Zeitschrift für Evaluation | 22. Jahrgang | 2023
Heft 1 | S. 12–38 | https://doi.org/10.31244/zfe.2023.01.02 |


# Herausforderungen und Erkenntnisse der organisationsübergreifenden Meta-Evaluierung von (Projekt-)Evaluierungen in der deutschen Entwicklungszusammenarbeit

*Kerstin Guffler*[1] | *Marian Wittenberg*[1] | *Laura Kunert*[1] | *Amélie Gräfin zu Eulenburg*[1]

**Zusammenfassung:** Die organisationsübergreifende Meta-Evaluierung des DEval untersuchte die Erfüllung von Qualitätsstandards bei 296 zentralen (Projekt-)Evaluierungen elf deutscher staatlicher und nichtstaatlicher Entwicklungszusammenarbeits-Organisationen. Dafür wurde ein Analyseraster erarbeitet, das aus der Überschneidung der OECD/DAC- und DeGEval-Qualitätsstandards bestand und somit für eine Vielzahl an Organisationen relevant ist. Die Ergebnisse zeigten, dass der überwiegende Teil der Qualitätsstandards durch die beteiligten Organisationen erfüllt wurde, dass der Erfüllung aber kein systematisches Qualitätsbewusstsein oder -verständnis zugrunde lag. Daher wird zukünftig eine explizite Verankerung der Qualitätsstandards in den Organisationsdokumenten sowie die Dokumentation der Erfüllung oder Nichterfüllung der Qualitätsstandards auf der Ebene einzelner Evaluierungen empfohlen.

**Schlagwörter:** Meta-Evaluierung, OECD/DAC-Standards, DeGEval-Standards, organisationsübergreifend

## Challenges and Findings of (Project) Evaluations Across German Organisations in the Area of Development Cooperation

**Abstract:** DEval's meta-evaluation examines the application of quality standards in 296 central (project) evaluations across eleven German governmental and non-governmental development cooperation organisations. For this purpose, an analysis grid was developed at the interface of quality standards defined by the OECD/DAC and the DeGEval, thus being relevant for a large number of organisations. The results show that the majority of given quality standards are applied by the organisations assessed, however application appears not to be based on a syste-

1 Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval)

matic awareness or understanding of evaluation quality. It is therefore recommended to explicitly define quality standards in organisational documents as well as to document application or non-application at the level of individual evaluations.

**Keywords:** Meta-Evaluation, OECD/DAC Standards, DeGEval Standards, Cross-organisational

## 1. Hintergrund

Dieser Beitrag stellt Ausschnitte aus der organisationsübergreifenden *Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen EZ* des DEval vor (Guffler/Kunert/Wittenberg/Herforth 2022a). So wie Evaluierungen von Maßnahmen der Entwicklungszusammenarbeit (EZ) deren Stärken und Schwächen untersuchen, reflektieren Meta-Evaluierungen systematisch die Stärken und Schwächen von Evaluierungen. Ziele sind hierbei, „den Evaluierungsprozess […] und die zukünftige Evaluierungsarbeit zu verbessern oder […] Informationen zur Glaubwürdigkeit der Ergebnisse für die Nutzenden zur Verfügung zu stellen" (Caracelli/Cooksy 2009: 2f.). In der EZ gibt es bereits eine Vielzahl an Meta-Evaluierungen, die diese Ziele verfolgten, beispielsweise Caspari (2010, 2011), FES (2015), Freimann/Krämer (2016, 2017), Hageboeck/Frumkin/Monschein (2013), HTSPE Limited (2011), Koy/Väth/Römling (2016), Krämer/Almqvist (2019), Mauthofer/Silvestrini (2018), Noltze/Euler/Verspohl (2018), Queiroz de Souza (2017), Silvestrini/Bäthge (2019), Silvestrini et al. (2018), UNFPA (2020) und Väth et al. (2022).

Meta-Evaluierungen untersuchen in der Regel die Qualität von Evaluierungen innerhalb einer Organisation. Die vorliegende Meta-Evaluierung des DEval hingegen war strategisch-systematisch ausgerichtet und untersuchte Evaluationen *mehrerer Organisationen*.[2] Wie im Mehrjährigen Evaluierungsprogramm (MEP) 2020–22 des DEval dargelegt, war das Ziel hierbei, den Fokus zu erweitern und die Stärken und Schwächen in der Evaluierungspraxis über Organisationen hinweg zu identifizieren (vgl. DEval 2020). Meta-Evaluierungen sind von sogenannten Meta-Analysen und Evaluierungssynthesen abzugrenzen. Meta-Analysen fassen Ergebnisse einer Vielzahl an Studien quantitativ zusammen und werten sie aus. In Evaluierungssynthesen wiederum wird eine Auswertung von Evaluierungen mit besonderem inhaltlichem Fokus (vgl. Caspari 2012) dargestellt.

Der Artikel ist entlang der Prozessphasen der DEval-Meta-Evaluierung aufgebaut und beschäftigt sich mit der Evaluierungsfrage: Inwieweit zeigen sich Stärken und Schwächen in der Erfüllung von international geltenden Qualitätskriterien in Evaluierungen der

2 *Die vorangegangene Meta-Evaluierung zur Nachhaltigkeit in der deutschen EZ* (vgl. Noltze et al. 2018) untersuchte ebenfalls Evaluierungen von mehr als einer Organisation, konkret der Deutschen Gesellschaft für internationale Zusammenarbeit (GIZ) und der KfW Entwicklungsbank (KfW).

deutschen EZ?[3] Anhand von Leitfragen wird im Folgendem zum einen die inhaltliche und methodische Herangehensweise inklusive der Ableitung der relevanten Qualitätskriterien aus international gültigen Qualitätsstandards beschrieben, konkret der Überschneidung von Standards des Entwicklungsausschusses der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD DAC) und der Gesellschaft für Evaluation (DeGEval). Zum anderen werden ausgewählte Herausforderungen und Erkenntnisse aufbereitet und Implikationen für Organisationen in der deutschen EZ abgeleitet.

## 2. Auswahl relevanter Organisationen und Evaluierungen

Im Folgenden wird die Identifikation der an der Meta-Evaluierung beteiligten Organisationen und die untersuchten Evaluierungen beschrieben.

*Wie wurden die beteiligten Organisationen identifiziert und ausgewählt?*
Die für die Meta-Evaluierung relevanten Organisationen wurden aus zwei grundlegenden Akteursgruppen der deutschen EZ ausgewählt, den staatlichen Durchführungsorganisationen und den nichtstaatlichen Organisationen. Durch das bestehende DEval-Mandat und die Möglichkeit einer Vollerhebung wurden die *vier staatlichen Durchführungsorganisationen* – GIZ, KfW, die Physikalisch-Technische Bundesanstalt (PTB) und die Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) – von Beginn an als beteiligte Organisationen gesetzt. In der Akteursgruppe der *nichtstaatlichen Organisationen* gestaltete sich die Identifizierung und Auswahl der beteiligten Organisationen komplexer. Um möglichst heterogene Organisationen zu untersuchen, wurde die diverse case method angewandt; hierbei wurden Auswahlkriterien herangezogen und sowohl Organisationen mit einer hohen und einer niedrigen Ausprägung der folgenden Kriterien aufgenommen. Dieses Vorgehen ermöglichte Aussagen zur Erfüllung hinsichtlich einer Bandbreite an unterschiedlichen Organisationen, jedoch keine Rückschlüsse auf die Verteilung der Erfüllung der Qualitätsstandards über die nichtstaatlichen Organisationen hinweg.

Für die Auswahl wurden die folgenden Kriterien angelegt: 1) *Höhe der BMZ-Zuwendungen* – unter der Annahme, dass verfügbare Ressourcen einen Einfluss auf die Anwendung von Qualitätsstandards haben, wurden zwei Organisationen mit den durchschnittlich höchsten und niedrigsten absoluten BMZ-Zuwendungen pro Jahr ausgewählt; 2) *relative Evaluierungstätigkeit* – unter der Annahme, dass Organisationen mit einer relativ hohen Evaluierungstätigkeit erfahrener waren und somit bessere Evaluierungen durchführten, wurde eine Organisation mit dem durchschnittlich niedrigsten und eine mit dem höchsten Verhältnis von BMZ-Förderungen zur Anzahl an Evaluierungen pro Jahr identifiziert; 3) *EZ-Haushaltstitel,* aus dem die Organisation

3 Einen Überblick über alle Analysen und Ergebnisse der organisationsübergreifenden Meta-Evaluierung finden sich im DEval-Bericht *Meta-Evaluierung von (Projekt-)Evaluierungen der deutschen Entwicklungszusammenarbeit* (vgl. Guffler et al. 2022a) und dem dazugehörigen Onlineanhang (Guffler et al. 2022b).

ihre EZ-Maßnahmen/Evaluierungen (mit-)finanziert und der über die für den Titel geltende Förderrichtlinie maßgeblich für ihre Evaluierungstätigkeit ist. Um das Spektrum an verschiedenen relevanten Haushaltstiteln abzudecken („Förderung entwicklungswichtiger Vorhaben der Sozialstruktur", „Förderung entwicklungswichtiger Vorhaben der politischen Stiftungen", „Förderung entwicklungswichtiger Vorhaben der Kirchen" und „Förderung entwicklungswichtiger Vorhaben privater deutscher Träger"), wurde mindestens eine Organisation je Haushaltstitel ausgewählt; 4) *Evaluierungshäufigkeit* – um über eine ausreichende Anzahl von Evaluierungen für die Meta-Evaluierung zu verfügen, wurde die Durchführung von circa zwei Evaluierungen pro Jahr und Organisation als Mindestanforderung definiert. Auf den Untersuchungszeitraum (Oktober 2016-Dezember 2020)[4] bezogen wurden dadurch ungefähr acht Evaluierungen pro Organisation für die Untersuchung gewährleistet.[5]

Entsprechend der Auswahlkriterien und einem anschließenden Konsultationsprozess mit den identifizierten Organisationen nahmen sieben nichtstaatliche deutsche EZ-Organisationen an der Meta-Evaluierung teil: CARE Deutschland e.V. (CARE), Deutscher Volkshochschulverband International (DVV), Deutsches Rotes Kreuz (DRK), Evangelisches Werk für Diakonie und Entwicklung e.V. (EWDE), Heinrich-Böll-Stiftung (hbs), Konrad-Adenauer-Stiftung (KAS) und MISEREOR.

*Wie wurden die Evaluierungen identifiziert und ausgewählt?*
Durch den im MEP des DEval festgelegten Fokus der Meta-Evaluierung auf *Projektevaluierungen*[6] waren diese als Gegenstand gesetzt. In der Klärungs- und Konzeptionsphase wurde darüber hinaus spezifiziert, dass *zentrale, das heißt in Deutschland von den Evaluierungseinheiten/-stellen (mit-)verantwortete Evaluierungen zwischen Oktober 2016 und Dezember 2020* untersucht werden, bei denen entweder die Evaluierung oder die EZ-Maßnahme vom *BMZ (mit-)gefördert* wurde. Dezentrale sowie strategische Evaluierungen, wie beispielsweise Förderbereichsevaluierungen, wurden ausgeschlossen.

Eine Herausforderung bestand in der Identifizierung aller relevanten Evaluierungen der beteiligten Organisationen im Untersuchungszeitraum. Nach einem zeitaufwändigen Prozess ergab sich das folgende Bild: Insgesamt wurden 849 Evaluierungen von den Organisationen (mit-)verantwortet, davon betrafen 576 vom BMZ (mit-)geförderte EZ-Maßnahmen (oder Evaluierungen). Der durchschnittliche Deckungsgrad lag somit bei 62 Prozent. Aus dieser *Grundgesamtheit* wurde im nächsten Schritt eine *nach Or-*

4 Bei CARE und der GIZ wurden Evaluierungen aus den Jahren 2018 bis 2020, bei der hbs Evaluierungen im Zeitraum von Januar 2016 bis Oktober 2020 aufgenommen.

5 Bei den Kriterien eins und zwei konnte eine Organisation nur einmal ausgewählt werden (das heißt, wenn eine Organisation über zwei Kriterien ausgewählt worden wäre, wurde sie nur über das erste Kriterium aufgenommen, beim zweiten Kriterium wurde die Organisation auf dem nächsten Platz ausgewählt). Kriterium eins wurde höher gewichtet als Kriterium zwei, da den Ressourcen einer Organisation ein großes Gewicht mit Bezug auf die Erfüllung der Qualitätsstandards zugesprochen wurde. Kriterium drei wurde ergänzend berücksichtigt und Kriterium vier wurde für alle beteiligten Organisationen herangezogen.

6 Als Projekte werden in diesem Zusammenhang als vom BMZ (mit-)geförderte EZ-Maßnahmen verstanden, die (zum Teil) in Partnerländern umgesetzt wurden. Da allerdings nicht alle beteiligten Organisationen eine Entwicklungsmaßnahme, sondern zum Teil auch Länderbüros oder Programme evaluierten, wurde auf das Präfix „(Projekt-)" im weiteren Verlauf verzichtet.

*ganisation und Jahr geschichtete, zufällige Stichprobe* gezogen.[7] Diese bestand aus 296 Evaluierungen und entsprach damit über die Organisationen hinweg durchschnittlich 75 Prozent der Grundgesamtheit. Die gewählten statistischen Auswahlparameter führten zu einer *disproportionalen Stichprobenziehung*, das heißt Organisationen mit einer höheren Anzahl an Evaluierungen gingen im Verhältnis mit weniger Evaluierungen in die untersuchte Stichprobe ein (Tabelle 1).

**Tabelle 1**: Anzahl an ausgewählten Evaluierungen je Organisation

| Organisation | GZ der (mit-) verantworteten Evaluierungen[a] | GG[b] | SP[c] | Anteil der GG je Organisation an GG aller Organisationen in % | Anteil der SP je Organisation an SP für alle Organisationen in % |
|---|---|---|---|---|---|
| BGR | 31 | 21 | 18 | 3,6 | 6,1 |
| CARE[d, e] | 6 | 6 | 6 | 1,0 | 2,0 |
| DRK | 64 | 20 | 17 | 3,5 | 5,7 |
| DVV | 56 | 20 | 17 | 3,5 | 5,7 |
| EWDE | 63 | 14 | 13 | 2,4 | 4,4 |
| GIZ[d] | 109 | 62 | 38 | 10,8 | 12,8 |
| hbs[f] | 27 | 22 | 19 | 3,8 | 6,4 |
| KfW | 239 | 230 | 68 | 39,9 | 23,0 |
| KAS | 39 | 20 | 17 | 3,5 | 5,7 |
| MISEREOR | 158 | 123 | 55 | 21,4 | 18,6 |
| PTB | 57 | 38 | 28 | 6,6 | 9,6 |
| | **849** | **576** | **296** | **100,0** | **100,0** |

*Anmerkungen:* GG=Grundgesamtheit; GZ=Gesamtzahl; SP=Stichprobe. [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 wurden; [c] Anzahl der untersuchten Evaluierungen; [d] Zahlen beziehen sich auf die Jahre 2018 bis 2020; [e] Vollerhebung, da weniger als 10 Evaluierungen vorlagen; [f] Zahlen beziehen sich auf den Zeitraum Januar 2016 bis Oktober 2020. Quelle: DEval, eigene Darstellung angelehnt an Guffler et al. (2022a: 22)

## 3. Identifikation relevanter Qualitätskriterien

Dieser Abschnitt stellt das Qualitätsverständnis der Meta-Evaluierung sowie die Identifikation und Messung der untersuchten Qualitätskriterien dar.

7 Die Stichprobe wurde nicht zufällig gezogen, da zum Zeitpunkt der Stichprobenziehung die Annahmen bestanden, dass die unterschiedliche Ressourcenausstattung der Organisationen (Organisation) und eine steigende Erfahrung mit der Anwendung von Qualitätsstandards über die Jahre hinweg (Jahre) Einfluss auf die Anwendung haben. Bei der Ziehung der Stichprobe wurde darauf geachtet, dass bei einem Konfidenzniveau von 95 Prozent eine Fehlerspanne von 10 Prozent (mögliche Abweichung der gefundenen Ergebnisse von ungefähr 10 Prozent vom realen Wert) nicht überschritten wird. Für Organisationen mit weniger als 10 Evaluierungen wurde eine Vollerhebung durchgeführt.

*Welches Qualitätsverständnis liegt der Meta-Evaluierung zugrunde?*
Um die Qualität der Evaluierungen umfassend zu untersuchen, befasste sich eine Evaluierungsfrage der Meta-Evaluierung von (Projekt-)Evaluierungen in der deutschen EZ mit dem Qualitätsverständnis der beteiligten Organisationen. Hier wurde deutlich, dass kein einheitliches oder übergeordnetes Qualitätsverständnis bestand. Allerdings verschriftlichten die meisten Organisationen in ihren Organisationsdokumenten (zum Beispiel Evaluierungskonzepte oder -leitlinien), durch Mitgliedschaften und/oder in den Vorgaben durch das BMZ (zum Beispiel Förderrichtlinien), sich an den OECD/DAC- und/oder DeGEval-Standards zu orientieren beziehungsweise diese einzuhalten.[8]

Diese beiden Qualitätsstandarddokumente wurden von Expertinnen und Experten in langjährigen Prozessen diskutiert, erarbeitet und herausgegeben und dienten in der Meta-Evaluierung als Grundlage für das Qualitätsverständnis. Sie vereinen teilweise inhaltlich überschneidende und teilweise unterschiedliche Aspekte. Somit wurde in der Meta-Evaluierung ein Analyseraster erstellt, das aus den oben genannten Qualitätsstandarddokumenten systematisch abgeleitete Qualitätskriterien enthält. Diese ließen sich drei Bereichen zuordnen: 1) der Überschneidung zwischen den OECD/DAC- und DeGEval-Standarddokumenten[9], 2) OECD/DAC-Standards ohne Überschneidung mit den DeGEval-Standards und 3) die OECD/DAC-Kriterien[10]. Der Fokus dieses Artikels liegt auf dem Bereich 1, der Überschneidung der Qualitätsstandarddokumente.

Die Messung der Qualität der untersuchten Evaluierungen erfolgte, indem die *Erfüllung der Qualitätsstandards* aus Bereich 1 untersucht und diese mit der Qualität gleichgesetzt wurde. Konkret wurde untersucht, ob und wenn ja inwieweit Belege dafür bestanden, dass die abgeleiteten Qualitätsstandards in den identifizierten Evaluierungen angewandt wurden. In diesem Zusammenhang ist darauf hinzuweisen, dass ein hohes Ausmaß der Erfüllung der Qualitätsstandards nicht zwangsläufig bedeutet,

8 Alle beteiligten Organisationen wurden anhand eines einheitlichen Analyserasters untersucht: sechs der elf beteiligten Organisationen dokumentierten, die OECD/DAC- und neun die DeGEval-Standards einhalten zu wollen. Zwei Organisationen bekannten sich im Untersuchungszeitraum nicht explizit zu einem der beiden Standarddokumente, waren aber bereit, sich entlang dieser Qualitätsstandards untersuchen zu lassen. Eine der beiden Organisationen trat im Untersuchungszeitraum der DeGEval als Mitglied bei. Einige Organisationen verfügten darüber hinaus über organisationsspezifische Qualitätsstandards und damit unabhängig von den OECD/DAC- oder DeGEval-Standards über weitere Qualitätsanforderungen an ihre Evaluierungen. Diese wurden nicht in den vorliegenden Artikel aufgenommen; Details zu ihren Inhalten und Ergebnissen finden sich in Guffler et al. (2022a, 2022b). Auch das DEval hat sich für seine Evaluierungen organisationsinternen Qualitätsstandards verschrieben (vgl. DEval 2018). Diese sind ebenfalls an die OECD/DAC- und DeGEval-Qualitätsstandards angelehnt, für die EZ-Organisationen jedoch nicht relevant und wurden daher bewusst nicht als Grundlage für die Meta-Evaluierung herangezogen.

9 Aus Sicht des Evaluierungsteams werden durch die identifizierten Teile der Überschneidung wesentliche Aspekte aus allen DeGEval-Standards und der meisten OECD/DAC-Standards abgebildet. Nicht dargestellt werden unter anderem die OECD/DAC-Qualitätskriterien *Berücksichtigung Kapazitätsentwicklung und Berücksichtigung von Gemeinschaftsevaluierungen.* Die Ergebnisse der beiden anderen Bereiche werden in Guffler et al. (2022a, 2022b) dargestellt.

10 Obwohl die Anwendung der OECD/DAC-Kriterien Teil der OECD/DAC-Standards ist (Standard 2.8), werden diese in der Meta-Evaluierung als separater Bereich aufgeführt, da alle Organisationen eine Erfüllung anstreben und dies schriftlich dokumentiert haben. Die OECD/DAC-Kriterien wurden nicht als eigener Bereich in den Artikel aufgenommen.

dass die Qualität der Evaluierung entlang alternativer Qualitätsverständnisse genauso (hoch oder niedrig) bewertet würde.

*Wie wurden die Qualitätskriterien identifiziert und gemessen?*
Die für das Analyseraster relevanten Qualitätskriterien wurden identifiziert, indem Textteile der Qualitätsstandards der beiden Standarddokumente mit den gleichen inhaltlichen Aspekten identifiziert, benannt und mit einem Qualitätskriterium hinterlegt wurden. Entsprechend wurden nicht alle Textstellen und damit inhaltlichen Aspekte der Standards in der Überschneidung berücksichtigt. Die Identifikation erforderte in vielen Fällen eine Interpretation durch das Evaluierungsteam, da der Detailgrad der formulierten Qualitätsstandards in den beiden Standarddokumenten unterschiedlich ist; während die OECD/DAC-Qualitätsstandards ausführlich formuliert sind, ist die Beschreibung in den DeGEval-Qualitätsstandards eher allgemein gehalten. Der Abgleich der Texte führte insgesamt zu 24 Qualitätskriterien, die aus Sicht des Evaluierungsteams nicht alle, aber wesentliche Aspekte der Qualitätsstandards aus den Standarddokumenten abbilden. Bei der Operationalisierung der Qualitätskriterien wurde – soweit möglich – auf die Inhalte der früheren DEval-Meta-Evaluierung Nachhaltigkeit (Noltze/Euler/Verspohl 2018) oder dem DEval-Monitoring der Systemprüfung (Lücking et al. 2015) zurückgegriffen. Ein Beispiel für die Identifikation und Messung eines Qualitätskriteriums ist in Tabelle 2 abgebildet.

**Tabelle 2:** Herleitung des Qualitätskriteriums *Beschreibung des Erkenntnisinteresses* aus OECD/DAC- und DeGEval-Qualitätsstandards

| Name | Standarddokument und Qualitätsstandard | Textteile mit inhaltlicher Überschneidung aus den beiden Standarddokumenten | Qualitätskriterium |
|---|---|---|---|
| Beschreibung Erkenntnisinteresse | OECD-DAC 2.1, 2.2, 2.7, 3.12 | „[…] Zweck […] der Evaluierung [wird] klar dargelegt" (OECD DAC 2010: 8)<br>„Die spezifischen Ziele der Evaluierung geben Aufschluss darüber, was mit der Evaluierung erreicht werden soll" (OECD DAC 2010: 8)<br>„Die Evaluierungsziele werden durch zweckdienliche und spezifische Evaluierungsfragen konkretisiert. […]" (OECD DAC 2010: 9)<br>„[…] Die ursprünglichen Evaluierungsfragen […] werden im Bericht dokumentiert" (OECD DAC 2010: 14) | Das Qualitätskriterium ist erfüllt, wenn<br>1. Zweck(e),<br>2. Ziel(e) und<br>3. Evaluierungsfrage(n)<br>der Evaluierung genannt werden. |
| | DeGEval G3 | „Zwecke, Fragestellungen […] der Evaluation […] sollen so genau dokumentiert und beschrieben werden, dass sie nachvollzogen und beurteilt werden können […]" (DeGEval 2016: 21) | |

*Anmerkungen:* Die Antwortmöglichkeiten für das Qualitätskriterium beinhalteten vier Stufen: Wenn keiner der drei im Qualitätskriterium genannten Punkte in der Evaluierung dargestellt wurde, wurde es als „nicht erfüllt" (1) eingestuft, wenn einer genannt wurde als „eher nicht erfüllt" (2), wenn zwei genannt wurden als „eher erfüllt" (3) und wenn alle drei genannt wurden als „vollständig erfüllt" (4). Details zu allen Überschneidungen und dem Kodierleitfaden mit allen Operationalisierungen finden sich im Onlineanhang von Guffler/Kunert/Wittenberg/Herforth (2022b).
Quelle: DEval, eigene Darstellung

*Wie können die identifizierten Qualitätskriterien inhaltlich gruppiert werden?*
In Anlehnung an die DeGEval-Standardgruppen wurden die Qualitätskriterien inhaltlich drei Standardclustern zugeordnet: 1) *Berichtslegung und Methoden*, 2) *Partizipation, Unabhängigkeit und Fairness* und 3) *Nutzbarkeit*. Die Namen der Standardcluster weisen Ähnlichkeiten mit der Benennung der DeGEval-Standardgruppen auf, stimmen mit ihnen aber nicht vollständig überein, da die untersuchten Qualitätskriterien inhaltlich auf der DeGEval- und OECD/DAC-Überschneidung aufbauten (Abbildung 1).

## 4. Datenerhebung und -analyse relevanter Qualitätskriterien

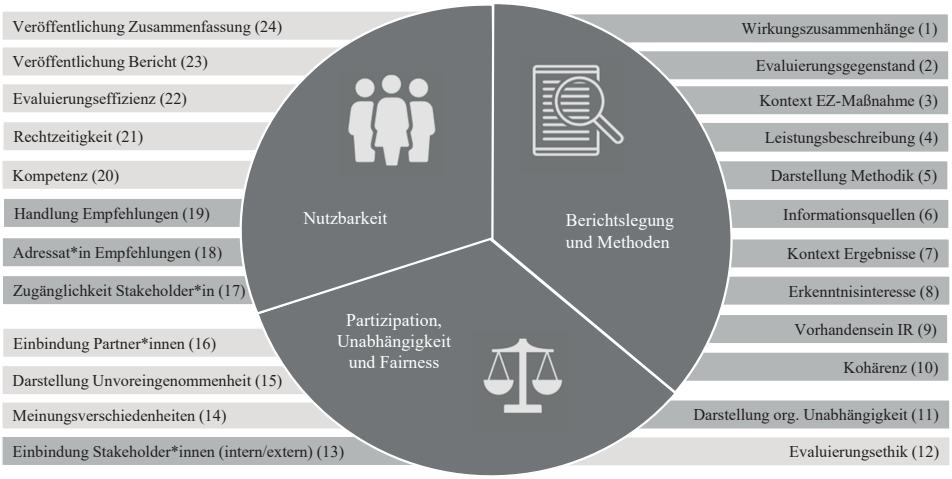
Nach Abschluss der Vorarbeiten – der Identifizierung und Operationalisierung der Qualitätskriterien sowie der Entwicklung eines umfangreichen Kodierleitfadens – erfolgte die Datenerhebung und -analyse der 24 Qualitätskriterien entlang der 296 Evaluierungen der beteiligten Organisationen. Dieses Vorgehen wird im Folgenden dargestellt.

*Wie konnten mehr als 1000 Dokumente einheitlich und intersubjektiv nachvollziehbar kodiert werden?*
Neben den Evaluierungsberichten und -anhängen wurden auch Leistungsbeschreibungen und Inception Reports kodiert (ungefähr 1000 Evaluierungsdokumente)[11]. Darüber hinaus wurden weitere relevante Organisationsdokumente (zum Beispiel Evaluierungskonzepte oder -leitlinien) untersucht. Die Herstellung eines einheitlichen Verständnisses der einzelnen Qualitätskriterien für eine Kodierung ist eine wichtige Voraussetzung. Im Kontext der Meta-Evaluierung war dies besonders komplex, da zum einen eine hohe Anzahl an Evaluierungen (beziehungsweise Evaluierungsdokumenten) vorlag und somit entsprechend viele Kodierende erforderlich waren, um im zeitlich gesetzten Rahmen zu bleiben. Insgesamt waren acht Kodierende beteiligt, wobei bis zu sechs Kodierende jeweils dasselbe Qualitätskriterium kodierten. Zum anderen erforderte die Heterogenität der beteiligten Organisationen mit ihren unterschiedlichen Arbeitsfeldern und Berichtsformaten einen anspruchsvollen und zeitaufwändigen Vorbereitungsprozess der Kodierung. Mithilfe von Kodierschulungen, Kodierleitfäden und -regeln sowie der Sicherstellung einer akzeptablen Interkoderreliabilität (Übereinstimmung der Kodierung zwischen den Kodierenden) wurde eine einheitliche Kodierung gewährleistet.[12] Die Qualitätskriterien wurden abschließend in

11 Eine Übersicht, welche Evaluierungsdokumente für die Kodierung welches Qualitätskriteriums herangezogen wurden, ist im Onlineanhang der Meta-Evaluierung zur Qualität der (Projekt-)Evaluierungen (Guffler et al. 2022b) einsehbar.

12 Die erste Phase der Kodierung stellte die Interkoderphase dar, in der jedes Qualitätskriterium in 30 Evaluierungen (circa 10 Prozent) von allen Kodierenden kodiert wurde. Über die Berechnung der Interkoderreliabilität am Ende der Interkoderphase wurde überprüft, ob unter den Kodierenden ein ausreichend einheitliches Verständnis der Qualitätskriterien vorlag (vgl. Döring/Bortz 2016). Für die Berechnung wurde der Koeffizient von Krippendorfs Alpha herangezogen, der durchschnittlich bei einem akzeptablen

**Abbildung 1**: Inhaltliche Zuordnung der Qualitätskriterien zu Standardclustern




*Anmerkungen:* IR=Inception Report; org.=organisationale; dunkler Balken=Qualitätskriterium wurde auf Ebene einzelner Evaluierungen untersucht; heller Balken=Qualitätskriterium wurde auf Organisationsebene über alle Evaluierungen hinweg untersucht.
Quelle: DEval, eigene Darstellung angelehnt an Guffler et al. (2022a: 13)

allen Evaluierungsdokumenten entlang von entweder ordinalen (1 = *nicht erfüllt*; 2 = *eher nicht erfüllt*; 3 = *eher erfüllt*; 4 = *vollständig erfüllt)* oder binären Bewertungsstufen (1 = *nicht erfüllt*; 4 = *vollständig erfüllt*) kodiert und mithilfe einer quantitativen Inhaltsanalyse (vgl. Döring/Bortz 2016) ausgewertet.

*Hätte es in der Dokumentenanalyse fälschlicherweise zu negativen Bewertungen kommen können?*
Das Evaluierungsteam bewertete, ob und inwieweit Belege erfasst werden konnten, die die Erfüllung eines Qualitätskriteriums darstellen. Dabei wurden zunächst nur die verschriftlichten Informationen aus den Evaluierungsdokumenten herangezogen. War die Erfüllung eines Qualitätskriteriums nicht belegbar beziehungsweise dokumentiert, wurde in einem ersten Schritt davon ausgegangen, dass dieses nicht erfüllt wurde. Es bestand allerdings die Möglichkeit, dass Organisationen Qualitätskriterien erfüllen, ohne dies in den einzelnen Evaluierungsdokumenten zu verschriftlichen. So gibt es Qualitätskriterien, deren Erfüllung erst nach Verschriftlichung des Evaluierungsberichtes stattfand und die folglich auch nicht im Evaluierungsbericht dokumentiert werden konnten. In anderen Fällen wurden Qualitätskriterien nicht für eine einzelne Evaluierung, sondern die gesamte Organisation festgelegt und somit nur auf Organisationsebene verschriftlicht.

Wert von 0,73 lag. Der niedrigste Wert lag bei 0,63, der höchste bei 0,89 (Werte von $1{,}00 \geq \alpha \geq 0{,}80$ waren als gut und Werte von $0{,}80 > \alpha \geq 0{,}67$ als akzeptabel anzusehen; vgl. Krippendorff 2012: 241).

Damit bestand das Risiko einer falsch-negativen Bewertung. Um dieses auszuschließen, wurden die Verantwortlichen der Evaluierungseinheiten/-stellen der beteiligten Organisationen zu den neun Qualitätskriterien online befragt, für die in der ersten Interkoderphase nicht ausreichend Informationen in den Evaluierungsdokumenten gefunden wurden (das heißt bei mehr als 24 der 30 Evaluierungen und mindestens neun Organisationen gab es durchgehend keine oder nur wenige schriftliche Informationen). Bei den neun Qualitätskriterien wurde davon ausgegangen, dass die Verantwortlichen der Evaluierungseinheiten/-stellen eine Einschätzung über die Häufigkeit der Erfüllung abgeben konnten, zum Beispiel, da ihnen Informationen aus Organisationsdokumenten vorlagen, die dem Evaluierungsteam nicht zur Verfügung standen. Dabei sollten die Befragten die durchschnittliche Häufigkeit der Erfüllung dieser Qualitätskriterien anhand einer fünfstufigen Skala bewerten (1=*nie*; 2=*selten*; 3=*teilweise*; 4=*überwiegend/häufig*; 5=*immer*).[13]

*Welche Gründe können dazu geführt haben, dass eine Erfüllung nicht auf Evaluierungsebene erfasst werden konnte?*
Eine fehlende nachvollziehbare Erfüllung der neun Qualitätsstandards auf Evaluierungsebene kann verschiedene Gründe haben: 1) die Qualitätskriterien bilden inhaltsschwere Themen ab (zum Beispiel „Evaluierungseffizienz"), die über alle Organisationen hinweg kaum angemessen operationalisiert werden konnten, da die Organisationen zum Teil stark abweichende Definitionen der Qualitätsstandards haben. Entsprechend würde eine Operationalisierung unter Umständen nur einer einzelnen Organisation gerecht werden. 2) Eine (begründete) Nichterfüllung auf Evaluierungs- und Organisationsebene ist aktuell in der Evaluierungspraxis noch nicht eingeführt. 3) Die Erfüllung wurde in Dokumenten verschriftlicht, die dem Evaluierungsteam nicht zugänglich waren (beispielsweise die „Kompetenz der Gutachtenden" in Bewerbungsunterlagen). 4) Eine Erfüllung wurde nur evaluierungsübergreifend in den Organisationsdokumenten festgehalten.

*Wie konnten begründete Nichterfüllungen von Qualitätskriterien berücksichtigt werden?*
Eine (begründete) Nichterfüllung eines Qualitätskriteriums ist grundsätzlich möglich, muss allerdings – den DeGEval-Standards entsprechend — „immer offen und nachvollziehbar etwa im Rahmen der Berichterstattung dokumentiert und begründet werden" (DeGEval 2016: 29). In der Meta-Evaluierung wurde infolgedessen anhand der Organisationsdokumente für die neun Qualitätskriterien der Onlinebefragung untersucht, ob eine aktive Auseinandersetzung der Organisation mit diesen Qualitätsstandards stattfand (da eine begründete Nichterfüllung auf Organisationsebene entsprechend für

13 Zusätzlich hatten die Verantwortlichen der Evaluierungseinheiten/-stellen die Möglichkeit, (explizit) keine Einschätzung über die Häufigkeit der Erfüllung einzelner Qualitätskriterien vorzunehmen. Eine fehlende Angabe wurde dabei nicht als eine Nichterfüllung gewertet, sondern als fehlender Wert aufgenommen. Die Verantwortlichen der Evaluierungseinheiten/-stellen wurden auch nach Organisationsdokumenten gefragt, in denen weitere Informationen verschriftlicht waren. Alle dem Evaluierungsteam zur Verfügung gestellten Dokumente wurden zusätzlich entlang der neun Qualitätskriterien kodiert.

jede einzelne Evaluierung gilt). Eine begründete Nichterfüllung von Qualitätskriterien wurde anschließend mit einer vollständigen Erfüllung gleichgesetzt und diese Logik gleichfalls auf die Bewertung der Erfüllung der OECD/DAC-Standards übertragen.[14] In den Organisationsdokumenten konnte in vier von 99 Fällen (neun Qualitätskriterien, die in der Onlinebefragung für elf Organisationen untersucht wurden) eine begründete Nichterfüllung festgestellt werden. Alle vier Fälle betrafen entweder die Veröffentlichung des Evaluierungsberichts oder der Zusammenfassung (dies entspricht ungefähr 4,0 Prozent der Fälle der Onlinebefragung).[15]

*Wann wurden die Qualitätskriterien als erfüllt bewertet?*
In der Meta-Evaluierung wurden die organisationsübergreifenden Mittelwerte je Qualitätskriterium auf Basis der jeweiligen Mittelwerte der Organisationen und unter Berücksichtigung der disproportionalen Stichprobenziehung mittels Designgewichten berechnet. Die Ergebnisse sind somit repräsentativ für die Organisationen und gemäß der jeweiligen Anzahl der Evaluierungen in der Stichprobe im Verhältnis zur Grundgesamtheit gewichtet (und nicht durch die Anzahl der Evaluierungen je Organisation verzerrt).[16] Für die Analyse der Erfüllung der einzelnen Qualitätskriterien je Organisation wurden die Mittelwerte (mit den Antwortmöglichkeiten 1 bis 4) mittels einer Normalisierung in Prozentwerte (auf einer Skala von 0 bis 100 Prozent) umgerechnet.[17]

Um die Erfüllung der Qualitätsstandards besser einordnen zu können, wurden im Austausch mit den beteiligten Organisationen Schwellenwerte festgelegt, die beschreiben, wann das jeweilige Qualitätskriterium als kaum, teilweise, größtenteils oder vollständig erfüllt gilt. Die Schwellenwerte[18] wurden in 25 Prozent-Schritten festgelegt (0 ≤ 25 Prozent = *kaum erfüllt*, > 25 ≤ 50 Prozent = *teilweise erfüllt*; > 50 ≤ 75 Prozent = *größtenteils erfüllt*; > 75 ≤ 100 Prozent = *vollständig erfüllt*). Wenn beispielsweise alle

14 Auch bei den OECD/DAC-Standards ist eine Dokumentation über eine begründete Nichterfüllung notwendig, um eine angemessene Untersuchung der Erfüllung sowie Querschnittsanalysen gewährleisten zu können.

15 Die Güte der Begründung wurde in den vier Fällen nicht bewertet. Vor dem Hintergrund der 2021 in Kraft getretenen BMZ-Leitlinien Evaluierung werden begründete Nichterfüllungen – insbesondere im Bereich der Veröffentlichung – zukünftig vermutlich stärker diskutiert werden, da diese laut BMZ vorzugsweise vollständig veröffentlicht werden sollen (BMZ 2021). Eine alternative Vorgehensweise wäre es, die (begründeten) Nichterfüllungen unabhängig von der Erfüllung zu analysieren und abzubilden.

16 Da Bewertungsstufen binär oder ordinal waren, bildeten die Mittelwerte Werte ab, die in den Bewertungsstufen nicht definiert waren.

17 So wurde beispielsweise der Mittelwert einer Organisation von 3,7 über die nachfolgende Formel in einen Prozentwert umgewandelt (Normalisierung): (3,7–1)/(4–1)*100=90 Prozent. Bei den Qualitätskriterien der Onlinebefragung bestanden Antwortmöglichkeiten zwischen 1 und 5, weshalb für sie die Normalisierung mittels (5–1) berechnet wurde. Die Normalisierung ermöglichte eine einheitliche Darstellung der Ergebnisse aus der Dokumentenanalyse und der Onlinebefragung und eine einfachere Einordnung der Ergebnisse entlang der Schwellenwerte.

18 Die Einteilung beruht auf der transparenten und nachvollziehbaren Herleitung, der Machbarkeit der Berechnung der Werte und letztlich der Akzeptanz durch die beteiligten Organisationen sowie den Bundesverband entwicklungspolitischer und humanitärer Nichtregierungsorganisationen (VENRO) und das BMZ.

Evaluierungsdokumente von einer beteiligten Organisation kodiert wurden und diese durchschnittlich bei 60 Prozent lagen, wurde der Schwellenwert für *größtenteils erfüllt* erachtet.[19] Dabei wurde berücksichtigt, dass die OECD/DAC- und DeGEval-Standards als *Maximalstandards* definiert sind. Maximalstandards stecken üblicherweise einen Referenzrahmen ab und beziehen sich auf den gesamten Bereich, in dem die Qualität einer Evaluierung sichergestellt werden kann. Es ist jedoch nicht gefordert, alle genannten Qualitätsstandards gleichzeitig und innerhalb einer jeden Evaluierung umzusetzen (vgl. DeGEval 2016). Es wird vielmehr davon ausgegangen, dass in der einzelnen Evaluierung eine (begründete) Auswahl getroffen wird.[20]

*Welche Stärken und Herausforderungen wies die Meta-Evaluierung auf?*
Aufgrund der oben beschriebenen kriterienbasierten Auswahl von möglichst heterogenen Organisationen für die Meta-Evaluierung, konnte bezüglich der Erfüllung von Qualitätskriterien eine große Bandbreite an Organisationen untersucht werden. Die Ergebnisse der nichtstaatlichen Organisationen sind nicht repräsentativ, dennoch lassen sich diese innerhalb dieser Bandbreite verorten. Darüber hinaus ist eine Übertragbarkeit der Erkenntnisse der Stichprobe der Evaluierungen auf die Grundgesamtheit je Organisation innerhalb der gewählten statistischen Kennwerte möglich.

Bei der Untersuchung der Qualitätskriterien mittels der Onlinebefragung bestanden Einschränkungen hinsichtlich der Triangulation der Daten und Methoden. So hätten zusätzlich zu den Verantwortlichen der Evaluierungseinheiten/-stellen die Einschätzungen ehemaliger Gutachtenden erhoben werden können. Hier hätte der Nutzen aber aufgrund der großen Anzahl der Evaluierungen in keinem angemessenen Verhältnis zum Aufwand gestanden.[21] Insgesamt wurden neun von 24 Qualitätskriterien mittels der Onlinebefragung erfasst (37,5 Prozent). Da die Qualitätskriterien aus der Onlinebefragung im Vergleich zu den Qualitätskriterien aus der Kodierung durchschnittlich rund sieben Prozent weniger erfüllt wurden und Organisationen sich auch mit einer 0-Prozent-Erfüllung (16 Prozent) bewertet haben, besteht kein Verdacht, dass sich die beteiligten Organisationen systematisch signifikant besser bewerteten. Eine unsystematische Überbewertung kann jedoch nicht ausgeschlossen werden. Daher ist eine Untersuchung der Qualitätskriterien auf Evaluierungsebene einer Einschätzung auf Organisationsebene vorzuziehen. Um das Vorge-

19 In Guffler et al. (2022a) wurden beteiligte Organisationen mit und ohne Verpflichtungsgrundlage zur Erfüllung der OECD/DAC- und/oder DeGEval-Qualitätsstandards getrennt voneinander dargestellt. Da es in diesem Artikel um die Darstellung der Erfüllung der Qualitätskriterien durch alle elf Organisationen insgesamt geht, und nicht um eine Bewertung der Erfüllung, wurden die Daten für den Artikel neu aufbereitet und alle elf Organisationen gemeinschaftlich abgebildet. Der Begriff „Erfüllung" wird in der Meta-Evaluierung und diesem Artikel unterschiedlich verwendet.

20 Womit erneut die Logik: „Erfüllung von Qualitätskriterien=begründete Nichterfüllung" zur Anwendung kommt.

21 Allein die Identifizierung der an den 296 Evaluierungen beteiligten Gutachtenden aus den letzten fünf Jahren wäre eine ressourcenintensive Aufgabe gewesen und hätte im Zeitrahmen der Meta-Evaluierung nicht umgesetzt werden können. Zudem hätten vermutlich aufgrund von Jobwechseln, Datenschutz und anderen Gründen nicht alle ehemaligen Gutachtenden kontaktiert werden können, so dass die Stichprobengröße reduziert worden wäre.

hen möglichst transparent darzustellen, wurden in den Abbildungen der Ergebnisse die Balken der Dokumentenanalyse und der Onlinebefragung in unterschiedlichen Schattierungen dargestellt.

Eine weitere Herausforderung bestand darin, dass die gewählten Operationalisierungen zwar wichtige Aspekte der Qualitätsstandards erfassten, den originären Qualitätsanspruch aber nicht abbilden konnten und somit an die Grenzen ihrer Messbarkeit stießen. So ist nur schwer einzuschätzen, wie viele Stakeholder(innen) in einer Evaluierung hätten eingebunden werden können oder ob die angewandten Methoden tatsächlich angemessen waren. Ebenso wäre bei den Qualitätskriterien „Darstellung Unvoreingenommenheit der Gutachtenden" und „Darstellung organisationale Unabhängigkeit der Gutachtenden" eine tiefergehende qualitative Analyse der Vergabeprozesse der Organisationen und/oder der Einstellungen der Gutachtenden erforderlich, um (den Anspruch an) die Qualitätskriterien vollständig zu erfassen. Zukünftig könnte diese Ausgangslage verbessert werden, wenn entsprechende Informationen direkt während der Evaluierung von den leitenden Gutachtenden oder Verantwortlichen der Evaluierungseinheiten/-stellen erhoben beziehungsweise festgehalten werden würden.

Da in der Meta-Evaluierung zum Teil Operationalisierungen herangezogen wurden, die nicht auf die Anwendungsform aller Organisationen zutrafen, wurde die Erfüllung für diese Qualitätsstandards bei einigen Organisationen unterschätzt (zum Beispiel bei dem Qualitätskriterium „Unvoreingenommenheit der Gutachtenden"). Um die herangezogenen Operationalisierungen der Meta-Evaluierung transparent darzustellen, wurde der Kodierleitfaden im Onlineanhang abgebildet (Guffler et al. 2022).

## 5. Ergebnisse

Im Folgenden werden die übergeordneten Ergebnisse und im Anschluss die Ergebnisse je Qualitätskriterium entlang der drei Standardcluster sowie daraus resultierende Schlussfolgerungen dargestellt. Details zu den Inhalten und der Messung der Qualitätskriterien finden sich im Kodierleitfaden des Onlineanhangs der Meta-Evaluierung (Guffler et al. 2022b).

*Übergreifende Ergebnisse*

In den elf untersuchten Organisationen wurden durchschnittlich 20 der 24 Qualitätskriterien (83 Prozent) *größtenteils* oder *vollständig* und 4 Qualitätskriterien (17 Prozent) *teilweise* erfüllt. Dies stellt eine Stärke in der Erfüllung der Qualitätskriterien von den Organisationen dar. Das Ausmaß, in dem die einzelnen Organisationen die Qualitätskriterien erfüllten, war sehr unterschiedlich. Konkret wichen bei 17 von 24 Qualitätskriterien (71 Prozent) die niedrigsten und höchsten durchschnittlichen Werte der beteiligten Organisationen über 50 Prozentpunkte voneinander ab. Besonders gut wurden die Qualitätskriterien *Beschreibung des Evaluierungsgegenstandes (2)* (96 Prozent), *Evaluierungsethik (12)* (94 Prozent) und *Nachvollziehbarkeit der Informa-*

*tionsquellen (6)*, *Zugänglichkeit für Stakeholder*innen (17)* und *Evaluierungseffizienz (22)* mit jeweils circa 83 Prozent erfüllt. Am unteren Ende – aber noch im Bereich *teilweise erfüllt* – befanden sich die Qualitätskriterien *Darstellung der Wirkungszusammenhänge (1)* (42 Prozent), *Darstellung der Angemessenheit des methodischen Vorgehens (5)* (35 Prozent) und *Darstellung der Unvoreingenommenheit der Gutachtenden (15)* (33 Prozent).

Schwächen zeigten sich insbesondere in der Nachvollziehbarkeit der Erfüllung von Qualitätskriterien in den Evaluierungsdokumenten. Bei neun von 24 Qualitätskriterien (38 Prozent) musste die Erfüllung über die Onlinebefragung auf Organisationsebene erfasst werden. Insbesondere war dies in den Standardclustern *Partizipation, Unabhängigkeit und Fairness* (4 von 6 Qualitätskriterien) und *Nutzbarkeit* (5 von 8 Qualitätskriterien) der Fall. Dies konnte – wie oben beschrieben – verschiedene Gründe haben, unter anderem, dass eine begründete Nichterfüllung oder eine Erfüllung nicht nachvollziehbar in den Evaluierungsdokumenten erfasst werden konnte. De facto ist die Verschriftlichung einer begründeten Nichterfüllung aktuell nicht Teil der Evaluierungspraxis. Ausnahmen bildeten in Einzelfällen begründete Nichterfüllungen der Qualitätskriterien *Veröffentlichung des Evaluierungsberichts (23)* und *Veröffentlichung der Zusammenfassung (24)* auf Ebene der Organisations- und nicht der Evaluierungsdokumente.

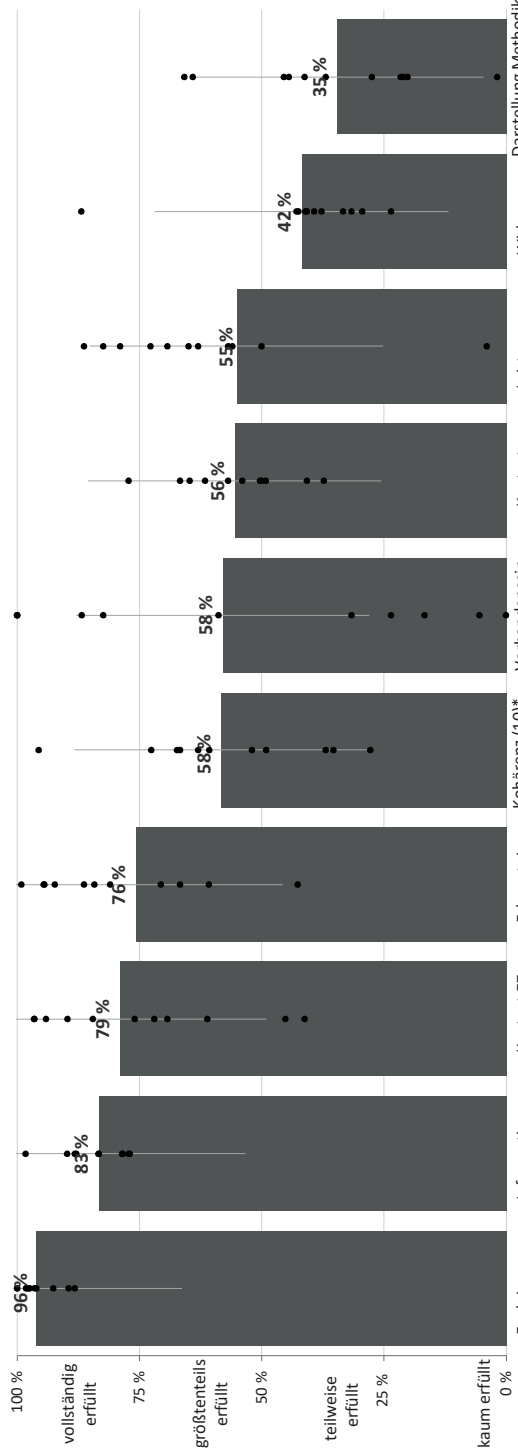
Als weitere Schwäche zeigte sich, dass in den Organisationsdokumenten kaum ein Bezug zu einzelnen Qualitätsstandards hergestellt wurde. Das heißt, dass die Organisationen die Qualitätsstandards nicht systematisch in den Organisationsdokumenten verankert und entsprechende Prozesse für ihre Erfüllung festgelegt haben. Konkret wurde bei den neun in der Onlinebefragung erhobenen Qualitätskriterien nur in 8,1 Prozent der Fälle in den Organisationsdokumenten auf konkrete Qualitätskriterien hingewiesen.

In der Onlinebefragung wurde darüber hinaus nach den Gründen für eine (potenzielle) Nichterfüllung von Qualitätskriterien gefragt. Diese lagen entweder auf Organisationsebene (zum Beispiel wurde die Nichterfüllung des Qualitätskriteriums *Transparenz von Meinungsverschiedenheiten [14]* mit der fehlenden Relevanz des Qualitätskriteriums für die Organisation begründet) oder auf Evaluierungsebene (beispielsweise wurde beim Qualitätskriterium *Veröffentlichung des Evaluierungsberichts (23)* der Schutz von durch die Evaluierung betroffenen Individuen beziehungsweise die Sicherstellung der Vertraulichkeit von mehreren beteiligten Organisationen angeführt). Zudem wurde eine Nichterfüllung auch auf die der Meta-Evaluierung zugrundeliegende Operationalisierung zurückgeführt, die bei einzelnen Organisationen nicht die organisationseigene Erfüllungsform widerspiegelte.

*Standardcluster Berichtslegung und Methoden*

Im Folgenden werden die Ergebnisse zur Erfüllung der Qualitätskriterien des Standardclusters *Berichtslegung und Methoden* dargestellt (Abbildung 2). Zu Beginn der Abschnitte zu den einzelnen Qualitätskriterien wird die durchschnittliche Erfüllung über die Organisationen hinweg in Klammern beschrieben, die auch in der Abbildung

**Abbildung 2:** Erfüllung der Qualitätskriterien im Standardcluster *Berichtslegung und Methoden*




*Anmerkungen:* Dunkler Balken=Qualitätskriterium wurde über Evaluierungsdokumente untersucht; *=Differenz zwischen dem niedrigsten Wert der Erfüllung einer beteiligten Organisation und dem höchsten Wert einer anderen betrug mehr als 50 Prozentpunkte.
Quelle: DEval, eigene Darstellung angelehnt an Güffler et al. (2022a: 39)

zu jedem Standardcluster über den Balken dargestellt ist.[22] Die Punkte auf den Säulen stellen das Ergebnis der einzelnen Organisationen dar. Darüber hinaus wird im Text Bezug auf die bestehenden Häufigkeiten der Bewertungsstufen (prozentualer Anteil an allen Evaluierungen in der Stichprobe) genommen.

- Beschreibung des *Evaluierungsgegenstands (2)* (durchschnittlich vollständig erfüllt): In fast allen Evaluierungen der elf Organisationen (91 Prozent) wurden Ziele, Zielgruppen und relevante Akteure der evaluierten EZ-Maßnahme beschrieben. Selten (9 Prozent) wurden entweder die Zielgruppen und/oder die relevanten Akteure nicht genannt.
- *Nachvollziehbarkeit der Informationsquellen (6)* (durchschnittlich vollständig erfüllt): In circa 92 Prozent der Evaluierungen wurde mindestens größtenteils nachvollziehbar beschrieben, welche Dokumente beziehungsweise Befragungen als Informationsquellen dienten. Der Detailgrad der verschiedenen Informationsquellen variierte jedoch stark – auch innerhalb einer Evaluierung. In acht Prozent der Evaluierungen wurden die Informationsquellen eher rudimentär beschrieben (zum Beispiel wurde aufgeführt, dass qualitative Daten erhoben wurden, aber nicht angegeben, welche Stakeholder(innen) im Rahmen welcher Evaluierungsfrage interviewt wurden).
- *Einbindung des Kontexts* beinhaltet die beiden Qualitätskriterien *Beschreibung des Kontextes der EZ-Maßnahme (3)* (durchschnittlich vollständig erfüllt) und *Berücksichtigung des Kontextes bei den Ergebnissen (7)* (durchschnittlich größtenteils erfüllt): Bei den elf Organisationen wurde in 69 Prozent der Evaluierungen der *Kontext der EZ-Maßnahme* als umfassend dargestellt, das heißt, es wurden mindestens zwei Kontextelemente (zum Beispiel politischer und wirtschaftlicher Kontext) umfänglich im Evaluierungsbericht und/oder den Evaluierungsberichtsanhängen beschrieben. Vor allem in den Evaluierungen der politischen Stiftungen wurde häufig eine umfangreiche und detaillierte Verortung des Kontextes der EZ-Maßnahme vorgenommen, vermutlich, da ihre EZ-Maßnahmen auf den politischen Kontext ausgerichtet sind. Eine vollständige Darstellung der fördernden und hemmenden Faktoren des Kontextes bei den Ergebnissen wurden dagegen (beispielsweise im Rahmen der Bewertung des OECD/DAC-Kriteriums Effektivität) deutlich weniger berücksichtigt (nur bei 30 Prozent). Als Good Practice wurde die Festlegung eines eigenen

22 Da Mediane robust gegenüber Ausreißern sind, wurden sie, zusätzlich zu den organisationsübergreifenden Mittelwerten, berechnet. Im Anschluss wurde untersucht, ob zwischen den beiden Berechnungen Abweichungen in den Erfüllungsstufen je Qualitätskriterium auftraten. Dabei wurde eine veränderte Erfüllungsstufe bei vier Qualitätskriterien identifiziert: Kontext EZ-Maßnahme (3) (Median=69,7 Prozent; die Erfüllungsstufe verringert sich auf „größtenteils erfüllt"), Meinungsverschiedenheiten (14) (Median=75,0 Prozent; die Erfüllungsstufe erhöht sich auf „vollständig erfüllt"), Darstellung Unvoreingenommenheit (15) (Median=0,0 Prozent; die Erfüllungsstufe verringert sich auf „kaum erfüllt") und Rechtzeitigkeit (21) (Median=75,0 Prozent; die Erfüllungsstufe erhöht sich auf „vollständig erfüllt"). Drei der vier Qualitätskriterien wurden auf Organisationsebene untersucht, daher lagen für sie entweder nur zwei oder fünf Antwortmöglichkeiten vor, so dass Abweichungen in den Erfüllungsstufen öfter auftraten. Bei dem Qualitätskriterium Kontext EZ-Maßnahme (3) hingegen lagen Mittelwert und Median nah aneinander (allerdings leicht über beziehungsweise unter dem Schwellenwert zu „vollständig erfüllt").

Textabschnitts im Rahmen einer vorgegebenen standardisierten Berichtsstruktur sowohl für die *Beschreibung des Kontextes der EZ-Maßnahme (3)* als auch für die *Berücksichtigung des Kontexts bei den Ergebnissen (7)* identifiziert.

- *Beschreibung des Erkenntnisinteresses (8)* (durchschnittlich vollständig erfüllt): In 51 Prozent der Evaluierungen wurden Zweck, Ziel und Evaluierungsfragen nachvollziehbar in den Evaluierungen beschrieben. Evaluierungsfragen und Ziele der Evaluierung (zum Beispiel die Wirksamkeit der EZ-Maßnahme zu untersuchen) wurden häufiger dargestellt als der Zweck der Evaluierung (zum Beispiel, dass untersucht wurde, ob eine EZ-Maßnahme verlängert werden soll).[23] Bei der Kodierung zeigte sich als Good Practice die Verankerung einer Ausformulierung der drei Punkte in der Leistungsbeschreibung.
- *Kohärenz von Daten-Ergebnissen-Schlussfolgerungen (10)* (durchschnittlich größtenteils erfüllt): In ungefähr 62 Prozent aller Evaluierungsberichte ergab sich die Mehrheit der Schlussfolgerungen kohärent aus den Ergebnissen der Datenanalyse; in rund 38 Prozent der Evaluierungsberichte war die Mehrheit der Schlussfolgerungen kaum oder nur teilweise nachvollziehbar (das heißt, es war unklar, aus welchen konkreten Daten/Ergebnissen die Schlussfolgerung abgeleitet wurde). Als Good-Practice-Beispiel konnten Evaluierungsberichte identifiziert werden, in denen die Schlussfolgerungen/Empfehlungen mit einer expliziten Referenz zu den relevanten Analyseergebnissen versehen waren.
- *Qualitätssicherung mit Inception Report (9)* (durchschnittlich größtenteils erfüllt): Das Qualitätskriterium bezieht sich auf das Vorhandensein eines Inception Reports und dieser lag für 59 Prozent aller Evaluierungen vor. Der Inception Report wurde unter anderem mit den beteiligten Organisationen im Rahmen einer Fokusgruppendiskussion als zentrales Merkmal des Qualitätssicherungsprozesses einer Evaluierung identifiziert, da über den Inception Report ein gemeinsames Verständnis zum Vorgehen und der Durchführung geschaffen wurde. Die Festlegung auf den Inception Report als Qualitätssicherungsprozess führte dazu, dass Organisationen, die andere Qualitätssicherungsprozesse verwendeten (wie beispielsweise die Kommentierung des Berichtsentwurfs durch Stakeholder(innen)), tendenziell zu niedrig bewertet wurden.
- *Informationsgehalt der Leistungsbeschreibung (4)* (durchschnittlich größtenteils erfüllt): Bei circa 62 Prozent der Leistungsbeschreibungen wurden vier oder mehr der acht nachfolgenden Aspekte beschrieben: Zweck, Nutzende, Ziele, Methoden, Zeitrahmen, verfügbare Mittel, Veröffentlichungsrechte (des Evaluierungsberichts) und mitwirkende Personen der Evaluierung. Vor allem die Aspekte Veröffentlichungsrechte und die Nutzenden der Evaluierung wurden häufig nicht genannt. Das Qualitätskriterium wurde von einer Organisation kaum erfüllt, da diese grundsätzlich keine Leistungsbeschreibung für ihre Evaluierungen anfertigte.
- *Darstellung der Wirkungszusammenhänge (1)* (durchschnittlich teilweise erfüllt): In vielen Evaluierungen wurden einzelne Elemente einer ‚Wirkungslogik'

23 Aus Gründen der Lesbarkeit wurden Werte in Grafiken und Text aufgrund von mehreren Nachkommastellen abweichend gerundet, das heißt die dargestellten Werte können zum Teil um maximal knapp ein Prozent von ihrem wahren Wert abweichen.

(Input-Output-Outcome-Impact) beschrieben. Die dargestellten Wirkungslogiken wurden allerdings in circa 72 Prozent der Evaluierungen unvollständig dargelegt, das heißt es fehlten entweder Elemente der Wirkungslogik oder es wurde nur für eines der Ziele der EZ-Maßnahme eine vollständige Wirkungslogik formuliert (19 Prozent). In 14 Prozent der Evaluierungen wurden die Wirkungslogiken für alle Ziele der EZ-Maßnahme vollständig ausformuliert, in weiteren circa 14 Prozent wurden keine Wirkungszusammenhänge beschrieben. Eine Nichterfüllung dieses Qualitätskriteriums wurde von manchen Organisationen durch fehlende Ressourcen beziehungsweise eine fehlende Verpflichtung durch den Auftraggebenden erklärt. In anderen Fällen wurde argumentiert, dass eine Wirkungslogik in weiteren – dem Evaluierungsteam nicht zur Verfügung gestellten – Dokumenten vorhanden sei. Die Verwendung einer Grafik für die Wirkungszusammenhänge und einer dazugehörigen textlichen Beschreibung einzelner Kettenelemente und Hypothesen wurde in diesem Zusammenhang als Good-Practice-Beispiel identifiziert.

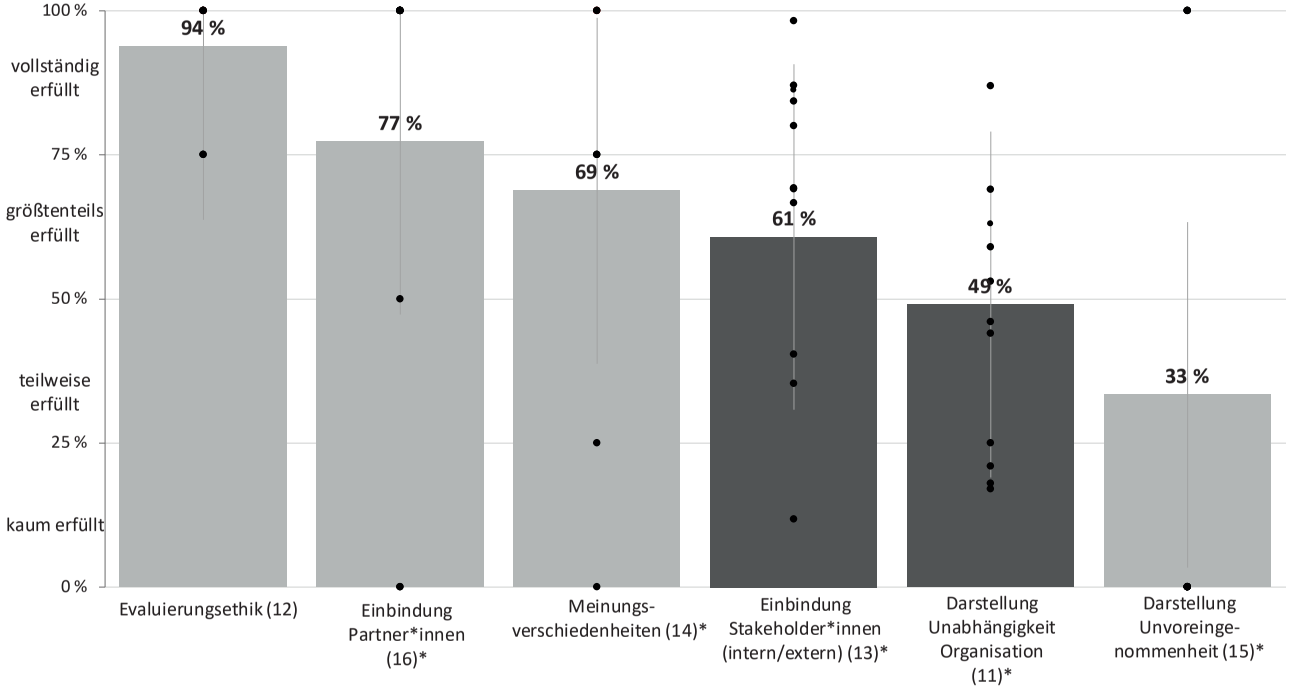
- *Darstellung der Angemessenheit des methodischen Vorgehens (5)* (durchschnittlich teilweise erfüllt): Mit diesem Qualitätskriterium wurde untersucht, ob beziehungsweise inwieweit eine nachvollziehbare Begründung vorlag, dass die genutzten Methoden angemessen und ob beziehungsweise inwieweit eine Diskussion der Limitationen des methodischen Vorgehens stattgefunden hatte. Die Darstellung einer nachvollziehbaren Begründung stellt den Nutzenden der Evaluierung wichtige Informationen zur Verfügung, um deren Validität und Reliabilität einschätzen zu können. Insgesamt wurde nur in 12 Prozent der Evaluierungen die Angemessenheit der Methoden ausführlich und nachvollziehbar begründet und ihre Limitationen dargestellt, in 14 Prozent wurden beide Aspekte knapp dargestellt. In 42 Prozent der Evaluierungen wurde nur einer der beiden Aspekte und in 32 Prozent keiner der beiden dargestellt. Ein Vergleich der Organisationen zeigte große Unterschiede. Als Good Practice konnte die systematische Bereitstellung von ausführlichen Berichtsabschnitten zu den Vor- und Nachteilen sowie den Limitationen der verwendeten Methoden angesehen werden.

*Standardcluster Partizipation, Unabhängigkeit und Fairness*

Im Folgenden werden die Ergebnisse der Erfüllung der Qualitätskriterien des Standardclusters *Partizipation, Unabhängigkeit und Fairness* dargestellt (Abbildung 3).

- *Evaluierungsethik (12)* (durchschnittlich vollständig erfüllt): Das Qualitätskriterium beschreibt, inwieweit Vorgaben zur Sicherheit und den Rechten der Evaluierungsbeteiligten erfüllt wurden. Da in der Onlinebefragung keine Spezifikation der Begriffe ‚Sicherheit und Rechte' vorgenommen wurde, konnte die Erfüllung verschiedenste inhaltliche Aspekte (beispielsweise der Schutz personenbezogener Daten oder Menschenrechte) und Formen der Umsetzung umfassen (zum Beispiel anhand von einer Schulung zu ethischem Handeln oder Dokumenten wie einem Verhaltenskodex). Bei acht Organisationen wurden in Organisationsdokumenten Vorgaben zur Sicherheit und den Rechten der Evaluierungsteilnehmenden verschriftlicht. Drei von ihnen verwiesen dabei direkt auf die Qualitätsstandards.

**Abbildung 3**: Erfüllung der Qualitätskriterien im Standardcluster *Partizipation, Unabhängigkeit und Fairness*




*Anmerkungen:* dunkler Balken=Qualitätskriterium wurde über Evaluierungsdokumente untersucht; heller Balken=Qualitätskriterium wurde über eine Onlinebefragung je Organisation untersucht. Bei den hellen Balken sind weniger als elf Punkte (Organisationen) abgebildet. Dies kann dadurch zustande kommen, dass in der Onlinebefragung mehrere Organisationen dieselbe Rückmeldung gegeben haben und somit ein Punkt mehrere Organisationen abbildet. *=Differenz zwischen dem niedrigsten Wert einer beteiligten Organisation und dem höchsten Wert einer anderen beteiligten Organisation betrug mehr als 50 Prozentpunkte.
Quelle: DEval, eigene Darstellung angelehnt an Guffler et al. (2022a: 45)

- *Einbindung der internen und externen Stakeholder*innen (13)* (durchschnittlich größtenteils erfüllt) und *Einbindung Partner*innen (16)* (durchschnittlich vollständig erfüllt): Das zuerst genannte Qualitätskriterium stellte die Einbindung von mindestens einem internen (zum Beispiel Auftraggebende) oder einem externen (zum Beispiel Partnerorganisationen im Partnerland) Stakeholder in verschiedenen Evaluierungsphasen dar.[24] In 39 Prozent der Evaluierungen wurde mindestens ein Stakeholder oder eine Stakeholderin in den gesamten Evaluierungsprozess einbezogen, in 40 Prozent der Evaluierungen in einzelnen beziehungsweise mehreren Evaluierungsphasen und in 21 Prozent wurden sie in keine Evaluierungsphasen einbezogen. Darüber hinaus ergab die Onlinebefragung, dass Partner(innen) in 76 Prozent der in der Stichprobe gezogenen Evaluierungen überwiegend/häufig oder immer einbezogen wurden (zweites Qualitätskriterium).[25] Als Gründe für eine Nichteinbindung von Partner(inne)n wurden von den Organisationen unter anderem ein zu hoher Koordinationsaufwand oder politische Sensibilitäten genannt. Als Good Practice zur Dokumentation dieses Qualitätskriteriums empfiehlt sich ein Zeitplan mit den einzelnen Evaluierungsphasen im Anhang des Berichts inklusive der Darstellung, ob und wenn ja, welche Stakeholder(innen) in die Evaluierung involviert waren.[26]
- *Transparenz von Meinungsverschiedenheiten (14)* (durchschnittlich größtenteils erfüllt): Da nur schwer ermittelt werden kann, ob Meinungsverschiedenheiten bestehen beziehungsweise bestanden, ist die Erfüllung dieses Qualitätskriteriums schwer einzuschätzen. Dennoch wurde von den Verantwortlichen der Evaluierungseinheiten/-stellen in durchschnittlich 69 Prozent der Fälle angenommen, dass Meinungsverschiedenheiten beschrieben wurden. Die transparente Darstellung von Meinungsverschiedenheiten innerhalb des Gutachtenden-Teams im Evaluierungsbericht wurde von den Organisationen allerdings als eher nicht erstrebenswert angesehen. Vielmehr wurde die Meinung vertreten, dass in der Zusammenarbeit mit externen Gutachtenden grundsätzlich Konsens angestrebt würde.
- *Unabhängigkeit der Gutachtenden* bestand aus der *Darstellung organisationale Unabhängigkeit der Gutachtenden (11)* (durchschnittlich teilweise erfüllt) und der *Darstellung Unvoreingenommenheit der Gutachtenden (15)* (durchschnittlich teilweise erfüllt): Die *Darstellung der organisationalen Unabhängigkeit (11)* war in der Hälfte aller Evaluierungen gegeben, da die Gutachtenden weder politisch, operativ noch beratend an der evaluierten EZ-Maßnahme beteiligt waren noch der Zielgruppe der EZ-Maßnahme angehörten. Die andere Hälfte wurde als organisational abhängig

24 Zum Beispiel, ob in der Konzeptionsphase Evaluierungsfragen gemeinsam erarbeitet wurden, in der Durchführungsphase diskutiert wurde, welche Zielgruppen befragt wurden, oder ob in der Berichtslegungsphase Einschätzungen zu den ersten Ergebnissen besprochen wurden. Stakeholder(innen) der Evaluierung waren Personen, Personengruppen oder Organisationen, die etwas zu verlieren oder zu gewinnen haben (vgl. Beywl/Niestroj 2009).

25 Da bei dem Qualitätskriterium *Einbindung der internen und externen Stakeholder*innen* (13) keine Unterscheidung zwischen internen und externen Stakeholder(inne)n gemacht wurde und überwiegend interne Stekeholder(innen) kodiert wurden, wurden zusätzliche Informationen über das Qualitätskriterium *Einbindung Partner*innen* (16) in der Onlinebefragung erhoben.

26 Bei der Darstellung der Stakeholder*innen ist der Schutz personenbezogener Daten zu berücksichtigen.

bewertet, wobei von diesen nur circa 17 Prozent der Evaluierungen schriftliche Rückschlüsse auf die (Un-)Abhängigkeit zuließen. Die restlichen 33 Prozent wurden als organisational abhängig bewertet, da keine Informationen bestanden. Die persönliche *Unvoreingenommenheit der Gutachtenden (15)* wurde von drei Organisationen über eine unterschriebene Erklärung zur Unabhängigkeit formal gewährleistet. Die übrigen Organisationen behalfen sich mit alternativen Anwendungsformen wie umfangreichen Einstellungsgesprächen oder Verträgen, die die Unvoreingenommenheit gewährleisten sollten. Bei diesen Qualitätskriterien bestanden Grenzen in der Messung, da nur schwer feststellbar gewesen wäre, ob Gutachtende beispielsweise für die evaluierte EZ-Maßnahme tätig waren.

*Standardcluster Nutzbarkeit*

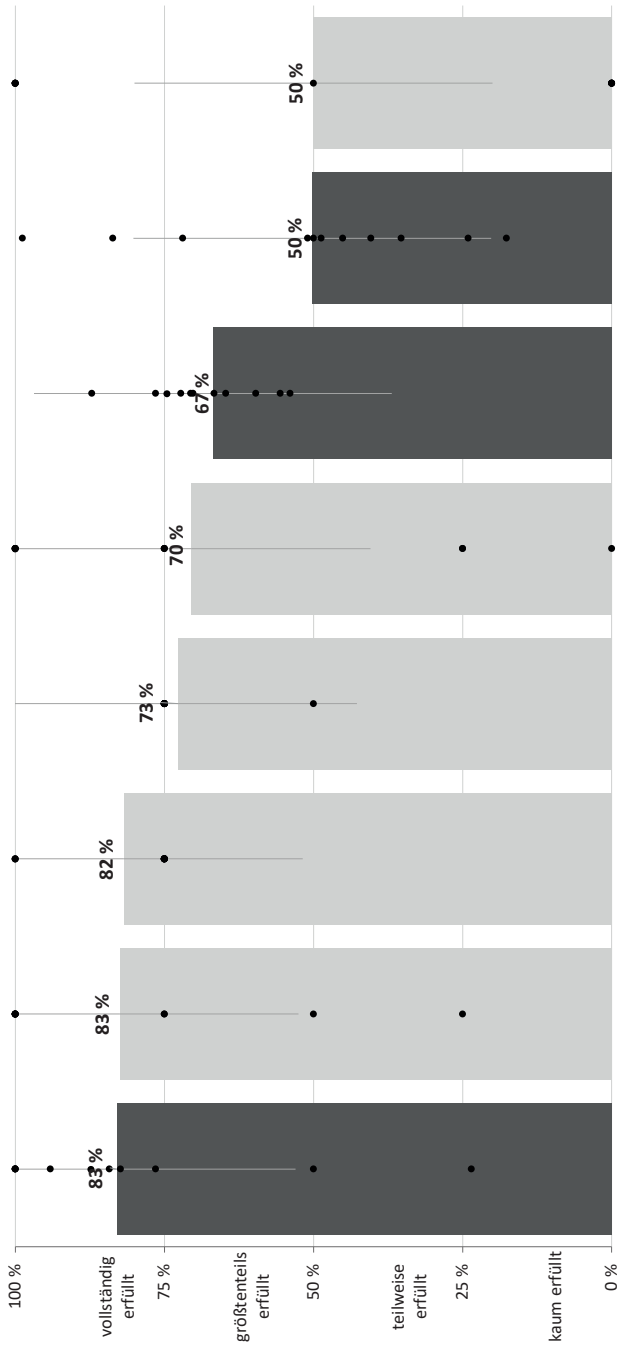
Im Folgenden werden die Ergebnisse zur Erfüllung der Qualitätskriterien des Standardclusters *Nutzbarkeit* dargestellt (Abbildung 4).

- *Zugänglichkeit* beinhaltete drei einzelne Qualitätskriterien: 1) die *Zugänglichkeit der Evaluierungsergebnisse für Stakeholder*innen (17)* (durchschnittlich vollständig erfüllt), 2) die *Veröffentlichung der Zusammenfassung des Evaluierungsberichts (24)* (durchschnittlich größtenteils erfüllt) und 3) die *Veröffentlichung des* (gesamten) *Evaluierungsberichts (23)* (durchschnittlich teilweise erfüllt). Die Zugänglichkeit der Evaluierungsergebnisse durch die Stakeholder(innen) erfolgte beispielsweise über den finalen Evaluierungsbericht, eine Zusammenfassung oder in Form einer abschließenden Präsentation. Fast alle Organisationen gaben die Rückmeldung, dass sie zum Schutz der Evaluierungsbeteiligten und Partner(innen) auf eine Veröffentlichung der vollständigen Berichte über eine extern zugängliche Webseite verzichteten,[27] eine Organisation erstellte zwei separate Teile des Berichtes und veröffentlichte den Teil ohne sensible Informationen. Die politischen Stiftungen veröffentlichen ihre Evaluierungen aufgrund der Sensibilität des politischen Kontextes, in dem sie arbeiten sowie der sich daraus ergebenden besonderen Schutzbedürftigkeit der Partner(innen) in der Regel nicht. Im Gegensatz zum Evaluierungsbericht wurde die Zusammenfassung der Evaluierung öfter über die Website der jeweiligen Organisation veröffentlicht. Obwohl bei diesen Qualitätskriterien eine begründete Nichterfüllung vorlag, ist dies qualitativ nicht mit einer tatsächlichen Erfüllung gleichzusetzen.[28]
- *Evaluierungseffizienz (22)* (durchschnittlich vollständig erfüllt): Die Mehrheit der Organisationen gab in der Onlinebefragung an, das Verhältnis zwischen Aufwand und Nutzen der Evaluierungen – häufig oder immer – zu berücksichtigen.

27 Bei den politischen Stiftungen wurde – wenn auch nicht mit eindeutigem Bezug zu den Standards – für dieses Qualitätskriterium in der BMZ-Förderrichtlinie eingeräumt, dass ein zurückhaltender Umgang mit dem DAC-Evaluierungsprinzip Transparenz möglich sei (vgl. BMZ 2016).

28 Die 2021 verabschiedeten Evaluierungsleitlinien des BMZ (2021) legen mittlerweile mit (implizitem) Bezug zu dem Qualitätskriterium *Veröffentlichung des Evaluierungsberichts (23)* fest, dass „die Berichte […] im Sinne der Transparenz vorzugsweise vollständig veröffentlicht [werden]" (BMZ 2021: 21).

**Abbildung 4:** Erfüllung der Qualitätskriterien im Standardcluster Nutzbarkeit




*Anmerkungen:* dunkler Balken=Qualitätskriterium wurde über Evaluierungsdokumente untersucht; heller Balken=Qualitätskriterium wurde über eine Onlinebefragung je Organisation untersucht. Bei den hellen Balken sind weniger als elf Punkte (Organisationen) abgebildet. Dies kann dadurch zustande kommen, dass in der Onlinebefragung mehrere Organisationen dieselbe Rückmeldung gegeben haben und somit ein Punkt mehrere Organisationen abbildet. *=Differenz zwischen dem niedrigsten Wert einer beteiligten Organisation und dem höchsten Wert einer anderen beteiligten Organisation betrug mehr als 50 Prozentpunkte.
Quelle: DEval, eigene Darstellung angelehnt an Guffler et al. (2022a: 50)

Einige Organisationen verwiesen in diesem Zusammenhang auf eine repräsentative Stichprobenziehung aus allen EZ-Maßnahmen, die eine wirtschaftliche Abdeckung der Evaluierungstätigkeit gewährleiste. Andere Organisationen entschieden sich für Schreibtischstudien, wenn der erwartete Nutzen einer Evaluierung die Durchführung von ressourcenintensiven Reisen nicht rechtfertigte. Einzelne Organisationen erklärten, dass die aus den Evaluierungen gewonnenen Erkenntnisse für das weitere Management der EZ-Maßnahmen sowie der Entwicklung neuer Maßnahmen immer nützlich seien und den Aufwand somit immer rechtfertigte. Keine Organisation ging auf die Evaluierungseffizienz während ihrer Durchführung ein, sondern nur auf die vorgelagerte Entscheidung, ob eine Evaluierung durchgeführt werden solle.[29]

- *Kompetenz der Gutachtenden (20)* (durchschnittlich vollständig erfüllt): Die Organisationen gaben in der Onlinebefragung an, dass bei 80 Prozent aller in der Stichprobe gezogenen Evaluierungen die Gutachtenden überwiegend/häufig oder immer über die notwendige und relevante Expertise verfügen. Dazu zählten Kenntnisse und Erfahrungen zu Land und Region, zur Durchführung und den Methoden von Evaluierungen sowie fach- und sektorspezifische Expertise. Einige Organisationen priorisierten Evaluierungs- sowie Fach- und Sektorkenntnisse und gaben an, Landes- oder Regionalkenntnisse auf anderen Wegen sicherzustellen, sollten diese bei den Gutachtenden fehlen. Mangelnde methodische Kenntnisse oder Erfahrung mit der Durchführung von Evaluierungen wurde hingegen versucht, durch eine stärkere Unterstützung der Evaluierungsverantwortlichen der Organisation auszugleichen.
- *Rechtzeitigkeit der Erkenntnisse (21)* (durchschnittlich größtenteils erfüllt): Nach Angabe der Organisationen aus der Onlinebefragung wurden bei 91 Prozent der Evaluierungen der vereinbarte Zeitplan überwiegend/häufig eingehalten. Ausnahmen wurden durch externe (zum Beispiel die Sicherheitslage im Partnerland) und/oder interne Faktoren (zum Beispiel die Verfügbarkeit von Partner(inne)n) begründet. Bei nur drei Organisationen konnten in Organisationsdokumenten Bezüge zur Rechtzeitigkeit von Evaluierungen gefunden werden.
- *Nützlichkeit der Empfehlungen* beinhaltete die *Adressat*innenorientierung der Empfehlungen (18)* (durchschnittlich größtenteils erfüllt) und die *Handlungsorientierung der Empfehlungen (19)* (durchschnittlich größtenteils erfüllt). Bei 37 Prozent der Evaluierungen waren nur für ein Viertel der Empfehlungen (oder weniger) Adressat*innen angegeben. Darüber hinaus war die Hälfte der Empfehlungen (oder mehr) nur bei 32 Prozent der Evaluierungen handlungsorientiert über konkrete Umsetzungshinweise dargestellt, so dass nachvollziehbar war, wann und wie die Empfehlungen umzusetzen sind. Das Qualitätskriterium *Handlungsorientierung der Empfehlungen (19)* wies im Gegensatz zum Qualitätskriterium *Adressat*innenorientierung der Empfehlungen (18)* eine große Streuung zwischen den Organisationen auf.

29 Im Rahmen einer weiteren Evaluierungsfrage – die nicht Gegenstand dieses Artikels ist – wurde versucht, Informationen zu den Kosten einer Evaluierung zu erhalten. Dies war aufgrund einer fehlenden einheitlichen Definition und fehlenden Daten nicht möglich. Schwierig dürfte somit auch die Bezifferung des Nutzen-Verhältnisses sein, den die Evaluierung einer Organisation gebracht hat.

## 6. Übergreifende Erkenntnisse, Implikationen und Ausblick

Die Meta-Evaluierung stellt organisationsübergreifende Erkenntnisse zur Evaluierungsqualität ausgewählter Organisationen aus der deutschen EZ unter Einbeziehung von staatlichen und nichtstaatlichen Organisationen zur Verfügung. Sie ermöglicht es den beteiligten Organisationen, ihre Evaluierungsarbeit im Vergleich zu anderen zu bewerten und die Ergebnisse zur Rechenschaftslegung nach außen wie zum organisationalen Lernen nach innen zu nutzen. Darüber hinaus hat sie durch die systematische Ableitung eines Analyserasters aus den international geltenden Qualitätsstandarddokumenten ein Instrument entwickelt und erprobt, das auch als Grundlage für ein zukünftiges Analyseraster der BMZ-Leitlinien Evaluierung (BMZ 2021) herangezogen werden kann.

Die Erkenntnisse und Implikationen der Meta-Evaluierung können in drei Bereiche eingeteilt werden: 1) Identifikation (nicht) relevanter Qualitätsstandards und Verschriftlichung dieser in Organisationsdokumenten, 2) Sicherstellung der Anwendung und Nachvollziehbarkeit der (Nicht-)Anwendung relevanter Qualitätsstandards auf Ebene der Evaluierung und 3) Gemeinsames Lernen. Die Implikationen können gegebenenfalls auch für andere Organisationen innerhalb oder außerhalb der EZ relevant sein.

*Identifikation (nicht) relevanter Qualitätsstandards und Verschriftlichung dieser in Organisationsdokumenten*

EZ-Organisationen werden durch die BMZ-Förderrichtlinien und Evaluierungsleitlinien beziehungsweise durch ihre eigens formulierten Ansprüche oder Mitgliedschaften in unterschiedlichem Maße dazu angehalten, die OECD/DAC- und DeGEval-Standards in ihren Evaluierungen zu erfüllen. Eine systematische Auseinandersetzung mit den verschiedenen Qualitätsstandards wäre notwendig, um deren Erfüllung in den Evaluierungen zu gewährleisten. Die Identifikation der einzelnen Qualitätskriterien und ihre Verschriftlichung hat bisher allerdings kaum Eingang in die Evaluierungspraxis der deutschen EZ gefunden. Dies stellt eine Schwäche der aktuellen Evaluierungspraxis dar.

Organisationen sollten ihr *Qualitätsbewusstsein* und *-verständnis* der für sie relevanten Qualitätskriterien schärfen. Dafür sollten die Organisationen in einem ersten Schritt diejenigen Qualitätsstandards identifizieren, die sie verbindlich einhalten wollen. Zweitens ist wichtig, diese Qualitätsstandards explizit in den Organisationsdokumenten einer Organisation (zum Beispiel Leitlinien, Konzeptionen oder Handreichungen) zu benennen und ihre Erfüllung in Organisationsprozessen angemessen festzulegen. Zudem sollten Qualitätsstandards, die systematisch nichterfüllt werden, identifiziert und ihre Nichterfüllung in Organisationsdokumenten sollte begründet dargestellt werden.

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

Insgesamt stellt die Erfüllung der Qualitätsstandards eine Stärke der beteiligten Organisationen dar, da 79 Prozent der Qualitätskriterien mindestens größtenteils oder vollständig erfüllt werden.

Aus Gründen der Transparenz und Nachvollziehbarkeit wäre in Fällen einer bewussten Nichterfüllung eines Qualitätsstandards auf Evaluierungsebene eine verschriftlichte Begründung auf Evaluierungsebene erforderlich. Dies geschieht aktuell nur in Ausnahmenfällen (wie beispielsweise im Qualitätskriterium *Veröffentlichung des Evaluierungsberichts (23)*. Obwohl dies mit dem Maximalprinzip der Qualitätsstandards gut vereinbar ist, zeigt sich hier eine Schwäche der aktuellen Evaluierungspraxis der beteiligten Organisationen.

Organisationen sollten in regelmäßigen Abständen die Qualität ihrer Evaluierungen durch *(organisationsinterne) Meta-Evaluierungen* überprüfen. Hierbei sollten sie sich auf die von ihnen in den Organisationsdokumenten beschriebenen Qualitätsstandards und gesetzten Anspruchsniveaus beziehen. Eine regelmäßige Überprüfung ist nützlich, um sich über die aktuellen beziehungsweise veränderten Stärken und Schwächen der eigenen Erfüllung der Qualitätsstandards zu informieren und Verbesserungspotenziale freizulegen.

*Gemeinsames Lernen*

Die Untersuchung strukturell heterogener Organisationen hat gezeigt, dass es eine große Variabilität bezüglich der Erfüllung der Qualitätskriterien gibt. Daher ist anzunehmen, dass die Organisationen über den Austausch von Good Practices voneinander lernen können und die Ergebnisse der Meta-Evaluierung für einen weiten Kreis an Organisationen (innerhalb und außerhalb der EZ) relevant sind und konkrete Lernmöglichkeiten beinhalten.

Organisationen sollten sich (regelmäßig) untereinander zur Identifikation, Sicherstellung der Erfüllung und Nichterfüllung sowie der Nachvollziehbarkeit der Erfüllung von Qualitätsstandards im Rahmen eines systematischen und fachübergreifenden Diskurses austauschen, um voneinander zu lernen und die eigene Evaluierungspraxis stetig zu verbessern. Der Austausch sollte sinnvollerweise über die an der Meta-Evaluierung beteiligten Organisationen hinausgehen und weitere Evaluierungstypen – beispielsweise dezentrale Evaluierungen – umfassen.

*Ausblick*

Inhaltliche und methodische Anregungen, die sich aus dieser Meta-Evaluierung für die Zukunft speisen, sind:

- Eine gezielte Untersuchung *verschiedener Evaluierungstypen* (zum Beispiel zentrale, dezentrale Evaluierungen etc.), um die Erfüllung der Qualitätsstandards in der Gesamtheit aller bestehenden Evaluierungen einer Organisation zu reflektieren.
- Die Datentriangulation zur Erfassung der Erfüllung der Qualitätsstandards auf Ebene der Evaluierung zu fördern, indem neben der Kodierung der Evaluierungsdokumente auch die an der Evaluierung *Beteiligten z.B. Gutachtende, Partner(innen) und/oder Verantwortliche der EZ-Maßnahme* nach ihrer Einschätzung der Erfüllung befragt werden.
- Die Ausweitung der Qualitätsstandards insbesondere im Bereich der Nutzung und des tatsächlichen Nutzens einer Evaluierung auf verschiedene Ebenen der beteilig-

ten Organisation (zum Beispiel Projektebene oder Ebene der Evaluierungseinheiten/-stellen), da diese aktuell wenig untersucht werden.

## Literatur


Beywl, Wolfgang/Niestroj, Melanie (2009): Das ABC der wirkungsorientierten Evaluation: Glossar – deutsch/englisch – der wirkungsorientierten Evaluation. Köln: Univation – Institut für Evaluation Dr. Beywl und Associates (2. Aufl.).

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

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

Caracelli, Valerie J./Cooksy, Leslie J. (2009): Metaevaluation in Practice. Journal of MultiDisciplinary Evaluation, 6 (11), S. 1–15.

Caspari, Alexandra (2010): Lernen aus Evaluierungen. Meta-Evaluation & Evaluationssynthese von InWEnt-Abschlussevaluierungen 2009. Bonn: Internationale Weiterbildung und Entwicklung.

Caspari, Alexandra (2011): Meta-Evaluation & Evaluationssynthese 2011 – Hauptbericht. o.V. Frankfurt am Main.

Caspari, Alexandra (2012): Meta-Evaluation, Evaluationssynthese, Evaluation Review and Systematic Review – eine Begriffsklärung. Frankfurt am Main: Fachhochschule Frankfurt am Main.

DeGEval (2016): Standards für Evaluation. Mainz: Gesellschaft für Evaluation e.V.

DEval (2018): Methoden und Standards 2018. Standards für Evaluierungen des DEval. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

DEval (2020): DEval-Evaluierungen 2020-2022. Themenschwerpunkte, laufende und geplante Evaluierungen des DEval. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin: Springer (5. Aufl.). https://doi.org/10.1007/978-3-642-41089-5

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

Freimann, Isabelle/Krämer, Matías (2016): Meta-Evaluierung der Projektevaluierungen (PEV). Bonn/Eschborn: Deutsche Gesellschaft für Internationale Zusammenarbeit.

Freimann, Isabelle/Krämer, Matías (2017): Querschnittsauswertung (QSA) von Projektevaluierungen (PEV) 2016 Meta-Evaluierung. Bonn/Eschborn: Deutsche Gesellschaft für Internationale Zusammenarbeit.

Guffler, Kerstin/Kunert, Laura/Wittenberg, Marian/Herforth, Nico (2022a): Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen Entwicklungszusammenarbeit. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

Guffler, Kerstin/Kunert, Laura/Wittenberg, Marian/Herforth, Nico (2022b): Meta-Evaluierung zur Qualität von (Projekt-)Evaluierungen in der deutschen Entwicklungszusammenarbeit – Onlineanhang. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

Hageboeck, Molly/Frumkin, Micah/Monschein, Stephanie (2013): Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012. Washington, D. C.: United States Agency for International Development.

HTSPE LTD. (2011): Mid-term Meta Evaluation of IPA Assistance Evaluation Report. Brüssel: EU-Kommission.

Koy, Jens/Väth, Susanne/Römling, Cornelia (2016): Meta-Evaluierung der Projektevaluierungen aus den Jahren 2014–2015. Aachen: Misereor.

Krämer, Matías/Almqvist, Olga (2019): Meta-Evaluierung und statistische Auswertung der Projektevaluierungen 2017 / 2018 – Teil I: Meta-Evaluierung. Bonn.

Krippendorff, Klaus (2012): Content Analysis: An Introduction to its Methodology. Thousand Oaks, CA: SAGE (2. Aufl.).

Lücking, Kim/Freund, Simon/Bettighofer, Simon (2015): Evaluierungspraxis in der deutschen Entwicklungszusammenarbeit. Umsetzungsmonitoring der letzten Systemprüfung und Charakterisierung wesentlicher Elemente. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

Mauthofer, Tatjana/Silvestrini, Stefan (2018): Meta-Evaluation of 33 Evaluation Reports of World Vision Germany. Saarbrücken: World Vision Germany.

Noltze, Martin/Euler, Michael/Verspohl, Ida (2018): Meta-Evaluierung von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

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

Queiroz de Souza, Andrea (2017): Meta-Evaluation and Analysis of Project Evaluations 2016. Bielefeld: Welthungerhilfe.

Silvestrini, Stefan/Bäthge, Sandra (2019): Meta-Evaluation of ADA Project and Programme Evaluations – Executive Summary. Saarbrücken/Wien: Agentur der Österreichischen Entwicklungszusammenarbeit.

Silvestrini, Stefan/Väth, Susanne/Römling, Cornelia/Lieckefett, Michael/Mikkolainen, Petra (2018): Meta-evaluation of Project and Programme Evaluations in 2015–2017. Saarbrücken/Helsinki: Ministry for Foreign Affairs of Finland.

UNFPA (2020): UNFPA Evaluation Office – Assessing the quality of developmental evaluations at UNFPA. New York: United Nations Population Fund.

Väth, Susanne Johanna/Silvestrini, Stefan/Gaus, Hansjörg/Mikkolainen, Petra/Flaig, Maja/Wicke, Janis (2022): Evaluation: Metaevaluation of MFA's Project and Programme Evaluations in 2017–2020. Saarbrücken/Helsinki: Ministry for Foreign Affairs of Finland.

**Dr. Kerstin Guffler** | Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) | Fritz-Schäffer-Str. 26 | D-53113 Bonn | E-Mail: kerstin.guffler@deval.org

**Marian Wittenberg** | Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) | Fritz-Schäffer-Str. 26 | D-53113 Bonn | E-Mail: marian.wittenberg@deval.org

**Laura Kunert** | Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) | Fritz-Schäffer-Str. 26 | D-53113 Bonn | E-Mail: laura.kunert@deval.org

**Amélie Gräfin zu Eulenburg** | Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) | Fritz-Schäffer-Str. 26 | D-53113 Bonn | E-Mail: amelie.eulenburg@deval.org