Prozent und damit etwas weniger als bei ihrer durchschnittlichen Anwendung der OECD-DAC- und der DeGEval-Standards. Dies zeigt, dass organisationspezifische Qualitätskriterien relevant sind und diese auch in den Evaluierungen angewandt wurden. Am häufigsten wurden die Qualitätskriterien im Bereich „Gender und Inklusion" sowie die drei GIZ-Qualitätskriterien „Methoden der Kontributionsanalyse Impact-Ebene", „Methoden der Kontributionsanalyse Outcome-Ebene" und „Effizienzanalyse (Follow-the-money-Ansatz)" angewandt (Abbildung 17). Zwei Qualitätskriterien wurden zu weniger als 50 Prozent in den Evaluierungsdokumenten der entsprechenden Organisation angewandt. Offen ist, inwieweit eine durchschnittliche Anwendung von circa 71 Prozent zukünftig die Ansprüche der Organisationen erfüllt oder ob eine Anpassung des Anspruchsniveaus nach oben sinnvoll wäre. Ähnlich wie bei den international geltenden Qualitätskriterien lagen keine Begründungen bei einer Nichtanwendung auf Evaluierungsebene vor.

Die GIZ[75] erfüllte die drei organisationspezifischen Qualitätskriterien durchschnittlich vollständig (circa 87 Prozent). Dieser Wert lag ungefähr beim durchschnittlichen Ergebnis der GIZ hinsichtlich der Anwendung der OECD/DAC- und der DeGEval-Qualitätskriterien (circa 86 Prozent). Damit ergab sich auch für die GIZ ein positives Bild. Eine Nichtanwendung ist in den Evaluierungsdokumenten nicht verschriftlicht worden. Hier besteht eine Schwäche.

75 Aus Gründen der Anonymität werden hier und in der Grafik nur die Ergebnisse der GIZ näher erläutert.

Abbildung 17 Erfüllung der organisationsspezifischen Qualitätskriterien



Quelle: DEval, eigene Darstellung

*Anmerkung: * = Qualitätskriterien der GIZ, ToR = Terms of References (Leistungsbeschreibung). Die Zuordnung der Indikatoren zu den anderen nichtstaatlichen Organisationen wird entsprechend den anderen Ergebnissen in Abschnitt 4.2 anonymisiert.*

4.2.4 Vergleich zur Meta-Evaluierung Nachhaltigkeit (GIZ und KfW)

Im vierten Unterabschnitt werden die Ergebnisse der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit für GIZ und KfW beschrieben.

Die vorliegende Meta-Evaluierung baut auf den in der Meta-Evaluierung Nachhaltigkeit angelegten Qualitätskriterien und generierten Ergebnissen auf. Die Meta-Evaluierung Nachhaltigkeit von Noltze et al. (2018) stellte eine organisationsübergreifende Meta-Evaluierung von GIZ- und KfW-Evaluierungen dar, in der die Qualität der Evaluierungspraxis von 513 zwischen 2010 und 2016 durchgeführten GIZ- und KfW-Evaluierungen untersucht wurde. Sie beschäftige sich – ebenso wie die vorliegende Meta-Evaluierung – mit dem Austausch zur und der Untersuchung der Qualität in der Evaluierungspraxis mit Fokus auf zwei staatliche Durchführungsorganisationen. Die 16 im Rahmen der Meta-Evaluierung Nachhaltigkeit untersuchten Qualitätskriterien fokussieren dabei auf den Bereich des methodischen Vorgehens und somit auf Aspekte des Standardclusters „Berichtslegung und Methoden“. In der vorliegenden Meta-Evaluierung wurden 15 dieser Qualitätskriterien aufgegriffen und für 106 zwischen 2016 und 2020 realisierte GIZ[76]- und KfW-Evaluierungen erneut untersucht. Damit konnte zum einen der aktuelle Stand der Anwendung dieser Qualitätskriterien, zum anderen aber auch die Differenz der Anwendung seit der letzten Erhebung analysiert werden. Im Kasten 7 werden das Fazit und die Hauptergebnisse der durchgeführten Untersuchungen präsentiert (die Namen der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit wurden beibehalten).

76 Bei der GIZ beziehen sich die Zahlen auf die Jahre 2018 bis 2020.

Kasten 7 Fazit zur Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit

Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW? (Evaluierungsfrage 2c)

Hinsichtlich der Anwendung der Qualitätskriterien zeichnete sich – mit wenigen Ausnahmen – ein positives Bild ab. Konkret wurden die Qualitätskriterien durchschnittlich zu circa 75 Prozent erfüllt. Dies entspricht einer etwas höheren Ausprägung der Anwendung als bei den OECD-DAC- und den DeGEval-Standards. Herausforderungen bestanden allerdings weiterhin in der Anwendung der Qualitätskriterien „Auswahlverfahren der Gesprächspartner beschrieben“ und „Kontroll-/Vergleichsgruppe einbezogen“. Darüber hinaus ist anzumerken, dass sich die Anwendung der Qualitätskriterien seit der Meta-Evaluierung Nachhaltigkeit in allen Qualitätskriterien – zum Teil bedeutsam – verbessert hat. Insgesamt zeigte sich ein durchschnittlicher Unterschied von 36 Prozent. Die Veränderungen deuten darauf hin, dass die nach der Meta-Evaluierung Nachhaltigkeit durchgeführten Maßnahmen zur Verbesserung der Evaluierungspraxis von GIZ und KfW einen Einfluss auf die Anwendung gehabt haben könnten. Dies ist vor dem Hintergrund der damit einhergehenden umfassenden Anstrengungen einer Vielzahl an Akteuren ein sehr positives Ergebnis. Anzumerken ist aber, dass Alternativerklärungen nicht ausgeschlossen werden konnten (beispielsweise relativ leicht zu erfüllende Operationalisierungen der Qualitätskriterien oder veränderte Dokumentationsweisen).

a) Stärken zeigten sich in der aktuellen Anwendung von 13 Qualitätskriterien und der positiven Veränderung der Anwendung seit der Meta-Evaluierung Nachhaltigkeit. (Ergebnis: 10)

- Durchschnittlich wurden die Qualitätskriterien zu ungefähr 75 Prozent erfüllt. Dabei wurden acht von 15 Qualitätskriterien vollständig und fünf größtenteils erfüllt. (Ergebnis: 10.1)
- Zwischen der Meta-Evaluierung Nachhaltigkeit (Zeitpunkt 1) und der vorliegenden Meta-Evaluierung (Zeitpunkt 2) hat sich die Anwendung der Qualitätskriterien durchschnittlich um 36 Prozent erhöht; statistisch signifikant und bedeutsam sind die Unterschiede in sieben Qualitätskriterien. (Ergebnis: 10.2)

b) Schwächen zeigten sich in der aktuellen Anwendung von zwei Qualitätskriterien. (Ergebnis: 11)

- Durchschnittlich wurden die Qualitätskriterien „Auswahlverfahren der Gesprächspartner beschrieben“ teilweise und „Kontroll-/Vergleichsgruppe einbezogen“ kaum erfüllt. (Ergebnis: 11.1)

Durchschnittlich wurden die Qualitätskriterien in der vorliegenden Meta-Evaluierung zu circa 75 Prozent erfüllt. Dabei wurden acht Qualitätskriterien vollständig, fünf größtenteils und je eines teilweise beziehungsweise kaum erfüllt; letztere sind „Auswahlverfahren der Gesprächspartner beschrieben“ und „Kontroll-/Vergleichsgruppe einbezogen“. Damit lag die durchschnittliche Anwendung dieser Qualitätskriterien etwas höher als die durchschnittliche Anwendung der OECD-DAC- und der DeGEval-Standards von GIZ und KfW. Die beiden am wenigsten angewandten Qualitätskriterien zeigten eine Erfüllung von weniger als 50 Prozent. Das Qualitätskriterium „Kontroll-/Vergleichsgruppe einbezogen“, mit dem untersucht wurde, ob die Wirkungen der EZ-Maßnahme anhand eines Vergleichs zwischen Kontroll- (außerhalb des Einflussbereichs der EZ-Maßnahme) und Vergleichsgruppe (innerhalb des Einflussbereichs der EZ-Maßnahme) ermittelt wurden, lag dabei mit einer Anwendung von 18 Prozent deutlich niedriger als die Anwendung des Qualitätskriteriums „Auswahlverfahren der Gesprächspartner beschrieben“ mit 43 Prozent.

Insgesamt wurden die Qualitätskriterien zum Zeitpunkt der vorliegenden Untersuchung im Vergleich zur Meta-Evaluierung Nachhaltigkeit durchschnittlich ungefähr 36 Prozent mehr angewandt. Dies zeigte einen deutlichen Unterschied in der Anwendung der methodisch ausgerichteten Qualitätskriterien durch GIZ und KfW; statistisch signifikant und bedeutsam ist diese Differenz für sieben Qualitätskriterien. Die deutlich positive Differenz der Ergebnisse zwischen der vorliegenden Meta-Evaluierung und der Meta-Evaluierung Nachhaltigkeit deutet auf eine Verbesserung der Anwendung der Qualitätskriterien hin (Abbildung 18). Diese Differenz variiert zwischen 8 Prozent beim Qualitätskriterium „Gegenstand (Vorhaben) beschrieben“ und 91 Prozent bei „Kausalität über Plausibilitäten hergeleitet“. Die unterschiedliche Ausprägung der Differenz bei den einzelnen Qualitätskriterien hängt dabei vermutlich mit dem Grad der Anwendung zum ersten Messzeitpunkt zusammen (beispielsweise wurde das Qualitätskriterium „Gegenstand [Vorhaben] beschrieben“

zum Zeitpunkt der Meta-Evaluierung Nachhaltigkeit bereits zu 92 Prozent angewandt, sodass eine mögliche Verbesserung nicht über 8 Prozent hätte hinausgehen können). Die sieben Qualitätskriterien, die sich statistisch bedeutsam verbessert haben, sind „Gegenstand (Vorhaben) beschrieben", „Vorher-Nachher-Vergleich", „Kausalität über Plausibilitäten hergeleitet", „Datengrundlage ausreichend hinsichtlich Schlussfolgerungen", „Wirkungslogik überwiegend durch Indikatoren operationalisiert", „Befragte Gesprächspartner identifiziert" und „Schlussfolgerungen aus Daten plausibel begründet".[77] Da das Qualitätskriterium „Kontroll-/Vergleichsgruppe einbezogen" weiterhin kaum angewandt wurde – die Anwendung veränderte sich von ungefähr 9 auf 18 Prozent –, kann eine systematische Nichtanwendung des Qualitätskriteriums in GIZ- und KfW-Projektevaluierungen nicht ausgeschlossen werden. Vor dem Hintergrund, dass die KfW Ex-post-Evaluierungen durchführte, es für beide Organisationen keine Verpflichtung gab, Kontroll-/Vergleichsgruppen in Evaluierungen (zum Beispiel in Form von *rigorous impact evaluations*) zu integrieren oder eine Integration zu prüfen, und zum Beispiel bei Politikberatungen ein Kontroll-/Vergleichsgruppen-Design kaum umzusetzen ist, könnte dies das niedrige Ergebnis der Anwendung des Qualitätskriteriums erklären.

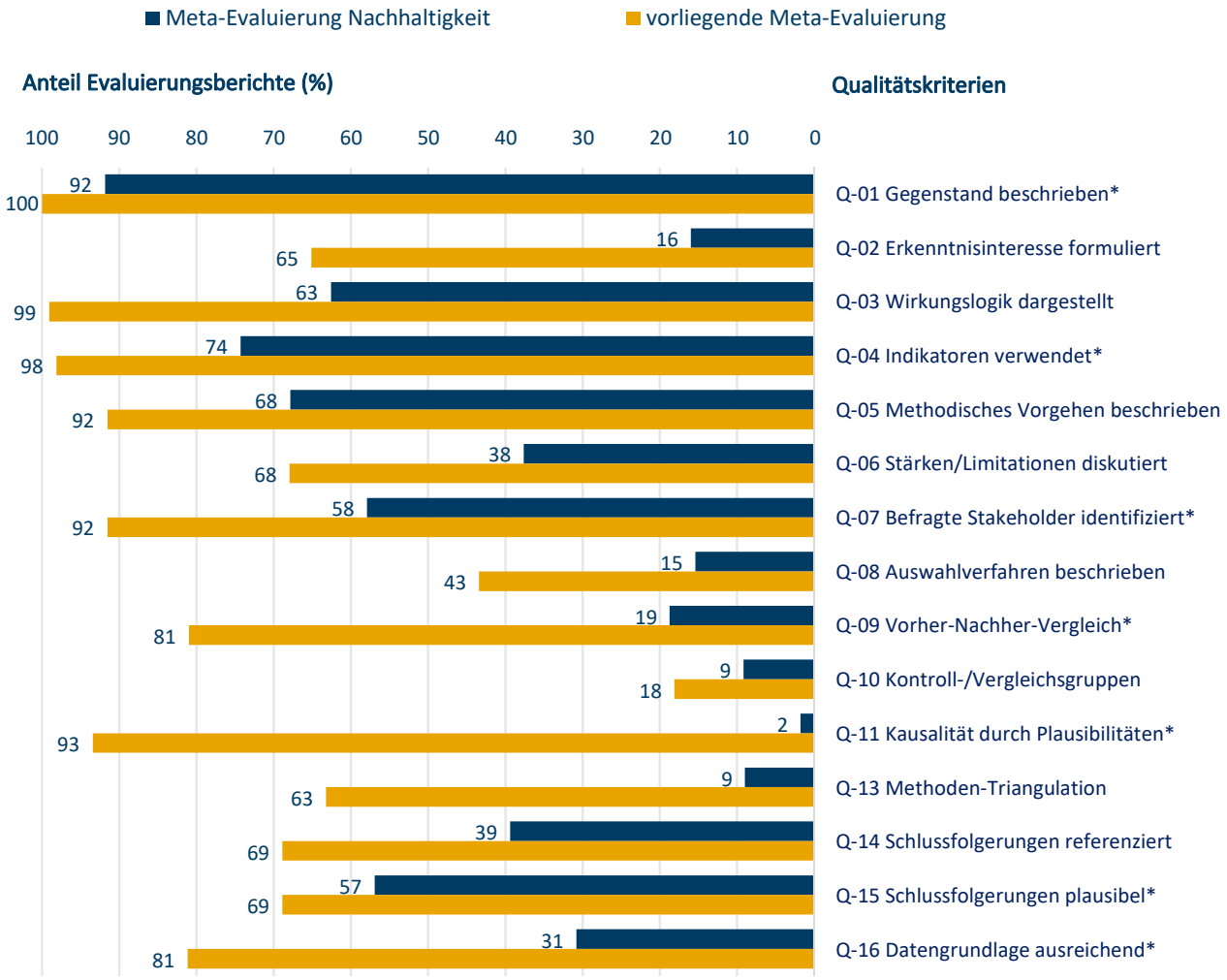
Die positive Differenz könnte mit Maßnahmen zur Erhöhung der Qualität von Projektevaluierungen von GIZ und KfW im Anschluss an die Meta-Evaluierung Nachhaltigkeit zusammenhängen, aber auch Alternativerklärungen sind möglich. Beide Durchführungsorganisationen hatten durch eigene Reformen ihrer Evaluierungssysteme, auf Grundlage der Empfehlungen aus der Meta-Evaluierung Nachhaltigkeit und mit Unterstützung von BMZ und DEval weitreichende Maßnahmen umgesetzt. Die GIZ hatte bereits Mitte 2017 eine Reform ihres Evaluierungssystems gestartet (GIZ, 2018a) und diese nach Übereinstimmung mit den Erkenntnissen und Empfehlungen der Meta-Evaluierung Nachhaltigkeit fortgeführt. Dazu gehörte unter anderem, dass mehr zuvor dezentral realisierte Projektevaluierungen heute zentral von der Stabsstelle Evaluierung koordiniert und verantwortet sowie von unabhängigen externen Expert*innen umgesetzt werden (GIZ, 2018a). Darüber hinaus werden Kontributionsanalysen inzwischen in den zentralen Projektevaluierungen standardmäßig durchgeführt (GIZ, 2018b). Dies könnte eine Erklärung für den Anstieg von 92 Prozent in der Anwendung des Qualitätskriteriums „Kausalität über Plausibilitäten hergeleitet" sein.[78] Die KfW hat währenddessen das Format der Ex-post-Evaluierungen entlang des „Rapid Appraisal 2.0" überarbeitet (KfW, 2019)[79], die Transparenz der methodischen Vorgehensweise erhöht und ergänzende Verfahren zur Auswertung von Schlussfolgerungen und Lernerfahrungen umgesetzt. Diese Maßnahmen könnten nachvollziehbar die positive Differenz in der Anwendung der Qualitätskriterien erklären. Nicht auszuschließen ist aber gleichfalls, dass der Unterschied auf die teilweise leicht zu erfüllenden Operationalisierungen der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit zurückzuführen ist. Möglich wäre zudem, dass GIZ und KfW bereits zum Zeitpunkt der ersten Meta-Evaluierung die Qualitätskriterien anwandten, dies aber nicht transparent und nachvollziehbar beschrieben war, sodass die Differenz auf eine systematischere Dokumentation zurückgeführt werden könnte. Schließlich – auch das ist nicht auszuschließen – könnte die Diskrepanz entlang einer Kombination der beschriebenen Erklärungen entstanden sein.

77 Für Details der Ergebnisse des Strukturgleichungsmodells siehe Onlineanhang, Abschnitt 4.1.3.

78 Das Qualitätskriterium lautet: „Das Kriterium ist erfüllt, wenn die Wirkungen des Entwicklungsvorhabens auf der Grundlage eines systematischen Verfahrens anhand von Plausibilitäten (insbesondere theoriebasierter Ansätze, zum Beispiel durch Kontributionsanalysen) ermittelt werden."

79 Dies bezieht sich insbesondere auf die DEval-Empfehlung 1 zur Weiterentwicklung der Evaluierungspraxis: „Vor dem Hintergrund zunehmender Anforderungen an die Evaluierung als Instrument für Lernen und Rechenschaftslegung sollten GIZ und KfW Maßnahmen entwickeln, die sicherstellen, dass weitere Potenziale zur Erhöhung der Evaluierungsqualität, insbesondere im Bereich des Wirkungs- und Nachhaltigkeitsnachweises, ausgeschöpft werden" (Noltze et al., 2018, S. 47).

Abbildung 18 Anteil Evaluierungsberichte je erfüllten Qualitätskriterien zu beiden Zeitpunkten



Quelle: DEval, eigene Darstellung angelehnt an Noltze et al. (2018, S. 28, Abbildung 4)
Anmerkung: Meta-Evaluierung Nachhaltigkeit = 513 Evaluierungen, vorliegende Meta-Evaluierung = 106 Evaluierungen; Q-12 Daten-Triangulation wurde in der vorliegenden Meta-Evaluierung nicht erhoben, da es über das Qualitätskriterium „Methoden-Triangulation" inhaltlich abgedeckt wurde. Die Namen der Qualitätskriterien wurden von der Meta-Evaluierung Nachhaltigkeit übernommen. Da einige OECD-DAC- und DeGEval-Qualitätskriterien der vorliegenden Meta-Evaluierung transformiert werden mussten, um sie in den Vergleich mit den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit aufzunehmen, sind die Ergebnisse nicht immer deckungsgleich mit denen in Abschnitt 4.2.1. Diese Qualitätskriterien haben sich statistisch signifikant ($p < 0,05$) und bedeutsam verbessert. Weitere Details finden sich im Onlineanhang, Abschnitt 4.1.3.*

4.3 Erklärung der Anwendung der Qualitätskriterien

In folgenden Abschnitt wird dargestellt, inwieweit die in der Literatur und den Fokusgruppendiskussionen identifizierten Faktoren die Anwendung der Qualitätskriterien erklären beziehungsweise mit ihnen zusammenhängen. Im Kasten 8 werden das Fazit und die Hauptergebnisse der Untersuchungen präsentiert.

Kasten 8 Fazit zur Erklärung der Anwendung der Qualitätsstandards

Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen? (Evaluierungsfrage 3)

a) Insgesamt gaben die Ergebnisse wenig Hinweise auf signifikante Zusammenhänge zwischen den identifizierten Faktoren und den Qualitätskriterien. Einige bedeutsame positive Zusammenhänge wurden in der evaluierungsspezifischen Dimension zwischen den Faktoren „Anzahl der internen und externen Gutachtenden", „Einbindung der internen und externen Stakeholder*innen" und „Jahr der Evaluierung" mit unterschiedlichen Qualitätskriterien identifiziert. Da in den Analysen viele Modelle mit zum Teil miteinander korrelierenden Qualitätskriterien geschätzt wurden, müssen die Ergebnisse vorsichtig interpretiert werden.

b) In den Regressionsanalysen wurden wenige bedeutsame Ergebnisse identifiziert.[80] (Ergebnis: 12)

- In der evaluierungsspezifischen Dimension konnte sowohl bei der „Kompetenz Gutachtende" (insbesondere des Proxys „Anzahl interne und externe Gutachtende") als auch bei den „Qualitätssicherungsprozessen" (insbesondere des Proxys „Einbindung Stakeholder*innen") ein positiver Zusammenhang mit der Anwendung verschiedener Qualitätskriterien ermittelt werden. Darüber hinaus zeigte das „Jahr der Evaluierung" positive Zusammenhänge mit zwei Qualitätskriterien. (Ergebnis: 12.1)
- Länderkontextspezifische Faktoren hingen nicht bedeutsam mit der Anwendung der Qualitätskriterien zusammen. (Ergebnis: 12.2)
- Die organisationsspezifischen Faktoren zeigten kein klares Bild hinsichtlich der Zusammenhänge mit der Anwendung der Qualitätskriterien. (Ergebnis: 12.3)

Die nachfolgenden Informationen geben Einblicke in die ermittelten Faktoren oder ihre Proxys, um die Ergebnisse der Analyse zu kontextualisieren. Ungefähr 81,1 Prozent der Evaluierungen wurden ausschließlich von externen Gutachtenden, 13,2 Prozent ausschließlich von internen Gutachtenden (Personen, die innerhalb der jeweiligen Organisation arbeiteten) durchgeführt. Bei 5,6 Prozent gab es zusätzlich zu externen eine oder mehrere interne Gutachtende. Insgesamt wurden die Evaluierungen von mindestens einem und bis zu zehn Gutachtenden durchgeführt (durchschnittlich: 1,9 Gutachtende). Der durchschnittliche Tagessatz der externen Gutachtenden lag bei ungefähr 440 Euro, das durchschnittliche Verhältnis zwischen den Gutachtenden-Tagen und dem Volumen der EZ-Maßnahme bei ungefähr 340.000 Euro/Tag. Bei 10 Prozent der Evaluierungen fand eine Remote-, bei 6 Prozent eine Semiremote-Datenerhebung statt, und bei 84 Prozent der Evaluierungen wurden die Daten vor Ort erhoben.[81] Tabelle 6 zeigt eine Übersicht über Ergebnisse der untersuchten Regressionsanalysen.[82]

Einige Faktoren konnten empirisch nur über Stellvertretervariablen (Proxys) untersucht werden, da entweder eine einheitliche organisationsübergreifende Definition oder nachvollziehbare Wirkungszusammenhänge oder die Datenverfügbarkeit nicht gegeben war. Zu diesen Faktoren gehörten „Kultureller Kontext", „Pandemie", „Kompetenz Gutachtende", „Kosten der Evaluierung", „Qualitätssicherungsprozesse", „Strukturierter Planungsprozess", „Evaluierungs- und Lernkultur" und „Evaluierungseinheit vorhanden".

Die Ergebnisse der Regressionsanalysen zeigten keine bedeutsamen empirischen Zusammenhänge zwischen dem Länderkontext und den untersuchten Qualitätskriterien oder dem Standardcluster „Berichtslegung und Methoden". Auf Basis der gewählten Modellspezifikationen wies nur der „Sozialkapital-Index" einen negativen aber nicht negativ bedeutsamen Zusammenhang mit der „Darstellung der Angemessenheit des methodischen Vorgehens" auf. Als mögliche Erklärung für den fehlenden Zusammenhang mit „Pandemie" könnte angenommen werden, dass der Zeitraum der Meta-Evaluierung ausschließlich den Beginn der Pandemie (das Jahr 2020) mit einschließt und somit potenzielle Auswirkungen auf die Qualität von Evaluierungen noch nicht erfasst werden konnten. Die Regressionsanalysen bestätigten die Erkenntnisse von

80 Weitere Details zu den Regressionsanalysen finden sich im Onlineanhang, Abschnitt 4.2.
81 Deskriptive Informationen zum Organisationskontext werden in Abschnitt 1.1 dargestellt.
82 Details zu den Regressionsanalysen finden sich im Onlineanhang, Abschnitt 4.2.

Wencker und Verspohl (2019), dass „Konflikte“ nicht mit Qualitätskriterien des Standardclusters „Berichtslegung und Methoden“ zusammenhängen. Die Ergebnisse zeigten zusätzlich, dass positive – wenn auch nicht bedeutsame – Zusammenhänge mit einzelnen Qualitätskriterien des Standardclusters „Partizipation, Unabhängigkeit und Fairness“ bestehen.

In der Evaluierungsdimension hingen die Faktoren „Kompetenz Gutachtende“ und „Qualitätssicherungsprozesse“ positiv mit einzelnen Qualitätskriterien zusammen. Zudem hing das „Jahr der Evaluierung“ positiv mit einigen Qualitätskriterien des Standardclusters „Berichtslegung und Methoden“ zusammen. Die „Kompetenz Gutachtende“, unter anderem operationalisiert über die „Anzahl der internen und externen Gutachtenden“, hing positiv und zum Teil bedeutsam mit ausgewählten Qualitätskriterien des Standardclusters „Berichtslegung und Methoden“ zusammen („Darstellung der Angemessenheit des methodischen Vorgehens“ und „Informationsgehalt der Zusammenfassung“), Aspekte des „Qualitätssicherungsprozesses“ einer Evaluierung (beispielsweise „Einbindung der Stakeholder*innen“) ebenfalls positiv und bedeutsam mit einzelnen Qualitätskriterien („Einbindung des Kontextes“, „Nachvollziehbarkeit der Informationsquellen“, „Summenindex Berichtslegung und Methoden“) und das „Jahr der Evaluierung“ mit mehreren Qualitätskriterien des Standardclusters „Berichtslegung und Methoden“ sowie mit dem Summen- und dem Faktorindex des Standardclusters positiv – zum Teil bedeutsam – zusammen, mit der „Nachvollziehbarkeit der Informationsquellen“ negativ. Ähnlich wie bei einer Meta-Evaluierung dezentraler Evaluierungen der Jahre 2017 bis 2020 aus Finnland von Väth et al. (2022) wurde der „Informationsgehalt der Leistungsbeschreibung“ auch als signifikanter, jedoch im Gegensatz zur finnischen Meta-Evaluierung nicht als starker Einflussfaktor ermittelt. Weitere signifikante aber nicht bedeutsame[83] Ergebnisse sind: Der „Informationsgehalt der Leistungsbeschreibung“ wies darüber hinaus einen negativen Zusammenhang mit dem „Informationsgehalt der Zusammenfassung“ auf. Der Proxy für die Kosten einer Evaluierung (Gutachtenden-Tage im Verhältnis zum Volumen der EZ-Maßnahme) stand negativ mit dem Qualitätskriterium „Kohärenz von Daten-Ergebnissen-Schlussfolgerungen“ in Verbindung. Während nicht wie bei Hundt und Bräuer (2021) eine Beziehung zwischen der Durchführung einer „Remote-Datenerhebung“ und der „Darstellung der Angemessenheit des methodischen Vorgehens“ festgestellt wurde, zeigte sich ein negativer Zusammenhang zwischen „Remote-Datenerhebung“ und „Kapazitätsentwicklung im Partnerland“ statt.

Die organisationsspezifischen Faktoren zeigten kein klares Bild hinsichtlich der Zusammenhänge mit der Anwendung der Qualitätskriterien. Die Ergebnisse zu den Faktoren des Organisationskontexts zeigten keine bedeutsamen Zusammenhänge mit einzelnen Qualitätskriterien. Einen positiven – nicht bedeutsamen – Zusammenhang zwischen der „Evaluierungstätigkeit“ und einer informativeren „Zusammenfassung“ und einen negativen mit der „Darstellung der Angemessenheit des methodischen Vorgehens“. Darüber hinaus bestand ein negativer Zusammenhang zwischen der Größe der „Evaluierungseinheiten/-stellen“ und dem „Informationsgehalt der Leistungsbeschreibung“. Die negativen Zusammenhänge lagen nur für vereinzelte Qualitätskriterien vor. Zudem hatten die einzelnen Organisationen zum Teil statistisch signifikante Zusammenhänge mit den Qualitätskriterien.

Für die Qualitätskriterien „Darstellung der Angemessenheit des methodischen Vorgehens“ und „Darstellung der Wirkungszusammenhänge“, die beide durchschnittlich nur „teilweise angewandt“ wurden (siehe Abschnitt 4.2), konnten signifikante, wenn auch nicht bedeutsame, Zusammenhänge ermittelt werden. Für das Qualitätskriterium „Darstellung der Wirkungszusammenhänge“ konnte ein positiver Zusammenhang mit dem „Jahr der Evaluierung“ ermittelt werden, das heißt, über die Jahre 2016 bis 2020 hinweg wurden die Wirkungszusammenhänge in den Evaluierungen der Organisationen besser dargestellt. Darüber hinaus hängen eine höhere „Anzahl an Gutachtenden“ und das „Jahr der Evaluierung“ je mit einer besseren „Darstellung der Angemessenheit des methodischen Vorgehens“ zusammen. Der Bezug ist am höchsten, wenn vier im Vergleich zu einem*r Gutachtenden die Evaluierung durchgeführt haben. Es zeigen sich auch negative Zusammenhänge mit dem „Sozialkapital-Index“ und der „Evaluierungstätigkeit“.

[83] Als bedeutsam gelten die Zusammenhänge zwischen Qualitätskriterien und Faktoren, die eine ausreichend große Effektstärke aufwiesen. Die Effektstärken wurden für die (geordneten) logistischen Modelle bestimmt (siehe Tabelle 6 sowie Onlineanhang, Abschnitt 4.2).

Tabelle 6 Zusammenhänge zwischen den Faktoren und der Anwendung der Qualitätskriterien

		SC B & M												SC N	SC P, U & F		
Faktor	**Proxy**	Evaluierungs-gegenstand	Kontext*	Wirkungszu-sammenhänge	Erkenntnis-interesse	Informations-quellen	Darstellung Methodik	Kohärenz	Leistungs-beschreibung	Zusammen-fassung	Vorhandensein IR	Summenindex SC B & M	Faktorindex SC B&M	Empfehlungen*	Kapazitäts-entwicklung	GAs Partnerland	Einbindung Stakeholder*innen
	Regressions-modell Nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Fragilität	Konflikt	–	–	–†	–	–	–	–	–	–	–	–	–		–	–	–
Kultur	Sozialkapital	+	+	+	+	+	–*	+	+	+	+	–†	–†				+
Pandemie	Jahr 2020	–	–	–	–	–	–	–	–	–	–	–	–		–	–	–
Kompetenz GAs	Anzahl GAs	+	+	+	+	+	+*	+	+	+***k	+	+†	+		+	+	
	Tagessatz GAs	+	+	+	+	+	+	+	+	+	+	+	+		+		+
Qualitäts-sicherung	Leistungs-beschreibung	+	+	+	+	+	+	+*		–*	+			+			
	Vorhandensein IR	+	+	+	+	+†k	+k	+	+	+			+	+*	–*		+
	Einbindung Stakeholder*innen	+	+**	+†	+	+***k	+	+	+†k	+	+	+*	+				
Remote-Datenerhebung							–								–*	+	
Kosten	GA-Tage/Kosten Maßnahme	+	+	+	+	+	+	–**	+	+	+	+	+				
Jahr der Evaluierung		+	+*	+***	+**k	–*	+*	+	+**k	+	+	+*	+*				
Evaluierungs-einheiten/-stellen absolut		+	+	+	+	+	–	+	–*	+	+	+	+	+	+	+	+

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

Quelle: DEval, eigene Darstellung

*Anmerkung: GAs = Gutachtende; SC = Standardcluster; B & M = Berichtslegung und Methoden; N = Nutzbarkeit; P, U & F = Partizipation, Unabhängigkeit und Fairness; blau = positiver Zusammenhang; orange = negativer Zusammenhang; † = marginaler Zusammenhang; grau = kein statistisch signifikanter Zusammenhang; leeres Feld = nicht getestet. † < 0,10; * = p < 0,05; ** = p < 0,01; *** = p < 0,001; k = kleine (*Odds Ratio *zwischen 1,68 und 3,45) und m = mittlere (*Odds Ratio *zwischen 3,46 und 6,7o) Effektstärke (Chen und Cohen, 2010). Die Ergebnisse der Kontrollvariablen sind nicht abgebildet. [1] Aus Gründen der Anonymität werden die Zusammenhänge zwischen den einzelnen Organisationen und den Qualitätskriterien nicht abgebildet, sondern die statistischen Zusammenhänge zeigen den Durchschnitt der Ergebnisse. * Für „Kontext" und „Nützlichkeit der Empfehlungen" wurden jeweils die beiden Qualitätskriterien zu einem Qualitätsstandard (als Summenindex) zusammengefasst.*

5. SCHLUSSFOLGERUNGEN UND EMPFEHLUNGEN

In der vorliegenden Meta-Evaluierung werden organisationsübergreifende Erkenntnisse zum Qualitätsverständnis und zur Anwendung ausgewählter Qualitätsstandards in Evaluierungen von staatlichen und nichtstaatlichen Organisationen aus der deutschen EZ zur Verfügung gestellt. Darüber hinaus werden Hinweise über mögliche Faktoren geliefert, die mit der Anwendung von Qualitätskriterien zusammenhängen. *Dies ermöglicht (übergeordnetes) Lernen. Aufgrund des organisationsübergreifenden Charakters können neben dem Grad der Anwendung der Qualitätsstandards verschiedenste Formen und Gründe für die Anwendung oder Nichtanwendung einzelner Qualitätsstandards erfasst und in nachfolgenden Studien, beispielsweise organisationsinternen Meta-Evaluierungen, berücksichtigt werden. Für alle beteiligten Organisationen bietet sie eine Orientierung für die Weiterentwicklung der eigenen Evaluierungspraxis nach innen, für die staatlichen Organisationen außerdem eine Möglichkeit, nach außen Rechenschaft über die Anwendung der Qualitätsstandards abzulegen. Schließlich stellen die OECD-DAC- und die DeGEval-Standards, die systematisch für das Analyseraster der Meta-Evaluierung abgeleitet wurden, ebenfalls für die mittlerweile in Kraft getretenen BMZ-Leitlinien Evaluierung zentrale Qualitätsstandards dar. Die Erkenntnisse und das Analyseraster der Meta-Evaluierung bilden somit eine Basis für ein potenziell noch zu entwickelndes Analyseraster hinsichtlich der BMZ-Leitlinien Evaluierung.*

Nachfolgend findet sich ein einleitender Abschnitt zur Einordnung der Schlussfolgerungen und Empfehlungen. Im Anschluss werden die Empfehlungen in fünf Themen gegliedert: 1) Identifikation relevanter Qualitätsstandards und systematische Verschriftlichung in Organisationsdokumenten, 2) Identifikation nicht relevanter Qualitätsstandards und systematische Verschriftlichung in Organisationsdokumenten, 3) Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung relevanter Qualitätsstandards auf Ebene der Evaluierung, 4) gemeinsames Lernen und 5) Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit. Da sowohl die Identifikation als auch die Sicherstellung und die Nachvollziehbarkeit der Anwendung verpflichtender Qualitätsstandards grundlegend sind, um gute Evaluierungen zu gewährleisten, besteht für alle Organisationen Handlungsbedarf. Dieser variiert für die einzelnen Organisationen entsprechend ihrem Grad der bisherigen Auseinandersetzung und Anwendung beziehungsweise Erfüllung. Wichtig ist dabei zu berücksichtigen, dass alle Qualitätsstandards des OECD DAC und der DeGEval als gleichwertig anzusehen sind und einbezogen werden sollten.

Die Empfehlungen der Meta-Evaluierung leiten sich aus den im vorherigen Kapitel dargestellten Ergebnissen ab und richten sich an das BMZ sowie an BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB. Darüber hinaus können die Empfehlungen auch für VENRO und weitere nichtstaatliche Organisationen angemessen und nützlich sein. Die Empfehlungen leiten sich vor allem aus den Evaluierungsfragen zum Qualitätsverständnis und zur Anwendung ausgewählter Qualitätsstandards und Qualitätskriterien ab (OECD-DAC-, DeGEval- und/oder organisationsspezifische Qualitätsstandards als auch Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit, Evaluierungsfragen 1 und 2a bis 2c). Die Antworten auf Evaluierungsfrage 3 (Faktoren zur Anwendung der Qualitätsstandards) werden als Hintergrundinformation oder im Rahmen der Umsetzungshinweise herangezogen. Die Empfehlungen sind allgemein formuliert, sodass alle beteiligten Organisationen sich entlang ihrer organisationsspezifischen Ergebnisse selbst innerhalb der Empfehlungen verorten müssen, da der in den Ergebnissen dargestellte organisationsübergreifende Mittelwert für die Bewertung der Anwendung der Qualitätsstandards einzelner Organisationen nicht ausreichend aussagekräftig ist. So gibt es Organisationen, die bei der Anwendung von Qualitätsstandards diese nur kaum oder teilweise angewandt haben, auch wenn der Durchschnittswert der Anwendung als größtenteils oder vollständig erfüllt bewertet wurde.

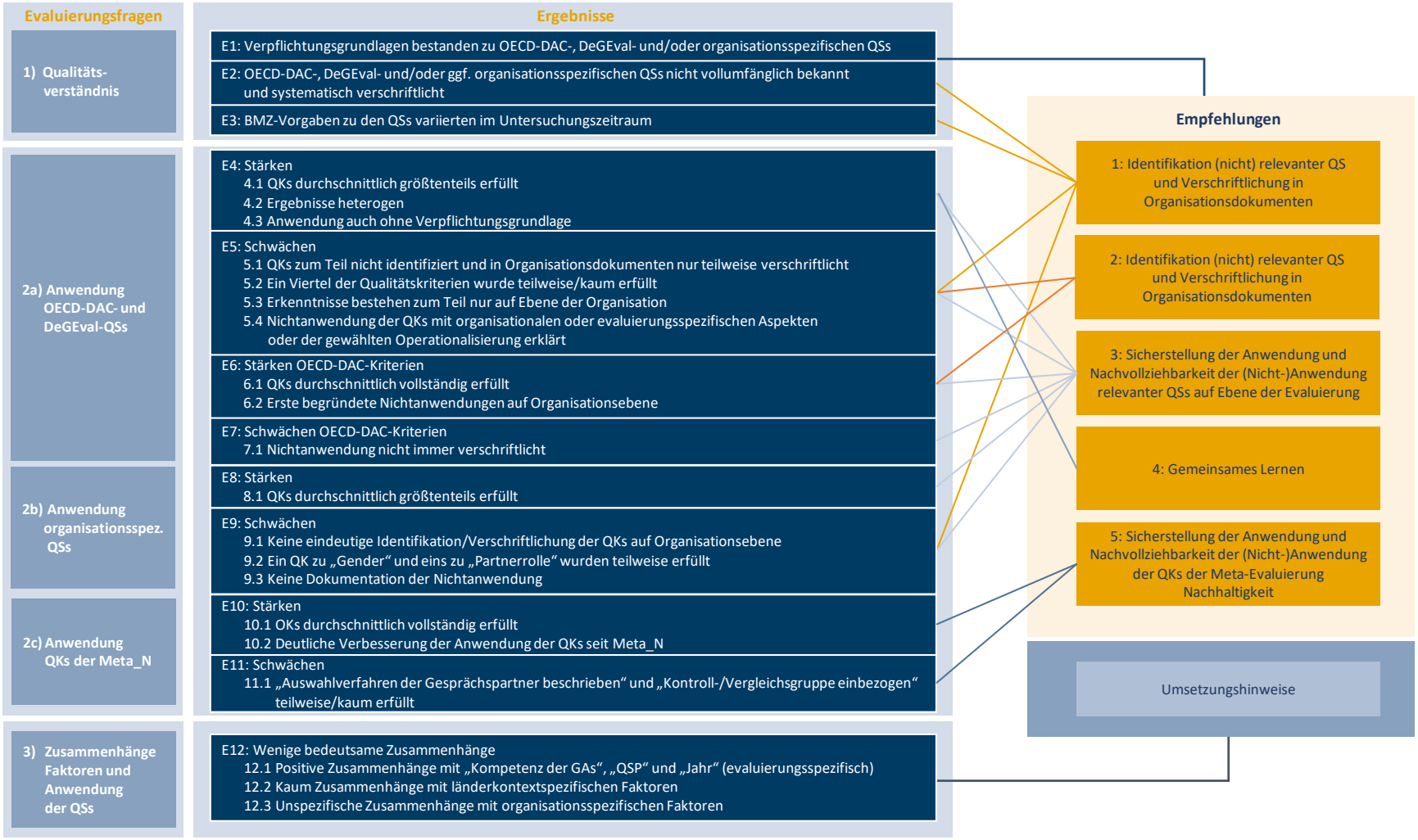
Die Erkenntnisse für die vier staatlichen Organisationen BGR, GIZ, KfW und PTB sind im Berichtsanhang (Abschnitt 7.1) gesondert dargestellt, sodass diese für die Rechenschaftslegung sowie die Umsetzungsplanung und das -monitoring für die Empfehlungen dieser Meta-Evaluierung herangezogen werden können. Die organisationsspezifischen Ergebnisse der nichtstaatlichen Organisationen wurden ausschließlich mit den jeweiligen Verantwortlichen der Evaluierungseinheiten/-stellen geteilt. Entlang der organisationsspezifischen Ergebnisse können die Organisationen folglich die für sie zutreffenden und

handlungsleitenden Empfehlungen und Umsetzungshinweise identifizieren.[84] Da in der kriterienbasierten Auswahl der nichtstaatlichen Organisationen ein Fokus auf ihre strukturelle Heterogenität gelegt und somit die Bandbreite möglicher Anwendungsgrade und -formen für unterschiedliche Organisationen abgebildet wurde, können sich auch nicht beteiligte nichtstaatliche Organisationen in den Ergebnissen verorten und damit an den Schlussfolgerungen und Empfehlung orientieren. Die an das BMZ gerichteten Empfehlungen beziehen sich auf das BMZ-Referat für Evaluierung.

Abbildung 19 gibt eine Übersicht, welche Ergebnisse zur Erarbeitung welcher Empfehlungen herangezogen wurden. Dabei wurden zugunsten der Übersichtlichkeit sowohl die Ergebnisse als auch die Empfehlungen verkürzt dargestellt.

84 Dass die Ergebnisse der Organisationen zum Teil anonym abgebildet werden, sollte die Notwendigkeit der Umsetzung der dargestellten Empfehlungen nicht beeinträchtigen. Das ist auch deshalb der Fall, da bei fast allen anonym abgebildeten Organisationen Verpflichtungsgrundlagen für die Anwendung der Qualitätsstandards bestehen und darüber hinaus die 2021 in Kraft getretenen BMZ-Leitlinien Evaluierung ebenfalls für die nichtstaatlichen Organisationen einen orientierenden Charakter haben.

Abbildung 19 Übersicht über die Herleitung der Empfehlungen aus den Ergebnissen



Quelle: DEval, eigene Darstellung
Anmerkung: E = Ergebnis; QS = Qualitätsstandard; QK = Qualitätskriterium; GA = Gutachtende; QSP = Qualitätssicherungsprozess; Meta_N = Meta-Evaluierung Nachhaltigkeit

Identifikation (nicht) relevanter Qualitätsstandards und Verschriftlichung dieser in Organisationsdokumenten

Um langfristig eine gute Evaluierungspraxis zu gewährleisten, ist es zielführend, relevante und nicht relevante Qualitätsstandards für die Evaluierungen einer Organisation zu identifizieren und systematisch in den Organisationsdokumenten zu verschriftlichen. Für Organisationen der deutschen EZ liegen unterschiedliche Verpflichtungen zur Umsetzung von Qualitätsstandards vor – unter anderem im Rahmen von Mitgliedschaften, Förderrichtlinien, Leitlinien und/oder Organisationsdokumenten. Dabei können für sie neben den in der Meta-Evaluierung untersuchten OECD-DAC- und DeGEval-Qualitätsstandards auch weitere Qualitätsstandards relevant sein (zum Beispiel Qualitätsstandards aus den neuen BMZ-Leitlinien Evaluierung, anderen Standarddokumenten, die beispielsweise in Partnerländern relevant sind, und/oder organisationsspezifische Qualitätsstandards). Im Rahmen der Meta-Evaluierung wurden die für eine Organisation verpflichtenden Qualitätsstandards (nicht Standarddokumente) von allen Organisationen nicht vollumfänglich in den Organisationsdokumenten identifiziert und verschriftlicht. Dies stellt eine Schwäche dar. Eine systematische Identifikation und Verschriftlichung aller verpflichtenden Qualitätsstandards in den Organisationsdokumenten ist zielführend, um 1) eine Identifikation aller relevanten Qualitätsstandards für eine Organisation sicherzustellen; 2) eine unbewusste Nichtanwendung einzelner Qualitätsstandards zu verhindern; 3) eine bewusste Nichtanwendung einzelner Qualitätsstandards zu verschriftlichen; 4) Prozesse zur Anwendung aller für die Organisation relevanten Qualitätsstandards in der Evaluierungspraxis zu etablieren und darauf aufbauend verbessern zu können; und 5) die Transparenz über die Relevanz verpflichtender international geltender sowie organisationsspezifischer Qualitätsstandards herzustellen.

Da Organisationen teilweise ihre inhaltlichen und strategischen Schwerpunkte ändern, ihre Evaluierungstätigkeit regional beziehungsweise sektoral anpassen, Standarddokumente in episodischen Abständen revidiert (zum Beispiel die DeGEval-Standards) oder neue entwickelt werden, ist die Identifikation und Verschriftlichung relevanter Qualitätsstandards in den Organisationen keine einmalige, sondern eine fortwährende Aufgabe. Oder wie im DeGEval-Standarddokument (DeGEval, 2016, S. 13) steht: „Schließlich ist es ein Gebot einer eigenen evaluativen Grundhaltung, auch Bewährtes immer wieder einer kritischen Prüfung zu unterziehen und auf mögliche Verbesserungen hin zu diskutieren“.

Hinweis: Die Empfehlungen können auch für nicht beteiligte Organisationen angemessen und nützlich sein.

Empfehlung 1

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten im Rahmen einer Revision ihrer Evaluierungspraxis – wenn noch nicht vorhanden – die für ihre Organisation verpflichtenden Qualitätsstandards identifizieren, in Organisationsdokumenten explizit benennen und ihre Anwendung in Evaluierungsprozessen festlegen. Die Identifikation und die systematische Verschriftlichung der Qualitätsstandards sollten in regelmäßigen Abständen überprüft werden. Dabei sollten die Organisationen ihren Anspruch an den Grad der Anwendung der einzelnen Qualitätsstandards konkret bestimmen. (Ergebnis: 2, 5.1, 9.1)

b) Das BMZ sollte im Rahmen anstehender Aktualisierungen von Förderrichtlinien oder Nebenbestimmungen für einzelne Haushaltstitel einen Beitrag dazu leisten, die BMZ-Leitlinien Evaluierung als Referenzdokument für Evaluierungen zu stärken. Im Rahmen der Aktualisierungen sollte das BMZ gemeinsam mit den betroffenen nichtstaatlichen Organisationen organisationale Besonderheiten (beispielsweise wie bei den Förderrichtlinien der politischen Stiftungen) festhalten und verschriftlichen. Das Maximalstandardprinzip sollte dabei erhalten bleiben. (Ergebnis: 3)

c) Das BMZ sollte basierend auf den BMZ-Leitlinien Evaluierung und im Austausch mit staatlichen und nichtstaatlichen Organisationen sowie unter Berücksichtigung des Analyserasters der vorliegenden Meta-Evaluierung ein Analyseraster für die Anwendung der Qualitätsstandards erarbeiten und den staatlichen und nichtstaatlichen Organisationen bereitstellen.

Empfehlung 2

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten in Organisationsdokumenten die generelle Nichtanwendung einzelner für sie verpflichtender Qualitätsstandards begründen und verschriftlichen. (Ergebnis: 5.1, 6.2)

b) Das BMZ sollte sich mit den staatlichen Organisationen bezüglich einer Anwendung und (begründeten) Nichtanwendung der in den BMZ-Leitlinien Evaluierung beschriebenen Qualitätsstandards verständigen, um eine Nichtanwendung auf Organisationsebene gemeinsam festzulegen oder Unstimmigkeiten zu dokumentieren.

Umsetzungshinweise:

- Für die Identifikation relevanter Qualitätsstandards sollten alle für die Organisation geltenden Qualitätsstandarddokumente (beispielsweise die 2021 in Kraft getretenen BMZ-Leitlinien Evaluierung, die Standarddokumente von OECD DAC und/oder DeGEval) sowie relevante interne Organisationsdokumente herangezogen werden. Dabei sollte berücksichtigt werden, dass bei verschiedenen Evaluierungstypen gegebenenfalls unterschiedliche Qualitätsstandards relevant werden können (zum Beispiel für dezentrale Evaluierungen).
- Um den Grad der angestrebten Erfüllung für die relevanten Qualitätsstandards festzulegen (Schwellenwerte eines Anspruchsniveaus), kann auf die organisationsspezifischen Ergebnisse oder das Anspruchsniveau dieser Meta-Evaluierung zurückgegriffen werden. Auch hierbei sollten die verschiedenen Evaluierungstypen berücksichtigt werden.
- Das BMZ sollte bei der Erarbeitung des Analyserasters unterschiedliche gleichwertige Formen der Anwendung von Qualitätsstandards berücksichtigen, um der Heterogenität der Organisation gerecht zu werden.

Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung relevanter Qualitätsstandards auf Ebene der Evaluierung

Die Anwendung der OECD-DAC- und der DeGEval-Standards ist in weiten Teilen bei den beteiligten deutschen EZ-Organisationen verankert, dabei variierte der Grad der Anwendung der Qualitätskriterien innerhalb einer und zwischen den Organisationen deutlich. Insgesamt wurden die OECD-DAC- und die DeGEval-Qualitätsstandards durchschnittlich größtenteils, die OECD-DAC-Kriterien vollständig, die organisationsspezifischen Qualitätsstandards größtenteils und die Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit vollständig erfüllt. Daran zeigt sich, dass die Anwendung der Qualitätsstandards Eingang in die Evaluierungspraxis gefunden hat. Dies stellte eine Stärke in der Anwendung der beteiligten Organisationen dar.

Die Nachvollziehbarkeit der Anwendung und vor allem der Nichtanwendung auf Ebene der Evaluierung ist verbesserungswürdig. Eine (begründete) Nichtanwendung fand nur in seltenen Fällen statt. Dies stellte eine Schwäche der beteiligten Organisationen hinsichtlich der Anwendung der Qualitätsstandards dar. Ohne die Nachvollziehbarkeit der Anwendung und Nichtanwendung einzelner Qualitätskriterien auf Ebene der einzelnen Evaluierung ist eine Analyse und Bewertung guter Evaluierungspraxis fehlerbehaftet, da eine Unterscheidung zwischen einem Qualitätskriterium, das nicht angewandt, angewandt, aber nicht dokumentiert, oder begründet nicht angewandt wurde, kaum möglich ist. Darüber hinaus wären Meta-Evaluierungen mit einer besseren Nachvollziehbarkeit besser durchzuführen. Da die international geltenden Standards Maximalstandards darstellen, ist eine Nichtanwendung von ausgewählten Standards bereits angelegt und stellt einen anerkannten Umgang in der Auseinandersetzung der Qualitätsstandards dar, der gelebt werden sollte. Entsprechend dem OECD DAC (OECD DAC, 2010, S. 5) sollten die Qualitätsstandards „umsichtig angewandt und an den lokalen beziehungsweise nationalen Kontext sowie die Ziele der jeweiligen Evaluierung angepasst werden“.

Empfehlung 3

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sollten – wenn noch nicht vorhanden – die Anwendung der auf Organisationsebene festgelegten Qualitätsstandards (Empfehlung 1) in den einzelnen Evaluierungen weiter verbessern, insbesondere die kaum oder teilweise angewandten Qualitätsstandards. Darüber hinaus sollte eine Anwendung oder (begründete) Nichtanwendung aller Qualitätsstandards auf Ebene jeder Evaluierung nachvollziehbar sein und von den Organisationen regelmäßig untersucht werden. (Ergebnis: 4.1, 4.2, 5.2, 5.3, 6.1, 7.1, 8.1, 9.2, 9.3)

b) Das BMZ sollte die staatlichen Organisationen zur Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung der relevanten Qualitätsstandards auf Ebene der Evaluierung anhalten.

Umsetzungshinweise:

- Sicherstellung der Anwendung der Qualitätsstandards in jeder Evaluierung:
  1) Da es Hinweise gibt, dass die Umsetzung von Qualitätssicherungsprozessen (unter anderem der „Einbindung der internen und externen Stakeholder*innen“) positiv mit der Anwendung ausgewählter Qualitätsstandards zusammenhängt, sollte gewährleistet werden, dass die Verantwortlichen der Evaluierungseinheiten/-stellen ausreichend geschult sind und genügend Ressourcen zur Verfügung haben, um diese Qualitätssicherungsprozesse angemessen durchzuführen. (Ergebnis: 12.1)
  2) Da Hinweise bestehen, dass die „Kompetenz der Gutachtenden“ – insbesondere die „Anzahl der Gutachtenden“ – mit einer besseren Anwendung ausgewählter methodischer Qualitätsstandards zusammenhängt, sollte von den Organisationen angedacht werden, für Evaluierungen – wenn möglich – Gutachtenden-Teams zu verpflichten. (Ergebnis: 12.1)
  3) Verschiedene Good-Practice-Beispiele der beteiligten Organisationen bezüglich ihrer Anwendung ausgewählter Qualitätsstandards finden sich im Bericht sowie im Onlineanhang (Abschnitt 4.1.1). Sie können als Vorlage für organisationsspezifische Maßnahmen herangezogen werden.
- Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätsstandards in jeder Evaluierung:
  1) Da einzelne Qualitätsstandards zu verschiedenen Zeitpunkten im Evaluierungsprozess und von unterschiedlichen Personen dokumentiert werden, sollten die Prozesse für die (Nicht-)Anwendung der einzelnen Qualitätsstandards auf Ebene der Evaluierung ausformuliert werden.
  2) Die Nachvollziehbarkeit sollte möglichst schlank und effizient aufgesetzt werden. Möglich wäre dies zum Beispiel über die Nachvollziehbarkeit der (Nicht-)Anwendung in den für jede Evaluierung nachvollziehbar umgesetzten Organisationsprozessen oder eine Dokumentation an einer übergeordneten Stelle in der Organisation (beispielsweise im Monitoring).[85]

Gemeinsames Lernen

Die strukturelle Heterogenität der teilnehmenden Organisationen spiegelt sich im unterschiedlichen Qualitätsverständnis und Grad der Anwendung einzelner Qualitätsstandards wider. Die dahinterliegenden unterschiedlichen Praktiken und Erfahrungen in der Identifikation, Verschriftlichung, Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätsstandards bieten den Organisationen die Möglichkeit, voneinander zu lernen. Da die Organisationen bei der Identifikation, Verschriftlichung, Sicherstel-

[85] Die Ergebnisse, Schlussfolgerungen und Empfehlungen der Evaluierungsfrage 2c für GIZ und KfW (Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW?) werden in Abschnitt 4.2.4 und Kapitel 5 dargestellt und an dieser Stelle nicht erneut aufgegriffen. Evaluierungsfrage 3 (Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen?) wurde für die staatlichen Durchführungsorganisationen nicht gesondert untersucht, da für sie keine abweichenden Wirkungszusammenhänge zwischen den untersuchten erklärenden Faktoren und den Qualitätsstandards angenommen wurden. Entsprechend werden an dieser Stelle ebenfalls keine weiteren Ergebnisse angeführt.

lung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätsstandards vielfältige Erfahrungen gesammelt haben, kann ein organisationsübergreifender Austausch (zum Beispiel entlang von Good-Practice-Beispielen) institutionelles Lernen und damit eine Verbesserung der Anwendung der Qualitätsstandards ermöglichen. Die beteiligten Organisationen erklärten bereits, dass der bisherige Erfahrungsaustausch eine gute Lernmöglichkeit sei, an den angeknüpft werden sollte. Auch vor dem Hintergrund eines zu entwickelnden BMZ-Leitlinien-Analyserasters ist ein systematischer Austausch der Organisationen mit dem BMZ zur Förderung eines gemeinsamen Standardbewusstseins sinnvoll.

Empfehlung 4

a) Die Evaluierungseinheiten/-stellen von BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR und PTB sowie Vertreter*innen von VENRO sollten ihre unterschiedlichen Erfahrungen in der Identifikation, Verschriftlichung, Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung aller Qualitätsstandards regelmäßig untereinander austauschen. Der Austausch sollte auch nicht beteiligte Organisationen integrieren und weitere Evaluierungstypen beinhalten – zum Beispiel dezentrale Evaluierungen –, um die Anwendung der Qualitätsstandards weiter zu verbessern. (Ergebnis: 4.2, 4.3)
b) Das BMZ sollte den Austausch zur Identifikation, Verschriftlichung, Sicherstellung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätsstandards mit und zwischen den Organisationen finanziell unterstützen.

Umsetzungshinweise:

- Ein regelmäßiger Austausch könnte zum Beispiel im Rahmen des jährlichen Treffens der Evaluierungseinheiten, in ausgewählten Netzwerken, Arbeitsgruppen und Foren erfolgen. Letztere können auch über den EZ-Kontext hinaus identifiziert werden. Dabei ist es gegebenenfalls sinnvoll, den Austausch in einem kleineren Rahmen durchzuführen, zum Beispiel nur mit Organisationen eines Haushaltstitels.
- Die finanzielle Unterstützung bezieht sich insbesondere auf Organisationen des Haushaltstitels „Private Träger“ und VENRO.

Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit

Die aktuelle Erfüllung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit zeigt ein positives Bild. Darüber hinaus hat sich die Anwendung dieser Qualitätskriterien seit der Meta-Evaluierung Nachhaltigkeit in GIZ- und KfW-Evaluierungen sehr deutlich erhöht. Die aktuelle Anwendung der Qualitätskriterien ist durchschnittlich vollständig erfüllt. Die Verbesserung der Anwendung der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit kann vermutlich unter anderem auf ihre Erkenntnisse und Empfehlungen sowie die mit Unterstützung des BMZ und des DEval umgesetzten Reformen der Evaluierungspraxis von GIZ und KfW zurückgeführt werden. Allerdings sind auch andere Erklärungen möglich, beispielsweise relativ leicht zu erfüllende Operationalisierungen der Qualitätskriterien oder eine veränderte Dokumentationsweise. Die untersuchten Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit fokussieren auf das methodische Vorgehen von GIZ- und KfW-Evaluierungen. Da das methodische Vorgehen auch in den BMZ-Leitlinien Evaluierung einen Schwerpunkt bildet, bietet sich ein Abgleich der Qualitätsstandards beider Dokumente und eine eventuelle Übernahme der Qualitätskriterien für das BMZ-Analyseraster an.

Die Empfehlungen können auch für nicht beteiligte Organisationen angemessen und nützlich sein.

Empfehlung 5

a) Das BMZ sollte im Zuge der Erarbeitung des Analyserasters für die in den BMZ-Leitlinien Evaluierung beschriebenen Qualitätsstandards (Empfehlung 1) die Übernahme der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit überprüfen und diese gegebenenfalls in das Analyseraster aufnehmen.
b) GIZ und KfW sollten (angelehnt an Empfehlung 5a) die Anwendung und (Nicht-)Anwendung der Qualitätskriterien aus der Meta-Evaluierung Nachhaltigkeit, die in ein BMZ-Analyseraster übernommen wurden, sicherstellen beziehungsweise verbessern und die Nachvollziehbarkeit der (begründeten) (Nicht-)Anwendung je Evaluierung gewährleisten. (Ergebnis: 10.1, 10.2, 11.1)

6. LITERATUR

AfrEA (2020), *The African Evaluation Guidelines 2020 Version*, African Evaluation Association, Washington, D. C.

Backhaus, K. et al. (2011), *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*, Springer, Berlin, 13., überarbeitete Auflage.

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

Beywl, W. und M. Niestroj (2009), *Das ABC der wirkungsorientierten Evaluation: Glossar - deutsch/englisch - der wirkungsorientierten Evaluation*, Univation - Inst. für Evaluation Dr. Beywl und Associates, Köln, 2., vollst. bearb. und erg. Aufl.

BMF (2020), *„Bundeshaushaltsplan 2020. Einzelplan 23. Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung“*, Bundesministerium der Finanzen, Berlin/Bonn.

BMZ (2006), *„Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen“*, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

BMZ (2007), *„Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit“*, BMZ Konzepte, Nr. 165, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

BMZ (2016), *„Richtlinien zu Förderung entwicklungswichtiger Vorhaben der politischen Stiftungen aus Kapitel 2303 Titel 68704“*, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

BMZ (2020), *„Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. BMZ-Orientierungslinie zum Umgang mit den OECD-DAC-Evaluierungskriterien in Evaluierungen der deutschen bilateralen Entwicklungszusammenarbeit“*, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

BMZ (2021), *„Evaluierung der* Entwicklungszusammenarbeit*: Leitlinien des BMZ“*, Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

Borrmann, A. et al. (1999), *Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit: Analyse, Bewertung, Reformen*, Nomos, Baden-Baden.

Borrmann, A. und R. Stockmann (2009), *Evaluation in der deutschen Entwicklungszusammenarbeit. Band 1: Systemanalyse*, Waxmann, Münster.

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

Caracelli, V. J. und L. J. Cooksy (2009), „Metaevaluation in Practice“, *Journal of MultiDisciplinary Evaluation*, Vol. 6, Nr. 11, S. 1–15.

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

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

Caspari, A. (2012), *„Meta-Evaluation, Evaluationssynthese, Evaluation Review and Systematic Review – eine Begriffsklärung“*, Fachhochschule Frankfurt am Main, Frankfurt am Main.

Church, C. und J. Shouldice (2002), *The Evaluation of Conflict Resolution Interventions: Framing the State of Play*, International Conflict Research, Derry/Londonderry.

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

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

DEval (2020), *„Evaluierungskriterien für Evaluierungen des Deutschen Evaluierungsinstituts der Entwicklungszusammenarbeit“*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, Bonn.

DEval (2021a), *„Ablauf einer DEval-Evaluierung - Zur Rolle der Referenzgruppe“*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval), Bonn.

DEval (2021b), *„DEval-Evaluierungen 2021 - 2023“*, Deutsche Evaluierungsinstitut der Entwicklungszusammenarbeit, Bonn.

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

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

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

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

GIZ (2018a), „Kontrolle ist gut: Höhere Standards für Evaluierungen“, *Deutsche Gesellschaft für Internationale Zusammenarbeit*, https://www.giz.de/de/mediathek/66304.html (zugegriffen 19.06.2022).

GIZ (2018b), *„Das Evaluierungssystem der GIZ: Zentrale Projektevaluierungen im BMZ-Geschäft“*, Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

von Gumppenberg, M.-C. et al. (2022), „Remote-Evaluation: Erfahrungen bei der Umsetzung von Remote-Evaluationen im Bereich der Entwicklungszusammenarbeit und Humanitären Hilfe – Genese, Stärken, Schwächen und Ausblick“, *Zeitschrift für Evaluation*, Nr. 1/22.

Hageboeck, M. et al. (2013), *„Meta-evaluation of quality and coverage of USAID evaluations 2009-2012“*, United States Agency for International Development, Washington, D. C.

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

Hundt, V. und B. Bräuer (2021), *„Remote und Semi-Remote – Erfahrungen bei der Durchführung zentraler Projektevaluierungen“*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

KfW (2019), *„15. Evaluierungsbericht 2017–2018“*, KfW Entwicklungsbank, Frankfurt am Main.

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

Krämer, M. und O. Almqvist (2019), *„Meta-Evaluierung und statistische Auswertung der Projektevaluierungen 2017 / 2018 - Teil II Statistische Auswertung“*, Bonn.

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

Kuckartz, U. (2014), *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*, Springer, Wiesbaden.

Lange, S. et al. (2020), *„Remote evaluation. Initial experience and recommendations“*, Physikalisch Technische Bundesanstalt, Braunschweig.

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

Mäder, S. (2020), *„Methoden als situierte Praxis: Die Gruppendiskussion in der Programmevaluation“*, Universität Hildesheim, Hildesheim.

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

Morgan, D. L. (1999), *The focus group guidebook*, SAGE, Thousand Oaks, CA.

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

OECD (2012), *„Evaluating Peacebuilding Activities in Settings of Conflict and Fragility - Improving Learning for Results“*, DAC Guidelines and Reference Series, Nr. 40, Organisation for Economic Co-operation and Development, Paris.

OECD (2013), *The DAC Network on Development Evaluation – 30 years of strengthening learning in development*, Organisation for Economic Co-operation and Development, Paris.

OECD (2021), *DAC-Prüfbericht über die Entwicklungszusammenarbeit: Deutschland 2021 (Kurzfassung): Wichtigste Ergebnisse und Empfehlungen*, Organisation for Economic Co-operation and Development, Paris.

OECD (2022), *„Recommendation of the Council on OECD Legal Instruments Public Policy Evaluation“*, Organisation for Economic Co-operation and Development, Paris.

OECD DAC (2002), *„Glossary of Key Terms in Evaluation and Results Based Management“*, Organisation for Economic Co-operation and Development, Development Assistance Committee, Paris.

OECD DAC (2010), *„Qualitätsstandards für die Entwicklungsevaluierung“*, Organisation for Economic Co-operation and Development, Development Assistance Committee, Paris.

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

Rodríguez Bilella, P. et al. (2016), *„Evaluation Standards for Latin America and the Caribbean“*, Renewable Energy for Latin America and the Caribbean, Fomento de Capacidades en Evaluación, Buenos Aires.

Seawright, J. und J. Gerring (2008), „Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options“, *Political Research Quarterly*, Vol. 61, Nr. 2, S. 294–308.

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

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

UNDAF (2017), *„UNDAC Companion Guidance: Theory of Change“*, United Nations Development Group, New York.

UNEG (2016), *„Norms and Standards for Evaluation“*, United Nations Evaluation Group, New York.

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

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

Weiber, R. und D. Mühlhaus (2010), *Strukturgleichungsmodellierung: eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*, Springer, Berlin.

Wencker, T. und I. Verspohl (2019), *„German Development Cooperation in Fragile Contexts“*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, Bonn.

World Bank (2020a), *„Summary of Lessons from an Informal Conversation: Challenges of Conducting Remote Evaluation Missions“*, World Bank, Washington, D. C.

World Bank (2020b), *„Summary of Lessons from the Knowledge Series on Using Technologies and Tools for Remote Data Collection: Experiences from Evaluation Offices of Multilateral Development Banks*“, World Bank, Washington, D. C.

7. ANHANG

7.1 Einordnung der Erkenntnisse für die Durchführungsorganisationen

Neben der Darstellung der Ergebnisse, Schlussfolgerungen und Empfehlungen entlang der Verpflichtungsgrundlagen der beteiligten Organisationen zur Anwendung der OECD-DAC- und/oder der DeGEval- sowie organisationsspezifischer Qualitätsstandards und den Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit (Kapitel 4 und 5) wurden die Erkenntnisse für die staatlichen Durchführungsorganisationen (DOs) BGR, GIZ, KfW und PTB im Folgenden gesondert ausgewertet und dargestellt.

Im Gegensatz zu den nichtstaatlichen Organisationen sind die Durchführungsorganisationen dazu verpflichtet, die BMZ-Leitlinien Evaluierung umzusetzen. Basierend auf dem Primat der instrumentellen Differenzierung der bilateralen staatlichen EZ müssen staatliche Organisationen BMZ-Leitlinien Evaluierung mittelbar umsetzen, für nichtstaatliche Organisationen gelten diese als Orientierung. In den Leitlinien (BMZ, 2021, S. 21) heißt es entsprechend: „Diese Leitlinien sind verbindlich für das BMZ und die staatlichen DOs (BGR, GIZ, KfW und PTB) […] Sie […] bieten Orientierung für deutsche zivilgesellschaftliche Organisationen – jeweils im Zusammenhang mit vertraglichen Vereinbarungen, Verwaltungsvorschriften oder Förderrichtlinien des BMZ hinsichtlich dieser Organisationen“. Darauf aufbauend sind die Ergebnisse, Schlussfolgerungen und Empfehlungen für die staatlichen Organisationen und das BMZ von einem besonderen Interesse. Aufgrund dessen werden die spezifischen Erkenntnisse für die DOs an dieser Stelle zusätzlich präsentiert (Kasten 9).

Kasten 9 Fazit zum Qualitätsverständnis und zur Anwendung der OECD-DAC- und der DeGEval-Qualitätsstandards durch die DOs

Der Fokus auf die Anwendung der OECD-DAC- und der DeGEval-Qualitätsstandards durch die vier staatlichen DOs ergab ein positives Bild. BGR, GIZ, KfW und PTB wandten die Qualitätskriterien jeweils zwischen 61 und 87 Prozent an. Die Anwendung konnte für drei DOs als „größtenteils erfüllt“ bewertet werden. Die GIZ kam als einzige Organisation auf einen durchschnittlichen Wert von 87 Prozent und erfüllt die Anwendung somit vollständig.

Die Anwendung der Qualitätskriterien unterschied sich sowohl innerhalb einer einzelnen als auch zwischen den staatlichen DOs zum Teil deutlich. Kritisch ist anzumerken, dass nicht alle Qualitätskriterien und ihre Anwendung in Organisationsdokumenten aufgeführt oder – im Falle einer bewussten Nichtanwendung – begründet dargelegt wurden. Die GIZ ist bislang die einzige Organisation, die in ihren internen Organisationsdokumenten auf die Anwendung einzelner (wenn auch nicht aller) Qualitätsstandards verweist, diese begründet und mit den Standarddokumenten des OECD DAC und der DeGEval in Verbindung bringt. Darüber hinaus wurden Nichtanwendungen auf Ebene der Evaluierung von allen Organisationen fast nie aufgeführt.

Wie ist das Qualitätsverständnis von Evaluierungen bei den beteiligten Organisationen in der deutschen EZ? (Evaluierungsfrage 1)[86]

- Das Qualitätsverständnis der vier staatlichen DOs beruhte auf den OECD-DAC- und den DeGEval-Standards.
- Die GIZ führte als einzige staatliche DO auch organisationsspezifische Qualitätsstandards an, die über die international geltenden Qualitätsstandards hinausgingen.

86 Die Ergebnisse, Schlussfolgerungen und Empfehlungen der Evaluierungsfrage 2c für GIZ und KfW (Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit in den Evaluierungen von GIZ und KfW?) werden in Abschnitt 4.2.4 und Kapitel 5 dargestellt und an dieser Stelle nicht erneut aufgegriffen. Evaluierungsfrage 3 (Inwieweit hängen länderkontext-, evaluierungs- und organisationsspezifische Faktoren mit der Anwendung der Qualitätsstandards zusammen?) wurde für die staatlichen Durchführungsorganisationen nicht gesondert untersucht, da für sie keine abweichenden Wirkungszusammenhänge zwischen den untersuchten erklärenden Faktoren und den Qualitätsstandards angenommen wurden. Entsprechend werden an dieser Stelle ebenfalls keine weiteren Ergebnisse angeführt.

Inwieweit zeigen sich Stärken und Schwächen bei der Anwendung der OECD-DAC- und der DeGEval-Standards und organisationsspezifischer Qualitätsstandards in den Evaluierungen der beteiligten deutschen EZ-Organisationen? (Evaluierungsfrage 2a und 2b)

a) Stärken zeigten sich in der Anwendung der Qualitätskriterien.

- Die 37 Qualitätskriterien wurden im Durchschnitt über alle staatlichen DOs hinweg zu 71 Prozent angewandt (damit liegt die Anwendung 3 Prozent über dem Durchschnitt aller Organisationen der Gruppe 1; Abschnitt 4.2.1). Die drei am häufigsten angewandten Qualitätskriterien waren „Evaluierungsethik (22)“, die durchschnittliche „Anwendung der OECD-DAC-Kriterien (33–37)“ und „Beschreibung des Evaluierungsgegenstands (1)“.
- Alle DOs wandten die Qualitätskriterien durchschnittlich mindestens zu über 60 Prozent an, wobei die GIZ sie als einzige Organisation durchschnittlich vollständig erfüllte (87 Prozent). Bei den drei anderen Organisationen war dies jeweils durchschnittlich größtenteils der Fall (KfW und PTB zu je 68 Prozent, BGR zu 61 Prozent).
- Bei allen DOs gab es Qualitätskriterien, deren Anwendung als „verfehlt“, „kaum erfüllt“ und/oder „teilweise erfüllt“ bewertet wurde.
- Die GIZ – als einzige staatliche Organisation mit organisationsspezifischen Qualitätsstandards – erfüllte diese vollständig.

b) Schwächen zeigten sich in der Identifikation und systematischen Verschriftlichung relevanter Qualitätskriterien in den Organisationsdokumenten und der Nachvollziehbarkeit der (Nicht-) Anwendung einiger Qualitätskriterien auf Evaluierungsebene.

- Die drei am wenigsten angewandten Qualitätskriterien waren „Berücksichtigung Kapazitätsentwicklung (26)“, „Darstellung der Angemessenheit des methodischen Vorgehens (7)“ und „Einbezug von Gutachtenden aus Partnerland (30)“.
- Die 15 in den Organisationsdokumenten untersuchten Qualitätskriterien wurden bei BGR (circa 53 Prozent) und GIZ (circa 73 Prozent) größtenteils, bei KfW und PTB (je zu circa 33 Prozent) teilweise thematisiert. Entsprechend waren die restlichen Qualitätskriterien nicht in den Organisationsdokumenten identifizierbar.
- In den Organisationsdokumenten wurde bei der Beschreibung der untersuchten Qualitätskriterien – mit Ausnahme der GIZ (bei fünf von 15 Qualitätskriterien) – von keiner Organisation ein expliziter Verweis auf die Standarddokumente gemacht; Evaluierungsprozesse zur Umsetzung der Qualitäts-standards wurden ebenfalls nicht immer verschriftlicht.
- Eine Nichtanwendung von Qualitätsstandards wurde in den Organisationsdokumenten ausschließlich von BGR und GIZ beim Qualitätskriterium „Veröffentlichung des Evaluierungsberichts (11)“ verschriftlicht. Eine begründete Nichtanwendung auf Ebene der Evaluierung wurde nur in Einzelfällen von einer DO dokumentiert.

Qualitätsverständnis der staatlichen Durchführungsorganisationen

Das Qualitätsverständnis der vier staatlichen DOs war angelehnt an die OECD-DAC- und/oder die DeGEval-Standards. Darüber hinausgehende organisationsspezifische Qualitätsstandards hatte nur die GIZ verschriftlicht. BGR, GIZ, KfW und PTB verpflichteten sich in ihren organisationsspezifischen Evaluierungsdokumenten beziehungsweise über eine Mitgliedschaft zur Anwendung der DeGEval-Standards und über die BMZ-Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit (BMZ, 2007) zur Anwendung der OECD-DAC-Standards. Die Verpflichtung zur Anwendung der OECD-DAC-Standards bestand bei BGR, KfW und PTB ausschließlich über die Verbindlichkeit dieser BMZ-Leitlinien.

Bewertung der Anwendung der Qualitätskriterien

Die Anwendung der Qualitätskriterien lag für alle vier DOs bei durchschnittlich 71 Prozent. Bei der GIZ waren es 87 Prozent, bei der KfW und der PTB jeweils 68 Prozent und bei der BGR 61 Prozent. Eine Nichtanwendung von Qualitätskriterien hingegen wurde – über Einzelfälle hinaus – auf Ebene der Evaluierung

bei keiner der vier DOs verschriftlicht. Dies kann als „kaum erfüllt" gewertet werden. Eine detaillierte Auswertung für die Anwendung der 37 Qualitätskriterien je Organisation findet sich in Tabelle 7. Insgesamt wurden bei der GIZ ungefähr 8 Prozent der Qualitätskriterien als „teilweise erfüllt" bewertet (drei von 31 Qualitätskriterien), bei der KfW und der PTB jeweils circa 26 Prozent (zehn von 31) und bei der BGR 43 Prozent teilweise oder schlechter erfüllt (16 von 31; Tabelle 8).

Tabelle 7 Überblick über die Ergebnisse der staatlichen DOs

QK	QS	SC	Ebene	Name Qualitätskriterium	BGR	GIZ	KfW	PTB	Durchschnitt
1	1	B & M	E	Evaluierungsgegenstand	93 %	100 %	98 %	96 %	97 %
/	2	B & M	E	Kontext	58 %	75 %	75 %	76 %	71 %
2	2a	B & M	E	Kontext der EZ-Maßnahme	76 %	72 %	97 %	85 %	82 %
3	2b	B & M	E	Kontext Ergebnisse	41 %	77 %	54 %	67 %	60 %
4	3	B & M	E	Wirkungs-zusammenhänge	41 %	87 %	38 %	43 %	52 %
5	4	B & M	E	Erkenntnisinteresse	67 %	99 %	43 %	81 %	72 %
6	5	B & M	E	Informationsquellen	83 %	98 %	77 %	79 %	84 %
7	6	B & M	E	Darstellung Methodik	2 %	66 %	20 %	37 %	31 %
8	7	B & M	E	Vorhandensein IR	0 %	100 %	87 %	100 %	72 %
9	8	B & M	E	Kohärenz	63 %	96 %	52 %	61 %	68 %
10	9	B & M	E	Leistungsbeschreibung	63 %	65 %	4 %	56 %	47 %
/	10	N	G	Zugänglichkeit	50 %	100 %	60 %	75 %	71 %
12	10a	N	O	Veröffentlichung Bericht	0 %	100 %	100 %	0 %	50 %
11	10b	N	O	Veröffentlichung Zusammenfassung	0 %	100 %	100 %	100 %	75 %
13	10c	N	E	Zugänglichkeit Stakeholder*innen	100 %	100 %	82 %	100 %	96 %
14	11	N	O	Kompetenz Gutachtende	75 %	100 %	75 %	100 %	88 %
15	12	N	O	Rechtzeitigkeit	50 %	75 %	75 %	75 %	69 %
/	13	N	E	Empfehlungen	48 %	71 %	36 %	83 %	59 %

QK	QS	SC	Ebene	Name Qualitätskriterium	BGR	GIZ	KfW	PTB	Durch-schnitt
16	13a	N	E	Adressat*innen Empfehlungen	24 %	72 %	18 %	99 %	53 %
17	13b	N	E	Handlung Empfehlungen	72 %	70 %	54 %	67 %	66 %
18	14	N	O	Evaluierungseffizienz	50 %	100 %	100 %	75 %	81 %
19	15	P, U & F	O	Meinungs-verschiedenheiten	100 %	75 %	/	0 %	58 %
/	16	P, U & F	G	Einbindung Stakeholder*innen	87 %	98 %	12 %	69 %	67 %
20	16a	P, U & F	E	Einbindung Stakeholder*innen (intern/extern)	87 %	98 %	12 %	69 %	67 %
21	16b	P, U & F	O	Einbindung Partner*innen*	100 %	100 %	100 %	50 %	88 %
22	17	P, U & F	O	Ethik	/	100 %	/	100 %	100 %
/	18	P, U & F	G	Unabhängigkeit	63 %	91 %	79 %	14 %	62 %
23	18a	P, U & F	E	Darstellung Unabhängigkeit Organisation	44 %	87 %	69%	21 %	55 %
24	18b	P, U & F	O	Darstellung Unvor-eingenommenheit	100 %	100 %	100 %	0 %	75 %
25	19	B & M	E	Zusammenfassung	89 %	76 %	44 %	89 %	75 %
26	20	P, U & F	E	Kapazitätsentwicklung	24 %	75 %	7 %	20 %	32 %
27	21	P, U & F	O	Gemeinschafts-evaluierungen	25 %	100 %	100 %	25 %	63 %
28	22	P, U & F	O	Partnerschaftliche Ansätze	25 %	50 %	50 %	75 %	50 %
/	23	P, U & F	G	Zusammensetzung Gutachtende	20 %	61 %	/	27 %	36 %
29	23a	P, U & F	O	Genderbalance	50 %	50 %	/	75 %	58 %
30	23b	P, U & F	E	GAs Partnerland	6 %	66 %	12 %	4 %	22 %

QK	QS	SC	Ebene	Name Qualitätskriterium	BGR	GIZ	KfW	PTB	Durch-schnitt
31	24	N	O	Ressourcen	100 %	100 %	75 %	75 %	88 %
32	25	N	O	Management-Response	50 %	50 %	75 %	100 %	69 %
/	26	K	E	OECD-DAC-Kriterien	98 %	100 %	100 %	100 %	99 %
33	26a	K	E	Relevanz-Kriterium	100 %	100 %	100 %	100 %	100 %
34	26b	K	E	Effektivitäts-Kriterium	100 %	100 %	100 %	100 %	100 %
35	26c	K	E	Effizienz-Kriterium	89 %	100 %	100 %	100 %	97 %
36	26d	K	E	Wirkungen-Kriterium	100 %	100 %	100 %	100 %	100 %
37	26e	K	E	Nachhaltigkeits-Kriterium	100 %	100 %	100 %	100 %	100 %
Durchschnitt					**61 %**	**87 %**	**68 %**	**68 %**	**71 %**

Quelle: DEval, eigene Darstellung
*Anmerkung: QK = Qualitätskriterium; QS = Qualitätsstandard; SC = Standardcluster; E = Evaluierungsebene; O = Organisationsebene; G = gemischt; B & M = Berichtslegung und Methoden; P, U & F = Partizipation, Unabhängigkeit und Fairness; N = Nutzbarkeit; K = kein Standardcluster. * Das Qualitätskriterium wurde nicht in die Berechnung für den Qualitätsstandard „Einbindung Stakeholder*innen" aufgenommen (siehe Abschnitt 4.2.1).*

Tabelle 8 zeigt die Ergebnisse zusammengefasst in Anzahl und Prozent je Qualitätskriterium je Bewertungs-maßstab.

Tabelle 8 Anzahl und Prozent der Qualitätskriterien je Bewertungsmaßstab und Organisation

	BGR		GIZ		KfW		PTB	
Bewertungsmaßstab	# QK	% QK	# QK	% QK	# QK	% QK	# QK	% QK
verfehlt	3	8 %	0	0 %	0	0 %	3	8 %
kaum erfüllt (> 0 % ≤ 25 %)	6	16 %	0	0 %	6	16 %	4	11 %
teilweise erfüllt (> 25 % ≤ 50 %)	7	19 %	3	8 %	4	11 %	3	8 %
größtenteils erfüllt (> 50 % ≤ 75 %)	5	14 %	9	24 %	8	22 %	10	27 %
vollständig erfüllt (> 75 % ≤ 99 %)	6	16 %	8	22 %	5	14 %	6	16 %
übertroffen (100 %)	9	24 %	17	46 %	11	30 %	11	30 %
keine Angabe in Onlinebefragung[a]	1	3 %	0	0 %	3	8 %	0	0 %
Summe	**37**	**100 %**	**37**	**100 %**	**37**	**100 %**	**37**	**100 %**

Quelle: DEval, eigene Darstellung

Anmerkung: QK = Qualitätskriterium. Bei der GIZ wurden alle drei organisationsspezifischen Qualitätskriterien vollständig erfüllt. [a] Im Gegensatz zur Dokumentenanalyse bestand in der Onlinebefragung für die Organisationen die Möglichkeit, keine Angaben zur Anwendung einzelner Qualitätskriterien zu machen, ohne dass dies als „kaum angewandt" gezählt wurde. Den Verantwortlichen der Evaluierungseinheiten/-stellen wurde dadurch ermöglicht, keine unangemessene Einschätzung zur Anwendung eines Qualitätskriteriums über alle Evaluierungen hinweg vornehmen zu müssen.

Die systematische Verankerung der Anwendung der Qualitätsstandards in den Organisationsdokumenten konnte bei KfW und PTB als „teilweise erfüllt", bei BGR und GIZ als „größtenteils erfüllt" bewertet werden. Eine systematische Verankerung zeichnete sich dadurch aus, dass jeder Qualitätsstandard in Organisationsdokumenten aufgenommen wurde und die jeweils für die Anwendung notwendigen Prozesse beschrieben waren, sodass die Anwendung aller relevanten Qualitätsstandards auf Ebene der Evaluierung sichergestellt werden kann. Diese Informationen können dann von den verschiedenen an der Konzeption, Durchführung und Verschriftlichung der Evaluierung Beteiligten herangezogen werden, um die Anwendung der Qualitätsstandards zu gewährleisten. Bei 15 der untersuchten Qualitätskriterien wurde eine systematische Verankerung untersucht. Dabei zeigte sich, dass bei bis zu zehn Qualitätskriterien keine Verankerung in den Organisationsdokumenten bestand (BGR: N = 7; GIZ: N = 4; KfW: N = 10; PTB: N = 10). Darüber hinaus wurde nur von zwei DOs eine Nichtanwendung des Qualitätskriteriums „Veröffentlichung des Evaluierungsberichts (11)" beschrieben und nur die GIZ hat bei fünf Qualitätskriterien einen Verweis auf die Standarddokumente ergänzt (BGR: N = 0; GIZ: N = 5; KfW: N = 0; PTB: N = 0; Tabelle 9). Die GIZ zeigte insgesamt die beste Verankerung der Qualitätsstandards in den Organisationsdokumenten.

Tabelle 9 Dokumentation der (Nicht-)Anwendung ausgewählter Qualitätskriterien in den Organisationsdokumenten der staatlichen Organisationen

	BGR	GIZ	KfW	PTB
Qualitätssicherung mit Inception Report (8)	A	A + B + V	A	A
Veröffentlichung des Evaluierungsberichts (11)	NA	A + B	NA	keine
Veröffentlichung der Zusammenfassung (12)	A	A + B	A	keine
Kompetenz der Gutachtenden (14)	A	A + B + V	A	A + B
Rechtzeitigkeit der Erkenntnisse (15)	keine	A + B + V	keine	keine
Evaluierungseffizienz (18)	keine	A + B + V	keine	keine
Transparenz von Meinungsverschiedenheiten (19)	keine	keine	keine	keine
Einbindung der Partner*innen (21)	A	A	A	keine
Evaluierungsethik (22)	A	A + B + V	keine	A + B
Darstellung Unvoreingenommenheit der Gutachtenden (24)	A	keine	keine	A
Berücksichtigung von Gemeinschaftsevaluierungen (27)	keine	A	keine	keine
Berücksichtigung partnerschaftlicher Ansätze (28)	keine	keine	keine	keine
Genderbalance im Team (29)	keine	keine	keine	keine
Ausreichende Ressourcen vorhanden (31)	keine	A	keine	keine
Vorhandensein einer Management-Response (32)	A	A	keine	A

Quelle: DEval, eigene Darstellung

Anmerkung: keine = keine Information zur Anwendung des Qualitätskriteriums gegeben; A = Information zur Anwendung des Qualitätskriteriums gegeben; NA = Information zur Nichtanwendung des Qualitätskriteriums gegeben; A + B = Information zur Anwendung des Qualitätskriteriums und Begründung gegeben; A + B + V = Information zur Anwendung des Qualitätskriteriums und Begründung gegeben, inklusive eines Verweises auf das Standarddokument

Die für alle Organisationen ausgesprochenen Empfehlungen (Kapitel 5) gelten auch für die vier staatlichen DOs, einzig der Umfang an Überarbeitung der eigenen Evaluierungspraxis unterscheidet sich zwischen BGR, GIZ, KfW und PTB. Da die Anwendung der Qualitätsstandards eine Grundlage der Evaluierungspraxis darstellt und in den Standarddokumenten bislang keinem Qualitätsstandard eine Priorität eingeräumt wird, ist die Dringlichkeit der Umsetzung der Empfehlungen 1 bis 3[87] für alle vier DOs hoch. Empfehlung 4 kann nachrangig bearbeitet werden, da sie eine – wenn auch wichtige – Option darstellt, die Anwendung der

87 Empfehlungen 1 (Identifikation und systematische Verschriftlichung der Anwendung von Qualitätsstandards auf Organisationsebene), 2 (Identifikation und systematische Verschriftlichung der Nichtanwendung von Qualitätsstandards auf Organisationsebene) und 3 (Anwendung der Qualitätsstandards auf Evaluierungsebene).

Qualitätsstandards zu verbessern. Da Empfehlung 5 die Empfehlungen 1 und 3 ergänzt, besteht für GIZ und KfW erst dann Handlungsbedarf, wenn entsprechende Qualitätskriterien der Meta-Evaluierung Nachhaltigkeit und weiterer Qualitätskriterien der BMZ-Leitlinien Evaluierung in ein BMZ-Analyseraster aufgenommen worden sind.

7.2 Auflistung der Qualitätskriterien

In Tabelle 10 werden die untersuchten Qualitätskriterien inklusive der ihnen zugeordneten Nummerierung und ihrer Namen in der Langversion dargestellt. Jedes Qualitätskriterium hat darüber hinaus einen Kurznamen, der zum Beispiel in den Abbildungen verwendet wird. Eine Liste der Kurznamen findet sich im Onlineanhang in Abschnitt 4.1.1.

Tabelle 10 Übersicht über die Nummerierung und Namen der untersuchten Qualitätskriterien

Nr. QK	Name QK (Name QS)	Nr. QK	Name QK (Name QS)
1	Beschreibung des Evaluierungsgegenstands	20	Einbindung der internen und externen Stakeholder*innen (Einbindung Stakeholder*innen)
2	Beschreibung des Kontexts der Entwicklungsmaßnahme (Einbindung des Kontexts)	21	Einbindung der Partner*innen (Einbindung Stakeholder*innen)
3	Berücksichtigung des Kontexts bei Ergebnissen (Einbindung des Kontexts)	22	Evaluierungsethik
4	Darstellung der Wirkungszusammenhänge	23	Darstellung organisationale Unabhängigkeit der Gutachtenden (Unabhängigkeit der Gutachtenden)
5	Beschreibung des Erkenntnisinteresses	24	Darstellung Unvoreingenommenheit der Gutachtenden (Unabhängigkeit der Gutachtenden)
6	Nachvollziehbarkeit der Informationsquellen	25	Informationsgehalt der Zusammenfassung
7	Darstellung der Angemessenheit des methodischen Vorgehens	26	Berücksichtigung Kapazitätsentwicklung
8	Qualitätssicherung mit Inception Report	27	Berücksichtigung von Gemeinschaftsevaluierungen
9	Kohärenz von Daten-Ergebnissen-Schlussfolgerungen	28	Berücksichtigung partnerschaftlicher Ansätze
10	Informationsgehalt der Leistungsbeschreibung	29	Genderbalance im Team (Zusammensetzung der Gutachtenden)
11	Veröffentlichung des Evaluierungsberichts (Zugänglichkeit)	30	Einbezug von Gutachtenden aus Partnerland (Zusammensetzung der Gutachtenden)
12	Veröffentlichung der Zusammenfassung (Zugänglichkeit)	31	Ausreichende Ressourcen vorhanden
13	Zugänglichkeit für Stakeholder*innen (Zugänglichkeit)	32	Vorhandensein einer Management-Response

Nr. QK	Name QK (Name QS)	Nr. QK	Name QK (Name QS)
14	Kompetenz der Gutachtenden	33	Anwendung OECD-DAC-Kriterium – Relevanz (Anwendung der OECD-DAC-Kriterien)
15	Rechtzeitigkeit der Erkenntnisse	34	Anwendung OECD-DAC-Kriterium – Effektivität (Anwendung der OECD-DAC-Kriterien)
16	Adressat*innen der Empfehlungen (Nützlichkeit der Empfehlungen)	35	Anwendung OECD-DAC-Kriterium – Effizienz (Anwendung der OECD-DAC-Kriterien)
17	Handlungsorientierung der Empfehlungen (Nützlichkeit der Empfehlungen)	36	Anwendung OECD-DAC-Kriterium – Wirkungen (Anwendung der OECD-DAC-Kriterien)
18	Evaluierungseffizienz	37	Anwendung OECD-DAC-Kriterium – Nachhaltigkeit (Anwendung der OECD-DAC-Kriterien)
19	Transparenz von Meinungsverschiedenheiten		

Quelle: DEval, eigene Darstellung
Anmerkung: QK = Qualitätskriterium; QS = Qualitätsstandard

7.3 Bewertungsskala für Evaluierungen des DEval

Kategorien	Verständnis
übertroffen	Befunde belegen eine über dem Anspruchsniveau liegende Anwendung des Qualitätskriteriums.
vollständig erfüllt	Befunde belegen die Erfüllung des Anspruchsniveaus bei der Anwendung des Qualitätskriteriums.
größtenteils erfüllt	Befunde, die die Erfüllung des Anspruchsniveaus bei der Anwendung des Qualitätskriteriums belegen, überwiegen.
teilweise erfüllt	Befunde, die die Erfüllung des Anspruchsniveaus bei der Anwendung des Qualitätskriteriums belegen, sind zum Teil vorhanden.
kaum erfüllt	Befunde, die die Erfüllung des Anspruchsniveaus bei der Anwendung des Qualitätskriteriums belegen, sind kaum vorhanden.
verfehlt	Befunde, die die Erfüllung des Anspruchsniveaus bei der Anwendung des Qualitätskriteriums belegen, sind nicht vorhanden.

7.4 Evaluierungsmatrix

	organisationsspez. Daten und Dokumente	wissenschaftliche und empirische Literatur	Interviews	Fokusgruppen-diskussionen
Evaluierungs-frage 1	X		X	
Evaluierungs-frage 2a	X		X	
Evaluierungs-frage 2b	X			
Evaluierungs-frage 2c	X			
Evaluierungs-frage 3	X	X		X

Quelle: DEval, eigene Darstellung

7.5 Zeitplan der Evaluierung

Zeitrahmen	Aufgaben
08/2020	Versand der Kurzmitteilung
12/2020	Referenzgruppensitzung zur Diskussion des Konzeptpapiers
12/2020–04/2021	Erstellung des Inception Reports
04/2021	Referenzgruppensitzung zur Diskussion des Inception Reports
04/2021–04/2022	Datenerhebung und Synthese der Ergebnisse
04/2022	Referenzgruppensitzung zu den Ergebnissen
04/2022–08/2022	Erstellung des Evaluierungsberichts
08/2022	Referenzgruppensitzung zu den Schlussfolgerungen und Empfehlungen
12/2022	Abschluss der Evaluierung nach Layout

7.6 Evaluierungsteam und Mitwirkende

Kernteam	
Dr. Kerstin Guffler	Teamleiterin
Dr. Nico Herforth	Teamleiter (Interim: 05/2021–02/2022)
Laura Kunert	Evaluatorin
Marian Wittenberg	Evaluator
Rebecca Maicher	Projektadministratorin
Mitwirkende	
Dr. Martin Noltze	interner Peer-Reviewer
Prof. Dr. Wolfgang Beywl	externer Peer-Reviewer
Prof Dr. Thomas Widmer	externer Gutachter für die Qualität von Evaluierungen
Lucia Citoler	studierende Beschäftigte
Rayan Doukali	studierender Beschäftigter
Annika Grotrian	studierende Beschäftigte
Christian Süper	studierender Beschäftigter
Maria Villa-Guillen	studierende Beschäftigte
Verantwortlich	
Amélie Gräfin zu Eulenburg	Abteilungsleiterin