Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit: Meta-Evaluierung

Noltze, Martin; Euler, Michael; Verspohl, Ida

Erstveröffentlichung / Primary Publication
Forschungsbericht / research report

NACHHALTIGKEIT IN DER DEUTSCHEN ENTWICKLUNGS-ZUSAMMENARBEIT

Meta-Evaluierung
2018

DEval
DEUTSCHES EVALUIERUNGSINSTITUT DER ENTWICKLUNGS-ZUSAMMENARBEIT

Die vorliegende Meta-Evaluierung „Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit“ ist Teil des DEval-Themenschwerpunktes Nachhaltigkeit. Die Meta-Evaluierung wird durch eine begleitende Evaluierungssynthese ergänzt. Im Rahmen eines integrierten Evaluierungsdesigns basieren beide Berichte auf einer gemeinsamen Datengrundlage und haben komplementäre Ziele.

	Meta-Evaluierung	Evaluierungssynthese
Inhalte	Auseinandersetzung mit der Evaluierungspraxis zur Nachhaltigkeit von Vorhaben der deutschen Entwicklungszusammenarbeit (EZ) Rekonstruktion des bisherigen Verständnisses von Nachhaltigkeit in der deutschen EZ und Abgleich mit dem modernen Verständnis der Agenda 2030 für nachhaltige Entwicklung Unterstützung der Ausgestaltung einer Agenda-2030-konformen Evaluierungspraxis	Analyse der Einflussfaktoren auf die Nachhaltigkeitsbewertung von Vorhaben Auseinandersetzung mit der Bewertung von Nachhaltigkeit deutscher EZ-Vorhaben Herausstellen von Ansatzpunkten zur Erhöhung der Nachhaltigkeit deutscher EZ-Vorhaben Unterstützung der strategischen und operativen Ausrichtung der deutschen EZ auf die Anforderungen der Agenda 2030 für nachhaltige Entwicklung
Methoden	Systematische Qualitätsanalyse und quantitative Inhaltsanalyse	Multivariate Regressionsanalysen
Datengrundlage	Evaluierungsberichte von Vorhaben der deutschen EZ und Sekundärdaten	
Integriertes Design	Die Ergebnisse der quantitativen Inhaltsanalyse der Meta-Evaluierung wurden als erklärende Variablen in die Regressionsanalysen der Evaluierungssynthese einbezogen. Die Ergebnisse der Qualitätsanalyse der Meta-Evaluierung wurden als Gewichtungsfaktor für die Aussagekraft der Beobachtungen in die Regressionsanalysen der Evaluierungssynthese einbezogen.	

NACHHALTIGKEIT IN DER DEUTSCHEN ENTWICKLUNGS-ZUSAMMENARBEIT

Meta-Evaluierung
2018

Impressum

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

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

Verfasst von
Dr. Martin Noltze
Dr. Michael Euler
Ida Verspohl

Verantwortlich
Prof. Dr. Jörg Faust (bis Juni 2016)
Dr. Sven Harten (ab Juni 2016)

Gestaltung
MedienMélange: Kommunikation!, Hamburg
www.medienmelange.de

Lektorat
Silvia Richter, mediamondi, Berlin
www.mediamondi.de

Bildnachweis
Gui Yongnian/123rf.com (Cover), Olaf Speier/Alamy Stock Foto (Kap. 1), dbimages/Alamy Stock Foto (Kap. 2 + 3), Dzianis Apolka/Alamy Stock Foto (Kap. 4), imageBROKER/Alamy Stock Foto (Kap. 5), Riccardo Lennart Niels Mayer/123rf.com (Kap. 6), Oleksandr Roslyak/123rf.com (Kap. 7)

Bibliografische Angabe
Noltze, M., M. Euler und I. Verspohl (2018), *Meta-Evaluierung von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit*, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval), Bonn.

Druck
Bonifatius,
Paderborn

FSC www.fsc.org MIX Papier aus verantwortungsvollen Quellen FSC® C011558

ISBN 978-3-96126-067-6 (gebundene Ausgabe)
ISBN 978-3-96126-068-3 (PDF)

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:
www.DEval.org/de/evaluierungsberichte.html

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

Danksagung

Das Evaluierungsteam wurde bei seiner Arbeit von zahlreichen Personen und Organisationen unterstützt. Für die wertvolle Unterstützung möchten wir uns an dieser Stelle recht herzlich bedanken.

Zentral für das Gelingen der vorliegenden Meta-Evaluierung und der begleitenden Evaluierungssynthese war zunächst die Unterstützung der Referenzgruppe. Besonderer Dank gilt hierbei den beteiligten Referaten des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ), dem Referat 105 (Michaela Zintl, Katrin von der Mosel und Berthold Hoffman) und dem Referat 300 (Gottfried von Gemmingen-Guttenberg, Dr. Ingolf Dietrich, Dr. Maya Schmaljohann, Cormac Ebken und Ruben Werchan), der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ; Dr. Ricardo Gómez, Dorothea Giesen-Thole, Valentin Dyckerhoff, Katrin Ladwig und Cornelia Skokov) und der KfW Entwicklungsbank (KfW; Prof. Dr. Eva Terberger, Martin Dorschel, Thomas Gietzen und Christian Schönhofen). Bedanken möchten wir uns dabei insbesondere für die vielen Anregungen und Kommentare im Rahmen der offenen und kritischen Diskussion. Besonderer Dank gilt der GIZ und der KfW für ihre tatkräftige Unterstützung im Rahmen der Datenerhebung – ohne die Übermittlung umfangreicher Daten und Dokumente wäre die Evaluierungsarbeit nicht möglich gewesen.

Weiterhin bedanken möchten wir uns bei unseren Kolleginnen und Kollegen am DEval, die den Evaluierungsprozess aufmunternd und kritisch begleitet haben. Unser Dank gilt dabei insbesondere unseren DEval-internen Gutachterinnen Dr. Kerstin Guffler und Solveig Gleser sowie unserem Institutsleiter Prof. Dr. Jörg Faust für ihre vielen Anregungen und Kommentare. Zudem danken wir Thomas Wencker für seine kritische Perspektive und die konstruktiven Vorschläge. Darüber hinaus bedanken wir uns bei Cornelia Michaels-Lampo und unserer Verwaltung für die administrative Unterstützung der Evaluierungstätigkeit. Besonderer Dank gilt auch der Öffentlichkeitsarbeit des DEval sowie der Lektorin dieses Berichts.

Ferner bedanken wir uns bei Jana Preiß, die uns im Rahmen einer assoziierten Masterarbeit bei der Durchführung der Kontextstudie der Meta-Evaluierung unterstützt hat.

Weiterer Dank gebührt unseren Praktikantinnen und studierenden Beschäftigten Helena Heberer, Niklas Witzig, Grisel Orozco, Sarah Stahlmann und Lea Smidt, deren Unterstützung für den Erfolg der Evaluierung von hohem Wert war. Wir bedanken uns herzlich für das große Engagement und den persönlichen Einsatz.

Ein besonderer Dank gilt weiterhin unserem externen Gutachter Prof. Dr. Sebastian Vollmer. Seine zahlreichen inhaltlichen und methodischen Anregungen haben entscheidend zur Qualität der vorliegenden Evaluierungsberichte beigetragen.

Abschließend möchten wir uns noch bei unseren Kolleginnen und Kollegen des Kompetenzzentrums Methoden bedanken, die uns über den gesamten Prozess der Evaluierungsarbeit mit kritischen Fragen und methodischen Anregungen zur Seite standen.

ZUSAMMENFASSUNG

Hintergrund, Ziele und Evaluierungsgegenstand

Die Agenda 2030 für nachhaltige Entwicklung betont die globale Bedeutung des Prinzips der Nachhaltigkeit. Mit ihr definiert sich Nachhaltigkeit nunmehr entlang von Kernprinzipien nachhaltiger Entwicklung: Ein universaler Geltungsanspruch, gemeinsame Verantwortung und Rechenschaftspflicht, das Zusammenspiel von sozialer, wirtschaftlicher und ökologischer Entwicklung und Inklusivität bilden die Prinzipien des modernen Nachhaltigkeitsverständnisses für Entwicklung.

Die deutsche Entwicklungszusammenarbeit (EZ) hat sich zu den Prinzipien der Agenda 2030 bekannt und zu ihrer Umsetzung verpflichtet. In der deutschen EZ ist der Begriff der Nachhaltigkeit bereits seit geraumer Zeit fester Bestandteil der entwicklungspolitischen Debatte. Prinzipiell wird dabei zwischen den Aspekten „nachhaltige Entwicklung“ und „Dauerhaftigkeit von Wirkungen“ unterschieden. Inwieweit sich diese beiden Aspekte nun in dem modernen Verständnis von Nachhaltigkeit nach der Agenda 2030 wiederfinden bzw. diesem entsprechen, ist bislang eine offene Frage. Weder das Verständnis von noch der praktische Umgang mit Nachhaltigkeit in der deutschen EZ wurden bis heute einer systematischen Analyse unterzogen. Die aktuelle Entwicklungsagenda gibt nun Anlass für eine umfassende Auseinandersetzung mit dem langjährigen Leitprinzip der deutschen EZ.

Ziel der vorliegenden Meta-Evaluierung ist eine erste umfassende und systematische empirische Auseinandersetzung mit der Evaluierungs- und Bewertungspraxis von Nachhaltigkeit in der deutschen EZ im Sinne einer Bestandsaufnahme. Die empirische Betrachtung der bisherigen Praxis dient der Rekonstruktion des bislang schwer greifbaren Verständnisses von Nachhaltigkeit in der deutschen EZ als Voraussetzung für einen Abgleich mit dem modernen Verständnis von Nachhaltigkeit nach den Prinzipien der Agenda 2030. Zweck der Meta-Evaluierung ist es somit, die Ausgestaltung einer Agenda-2030-konformen Evaluierungs- und Bewertungspraxis zu unterstützen.

Den Gegenstand der Meta-Evaluierung bildet die Evaluierungs- und Bewertungspraxis der Nachhaltigkeit von Vorhaben der deutschen EZ, dargestellt in den Evaluierungsberichten der beiden großen deutschen staatlichen Durchführungsorganisationen (DO), der KfW Entwicklungsbank (KfW) und der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ). Beide DO bewerten die Nachhaltigkeit von Vorhaben entlang der internationalen Evaluierungskriterien des Entwicklungsausschusses der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD-DAC). Auf der Grundlage einer Orientierungshilfe des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) aus dem Jahr 2006 bildet der Aspekt der Dauerhaftigkeit von Wirkungen über die Zeit den Kern des Evaluierungskriteriums Nachhaltigkeit. Zu Beginn der Meta-Evaluierung wurde ferner angenommen, dass sich hinter dem Wirkungsbegriff und im Zusammenspiel mit den anderen Evaluierungskriterien – Relevanz, Effektivität, Effizienz und übergeordnete entwicklungspolitische Wirkungen (Impact) – auch der Aspekt der nachhaltigen Entwicklung verbirgt.

Methodisches Vorgehen

Bei der vorliegenden Untersuchung handelt es sich um eine thematische Meta-Evaluierung. Dabei wurde das klassische Meta-Evaluierungsdesign einer reinen Qualitätsbewertung um die systematische Auseinandersetzung mit dem inhaltlichen Bewertungskriterium der Nachhaltigkeit von EZ-Vorhaben erweitert. Die Datengrundlage der Meta-Evaluierung bildet eine repräsentative Stichprobe von 513 Evaluierungsberichten von Vorhaben der deutschen technischen und finanziellen EZ. Die Ergebnisse der Meta-Evaluierung fließen im Rahmen eines integrierten Forschungsdesigns zudem auch in die begleitende Evaluierungssynthese ein, die sich mit den Einflussfaktoren der Nachhaltigkeit auseinandersetzt.

Zentrale Ergebnisse, Schlussfolgerungen und Empfehlungen zur Bewertung von Nachhaltigkeit in der deutschen EZ

Die Ergebnisse der vorliegenden Meta-Evaluierung bestätigen die Vorabannahme, dass sich in den Evaluierungskriterien neben dem Aspekt der Dauerhaftigkeit auch der Aspekt der nachhaltigen Entwicklung verbirgt; damit belegen sie erstmals empirisch, dass Nachhaltigkeit in der deutschen EZ-Evaluierungspraxis bereits umfassend verstanden, evaluiert und bewertet wird. Gleichzeitig besteht eine deutliche Abweichung von den Ansprüchen der Agenda 2030. Wesentliche Prinzipien der Agenda 2030, etwa das Zusammenspiel der Dimensionen der Nachhaltigkeit, sind noch kein systematischer Bestandteil der Bewertungspraxis. Somit widerlegen die Ergebnisse zwar

die mögliche Annahme, in den DAC-Evaluierungskriterien sei ausschließlich ein enges Nachhaltigkeitsverständnis im Sinne der Dauerhaftigkeit von Wirkungen angelegt; sie weisen jedoch auf deutliche Diskrepanzen zum modernen Nachhaltigkeitsverständnis im Sinne der Agenda 2030 hin.

Die Ergebnisse zeigen auch, dass die Evaluierung und Bewertung von Nachhaltigkeit in der Praxis aufgrund eines fehlenden konzeptionellen Rahmens für ein umfassendes Nachhaltigkeitsverständnis bislang unsystematisch und uneinheitlich erfolgt. Auch die seit 2006 vorgeschlagenen Prüffragen der Orientierungshilfe des BMZ werden bislang nicht systematisch berücksichtigt. In der Gesamtschau zeigt sich, dass die derzeitige Konzeption der DAC-Kriterien die Evaluierung von Nachhaltigkeit im umfassenden Sinne zwar zulässt, jedoch keinesfalls systematisch und verbindlich vorgibt. Für die aggregierte Betrachtung der Nachhaltigkeitsnote über verschiedene Vorhaben hinweg bedeutet dies – aufgrund der fehlenden Systematik – eine eingeschränkte Vergleichbarkeit, die dem strategischen Lernen aus Evaluierungen entgegensteht. Eine rigorose vergleichende Perspektive auf die Nachhaltigkeit von Vorhaben ist derzeit nur unter erheblichem Aufwand – wie mit der vorliegenden erweiterten Meta-Evaluierung und der begleitenden Evaluierungssynthese verbunden – möglich.

Der zukünftige Umgang mit der Agenda 2030 und der Nachhaltigkeit von EZ-Vorhaben in Evaluierungen ist eine globale Aufgabe. Mit Blick auf die deutsche EZ hat die vorliegende Meta-Evaluierung konkreten Handlungsbedarf identifiziert. Die Schlussfolgerungen rufen nach einer Reform der bisherigen Evaluierungs- und Bewertungspraxis. Neben dem Harmonisierungs- und Koordinierungsgedanken der Erklärung von Paris zur Effektivität der EZ und dem Aktionsplan von Accra verlangt der universelle Charakter der Agenda 2030 dabei auch nach Austausch und Abstimmung auf internationaler Ebene. Die nun folgenden Empfehlungen zielen darauf ab, die laufenden Reformprozesse auf Ebene der deutschen EZ zu unterstützen und die Diskussionen auf internationaler Ebene zu bereichern. Zunächst werden die wesentlichen Empfehlungen zur Weiterentwicklung der Evaluierungspraxis dargestellt. Anschließend folgen die grundlegenden Empfehlungen zur Weiterentwicklung des Evaluierungssystems.

Empfehlungen zur Weiterentwicklung der Evaluierungspraxis:

Dem BMZ und den DO wird empfohlen, die Nachhaltigkeit von Vorhaben im Sinne der Prinzipien der Agenda 2030 für nachhaltige Entwicklung künftig im Rahmen eines zusätzlichen Bewertungskriteriums zu evaluieren.

Einhergehend mit dem Einbeziehen von Nachhaltigkeit im Sinne der Agenda 2030 als zusätzliches Bewertungskriterium werden dem BMZ die konzeptionelle Schärfung der DAC-Kriterien und eine höhere Verbindlichkeit der BMZ-Orientierungshilfe für den Umgang mit den DAC-Kriterien empfohlen.

Im Rahmen der Reform der Evaluierungskriterien für die Erfolgsbewertung von EZ-Vorhaben wird dem BMZ empfohlen, das bisherige Evaluierungskriterium der Nachhaltigkeit nach OECD-DAC im Sinne der Dauerhaftigkeit von Wirkungen zu erhalten und die entsprechenden Prüffragen auf diesen Aspekt auszurichten.

Mit Blick auf die Prinzipien der Agenda 2030 sollten GIZ und KfW untersuchen, wie in Evaluierungen künftig die nicht intendierten Wirkungen eines Vorhabens und die Wechselwirkungen zwischen den Dimensionen der Nachhaltigkeit identifiziert und geprüft werden können.

Die Umsetzung und konzeptionelle Ausgestaltung der Empfehlungen zur Evaluierungspraxis sollten in Deutschland auf der Grundlage eines gemeinsamen Prozesses unter Federführung des BMZ und unter Beteiligung der DO und des DEval erfolgen. Es wird empfohlen, diesen Prozess inklusive einer Pilotphase bis Ende 2018 abzuschließen, um eine Agenda-2030-konforme Evaluierungs- und Bewertungspraxis der deutschen EZ ab 2019 zu gewährleisten. Gleichzeitig sollten die laufenden Reformbemühungen innerhalb der deutschen EZ auf internationale Anschlussfähigkeit geprüft und in die entsprechenden Foren eingebracht werden.

Empfehlungen zur Weiterentwicklung des Evaluierungssystems

Dem BMZ wird empfohlen, eine übergeordnete Evaluierungsstrategie zu entwickeln, die sich über die Zeit thematische Schwerpunkte setzt.

In der Evaluierungsstrategie sollte das BMZ definieren, welche Anforderungen sich aus den Fragestellungen um die Agenda 2030 für die jeweiligen Evaluierungen – also auf Ebene der Module, der Programme und der Länderstrategien – ergeben.

Zentrale Ergebnisse, Schlussfolgerungen und Empfehlungen zur Qualität der Evaluierungspraxis

Im Rahmen der Meta-Evaluierung wurde die Auseinandersetzung mit der Bewertung von Nachhaltigkeit in der deutschen EZ von der Analyse der Evaluierungsqualität begleitet. Dabei geben die Ergebnisse der Qualitätsbewertung Hinweise auf die Belastbarkeit der Ergebnisse und der Schlussfolgerungen der Evaluierungen hinsichtlich der Nachhaltigkeit deutscher EZ-Vorhaben.

Die Ergebnisse zeigen, dass GIZ und KfW die aus Modulevaluierungen hervorgehenden Ergebnisse und Schlussfolgerungen auf eine dem Umfang dieser Evaluierungen angemessene evaluatorische Grundlage stellen. Neben der Gegenstandsbeschreibung enthält die Mehrheit der Evaluierungsberichte eine nachvollziehbare Darstellung der zu überprüfenden Wirkungszusammenhänge und der methodischen Vorgehensweise. Die deutsche EZ zeichnet sich zudem durch einen hohen Deckungsgrad in der Evaluierung aus: Die GIZ unterzieht nahezu alle Module einer systematischen Erfolgsbewertung, die KfW arbeitet mit einer repräsentativen Stichprobe.

Es hat sich jedoch auch gezeigt, dass die Evaluierungsqualität auf Modulebene verbessert werden kann. Dabei sollten durch systematische Analyse- und Triangulationsverfahren vor allem die Anstrengungen zur Aufdeckung von Ursache-Wirkungs-Beziehungen erhöht werden. Dasselbe gilt für die Nachvollziehbarkeit von Ergebnissen und Schlussfolgerungen in den Evaluierungsberichten. Dabei gilt es auch, die zur Verfügung stehenden Ressourcen auf den Zweck einer Evaluierung auszurichten. Bei den dezentralen Evaluierungen umfasste das Erkenntnisinteresse bisher neben der Evaluierung selbst auch den Aspekt der Prüfung. Ein belastbarer Wirkungs- und Nachhaltigkeitsnachweis lässt sich ferner durch die geeignete Wahl des Evaluierungszeitpunktes erreichen: Im Rahmen von Ex-Post-Evaluierungen besteht die Möglichkeit, Wirkungen und deren Nachhaltigkeit in gewissem zeitlichem Abstand zum Ende der Vorhaben tatsächlich zu beobachten. Bei den dezentralen Evaluierungen, die im Verlauf eines Vorhabens durchgeführt werden, erfolgt der Nachhaltigkeitsnachweis hingegen auf einer reinen Zukunftseinschätzung. Vor dem Hintergrund eingeschränkter Datenverfügbarkeit im Kontext der EZ bieten Monitoringdaten eine relevante Datenquelle. Deren Potenzial für einen belastbaren Wirkungs- und Nachhaltigkeitsnachweis wird allerdings bislang nicht ausgeschöpft.

Die Ergebnisse der Meta-Evaluierung haben darüber hinaus einen interessanten Zusammenhang zwischen evaluatorischer Qualität und inhaltlichem Erkenntnisgewinn aufgedeckt: Mit zunehmender Qualität der Evaluierungen erhöht sich die Zahl der für die Nachhaltigkeitsbewertung hinzugezogenen Kriterien. Anspruchsvollere Evaluierungen stellen die Bewertung von Nachhaltigkeit auf eine breitere Basis und sind zudem auch der Belastbarkeit der Aussagen zuträglich. Ein direkter Zusammenhang zwischen der Evaluierungsqualität und der Einzelbewertung eines Kriteriums oder der Gesamtbewertung der Nachhaltigkeit eines Vorhabens besteht nicht.

Aufgrund des Zusammenhangs zwischen der Qualität und der Ausführlichkeit, mit der das Thema Nachhaltigkeit in Evaluierungen behandelt wird, sowie des engen Zusammenhangs von Wirkungs- und Nachhaltigkeitsnachweis ergeben sich eine Reihe von Empfehlungen, die sich auf die Qualität von Evaluierungen und das zugrunde liegende Evaluierungssystem beziehen. Auch hier werden zunächst Empfehlungen zur Weiterentwicklung der Evaluierungspraxis dargestellt. Anschließend folgen die Empfehlungen zur Weiterentwicklung des Evaluierungssystems.

Empfehlungen 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.

Aufgrund der anhaltend geringen Bedeutung, die Monitoringdaten in Modulevaluierungen beigemessen wird, sollten die DO systematisch untersuchen, welche Hindernisse hier bestehen und wie diese überwunden werden können. Dabei sollten sie prüfen, inwieweit sich die Monitoringsysteme der Vorhaben über die Zielsysteme der Vorhaben mit dem Zielsystem der nachhaltigen Entwicklungsziele (SDGs) verknüpfen lassen.

Im Sinne der Transparenz und als Anreiz für eine nachvollziehbare Berichtslegung sollten GIZ und KfW unter Abwägung der Chancen und Risiken die Möglichkeit prüfen, die Evaluierungsberichte – gegebenenfalls zunächst in einer Pilotphase – vollständig zu veröffentlichen und das BMZ über die Erfahrungen hierzu in Kenntnis setzen.

Um die Evaluierungsqualität zu steigern, wird der GIZ empfohlen, die Funktion der Qualitätssicherung langfristig in der Stabsstelle Evaluierung zu verankern. Alle Modulevaluierungen sollten künftig durch die Stabsstelle gesteuert werden.

Die Erhöhung der Qualität von Evaluierungen sollte in der GIZ durch eine Trennung zwischen Prüfung und Evaluierung unterstützt werden.

Im Hinblick auf den geeigneten Zeitpunkt für einen aussagekräftigen Wirkungs- und Nachhaltigkeitsnachweis sollte das Format von Ex-post-Evaluierungen in der GIZ erneut an Bedeutung gewinnen. Bei der Durchführung von Ex-post-Evaluierungen sollten sowohl GIZ als auch KfW darauf achten, die Steuerungsrelevanz sicherzustellen. Dies kann beispielsweise durch thematische Fokussierung oder durch die geeignete Wahl des Evaluierungszeitpunktes erfolgen.

Empfehlungen zur Weiterentwicklung des Evaluierungssystems:

Im Sinne des gemeinsamen Lernens und der Rechenschaftslegung wird dem BMZ empfohlen, die Evaluierungspraxis von GIZ und KfW auf der Grundlage der Gemeinsamen Verfahrensreform (GVR) und der Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit zu harmonisieren. Dabei sollte das BMZ verbindliche Vorgaben zu Zeitpunkt, Umfang und Benotungssystem schaffen, um die Evaluierungstypen für Modulevaluierungen zu vereinheitlichen.

Dem BMZ wird empfohlen, durch die Festlegung einheitlicher Mindeststandards das Ausschöpfen von Potenzialen zur Erhöhung der Evaluierungsqualität in Modulevaluierungen zu unterstützen.

Das BMZ sollte die DO dazu anhalten, die Evaluierungsberichte in sich nachvollziehbar zu gestalten, sodass sie für sich stehen können. Je nach Ausgang einer entsprechenden Prüfung sollte das BMZ die DO zu einer vollständigen Veröffentlichung der Evaluierungsberichte anhalten.

Das BMZ sollte dafür sorgen, dass neben der Qualitätssicherung der Modulevaluierungen durch die Evaluierungseinheiten von GIZ und KfW regelmäßig eine externe, organisationsübergreifende Meta-Evaluierung zu einer Stichprobe von Evaluierungen stattfindet.

INHALT

Danksagung v
Zusammenfassung vii
Abkürzungen und Akronyme 2

1. Einleitung 3

1.1 Hintergrund 4
1.2 Ziel der Meta-Evaluierung 5
1.3 Gegenstand 6
1.4 Evaluierungsfragen 6
1.5 Aufbau des Evaluierungsberichtes 6

2. Evaluierung von Nachhaltigkeit in der deutschen EZ 8

2.1 Nachhaltigkeit in der Wirksamkeitsdebatte der deutschen EZ 9
2.2 Der konzeptionelle Rahmen der Meta-Evaluierung im Hinblick auf die Bewertung der Nachhaltigkeit 10
2.3 Die Evaluierungspraxis in der deutschen finanziellen und technischen Zusammenarbeit 11

3. Methodische Vorgehensweise 14

3.1 Datengrundlage 15
3.2 Evaluierungspraxis 16
3.3 Bewertungspraxis 17
3.4 Kontextstudie 20
3.5 Limitationen 21

4. Ergebnisse 22

4.1 Qualität der Evaluierungsberichte 23
4.2 Die Bewertung von Nachhaltigkeit in Evaluierungen der GIZ und KfW 27
4.2.1 Übergreifende Erkenntnisse 28
4.2.2 Kontext 32
4.2.3 Implementierung 35
4.2.4 Outcome 36
4.2.5 Kapazitäten vor Ort 38
4.2.6 Impact 39
4.2.7 Absehbarkeit von Wirkungen 40
4.2.8 Zusammenspiel der Dimensionen der Nachhaltigkeit 41
4.3 Zusammenhang zwischen Qualität und der Bewertung von Nachhaltigkeit 42
4.4 Evaluierung und Bewertung von Nachhaltigkeit im internationalen Vergleich 42

5. Schlussfolgerungen und Empfehlungen 45

5.1 Die Qualität der deutschen Evaluierungspraxis 46
5.2 Die Bewertung von Nachhaltigkeit in der deutschen EZ 48

6. Literatur 52

7. Anhang 56

7.1 Abbildungen 57
7.2 Tabellen 74
7.3 Evaluierungsteam und Mitwirkende 80
7.4 Zeitplan 81

Abbildungen

Abbildung 1 Anzahl Evaluierungsberichte nach Anzahl erfüllter Qualitätskriterien 24
Abbildung 2 Anteil Evaluierungsberichte nach erfüllten Qualitätsbereichen 25
Abbildung 3 Anteil Evaluierungsberichte nach verwendeten Datenerhebungsmethoden 27
Abbildung 4 Anteil Evaluierungsberichte nach erfüllten Qualitätskriterien 28
Abbildung 5 Qualitäts-Index nach Evaluierungstyp 29
Abbildung 6 Anteil Evaluierungsberichte mit Bezug zu Bewertungskriterien und -bereichen 31
Abbildung 7 Einfluss der Nachhaltigkeitskriterien und -bereiche auf die Nachhaltigkeitsbewertung 33
Abbildung 8 Relativer Anteil Evaluierungsberichte nach differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung 34
Abbildung 9 Anteil Evaluierungsberichte nach angestrebten und erreichten Oberzielen und Dimensionen der Nachhaltigkeit 41
Abbildung 10 Qualitätsindex nach Anzahl differenzierter Nachhaltigkeitskriterien und nach aggregiertem Einfluss auf die Nachhaltigkeitsbewertung 43
Abbildung 11 Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien 57
Abbildung 12 Relativer Anteil Evaluierungsberichte nach Nachhaltigkeitsbereich und Einfluss auf Nachhaltigkeitsbewertung 58
Abbildung 13 Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien nach Durchführungsorganisation 59
Abbildung 14 Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien nach Durchführungsorganisation 60
Abbildung 15 Relativer Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation 61
Abbildung 16 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation 62
Abbildung 17 Relativer Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Evaluierungstyp 63
Abbildung 18 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Evaluierungstyp 64
Abbildung 19 Relativer Anteil Ex-post-Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation 65
Abbildung 20 Relativer Anteil Ex-post-Evaluierungen mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation 66
Abbildung 21 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Sektor 67
Abbildung 22 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitsbereichen und Einfluss auf Nachhaltigkeitsbewertung nach Sektor 68
Abbildung 23 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Region 69

Abbildung 24 Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitsbereichen und Einfluss auf Nachhaltigkeitsbewertung nach Region 70
Abbildung 25 Anteil Evaluierungsberichte nach angestrebten und erreichten Oberzielen nach Durchführungsorganisation 71
Abbildung 26 Anteil Evaluierungsberichte nach angestrebtem und erreichtem Oberziel, Evaluierungstyp und Nachhaltigkeitsdimension 72
Abbildung 27 Anteil Evaluierungsberichte nach angestrebtem und erreichtem Oberziel, Sektor und Nachhaltigkeitsdimension 73

Tabellen

Tabelle 1 Vorstellung der Datengrundlage 15
Tabelle 2 Übersicht der Qualitätskriterien 17
Tabelle 3 Übersicht der Nachhaltigkeitskriterien 18
Tabelle 4 Analyseraster der Qualitätsbewertung 74
Tabelle 5 Analyseraster der Nachhaltigkeitsbewertung 76

ABKÜRZUNGEN UND AKRONYME

BMZ
Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung

DAC
Entwicklungsausschuss (Development Assistance Committee) der OECD

DO
Durchführungsorganisation

EZ
Entwicklungszusammenarbeit

FZ
Finanzielle Zusammenarbeit

GIZ
Deutsche Gesellschaft für Internationale Zusammenarbeit

GVR
Gemeinsame Verfahrensreform

KfW
KfW Entwicklungsbank

OECD
Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (Organisation for Economic Co-operation and Development)

OZ
Oberziel

PEV
Projektevaluierung

PFK
Projektfortschrittskontrolle

SDGs
Nachhaltige Entwicklungsziele (Sustainable Development Goals)

TZ
Technische Zusammenarbeit

UE
Unabhängige Evaluierungen der GIZ

1.

EINLEITUNG

Bei der vorliegenden rigorosen Meta-Evaluierung handelt es sich um eine erste umfassende und systematische empirische Auseinandersetzung mit der Evaluierungs- und Bewertungspraxis der Nachhaltigkeit von Vorhaben der deutschen bilateralen Entwicklungszusammenarbeit. Grundlage der Betrachtung bilden Evaluierungen von Vorhaben der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) und der KfW Entwicklungsbank (KfW), die durch öffentliche Mittel des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) finanziert werden.

1.1 Hintergrund

Mit der Agenda 2030 für nachhaltige Entwicklung erlangt das Prinzip der Nachhaltigkeit globale Bedeutung. Die prominente Herausstellung des Begriffs „Nachhaltigkeit" ist dabei Konsequenz der langjährigen Diskussion in der internationalen Entwicklungsdebatte, die in den 1980er-Jahren von den Vereinten Nationen angestoßen und anschließend über verschiedene Weltentwicklungskonferenzen weitergeführt wurde. In der jüngsten Vergangenheit gipfelte diese Debatte in der Einführung der Agenda 2030 für nachhaltige Entwicklung. Die anhaltende Diskussion um die Nachhaltigkeit steht dabei stellvertretend für nichts weniger als für die existenzielle Frage nach der Zukunftsfähigkeit der Entwicklung von Mensch und Umwelt. Doch so entscheidend, wie das Prinzip Nachhaltigkeit allseits für Entwicklung hervorgehoben wird, so umfassend und vielfältig sind die ihm zugrunde liegenden Konzepte.

Die Mehrdimensionalität des Nachhaltigkeitsbegriffs kommt auch in der Entwicklungszusammenarbeit (EZ) zum Ausdruck: Hier wird sprachlich gemeinhin zwischen den Aspekten einer „nachhaltigen Entwicklung" und der „Dauerhaftigkeit von Wirkungen" unterschieden. Eine konzeptionelle Klärung des Begriffs **„Nachhaltigkeit"** ergibt sich aus dieser Differenzierung jedoch nicht. Letztlich bleibt unklar, wie Nachhaltigkeit im Politikfeld der EZ in der Praxis tatsächlich verstanden wird. Mit dem Bedeutungszuwachs des Prinzips der Nachhaltigkeit durch die Agenda 2030 kann eine solche Unschärfe allerdings nicht mehr akzeptiert werden. Eine umfassende Auseinandersetzung mit dem Nachhaltigkeitsverständnis ist dringend erforderlich. Wie wird Nachhaltigkeit verstanden? Wie lässt sich Nachhaltigkeit messen und bewerten? Wie belastbar ist bereits vorhandenes Wissen? Diese Fragen lassen sich nicht allein theoretisch beantworten, sondern erfordern zudem eine fundierte empirische Auseinandersetzung mit einem langjährigen Leitprinzip der EZ. Mit dem Ziel einer möglichst offenen Herangehensweise an dieses Thema nimmt die vorliegende Meta-Evaluierung einen umfassenden und unvoreingenommenen Blick auf die Nachhaltigkeit ein, welcher – wenn notwendig – eine nach den Aspekten einer nachhaltigen Entwicklung und der Dauerhaftigkeit entwicklungspolitischer Wirkungen differenzierte Betrachtung erlaubt. Der Hintergrund dieser beiden Aspekte von Nachhaltigkeit wird im Folgenden kurz vorgestellt und später an verschiedenen Stellen des Berichts diskutiert.

In der internationalen Debatte hat der Aspekt der **„nachhaltigen Entwicklung"** als Teil des Prinzips Nachhaltigkeit eine lange Geschichte. Bereits im 17. Jahrhundert wurde Nachhaltigkeit in der Forstwirtschaft als handlungsleitendes Prinzip der Ressourcennutzung herausgestellt: Demnach sollten immer nur so viele Bäume abgeholzt werden, wie unter dem Einsatz verfügbarer Ressourcen auch wieder nachwachsen können. In der jüngeren Geschichte wurde dieses Grundprinzip in den 1970er-Jahren in die Diskussion um die „Grenzen des [ökonomischen] Wachstums" (Meadows et al., 1972) aufgenommen. In den 1980er-Jahren entwickelte sich daraufhin ein (mehr)dimensionales Konzept der sozialen, wirtschaftlichen und ökologischen Nachhaltigkeit (Grunwald und Kopfmüller, 2006). Seit dem „Brundtland-Bericht" von 1987 steht zudem die Sicherung der Bedürfnisse zukünftiger Generationen im Zentrum des Nachhaltigkeitsgedankens (World Commission on Environment and Development, 1987); sie ist seit der UN-Konferenz für Umwelt und Entwicklung in Rio de Janeiro 1992 international akzeptiert. Heute ist die Agenda 2030 für nachhaltige Entwicklung die logische Konsequenz eines zunehmend integrierten und komplexer werdenden Nachhaltigkeitsverständnisses: Ein universaler Geltungsanspruch, gemeinsame Verantwortung und Rechenschaftspflicht, Inklusivität sowie das synergetische Zusammenspiel sozialer, wirtschaftlicher und ökologischer Entwicklung gehören zu den Grundprinzipien der Agenda 2030 (UN, 2015). Zudem bilden 17 globale Nachhaltigkeitsziele (die Sustainable Development Goals, SDGs) mit 169 Unterzielen das begleitende Zielsystem. Die Relevanz und der Einfluss der internationalen Debatte auf das konzeptionelle Verständnis

von Nachhaltigkeit als Leitprinzip der EZ in Deutschland sind unbestritten (König und Thema, 2011). Unklar bleibt jedoch, inwieweit es der EZ in der Praxis gelungen ist, dem zunehmend komplexer werdenden Verständnis von Nachhaltigkeit zu entsprechen bzw. ob dies überhaupt möglich ist. Skeptiker vermuten, dass der Grad an Komplexität die Kapazitäten der EZ übersteigt und daher die Wahrscheinlichkeit, die damit einhergehenden Ziele zu erreichen, stetig abnimmt (Klasen, 2015; Nuscheler, 2007). Dieses Risiko erscheint hinsichtlich der Komplexität der Agenda 2030 relevanter denn je. Der Umgang mit Nachhaltigkeit in der EZ steht somit stellvertretend auch für die Tendenz, dass die EZ auf komplexe Herausforderungen gerne mit komplexen Lösungen reagiert, die dann in den Kontexten vor Ort schwer umsetzbar sind.

Auch der zweite Aspekt des Nachhaltigkeitsbegriffs – **„Dauerhaftigkeit von Wirkungen“** – ist seit langer Zeit mit der EZ verbunden. Dieser Aspekt wurde 1991 durch den Entwicklungsausschuss (DAC) der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) für die Erfolgsbewertung von Vorhaben der EZ herausgestellt (OECD, 1991). 2006 nahm das Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) das OECD-DAC Verständnis in die „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen“ (BMZ, 2006) auf. Seither wird in Evaluierungen und Prüfungen neben Relevanz, Effektivität, Effizienz und entwicklungspolitischen Wirkungen (Impact) immer auch die Nachhaltigkeit bewertet. Dabei erfolgt die Nachhaltigkeitsbewertung entlang von drei zentralen Prüfaspekten: Erstens wird die Dauerhaftigkeit der entwicklungspolitischen Wirkungen bewertet, zweitens die Stabilität des Umfelds im Hinblick auf soziale Gerechtigkeit, wirtschaftliche Leistungsfähigkeit, politische Stabilität und ökologisches Gleichgewicht geprüft, und drittens werden die vorhandenen Risiken und Potenziale für die (fortdauernde) Wirksamkeit des Vorhabens abgeschätzt. Anhand dieser drei Aspekte wird jedoch deutlich, dass sich das Nachhaltigkeitsverständnis nach OECD-DAC keinesfalls rein auf die Dauerhaftigkeit beschränkt, sondern über den Wirkungsbezug eng mit dem Aspekt der nachhaltigen Entwicklung verbunden ist.

Aufgrund des konzeptionellen Zusammenhangs zwischen den Aspekten der nachhaltigen Entwicklung und der Dauerhaftigkeit von Wirkungen wird in der vorliegenden Meta-Evaluierung angenommen, dass Nachhaltigkeit in der Praxis bereits umfassender verstanden wird, als die vorhandenen Vorgaben und Leitlinien von BMZ, GIZ und KfW zunächst vermuten lassen. Somit wird davon ausgegangen, dass sich ein umfassendes Verständnis von Nachhaltigkeit bereits in der Evaluierungs- und Bewertungspraxis niederschlägt und umgekehrt auch anhand dieser nachvollziehbar ist. Allerdings wird erwartet, dass die Komplexität des Nachhaltigkeitsverständnisses und fehlende Vorgaben in der Vergangenheit dazu geführt haben, dass Nachhaltigkeit in Evaluierungen sehr uneinheitlich verstanden und bewertet wurde.

1.2 Ziel der Meta-Evaluierung

Die vorliegende rigorose Meta-Evaluierung ist die erste umfassende und systematische empirische Auseinandersetzung mit der Evaluierungs- und Bewertungspraxis von Nachhaltigkeit in der deutschen EZ. Den Anlass hierfür bildet die Agenda 2030, durch die das Prinzip der Nachhaltigkeit für Entwicklung an Bedeutung gewinnt. Das erklärte Ziel der Meta-Evaluierung liegt dabei in einer Bestandsaufnahme der bisherigen Evaluierungs- und Bewertungspraxis von Nachhaltigkeit in der EZ. Die empirische Auseinandersetzung mit der bisherigen Praxis erlaubt somit, das bislang schwer greifbare Verständnis von Nachhaltigkeit in Evaluierungen der deutschen EZ zu konkretisieren und dieses schließlich mit dem modernen Verständnis von Nachhaltigkeit nach der Agenda 2030 abzugleichen. Dementsprechend umfasst der zentrale Beitrag dieser Meta-Evaluierung zwei Aspekte: Erstens wird die oftmals rein theoretisch geführte Diskussion um Nachhaltigkeit auf eine breite empirische Basis gestellt; zweitens wird auf der Basis der Ergebnisse ein Vorschlag dafür erarbeitet, wie Nachhaltigkeit zukünftig evaluiert und bewertet werden sollte. Zweck der Meta-Evaluierung ist schließlich, die Ausgestaltung einer Agenda-2030-konformen Evaluierungs- und Bewertungspraxis zu unterstützen.

1.3 Gegenstand

Den unmittelbaren Gegenstand der Meta-Evaluierung bildet die Evaluierungs- und Bewertungspraxis der Nachhaltigkeit von Vorhaben in der deutschen EZ, dargestellt in den Evaluierungsberichten der Durchführungsorganisationen (DO). Den mittelbaren Gegenstand bildet die Nachhaltigkeit von Vorhaben der deutschen finanziellen und technischen Entwicklungszusammenarbeit. Die Bearbeitung des Evaluierungsgegenstandes dient der fundierten Auseinandersetzung mit dem Nachhaltigkeitsverständnis der deutschen EZ.

Da der Evaluierungsgegenstand möglichst umfassend bearbeitet werden soll, beschränkt er sich weder auf bestimmte Sektoren noch auf bestimmte Regionen oder Typen von Vorhaben. Neben rein bilateralen Vorhaben in bestimmten Ländern sind auch Regional-, Sektor- und Globalvorhaben Teil der Untersuchung. Um dennoch die Durchführbarkeit dieser ersten rigorosen thematischen Meta-Evaluierung zu gewährleisten, wurde der Evaluierungsgegenstand folgendermaßen eingegrenzt:

Erstens beschränkt sich die Untersuchung auf die Evaluierungs- und Bewertungspraxis der beiden großen staatlichen Durchführungsorganisationen KfW und GIZ.[1] Diese beiden DO setzen jährlich einen wesentlichen Anteil der öffentlichen Entwicklungsfinanzierung um und verfügen jeweils über ein hoch diversifiziertes Portfolio von Vorhaben über alle Sektoren und Regionen der deutschen EZ hinweg. Gleichzeitig weisen die beiden DO einen hohen Deckungsgrad an Evaluierungen von Einzelvorhaben (den heutigen Modulen) auf. Bei allen Evaluierungen wurde stets auch die Nachhaltigkeit bewertet.

Zweites erfolgt eine zeitliche Eingrenzung: Die systematische und weitestgehend einheitliche Bewertung von Nachhaltigkeit als eines der Erfolgskriterien der deutschen EZ begann 2006 mit der Verabschiedung der BMZ-Orientierungshilfe zum Umgang mit den DAC-Kriterien. Deshalb werden nur Evaluierungen berücksichtigt, die zwischen Juli 2006 und dem Zeitpunkt der Datenerhebung im Oktober 2017 durchgeführt und abgeschlossen wurden.

1.4 Evaluierungsfragen

Die Evaluierungsziele werden durch fünf Evaluierungsfragen operationalisiert:

Evaluierungsfrage 1: Anhand welcher Kriterien wird Nachhaltigkeit im Rahmen von Evaluierungen bewertet?

Evaluierungsfrage 2: Inwieweit ist die Evaluierungspraxis der deutschen EZ angemessen, um Nachhaltigkeit zu bewerten?

Evaluierungsfrage 3: Inwieweit deckt sich die Evaluierungspraxis zu Nachhaltigkeit in der deutschen EZ mit internationalen Standards und heutigen Anforderungen?

Evaluierungsfrage 4: Wie ist die methodische Qualität der Evaluierungen?

Evaluierungsfrage 5: Inwieweit und in welchem Maße beeinflusst die methodische Qualität der Evaluierungen die Nachhaltigkeitsbewertung?

1.5 Aufbau des Evaluierungsberichtes

Der Bericht der Meta-Evaluierung gliedert sich wie folgt:

Kapitel 2 beginnt mit der Darstellung der Nachhaltigkeit als Erfolgskriterium in der Wirksamkeitsdebatte der deutschen EZ (Kapitel 2.1). Darauf aufbauend wird der konzeptionelle Rahmen der Meta-Evaluierung beschrieben (Kapitel 2.2). Das Kapitel schließt mit einem Einblick in die Evaluierungspraxis der deutschen technischen und finanziellen Zusammenarbeit ab (Kapitel 2.3).

Die methodische Vorgehensweise der Meta-Evaluierung wird in Kapitel 3 dargelegt. Das Kapitel beginnt mit der Beschreibung der Datengrundlage (Kapitel 3.1). Anschließend wird die methodische Vorgehensweise der Meta-Evaluierung hinsichtlich der Analyse der Evaluierungsqualität (Kapitel 3.2) und der Bewertungspraxis (Kapitel 3.3) vorgestellt. Die methodische Vorgehensweise der Kontextstudie findet sich in Kapitel 3.4.

[1] *Andere staatliche DO, wie die Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) und die Physikalisch-Technische Bundesanstalt (PTB), sind nicht Teil der Betrachtung.*

Abschließend folgt eine Diskussion zu den Limitationen der Meta-Evaluierung (Kapitel 3.5).

Die Ergebnisse der Untersuchung werden in Kapitel 4 vorgestellt. Das Kapitel beginnt mit den Ergebnissen zur Qualität der Evaluierungen (Kapitel 4.1) und widmet sich anschließend den Ergebnissen der Nachhaltigkeitsbewertung (Kapitel 4.2), welche entlang des konzeptionellen Rahmens der Meta-Evaluierung diskutiert werden. Abschließend werden in Kapitel 4.3 die Ergebnisse zu möglichen Zusammenhängen zwischen der Qualität von Evaluierungen und der Bewertungspraxis sowie in Kapitel 4.4 die Ergebnisse der Kontextstudie dargestellt.

Die Schlussfolgerungen und Empfehlungen der Meta-Evaluierung finden sich in Kapitel 5.

2.

EVALUIERUNG VON NACHHALTIGKEIT IN DER DEUTSCHEN EZ

2.1
Nachhaltigkeit in der Wirksamkeitsdebatte der deutschen EZ

Wegweisend für die Entwicklung des Nachhaltigkeitsverständnisses in der deutschen EZ war der internationale Diskurs um das Prinzip der Nachhaltigkeit seit den 1970er-Jahren (siehe Kapitel 1.1). Die Umsetzung der Diskussion in eine Auseinandersetzung mit der Nachhaltigkeit von EZ-Vorhaben erfolgte in der Praxis jedoch deutlich zeitversetzt. So rückte Nachhaltigkeit als Evaluierungskriterium erst ab Ende der 1980er-Jahre in den Fokus der deutschen Wirksamkeitsdebatte (Stockmann und Gaebe, 1993). Zu dieser Zeit bewegte sich das Nachhaltigkeitsverständnis zwischen den Aspekten der nachhaltigen Entwicklung einerseits und dem Aspekt der Dauerhaftigkeit von Entwicklungserfolgen über die Zeit andererseits.

Den Anstoß, Nachhaltigkeit als Erfolgskriterium in Projektevaluierungen aufzunehmen, gab es 1986 durch eine Empfehlung des OECD-DAC, auf deren Grundlage später auch das BMZ Nachhaltigkeit als relevanten Maßstab für die Erfolgsbewertung der deutschen EZ ausrief. Einzug in die Evaluierungspraxis der staatlichen EZ hielt das Kriterium Nachhaltigkeit Ende der 1980er-Jahre zunächst über die Ex-post-Evaluierungen der KfW (Stockmann und Gaebe, 1993). Später wurde Nachhaltigkeit zunehmend auch in Evaluierungen der GIZ berücksichtigt.

Den Auslöser für die systematische Auseinandersetzung mit Nachhaltigkeit als Erfolgskriterium der EZ bildete schließlich eine 1998/99 durch das BMZ in Auftrag gegebene Wirkungsuntersuchung. In der Studie wurde in 32 ausgewählten Ex-post-Evaluierungen der deutschen staatlichen technischen und finanziellen Zusammenarbeit (TZ und FZ) auch die langfristige Wirksamkeit untersucht. Neben der Aggregation der inhaltlichen Ergebnisse stand anschließend die Evaluierungs- und Bewertungspraxis im Fokus einer begleitenden Querschnittsauswertung durch Caspari (2004). Die damit einhergehende Debatte machte deutlich, dass Nachhaltigkeit zu diesem Zeitpunkt in der deutschen EZ sehr heterogen verstanden und bewertet wurde, was die Möglichkeiten einer inhaltlichen Querschnittsauswertung deutlich einschränkte.

Aufgrund dieser Ergebnisse und vor dem Hintergrund der Empfehlungen des OECD-DAC, die Evaluierungssysteme der Mitgliedsstaaten zu harmonisieren, widmete sich daraufhin eine Arbeitsgruppe unter Federführung des BMZ und Beteiligung der Durchführungsorganisationen dem Thema „Evaluierung aus einem Guss". Ziel war es, Evaluierungen in der bilateralen deutschen EZ stärker zu vereinheitlichen und an internationalen Standards zu orientieren. Die Arbeit dieser Gruppe führte schließlich zur verbindlichen Festlegung der OECD-DAC-Evaluierungskriterien als Orientierungsrahmen für die Erfolgsbewertung der deutschen EZ (OECD, 1991). Neben den Kriterien Relevanz, Effektivität, Effizienz und übergeordnete entwicklungspolitische Wirkungen gehört seither auch die Nachhaltigkeit zu den verbindlichen Evaluierungskriterien. Im Rahmen des Leitfadens wurde das Evaluierungskriterium Nachhaltigkeit durch drei zentrale Prüffragen operationalisiert (BMZ, 2006):

- Inwieweit sind die positiven Veränderungen durch die Entwicklungsmaßnahme und ihre Wirkungen mit Blick auf die entwicklungspolitischen Zielsetzungen (summarisch) als dauerhaft einzuschätzen?
- Wie stabil ist die Situation im Umfeld der Entwicklungsmaßnahme bezüglich der Faktoren „soziale Gerechtigkeit", „wirtschaftliche Leistungsfähigkeit", „politische Stabilität" und „ökologisches Gleichgewicht"?
- Welche Risiken und Potenziale zeichnen sich für die nachhaltige Wirksamkeit der Entwicklungsmaßnahme ab, und wie wahrscheinlich ist das Eintreten dieser Faktoren?

Wie in Kapitel 1.1 bereits herausgestellt, beinhaltet das zugrunde liegende Nachhaltigkeitsverständnis dabei sowohl den Aspekt der Dauerhaftigkeit als auch – über den Wirkungsbegriff – den Aspekt der nachhaltigen Entwicklung. Insofern handelt es sich bei dem Evaluierungsverständnis von Nachhaltigkeit in der deutschen EZ seit 2006 um ein umfassendes und vielfältiges Konzept (siehe Kapitel 2.2). In der Evaluierungspraxis ermöglicht erst die Gesamtbetrachtung der Ergebnisse zu den fünf Evaluierungskriterien Relevanz, Effektivität, Effizienz, entwicklungspolitische Wirkungen und Nachhaltigkeit demnach eine abschließende Diskussion und Bewertung der Nachhaltigkeit mit den Aspekten nachhaltiger Entwicklung und Dauerhaftigkeit.

Mit der Festlegung der zentralen Prüffragen im BMZ-Leitfaden von 2006 wurde schließlich im Rahmen einer Vereinbarung zwischen GIZ und KfW auch eine Bewertungsskala festgelegt. Seither wird Nachhaltigkeit entlang von vier Nachhaltigkeitsnoten[2] bewertet, welche in jedem Evaluierungsbericht dargestellt werden: Note 1 wird vergeben, wenn die (bisher positive) entwicklungspolitische Wirksamkeit des Vorhabens mit hoher Wahrscheinlichkeit unverändert fortbestehen oder zunehmen wird. Ein Vorhaben erhält die Note 2, wenn dessen (bisher positive) entwicklungspolitische Wirksamkeit mit hoher Wahrscheinlichkeit nur geringfügig zurückgehen wird. Die Note 3 bedeutet, dass die (bisher positive) entwicklungspolitische Wirksamkeit mit hoher Wahrscheinlichkeit deutlich zurückgehen, aber positiv bleiben wird, oder aber, dass sie zum Evaluierungszeitpunkt als nicht ausreichend eingeschätzt, sich aber mit hoher Wahrscheinlichkeit positiv entwickeln wird. Die Note 4 wird vergeben, wenn die entwicklungspolitische Wirksamkeit zum Evaluierungszeitpunkt als nicht ausreichend eingeschätzt und sich mit hoher Wahrscheinlichkeit auch nicht verbessern wird. Bei der abschließenden Betrachtung der Nachhaltigkeit eines Vorhabens beschreiben die Noten 1 bis 3 ein Vorhaben als „nachhaltig", die Note 4 als „nicht nachhaltig". Gleichzeitig wird Nachhaltigkeit bei der Gesamtbewertung eines Vorhabens ein vergleichsweise hohes Gewicht beigemessen: So kann ein Vorhaben insgesamt nur als „erfolgreich" (Note 1 bis 3 von 6) bewertet werden, wenn auch das Kriterium Nachhaltigkeit als „erfolgreich" eingestuft wurde. Nur den Kriterien „Effektivität" und „entwicklungspolitische Wirkungen" kommt ebenfalls ein solches Gewicht zu.

2.2 Der konzeptionelle Rahmen der Meta-Evaluierung im Hinblick auf die Bewertung der Nachhaltigkeit

Aufgrund der breit geführten Debatte zum Prinzip der Nachhaltigkeit in der EZ und zur systematischen Erfolgsbewertung von Vorhaben der deutschen EZ entlang der DAC-Kriterien wird bei der vorliegenden Meta-Evaluierung davon ausgegangen, dass Nachhaltigkeit bereits seit einiger Zeit umfassend und vielschichtig verstanden wird. Ferner wird davon ausgegangen, dass das Nachhaltigkeitsverständnis dabei einerseits auf dem Aspekt der nachhaltigen Entwicklung im Sinne zukunftsfähiger Wirksamkeit und andererseits auf dem Aspekt der Dauerhaftigkeit der Wirkungen basiert (siehe Kapitel 1.1 und 2.1). Ein solch umfassendes Nachhaltigkeitsverständnis in der Evaluierungs- und Bewertungspraxis wird demnach erst in der Gesamtschau aller DAC-Kriterien nachvollziehbar, da sich wesentliche Teile erst über den Nachweis entwicklungspolitischer Wirksamkeit ergeben. Der konzeptionelle Rahmen für die empirische Auseinandersetzung mit der Nachhaltigkeit im umfassenden Sinne muss somit nachhaltigkeitsbezogene Aspekte aus allen fünf Bereichen der Erfolgsbewertung (den fünf Evaluierungskriterien nach OECD-DAC) einbeziehen. In der Gesamtschau der Prüffragen aller DAC-Kriterien konnten insgesamt sieben (abgrenzbare) Bereiche mit konkretem Nachhaltigkeitsbezug identifiziert werden, die auch in der Literatur immer wieder mit Nachhaltigkeit in Zusammenhang gebracht werden. Die Bereiche werden im Folgenden einzeln vorgestellt.

Laut dem BMZ-Leitfaden soll bei der Bewertung des Evaluierungskriteriums „Nachhaltigkeit" nach OECD-DAC die Stabilität des Umfelds einer Entwicklungsmaßnahme berücksichtigt werden (BMZ, 2006). Die Analyse des **1) Kontextes einer Entwicklungsmaßnahme** soll dabei anhand der Faktoren „soziale Gerechtigkeit", „wirtschaftliche Leistungsfähigkeit", „politische Stabilität" und „ökologisches Gleichgewicht" erfolgen. Die Untersuchung von Kontextfaktoren ermöglicht schließlich die fundierte Auseinandersetzung mit den externen Risiken und Potenzialen im Hinblick auf die Dauerhaftigkeit der positiven Wirkungen über die Zeit.

Nach der Logik der DAC-Kriterien sind weiterhin Kriterien aus dem Bereich der **2) Implementierung** von Maßnahmen für die Einschätzung nachhaltiger Entwicklungserfolge von Bedeutung, beispielsweise die Partizipation der Partner und Zielgruppen an den Umsetzungsprozessen oder der Grad der Anpassung an die Prioritäten des Partnerlandes. Solche Aspekte der internationalen Wirksamkeitsagenda der EZ sind wesentliche Bestandteile des Relevanz- bzw. des Effektivitätskriteriums (BMZ, 2006).

Des Weiteren sind auch Ergebnisse der **3) Outcomes**, d. h. der kurz- und mittelfristigen Wirkungen einer Entwicklungsmaß-

[2] *Die anderen vier DAC-Kriterien werden durch eine Notenskala von 1 bis 6 bewertet. Seit 2014 bewertet die GIZ die Nachhaltigkeit anhand einer sechsstufigen Skala entlang eines Punktesystems: „sehr erfolgreich" (14–16 Punkte), „erfolgreich" (12–13 Punkte), „eher erfolgreich" (10–11 Punkte), „eher unbefriedigend" (8–9 Punkte), „unbefriedigend" (6–7 Punkte) und „sehr unbefriedigend" (4–5 Punkte).*

nahme, für die Analyse nachhaltiger Entwicklung von Relevanz (Ashoff, 2015). Neben der Quantität und Qualität von Maßnahmen gehören dazu auch die angestoßenen Veränderungen, etwa im Hinblick auf Eigenverantwortung, Bewusstsein oder Resilienz der Akteure vor Ort oder auch die dadurch erzielte Breitenwirksamkeit eines Vorhabens (Boone, 1995). Die Diskussion der Outcomes erfolgt im BMZ-Leitfaden vornehmlich über das Effektivitätskriterium, teilweise aber auch bereits über das Impact-Kriterium.

Unter dem Nachhaltigkeitskriterium wird hinsichtlich der Prüffrage zu den Risiken und Potenzialen auch empfohlen, die **4) Kapazitäten vor Ort** zu berücksichtigen. Dies gibt Hinweise darauf, inwieweit es den lokalen Partnern, Trägern und Zielgruppen gelingt, die Leistungen und Wirkungen auch ohne externe Unterstützung aufrechtzuerhalten. Während der direkte Einfluss der Leistungen von Vorhaben der deutschen EZ über die Zeit abnimmt und schließlich mit dem Auslaufen der Förderung endet, gewinnen die Kapazitäten vor Ort dabei im Zeitverlauf an relativer Bedeutung (van Tulder und Pfisterer, 2008). Die Kapazitäten vor Ort werden bislang vornehmlich unter dem Evaluierungskriterium „Nachhaltigkeit" diskutiert.

Entlang der Wirkungszusammenhänge bilden die Beiträge einer Maßnahme zu den übergeordneten beabsichtigten entwicklungspolitischen Wirkungen – dem **5) Impact** – einen weiteren integralen Bestandteil des Nachhaltigkeitsverständnisses. Dazu gehören die positiven und negativen, primären und sekundären Langzeiteffekte, die direkt oder indirekt, beabsichtigt oder unbeabsichtigt durch ein Vorhaben hervorgerufen werden. Die intendierten Wirkungen werden in der Regel bewertet, indem die geplanten und erreichten Wirkungen eines Vorhabens im Hinblick auf formulierte Oberziele (OZ) sowie auf globale Agenden (etwa Armutsbekämpfung) miteinander verglichen werden. Auch die nicht intendierten Wirkungen fließen in die Bewertung ein. Der BMZ-Leitfaden verweist hier zur Einschätzung von Nachhaltigkeit auf die Feststellung „übergeordneter entwicklungspolitischer Wirkungen", hebt also explizit auf die Impact-Ebene ab. Impact bildet nach OECD-DAC ein eigenes Evaluierungskriterium.

Ferner bildet die **6) Absehbarkeit von Wirkungen** einen zentralen Aspekte des Nachhaltigkeitsverständnisses in der deutschen EZ (Caspari, 2004; OECD, 1991; Stockmann und Gaebe, 1993; Stockmann und Silvestrini, 2012). Laut der Orientierungshilfe des BMZ soll dabei abgeschätzt werden, inwieweit die positiven Wirkungen der Entwicklungsmaßnahme über das Ende der Unterstützung hinaus fortbestehen (BMZ, 2006). Die Absehbarkeit von Wirkungen ist der zentrale Aspekt des Nachhaltigkeitskriteriums nach OECD-DAC.

Schließlich umfasst eine Auseinandersetzung mit den Wirkungen (unter Impact) neben den Nachhaltigkeits-Dimensionen soziale Gerechtigkeit, wirtschaftliche Leistungsfähigkeit, politische Stabilität und ökologisches Gleichgewicht auch die Betrachtung potenzieller Synergien und/oder Konflikte zwischen den Dimensionen (BMZ, 2006). Es wird davon ausgegangen, dass sich durch die Berücksichtigung aller Dimensionen Synergien erreichen und somit nachhaltigere Wirkungen erzielen lassen (OECD, 2016a). In der Nachhaltigkeitsdebatte geht es dabei um das **7) Zusammenspiel der Dimensionen der Nachhaltigkeit**, die folglich auch in Evaluierungen der EZ Berücksichtigung finden sollten (Cutter, 2014; Dietz und Hanemaaijer, 2012; Islam und Clarke, 2005). Aufgrund der Bedeutung für die Nachhaltigkeit wurde das Zusammenspiel der Dimensionen auch in die Kernprinzipien der Agenda 2030 aufgenommen (UN, 2015).

2.3 Die Evaluierungspraxis in der deutschen finanziellen und technischen Zusammenarbeit

Ziel der Evaluierung von Entwicklungsmaßnahmen ist es, die nachhaltige entwicklungspolitische Wirksamkeit der EZ zu beurteilen. Laut den „Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit" führen die DO eigenverantwortlich Evaluierungen für eine aussagekräftige Stichprobe von abgeschlossenen und gegebenenfalls laufenden Entwicklungsmaßnahmen durch. Dies erfolgt „in Abstimmung mit von der Bundesregierung festgelegten Verfahren und in Anlehnung an die OECD-DAC-Kriterien und -Standards für unabhängige Evaluierung" (BMZ, 2008).

Gemäß der Vorgaben weist die deutsche staatliche EZ auf Modulebene insgesamt einen hohen Deckungsgrad an Evaluierungen auf: Bei der GIZ wurden in den vergangenen zehn Jahren nahezu alle Vorhaben (bzw. Module/Phasen) mindestens einer Evaluierung unterzogen. Bei der KfW wird mindestens die Hälfte aller Vorhaben in jedem Sektor evaluiert. Zum Zeitpunkt der Datenerhebung der vorliegenden Meta-Evaluierung im Oktober 2016 lagen 1.081 abgeschlossene Evaluierungen vor, mit denen seit 2006 die Nachhaltigkeit von insgesamt 1.269 Vorhaben bewertet wurde. Dabei kamen verschiedene Evaluierungstypen zum Einsatz: Einige Formate werden im Verlauf von Vorhaben eingesetzt und dabei zum Teil auch für die Steuerung und Planung von Folgemaßnahmen verwendet; Ex-post-Evaluierungen hingegen blicken in einem gewissen zeitlichen Abstand auf den Erfolg abgeschlossener Maßnahmen zurück.

Die KfW setzt für die Erfolgsbewertung von Vorhaben der finanziellen Zusammenarbeit ausschließlich Ex-post-Evaluierungen ein, die in der Regel drei bis fünf Jahre nach Ende der Vorhaben durchgeführt werden. Die Auswahl der Vorhaben erfolgt seit 2006 anhand eines festgelegten Stichprobenplans, welcher jedes Jahr 50 Prozent der „evaluierungsreifen“[3] Vorhaben eines jeden Sektors einbezieht. Gesteuert werden die Evaluierungen von der unabhängigen Evaluierungseinheit der KfW Entwicklungsbank (FZ-E). Die Evaluierungen werden von Mitarbeiterinnen und Mitarbeitern der FZ-E gemeinsam mit sogenannten Abgeordneten, d. h. Mitarbeiterinnen und Mitarbeitern anderer Unternehmensbereiche, sowie mit externen Gutachterinnen und Gutachtern durchgeführt. Die Ex-post-Evaluierungen der KfW folgen in der Regel einem standardisierten Prozess: Nach der Erstellung eines Evaluierungskonzepts wird ein Fragebogen an den Träger der Entwicklungsmaßnahme versandt. Anschließend werden verfügbare Monitoring- und Abschlusskontrollberichte ausgewertet. Es folgen Vor-Ort-Missionen, teilweise unterstützt durch unabhängige technische Sachverständige, und schließlich die Erstellung eines Evaluierungsberichtes. Nach Angaben der KfW umfasst der Gesamtprozess einer Ex-post-Evaluierung etwa 37 Arbeitstage, davon 27 Arbeitstage für die eigentliche Evaluierung und 10 Arbeitstage für die Qualitätssicherung durch die Evaluierungseinheit.

Die Evaluierung von Vorhaben der TZ erfolgt in der GIZ seit 2006 in Form zentraler und dezentraler Evaluierungen. Zu den zentralen Evaluierungstypen der GIZ gehören Schluss- und Ex-Post-Evaluierungen. Die dezentralen Evaluierungstypen sind die heutigen Projektevaluierungen (PEV) und die früheren Projektfortschrittskontrollen (PFK). Die vier betrachteten Evaluierungsformate der GIZ werden im Folgenden kurz vorgestellt.

Bei den PEV handelt es sich um ein Evaluierungsformat, das, wenn eine Folgemaßnahme geplant ist, mit der Projektprüfung verbunden ist. Die PEV wurden im April 2014 eingeführt und bilden heute das Evaluierungsformat der GIZ auf Ebene der Module. Sie kommen in der Regel zwölf bis sechs Monate vor Ende der Vorhaben zum Einsatz. Eine PEV beginnt mit der Festlegung des Evaluierungsgegenstandes und der Ausgestaltung des Evaluierungsdesigns. Dabei wird definiert, welche Aktivitäten dem Zweck der Evaluierung und welche Aktivitäten dem Zweck der Prüfung dienen. Anschließend beginnt die Datenerhebung (vor Ort) mit einem Auftakt-Workshop mit den Stakeholdern einer Evaluierung. In einem Abschlussworkshop werden die vorläufigen Ergebnisse nach der Systematik der OECD-DAC-Kriterien vorgestellt. Die Abnahme des Evaluierungsberichtes liegt in der Verantwortung der Auftragsverantwortlichen der Vorhaben. Zuvor wird der Bericht selbst einem Qualitätscheck durch die Stabsstelle Evaluierung der GIZ unterzogen. Auch die Abnahme des zu veröffentlichenden Kurzberichtes obliegt der Stabsstelle. Für PEV ohne Folgemaßnahme werden durchschnittlich 49 und für PEV mit Folgemaßnahmen 74 Arbeitstage benötigt, wobei nicht klar ist, wie viele Tage davon für die Evaluierung und wie viele für die Prüfung zur Planung des Folgevorhabens verwendet werden. Das Mengengerüst ist flexibel gestaltbar, je nachdem, ob es sich um ein besonders komplexes Vorhaben handelt oder ob eher ein mittlerer oder einfacher Aufwand erwartet wird. Für die Qualitätssicherung des Kurzberichtes veranschlagt die Stabsstelle Evaluierung ca. einen Arbeitstag.

Die früheren PFK – die von den PEV abgelöst wurden – vereinten immer Evaluierungs- und Prüfungsaspekte in einem Format. PFK ohne Folgephase gab es nicht. Evaluierungsgegenstand war die jeweilige Phase der Entwicklungsmaßnahme. Der Ablauf einer PFK ähnelte bereits dem Prozess der PEV.

3 Als „evaluierungsreif“ gelten die Vorhaben der KfW, die zum Zeitpunkt der Stichprobenziehung bereits seit mindestens drei Jahren beendet sind.

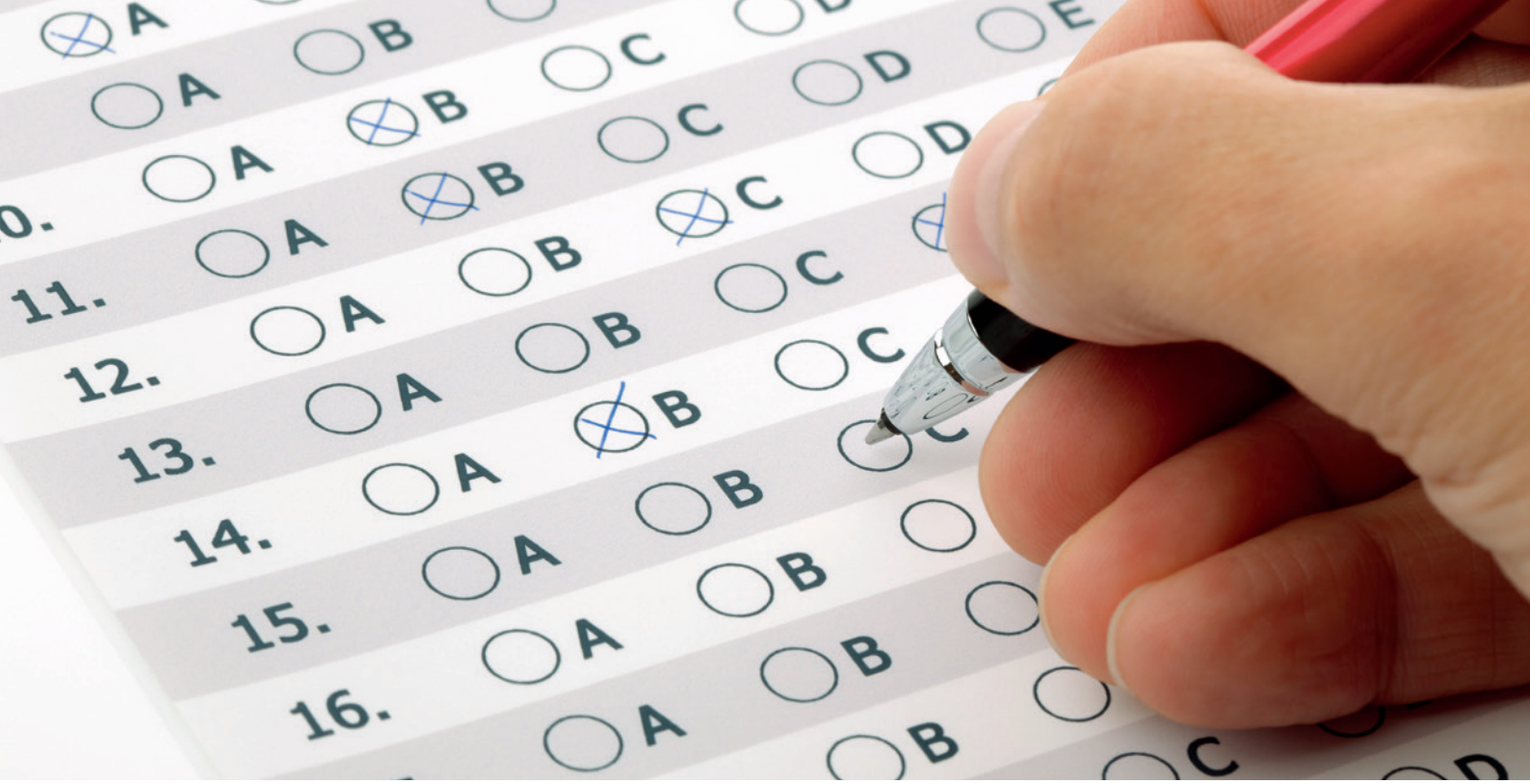
Auch die PFK wurden in der Regel zwölf bis sechs Monate vor Ende der Vorhaben durchgeführt, und auch bei ihnen ging ein Abstimmungsprozess mit den Partnern vor Ort voraus. Wie heute bei den PEV lag die Steuerungsverantwortung der PFK bei dem Auftragsverantwortlichen der Vorhaben. Eine Qualitätsüberprüfung der PFK durch die Stabsstelle Evaluierung wurde stichprobenhaft im Rahmen von GIZ-Meta-Evaluierungen vollzogen. Es gab keine festen Vorgaben für die Gesamtarbeitstage, im Schnitt wurden jedoch etwa 23 Arbeitstage für Vorbereitung, Durchführung und Auswertung benötigt.

Die Schluss- und Ex-post-Evaluierungen der GIZ wurden im Rahmen des ehemaligen Unabhängigen Evaluierungsprogramms (UE-Programms)[4] von der Stabsstelle Evaluierung verantwortet, konzipiert und gesteuert. Hierbei wurden jeweils einzelne Sektoren über einen bestimmten Zeitraum betrachtet. Mit der Durchführung wurden in der Regel unabhängige Institute und Consultingfirmen beauftragt. Die Schluss-Evaluierungen fanden in der Regel im Zeitraum zwischen sechs Monaten vor und sechs Monaten nach Vorhabenende statt; Ex-post-Evaluierungen erfolgten zwei bis fünf Jahre nach Ende der Vorhaben. Für die Durchführung der Evaluierung waren 42 Tage für den internationalen Gutachter und 30 für den nationalen Gutachter vorgesehen, bei einzelnen, methodisch anspruchsvollen Evaluierungen konnten es mehr sein. Management und Qualitätssicherung durch die Stabsstelle nahmen ca. 12 Arbeitstage in Anspruch. Im Unterschied zu den PFK und PEV war Evaluierungsgegenstand die Entwicklungsmaßnahme über ihre gesamte Laufzeit mit allen Phasen; zudem wurde ein Inception Report erstellt.

Letztlich sind die Angaben zum Aufwand der einzelnen Evaluierungsinstrumente allerdings nur eingeschränkt miteinander vergleichbar, da es sich bei PFK und UE um vorgeschlagene Planwerte und bei den Angaben zu PEV um Ist-Werte handelt (GIZ, 2016).

Die Gegenüberstellung der verschiedenen Evaluierungsformate von KfW und GIZ macht Unterschiede deutlich. So wurden die ehemaligen Schluss- und Ex-post-Evaluierungen der GIZ vergleichsweise aufwändig durchgeführt und von der Stabsstelle gesteuert. Die Ex-post-Evaluierungen der KfW haben einen geringeren Umfang und werden im Rahmen des Abgeordnetensystems von Mitarbeiterinnen und Mitarbeitern begleitet. Auch hier erfolgt die Qualitätssicherung zentral durch die Evaluierungseinheit. Damit unterscheiden sich diese drei Formate von den dezentralen Evaluierungen der GIZ (PEV und PFK), die in der Verantwortung der Auftragsverantwortlichen der Vorhaben liegen und bislang nur stichprobenhaft oder zu Teilen eine Qualitätssicherung durch die Stabsstelle erfahren. Der Aufwand für die ehemaligen PFK lag tendenziell unter dem Aufwand für die übrigen Evaluierungsformate.

4 Zu den unabhängigen Evaluierungen gehörten auch Ex-ante- und Zwischenevaluierungen. Aufgrund der geringen Aussagekraft hinsichtlich der Nachhaltigkeit werden diese Evaluierungstypen hier nicht berücksichtigt.



3.

METHODISCHE VORGEHENSWEISE

3.1 Datengrundlage

Datengrundlage bilden die evaluierten Vorhaben der KfW und der GIZ seit Verabschiedung des BMZ-Leitfadens zum einheitlichen Umgang mit den DAC-Kriterien 2006. In die Grundgesamtheit flossen grundsätzlich alle Vorhaben ein, die anhand von dezentralen und zentralen Evaluierungen unabhängig voneinander auf ihre Nachhaltigkeit hin bewertet wurden.

Bei der Bestimmung der Grundgesamtheit galt es zu beachten, dass Vorhaben von GIZ und KfW häufig aus chronologisch und inhaltlich aufeinander aufbauenden Phasen (bzw. Modulen) bestehen. Während auf Schluss- und Ex-post-Evaluierungen keine weitere Phase (bzw. Modul) eines Vorhabens folgt, kann es nach Durchführung einer PFK oder PEV eine weitere Phase bzw. ein weiteres Modul eines Vorhabens und somit eine zeitlich spätere Evaluierung geben. Im Sinne einer möglichst späten Bewertung der Nachhaltigkeit enthält die Grundgesamtheit nur die jeweils jüngste Evaluierung einer Maßnahme.

Zum Zeitpunkt der Datenerhebung im Oktober 2016 erfüllten 1.015 Vorhaben diese Voraussetzungen. Aus dem Bereich der finanziellen Zusammenarbeit wurden 462 Ex-post-evaluierte Vorhaben der KfW in die Grundgesamtheit aufgenommen. Von den zentralen Evaluierungen der GIZ flossen 56 Ex-post- und 44 Schluss-evaluierte Vorhaben ein. Aus dem Bereich der dezentralen Evaluierungen wurden 110 Vorhaben berücksichtigt, die Projektevaluierungen unterzogen worden waren, sowie 343 Vorhaben, die durch Projektfortschrittskontrollen überprüft worden waren (siehe Tabelle 1).

Für die vorliegende Meta-Evaluierung wurde eine aussagekräftige Stichprobe der vorgestellten Grundgesamtheit untersucht. Dabei wurden unterschiedliche Evaluierungstypen sowie die Verteilung der Nachhaltigkeitsnoten berücksichtigt. Formal erfolgte eine nach Evaluierungstyp geschichtete, randomisierte Stichprobenziehung, die pro Typ sowohl für den Mittelwert der vierstufigen Notenverteilung (1–4) als auch für die binäre Verteilung zwischen „erfolgreichen“ (Note 1–3) und „nicht erfolgreichen“ (Note 4) Vorhaben repräsentativ ist (siehe Tabelle 1). Insgesamt hielten damit 513 Vorhaben Einzug in die Stichprobe.

Entsprechend der Zweiteilung dieser Meta-Evaluierung werden die Analyse zur Qualität der Berichte und die Analyse zur Nachhaltigkeitsberichterstattung mit dem jeweils zugehörigen methodischen Vorgehen in zwei Kapiteln dargestellt.

Tabelle 1: Vorstellung der Datengrundlage

	Evaluierungstyp	Zeitpunkt im Vorhaben	Anzahl der evaluierten Vorhaben	Anzahl der evaluierten Vorhaben in der Stichprobe
GIZ	PFK	12 bis 6 Monate vor Ende	343	174
	Schluss-Evaluierungen	± 6 Monate vor/nach Ende	44	38
	PEV	12 bis 6 Monate vor Ende	110	82
	Ex-post-Evaluierungen	2 bis 5 Jahre nach Ende	56	47
KfW	Ex-post-Evaluierungen	3 bis 5 Jahre nach Ende	462	172
Gesamt			**1.015**	**513**

Quelle: eigene Darstellung

3.2 Evaluierungspraxis

Für eine fundierte inhaltliche Analyse der Bewertung von Nachhaltigkeit auf der Grundlage von Evaluierungen muss zunächst die Belastbarkeit der Evaluierungsergebnisse überprüft werden. Dabei wird davon ausgegangen, dass die Überprüfung der Qualität einer Evaluierung insgesamt es auch erlaubt, Rückschlüsse auf die Qualität der darin enthaltenen Evaluierung zum Kriterium Nachhaltigkeit zu ziehen. Die Überprüfung der Evaluierungsqualität bildet den ersten Teil der vorliegenden Meta-Evaluierung.

Eine Meta-Evaluierung wird auch als „Evaluierung von Evaluierungen" bezeichnet wird (Patton, 2008; Scriven, 1991, 2009). Ziel einer Meta-Evaluierung ist die systematische Auseinandersetzung mit der Qualität von Evaluierungsprozessen und mit der Belastbarkeit der gezogenen Schlussfolgerungen (Leeuw und Cooksy, 2005). Um die Qualität einzelner Evaluierungen vergleichen zu können, müssen zunächst einheitliche Qualitätskriterien an die untersuchten Evaluierungsberichte angelegt werden.

Bei der Entwicklung des Evaluierungsrasters (Tabelle 2) zur Qualitätsbewertung wurden Erkenntnisse der Evaluierungsforschung (Patton, 2008; Scriven, 2009; Stufflebeam, 2001; Widmer, 2006) sowie Anwendungsbeispiele aus dem Bereich der EZ (Carlsson und Wohlgemuth, 1996; Hageboeck et al., 2013; Leeuw und Cooksy, 2005) herangezogen. Darüber hinaus wurden die internen Vorgaben der Evaluierungspraxis von KfW und GIZ berücksichtigt. Weitere Orientierung boten die bereits vorliegenden – wenn auch unveröffentlichten – Meta-Evaluierungen im Bereich der deutschen TZ. Anschließend erfolgte ein Pretest des Evaluierungsrasters anhand ausgewählter Berichte nach Evaluierungstypen.

Das finale Evaluierungsraster gliedert sich in sechs Analysebereiche mit insgesamt 16 Qualitätskriterien. Für jeden Bericht in der Stichprobe wurden alle Kriterien jeweils als „erfüllt" oder „nicht erfüllt" gewertet. Mit der Aufnahme der einzelnen Kriterien in das Evaluierungsraster wurde für jedes Bewertungskriterium eine Definition erarbeitet. Eine ausführliche Beschreibung der angelegten Bewertungskriterien inklusive der Definitionen befindet sich im Anhang in Tabelle 4. Die Basis für die Überprüfung der Evaluierungsqualität bildete dabei stets der Gesamtbericht, d. h. sämtliche schriftlichen Ausführungen der Evaluierung, inklusive Anhänge.

Im Rahmen der Untersuchung der Berichtsqualität wurden zunächst das entwickelte Kriterienraster sowie die zu untersuchenden Berichte als PDF-Dateien in die qualitative Analysesoftware „MAXQDA" überführt. In einem weiteren Schritt wurde anhand der gelesenen Berichte die Einschätzung, ob ein Kriterium erfüllt ist, mit Hilfe der Datenmanagementsoftware „Microsoft Access" in einer Datenbank festgehalten. Die jeweilige Textstelle, auf der die Einschätzung beruht, wurde mittels Softwareunterstützung referenziert, sodass die Einschätzungen auch im Nachhinein nachvollziehbar sind.

Um die intersubjektive Vergleichbarkeit der Einschätzungen innerhalb des Evaluationsteams zu überprüfen, wurden 10 Prozent der untersuchten Berichte – geschichtet nach Evaluierungstyp – mehrfach kodiert, d. h. von verschiedenen Personen gelesen und bewertet. Anschließend wurde mit Hilfe des Kappa-Inter-Kodierer-Reliabilitätskoeffizienten nach Cohen der Grad der Übereinstimmung des Kodierverhaltens über die Evaluatorinnen und Evaluatoren hinweg errechnet. Der Kappa-Wert für die Qualitätsbewertung liegt bei 0,62 und steht für eine substanzielle bzw. starke Übereinstimmung (Landis und Koch, 1977).[5]

Neben einer beschreibenden Analyse der Qualitätskriterien ermöglicht ein aggregierter „Qualitäts-Index" den direkten Vergleich zwischen Evaluierungsberichten. Zur Bildung des Index wurde zunächst die Anzahl erfüllter Kriterien summiert. Da der Fokus der Untersuchung auf der Qualität der Aussagen zur Einschätzung der Nachhaltigkeit von Vorhaben lag, wurden Kriterien, die Aussagen über die Belastbarkeit der Ergebnisse liefern (Q-9 bis Q-16), doppelt gewichtet. Eine einzelne Evaluierung konnte demnach maximal 24 Punkte erreichen. Um die Interpretation zu erleichtern, wurde schließlich der jeweils erreichte Wert durch den Maximalwert von 24 Punkten geteilt und somit der Index auf den Wertebereich von 0 bis 1 skaliert.

[5] *Die übliche bzw. meistzitierte Interpretation des Cohen-Kappa-Koeffizienten geht auf eine Arbeit von Landis und Koch (1977) zurück, welche folgende Interpretationsstufen vorschlagen: „0,01–0.20 = geringe Übereinstimmung (slight agreement), 0,21–0,40 = moderate mittlere Übereinstimmung (fair agreement), 0,41–0,60 = moderate Übereinstimmung (moderate agreement), 0,61–0,80 = starke Übereinstimmung (substantial agreement), 0,81–0,99 = fast perfekte Übereinstimmung (almost perfect agreement).*

Tabelle 2: Übersicht der Qualitätskriterien

Bereiche	Kriterien
1) Hintergrund der Evaluierung	Q-01 Gegenstand beschrieben
	Q-02 Erkenntnisinteresse formuliert
2) Darstellung der Wirkungszusammenhänge	Q-03 Wirkungslogik dargestellt
	Q-04 Indikatoren formuliert
3) Methodisches Vorgehen	Q-05 Methodisches Vorgehen beschrieben
	Q-06 Stärken und Limitationen der Evaluierung diskutiert
	Q-07 Befragte Stakeholder identifiziert
	Q-08 Auswahlverfahren beschrieben
4) Evaluierungsdesign	Q-09 Vorher-Nachher-Vergleich
	Q-10 Kontroll-/Vergleichsgruppen
	Q-11 Kausalität durch Plausibilitäten hergeleitet
5) Robustheit der Ergebnisse	Q-12 Daten-Triangulation
	Q-13 Methoden-Triangulation
6) Auswertung/Schlussfolgerungen	Q-14 Schlussfolgerungen referenziert
	Q-15 Schlussfolgerungen plausibel
	Q-16 Datengrundlage ausreichend

Quelle: eigene Darstellung

3.3 Bewertungspraxis

Den zweiten Teil der Meta-Evaluierung bildet die Untersuchung der Bewertungskriterien für Nachhaltigkeit. In Übereinstimmung mit der Logik der Qualitätsprüfung (dargestellt in Kapitel 3.2) gingen einzelne Bewertungskriterien in ein Bewertungsraster ein, welches anschließend den Rahmen für die quantitative Inhaltsanalyse darstellte (siehe Tabelle 3). Damit wurde das klassische Design einer Meta-Evaluierung im Sinne einer Qualitätsbewertung um die Auseinandersetzung mit den konkreten Bewertungskriterien erweitert. Erst diese Erweiterung auf eine thematische Meta-Evaluierung erlaubt schließlich die umfassende Auseinandersetzung mit der Evaluierungs- und Bewertungspraxis von Nachhaltigkeit in der deutschen EZ.

Den konzeptionellen Rahmen des Evaluierungsrasters zur Analyse der Bewertungskriterien von Nachhaltigkeit bildete der BMZ-Leitfaden zum Umgang mit den OECD-DAC-Evaluierungskriterien (siehe Kapitel 2.2). Die daraus abgeleiteten Bereiche gaben Orientierung bei der Identifizierung von spezifischen Kriterien, die erwartungsgemäß für die Bewertung der Nachhaltigkeit von Vorhaben herangezogen werden. Zudem wurden theoretisch möglichen Kriterien anhand der Auswertung der Leitlinien von KfW und GIZ gesammelt. Ferner erfolgte ein Abgleich mit der aktuellen Literatur zur Evaluierung von Nachhaltigkeit. Allerdings lässt die hohe Komplexität des Nachhaltigkeitskonzepts vermuten, dass eine rein deduktive Herangehensweise das zugrunde liegende Nachhaltigkeits-Verständnis in der Praxis nicht vollständig abbilden kann. Daher wurde das deduktive Vorgehen durch eine explorative Studie von 40 Evaluierungsberichten von KfW und GIZ ergänzt. Die Studie diente dem Abgleich zwischen Theorie und

Praxis, wobei Besonderheiten von FZ- und TZ-Projekten, von unterschiedlichen Evaluierungstypen sowie der Evaluierungs- und Bewertungspraxis über die Zeit berücksichtigt wurden. Das Analyseraster wurde anschließend nochmals anhand einiger ausgewählter Berichte getestet.

Das Ergebnis dieses Vorgehens findet sich in Tabelle 3. Die Kriterien für die Nachhaltigkeitsbewertung gliedern sich nach den in Kapitel 2.2 dargestellten Bereichen: 1) Kontext, 2) Implementierung, 3) Outcome, 4) Kapazitäten vor Ort, 5) Impact (nicht intendierte Wirkungen) 6) Dauerhaftigkeit von Wirkungen sowie 7) Zusammenspiel der Dimensionen der Nachhaltigkeit. Diese Bereiche bilden den konzeptionellen Rahmen für 18 Nachhaltigkeitskriterien, die weiterhin nach Akteuren, Nachhaltigkeitsdimensionen und Kapazitätstypen in 48 Kriterien unterteilt wurden. Zusätzlich wurden bei der Auswertung auch die Ober- und Programmziele erfasst und den Dimensionen der Nachhaltigkeit sowie den SDGs zugeordnet.

Tabelle 3: Übersicht der Nachhaltigkeitskriterien

Bereiche	Kriterien	Differenzierte Kriterien
1) Kontext	1. Kontext nach Dimensionen	N-01 Soziale Dimension
		N-02 Wirtschaftliche Dimension
		N-03 Politische Dimension
		N-04 Ökologische Dimension
2) Implementierung	2. Anpassung (Alignment)	N-05 Anpassung an nationale Regelungen
		N-06 Anpassung an soziokulturellen Kontext auf Ebene der Zielgruppen
	3. Partizipation	N-07 Partizipation des entwicklungspolitischen Partners
		N-08 Partizipation der Zielgruppe(n)/Bevölkerung
	4. Steuerung	N-09 Nutzung der (institutionellen) Strukturen vor Ort
		N-10 Management response/Lernen aus M&E, lessons learned
		N-11 Upscaling umgesetzt
		N-12 Exit-Strategie vorhanden
3) Outcome	5. Akzeptanz und Eigenverantwortung (Ownership)	N-13 Akzeptanz und Eigenverantwortung des privatwirtschaftlichen Trägers
		N-14 Akzeptanz und Eigenverantwortung des politischen Partners
		N-15 Akzeptanz und Eigenverantwortung der Zielgruppe
	6. Leistung (Output) des Trägers/Partners	N-16 Service-/Produkt-Qualität
		N-17 Service-/Produkt-Quantität
	7. Nutzung der Leistungen (Outputs)	N-18 Nutzung der Leistungen durch Partner/Träger
		N-19 Nutzung der Leistungen durch Zielgruppe
	8. Bewusstseinsveränderung	N-20 Bewusstseinsveränderung bei Partner/Träger
		N-21 Bewusstseinsveränderung bei Zielgruppe
	9. Resilienz und Anpassungsfähigkeit	N-22 Resilienz und Anpassungsfähigkeit des Partners/Trägers
		N-23 Resilienz und Anpassungsfähigkeit der Zielgruppe
	10. Reichweite und Breitenwirksamkeit	N-24 Strukturbildung
		N-25 Diffusion

Bereiche	Kriterien	Differenzierte Kriterien
4) Kapazitäten vor Ort	11. Kapazitäten des politischen Partners	N-26 Finanzielle Kapazitäten
		N-27 Fachlich-personelle Kapazitäten
		N-28 Institutionelle Kapazitäten
	12. Kapazitäten des Trägers	N-29 Finanzielle Kapazitäten
		N-30 Fachlich-personelle Kapazitäten
		N-31 Institutionelle Kapazitäten
	13. Kapazitäten der Zielgruppe	N-32 Finanzielle Kapazitäten
		N-33 Fachlich-personelle Kapazitäten
		N-34 Institutionelle Kapazitäten
5) Impact[6]	14. Nicht intendierte Wirkungen nach Dimensionen	N-35 Soziale Dimension
		N-36 Wirtschaftliche Dimension
		N-37 Politische Dimension
		N-38 Ökologische Dimension
6) Absehbarkeit des Erhalts von Wirkungen	15. Absehbarkeit des Erhalts von Wirkungen nach Dimensionen	N-39 Soziale Dimension
		N-40 Wirtschaftliche Dimension
		N-41 Politische Dimension
		N-42 Ökologische Dimension
7) Zusammenspiel der Dimensionen der Nachhaltigkeit	16. Dimensionen-Synergien	N-43 Schaffung von Synergien durch Vorhaben
		N-44 Identifizierung von Synergien durch Evaluierung
	17. Dimensionen-Konflikte	N-45 Identifizierung von Zielkonflikten durch Vorhaben
		N-46 Identifizierung von Zielkonflikten durch Evaluierung
	18. Nebenwirkungen hinnehmbar	N-47 Einstufung eventueller Kompensationsmaßnahmen durch Vorhaben als ausreichend und/oder von möglichen Nebenwirkungen als „hinnehmbar"
		N-48 Einstufung von eventuellen Nebenwirkungen durch Evaluierung als „hinnehmbar"

Quelle: eigene Darstellung

Im Rahmen der quantitativen Inhaltsanalyse wurden nur diejenigen Kriterien berücksichtigt, die laut Evaluierungsbericht in einen direkten Zusammenhang mit Nachhaltigkeit gebracht wurden.[7] Die Basis bildete dabei stets der Gesamtbericht, d. h. sämtliche schriftlichen Ausführungen der Evaluierung. Für die nachhaltigkeitsbezogenen Aussagen wurde anschließend überprüft, ob durch den Evaluierungsbericht das Vorhandensein oder das Nicht-Vorhandensein eines Kriteriums festgestellt wurde (also z. B., ob es Ownership gab oder nicht). Kam ein Bericht bezüglich eines Kriteriums zu keiner klaren Aussage, wurde das Vorhandensein bzw. Nicht-Vorhandensein eines Kriteriums als „unklar" definiert. Ferner wurde in der

[6] *Der Bereich „Impact" enthält sowohl „intendierte" als auch „nicht intendierte" Wirkungen (siehe Kapitel 2.2). Da die „intendierten Wirkungen" jedoch integraler Bestandteil der Bewertung des OECD-DAC Kriteriums „Impact" sind, erhielten nur Kriterien um die „nicht-intendierten" Wirkungen Eingang in das Bewertungsraster der Nachhaltigkeit. Die „intendierten Wirkungen" wurden zusätzlich erhoben; die Ergebnisse werden in Kapitel 4.2.6 vorgestellt.*

[7] *Dies bedeutet, dass nur solche Textstellen als Bewertungsbasis dienten, in denen das jeweilige Kriterium 1) mit dem Wort „Nachhaltigkeit", 2) mit übergeordneten Wirkungen (Impact), 3) mit deren Dauerhaftigkeit, 4) mit einer Risikobetrachtung oder 5) mit dem Zusammenspiel der Dimensionen der Nachhaltigkeit in Zusammenhang gebracht wurde.*

Datenerhebung erfasst, inwiefern das Vorhandensein bzw. Fehlen des jeweiligen Kriteriums als für Nachhaltigkeit förderlich, hinderlich oder als unklar in seinen Auswirkungen betrachtet wurde.[8]

Im Rahmen der Überprüfung der Interkodierer-Reliabilität der Nachhaltigkeitsbewertung ergibt sich ein Kappa-Wert von 0,63. Damit liegt der Gesamt-Kappa-Wert für die Qualitäts- und Nachhaltigkeitsbewertung[9] bei 0,63; folglich liegt auch hier eine substanzielle Übereinstimmung vor (Landis und Koch, 1977).

3.4 Kontextstudie

Die bisher vorgestellte methodische Vorgehensweise ermöglicht es, die Evaluierungs- und Bewertungspraxis der deutschen EZ systematisch zu überprüfen. Ob diese auch angemessen ist, lässt sich jedoch nur im internationalen Vergleich bewerten (siehe Evaluierungsfrage 3). Eine solche Gegenüberstellung erfolgt in der vorliegenden Meta-Evaluierung in Form einer Kontextstudie, die sich der Evaluierungs- und Bewertungspraxis anderer bi- und multilateraler Entwicklungsorganisationen widmet. Da für die Evaluierungseinheiten der OECD-Länder die DAC-Kriterien von 1991 die Grundlage der Nachhaltigkeitsbewertung bilden (OECD, 1991), wurde in der Kontextstudie untersucht, wie diese mit dem Bewertungskriterium Nachhaltigkeit umgehen. Darüber hinaus wurden ausgewählte multilaterale Organisationen mit differenzierten Ansätzen zur Evaluierung von Nachhaltigkeit in die Analyse einbezogen.

Die Grundgesamtheit der Studie bestand aus 40 Evaluierungseinheiten aus 37 Mitgliedsstaaten des OECD-DAC-Netzwerks der Evaluierung in der Entwicklungszusammenarbeit (EvalNet) sowie aus neun multilateralen Organisationen, deren Evaluierungssysteme in der aktuellen Runde des DAC-Peer-Review-Prozesses ausführlich untersucht wurden (OECD, 2016b).[10]

Die Datengrundlage der Untersuchung bildeten die online verfügbaren Leitlinien der Evaluierungseinheiten zum Umgang mit dem Evaluierungskriterium Nachhaltigkeit. Dabei wurde ein schrittweises Vorgehen gewählt: Zunächst wurden die Webseiten auf ihre transparente Darstellung der konkreten Bewertungspraxis hin untersucht. Anschließend wurden 24 Evaluierungseinheiten, zu deren Bewertungssystem ausreichend Informationen verfügbar waren, in die vergleichende Betrachtung der Bewertungssysteme einbezogen, darunter 18 bi- und sechs multilaterale Evaluierungseinheiten: die Länder Australien, Belgien, Dänemark, Frankreich, Japan, Kanada, Luxemburg, Neuseeland, Niederlande, Norwegen, Österreich, Schweden, Schweiz, das Vereinigte Königreich, die Vereinigten Staaten und Deutschland sowie die Afrikanische Entwicklungsbank, die Asiatische Entwicklungsbank, die Europäische Investitionsbank, die Europäische Kommission, das Entwicklungsprogramm der Vereinten Nationen und die Weltbank-Gruppe. Im Zentrum der Gegenüberstellung stand neben der Definition des Nachhaltigkeitskriteriums die Darstellung der Bewertungspraxis. Dabei zeigte sich, dass insbesondere in der Schweiz und in den USA sowie in Evaluierungen der Weltbank und der Afrikanischen Entwicklungsbank ein umfassender Nachhaltigkeitsbegriff zugrunde gelegt wird; zur Evaluierung von Nachhaltigkeit wurden hier mindestens 8 von 39 Kriterien herangezogen. Zu den wesentlichen Kriterien gehören die finanzielle, politische, technische und soziale Nachhaltigkeit sowie das Kriterium der Eigenverantwortung. Abschließend erfolgte eine vertiefende Analyse der Nachhaltigkeitsbewertung von drei bi- und zwei multilateralen Evaluierungseinheiten, die neben einzelnen Bewertungskriterien auch Bewertungsskalen (Noten- oder Punktevergabe) verwenden und in dieser Hinsicht eine hohe Vergleichbarkeit mit der Praxis der deutschen EZ aufweisen.[11]

[8] *Im Rahmen der Entwicklung des Kriterienrasters wurden ähnlich der Qualitätsbetrachtung zunächst sämtliche zu untersuchenden Evaluierungsberichte und Nachhaltigkeitskriterien in MAXQDA angelegt und anschließend mit Hilfe einer Microsoft-Access-Datenbank kodiert. Auch hier wurde die Interkodierer-Reliabilität durch das doppelte Kodieren von 10 Prozent der Evaluierungsberichte, geschichtet nach Evaluierungstyp, geprüft (siehe Kapitel 3.2).*

[9] *Die Gesamtbetrachtung basiert auf der aggregierten Auswertung der Übereinstimmung der Qualitäts- sowie der Nachhaltigkeitsbewertung durch drei Evaluatorinnen und Evaluatoren. Der Gesamt-Kappa-Wert ist der Mittelwert aus den Kappa-Werten der Qualitäts- und Nachhaltigkeitskriterien. Ein Wert von 0 bedeutet maximale Divergenz zwischen den Evaluatorinnen und Evaluatoren; ein Wert von 1 entspricht der maximalen Übereinstimmung zwischen den Evaluatorinnen und Evaluatoren (für weitere Erläuterungen siehe Kapitel 3.5).*

[10] *Eine erste Studie zum Stand der Evaluierungssysteme fand 2010 statt (OECD, 2010a).*

[11] *Die Schritte „Screening der Webseiten der Evaluierungseinheiten" und „Vergleichende Betrachtung von 24 Evaluierungseinheiten" wurden vom DEval durchgeführt. Die Vertiefungsstudie wurde in Kooperation mit Jana Preiß im Rahmen ihrer Masterarbeit an der Freien Universität Berlin durchgeführt (Preiß, 2017).*

3.5 Limitationen

Bei der vorliegenden Meta-Evaluierung handelt es sich um eine Schreibtischstudie auf der Grundlage von Sekundärdaten. Die Analysetiefe war damit durch die Berichterstattung nach den Vorgaben der GIZ und KfW zu den jeweiligen Evaluierungsformaten vorgegeben. Da Nachhaltigkeit immer nur einen Teil der Erfolgsbewertung einer Evaluierung bildet und die dazugehörigen Ausführungen entsprechend knapp ausfielen, war es nicht immer möglich, ein Kriterium als Erfolgs- oder Misserfolgsfaktor für die Nachhaltigkeit eines Vorhabens zuzuordnen. Als Konsequenz musste der positive bzw. negative Einfluss eines Kriteriums oftmals als „unklar" kodiert und in der Folge von der Analyse ausgeschlossen werden. Allerdings überstieg die Anzahl der als „unklar" kodierten Textstellen bei keinem Kriterium die „eindeutig als positiv oder negativ" bestimmbaren Fälle, sodass dieser Umstand vermutlich nur eine geringe Rolle spielt. Bei Kriterien, zu denen nur wenig berichtet wurde, kann er jedoch durchaus ins Gewicht fallen.

Die Ausführlichkeit der Berichtslegung spielt auch für die Qualitätsbewertung eine Rolle. Diese basierte allein auf der Auswertung der Evaluierungsberichte. Aus den unterschiedlichen Vorgaben zur Ausführlichkeit der Berichtslegung ergeben sich möglicherweise Diskrepanzen zwischen der tatsächlichen Qualität einer Evaluierung und der nachvollziehbaren Qualität auf Grundlage der Evaluierungsberichte. Um diese möglichst gering zu halten, wurde bei der Auswahl der Evaluierungskriterien darauf geachtet, dass nur solche Kriterien Eingang in die Qualitätsbewertung erhielten, die theoretisch von einem Großteil der Evaluierungstypen erfüllt werden müssten, da sie hinsichtlich der Qualität einen eher grundsätzlichen Charakter aufweisen.

Bei der Analyse des Nachhaltigkeitsverständnisses entlang der berichteten Kriterien ergaben sich schließlich eine Reihe von Herausforderungen bezüglich der Endogenität: Möglicherweise werden bestimmte Nachhaltigkeitskriterien häufiger oder ausführlicher in Evaluierungen diskutiert, da sie einen besonders positiven oder besonders negativen Einfluss haben, während neutrale Einflüsse hingegen seltener hervorgehoben werden. Möglicherweise sind negative/positive Ausprägungen eines Kriteriums auch methodisch einfacher bzw. schwieriger nachzuweisen. Zudem könnten auch die unterschiedlichen Vorgaben und Erwartungen an einen spezifischen Evaluierungstyp die Evaluierungsergebnisse beeinflussen. In der Diskussion der Ergebnisse (Kapitel 4) wird daher auf systematische Unterschiede in den genannten Aspekten geachtet und wenn nötig auf mögliche Einschränkungen in der Aussagekraft der Ergebnisse hingewiesen.

Eine weitere Einschränkung in quantitativen Inhaltsanalysen ist die intersubjektive Vergleichbarkeit des Kodierverhaltens zwischen mehreren Evaluatorinnen und Evaluatoren. Es besteht die Gefahr, dass unterschiedliche Personen ein und denselben Sachverhalt unterschiedlich interpretieren und dementsprechend zu unterschiedlichen Ergebnissen kommen können. Aus diesem Grund wurde die Interkodierer-Reliabilität des Evaluierungsteams durch den Kappa-Wert nach Cohen überprüft (siehe Kapitel 3.2 und 3.3). Der Gesamt-Kappa-Wert von 0,63 belegt eine moderate Übereinstimmung zwischen den Evaluatorinnen und Evaluatoren bei der Qualitäts- und Nachhaltigkeitsbewertung. Da dieser Wert zwar als starke, jedoch nicht als sehr starke Übereinstimmung interpretiert werden kann, wurde für die Qualitätsbewertung zusätzlich noch eine externe Perspektive einbezogen. Dabei wurden die Qualitätsbewertungen von Berichten, die sowohl in dieser Meta-Evaluierung als auch in den Meta-Evaluierungen der GIZ analysiert wurden, miteinander verglichen. Dabei zeigte sich, dass sich die Einschätzungen dazu, wieviel Prozent der maximalen Punktzahl erreicht wurden, ähneln. Im Vergleich mit der GIZ-Meta-Evaluierung im Bereich Gesundheit (Raetzell und Krämer, 2013) weicht die Einschätzung in der Mehrzahl der untersuchten Vorhaben um weniger als zehn Prozent ab (wobei 100 Prozent die Erfüllung aller Kriterien darstellen) und übersteigt nur in einem Fall 20 Prozent. Da bislang keine Meta-Evaluierungen zu Evaluierungen der KfW vorliegen, konnte ein solcher Abgleich nur für Beobachtungen der GIZ durchgeführt werden.

4.

ERGEBNISSE

Im Folgenden werden die Ergebnisse der Meta-Evaluierung im Detail dargestellt. Dabei widmet sich Kapitel 4.1. den Ergebnissen zur Evaluierungsqualität. In Kapitel 4.2 werden die Ergebnisse zum Nachhaltigkeitsverständnis entlang der Bewertungskriterien diskutiert. Abschließend erfolgt in Kapitel 4.3 die Darstellung der Ergebnisse aus der Kontextstudie.

4.1 Qualität der Evaluierungsberichte

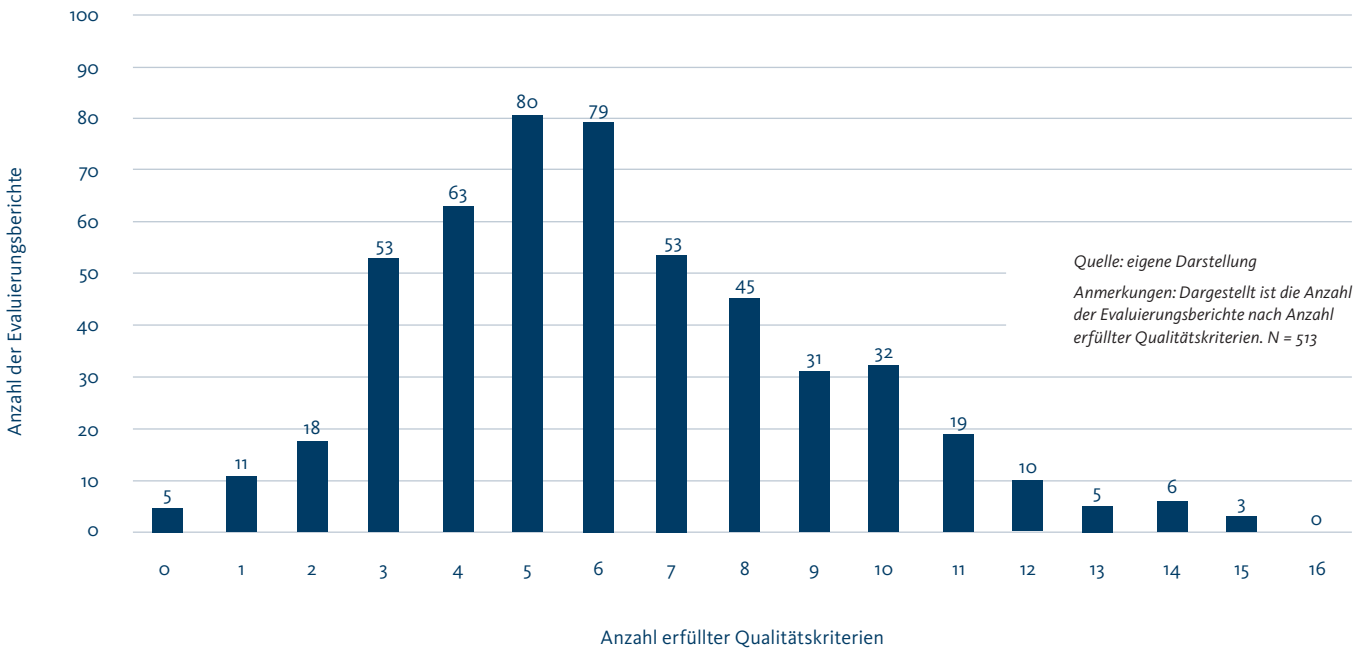
Durch die Analyse der Evaluierungsqualität konnte im Rahmen der Meta-Analyse die Belastbarkeit der Evaluierungs-Ergebnisse zum Thema Nachhaltigkeit beurteilt werden. Entsprechend der Ausführungen zu den methodischen Limitationen (siehe Kapitel 3.5) werden die Ergebnisse mit Vorsicht diskutiert, denn Diskrepanzen zwischen der tatsächlichen Qualität einer Evaluierung und der anhand der Evaluierungsberichte nachvollziehbaren Qualität können nicht gänzlich ausgeschlossen werden. Die Ergebnisse zur methodischen Qualität stehen somit immer in Zusammenhang mit der Nachvollziehbarkeit der Qualität auf der Grundlage der vorliegenden Evaluierungsberichte. Zusätzliche Dokumente, zum Beispiel Evaluierungskonzeptionen oder *terms of references*, wurden nur dann analysiert, wenn sie Teil der Anhänge der Evaluierungsberichte waren. Kontextualisierende Informationen, beispielsweise zu den für die jeweilige Evaluierung aufgewendeten Ressourcen, waren nur selten verfügbar und flossen somit nur allgemein in die Analyse ein. Die Ergebnisse der Meta-Evaluierung lassen jedoch den Schluss zu, dass das dabei gewählte Vorgehen hinsichtlich der Vergleichbarkeit der einzelnen Evaluierungsberichte durchaus angemessen war: Die Normalverteilung der Anzahl der Qualitätskriterien über alle Berichte hinweg (Abbildung 1) belegt, dass die untersuchten Berichte die gesamte Bandbreite des angelegten Qualitätsrasters abdecken. Im Durchschnitt wurden 6,2 der 16 möglichen Qualitätskriterien erfüllt.

Die Analyse der Qualitätsbereiche (Abbildung 2) zeigt, dass der überwiegende Teil der Evaluierungsberichte den Hintergrund einer Evaluierung (93 %), die Wirkungszusammenhänge (85 %) und das methodische Vorgehen (84 %) nachvollziehbar darstellt. Dabei wurde ein Qualitätsbereich als erfüllt gewertet, wenn mindestens eines der zugehörigen Qualitätskriterien erfüllt wurde. Ein weitaus geringerer Erfüllungsgrad findet sich beim Evaluierungsdesign (25 %) und bei der Robustheit der Ergebnisse (33 %). Die Ergebnisse dieser aggregierten Betrachtung geben einen ersten Eindruck davon, in welchen Bereichen die Evaluierungen besonders gut beziehungsweise weniger gut abschneiden.

Die Ergebnisse zur Qualität der Evaluierungsberichte zeigen, dass nahezu alle Evaluierungen (92 %) eine Gegenstandsbeschreibung (Q-01) vornehmen (siehe Abbildung 4). Dies bedeutet aber auch, dass letztlich nicht alle Evaluierungen ausreichend Informationen bereitstellen, die den Leserinnen und Lesern zeigen, womit sich die Evaluierung genau beschäftigt. Eine geringe Nachvollziehbarkeit der Informationen wird insbesondere im Hinblick auf das Bewertungskriterium der Operationalisierung des Erkenntnisinteresses (Q-02) entlang der standardisierten Prüffragen nach OECD-DAC-Kriterien deutlich: Lediglich in 16 Prozent der Fälle ist anhand der Evaluierungsberichte ein gegenstandsspezifisches Erkenntnisinteresse erkennbar, d. h. es wurden hier Evaluierungsfragen entlang der DAC-Kriterien aufgeführt, die auf den spezifischen Gegenstand ausgerichtet sind (Q-02). Eine ergänzende Auswertung ausgewählter zusätzlicher Dokumente der KfW und GIZ zeigt, dass das Erkenntnisinteresse einer Evaluierung zwar aus zusätzlichen Dokumenten – wie dem Konzeptpapier einer Evaluierung oder den *terms of reference* – rekonstruiert werden kann, aus dem eigentlichen Evaluierungsbericht allein jedoch nicht hervorgeht. Daher liegt die Vermutung nahe, dass die Evaluierungsberichte seitens der DO nicht als für sich stehende Produkte gesehen werden, die ohne zusätzliche Dokumente verstanden werden können.

Bei der Mehrheit der Evaluierungen von KfW und GIZ wird der Wirkungsnachweis über einen Soll-Ist-Abgleich entlang ausgewählter Indikatoren der Wirkungslogik erarbeitet. Die Ergebnisse der vorliegenden Meta-Evaluierung zeigen, dass die Voraussetzungen für ein solches Vorgehen bei den meisten Evaluierungen grundsätzlich geschaffen wurden. In der überwiegenden Anzahl der Berichte wurden die Wirkungslogik (Q-03, 63 %) und die dazugehörigen Wirkungsindikatoren (Q-04, 74 %) dargestellt. In rund einem Drittel der analysierten Vorhaben wurden Wirkungslogiken in den Evaluierungsberichten nicht transparent gemacht, was jedoch nicht ausschließt,

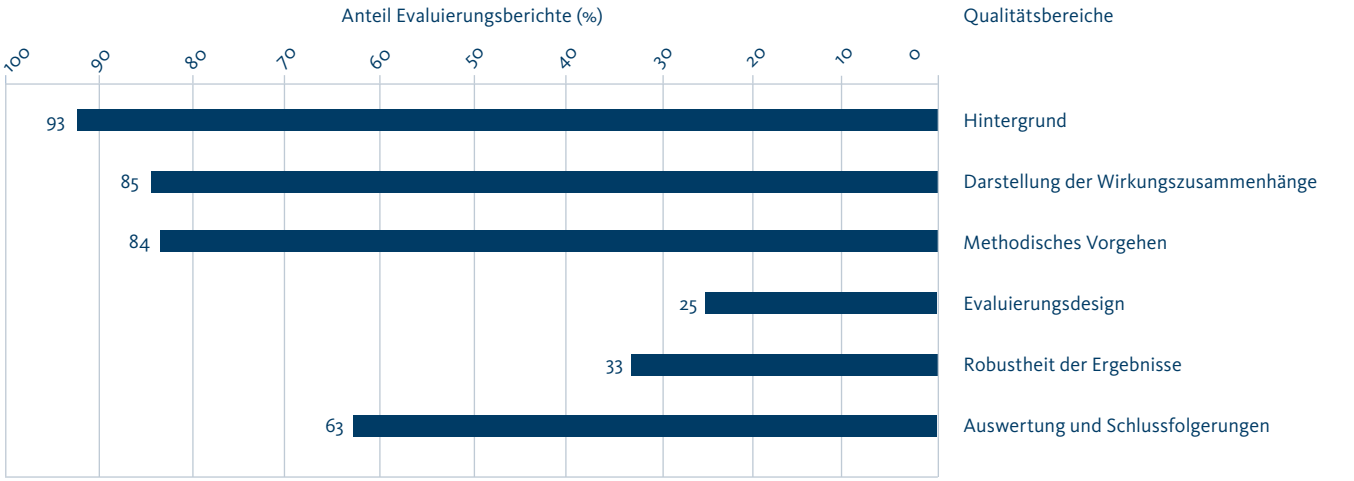
Abbildung 1: Anzahl Evaluierungsberichte nach Anzahl erfüllter Qualitätskriterien



dass solche verwendet wurden. Andererseits ist die Darstellung von Wirkungslogiken und Indikatoren keine hinreichende Bedingung dafür, dass im Rahmen von Soll-Ist-Vergleichen kausale Schlussfolgerungen getroffen werden. Nur in wenigen Fällen wurde der Herausforderung der kausalen Zuordnung durch das Hinzuziehen aufwändigerer Verfahren der Wirkungsanalyse begegnet. Vorher-Nachher-Vergleiche wurden lediglich in 19 Prozent der Evaluierungen einbezogen (Q-09). Eine mögliche Ursache hierfür könnte darin liegen, dass sich für Vorhaben mit besonders innovativen Ansätzen oder Maßnahmen, die mit Neugründungen verbunden sind, kaum Daten zur Ausgangslage finden lassen. In solchen Fällen können Baseline-Daten lediglich durch Sekundärdaten rekonstruiert werden, die sich im Rahmen von Evaluierungen nur schwerlich dem Ist-Status gegenüberstellen lassen. Mit Kontrollgruppen arbeiteten nur 9 Prozent. Aufwändigere theoriebasierte Verfahren, zum Beispiel Kontributionsanalysen, die sich dem Attributionsproblem durch systematische Verfahren der plausiblen Assoziation von Ursache-Wirkungsbeziehungen widmen, kommen bisher kaum zum Einsatz. Dieses Ergebnis ist auch im Hinblick auf die Nachhaltigkeit von Bedeutung, da der Wirkungsnachweis die wesentliche Grundlage für die Nachhaltigkeitsbewertung bildet. Insofern muss auch überprüft werden, inwieweit sich die Qualität einer Evaluierung auch empirisch auf die Bewertung von Nachhaltigkeit auswirkt. Die Ergebnisse sind später in Kapitel 4.3 dargestellt.

Die Arbeit mit Baseline-Daten, Kontrollgruppen oder systematischen Verfahren der plausiblen Assoziation von Ursache-Wirkungsbeziehungen ist für eine belastbare Wirkungsanalyse unverzichtbar. Ohne solche Verfahren ist die Darstellung von Wirkungszusammenhängen unzulässig. In diesen Fällen kann die Unsicherheit über Ursache-Wirkungsbeziehungen nur durch den Einsatz systematischer Triangulationsverfahren reduziert werden. Bei rund einem Drittel der Evaluierungen wurden systematische Verfahren der Datentriangulation genutzt. Die Gegenüberstellung unterschiedlicher Methoden ist nur in knapp einem von zehn Fällen hinreichend belegt. Erstaunlich ist in diesem Zusammenhang, dass im Rahmen von Soll-Ist-Vergleichen nur selten auf nachvollziehbare Weise auf Monitoringdaten zurückgegriffen wird. Lediglich in 31 Prozent der Evaluierungen gingen Informationen aus den

Abbildung 2: Anteil Evaluierungsberichte nach erfüllten Qualitätsbereichen



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte, die mindestens ein Kriterium des angegebenen Bereiches der Qualitätskriterien erfüllen. N = 513

Monitoringsystemen der DO und/oder der Partner und Träger explizit in die Analysestrategie ein (siehe Abbildung 4). Dies bedeutet nicht, dass den Evaluierungen nur in rund einem Drittel der Fälle Monitoringdaten zur Verfügung standen – im Gegenteil: Es kann davon ausgegangen werden, dass die Gutachterinnen und Gutachter von den Vorhaben bzw. Trägern stets auch Monitoringdaten zur Verfügung gestellt bekommen. Dass in den Schlussfolgerungen der Evaluierungen kaum auf diese Daten eingegangen wird, deutet vielmehr darauf hin, dass sie oftmals nicht dem Zweck bzw. dem Bedarf einer Evaluierung entsprechen. Dies erklärt auch, dass sich in einigen der analysierten Evaluierungen Hinweise darauf finden, dass die Vorhaben zukünftig mehr in den Aufbau und die Pflege von wirkungsorientierten Monitoringsystemen investieren sollten.

Die Ergebnisse der Meta-Evaluierung belegen, dass die Belastbarkeit des Wirkungsnachweises über den Einsatz einschlägiger kausalanalytischer Ansätze und Triangulationsverfahren verbessert werden könnte. Bislang ist nur bei einem Drittel der Evaluierungen ersichtlich, dass ihre Schlussfolgerungen auf einer ausreichenden Datengrundlage (Q-16) beruhen. Basis für diese Einschätzung bilden die Angaben zum Evaluierungsdesign und zu den Datenerhebungsmethoden. Ein Grund hierfür ist sicherlich die geringe Datenverfügbarkeit, insbesondere wenn Vorhaben in fragilen Kontexten evaluiert werden. Allerdings könnten die Schwächen einer Datenquelle durch systematische Triangulationsverfahren vermindert werden; diese kommen bislang nur in ca. 30 Prozent der Evaluierungen zum Einsatz. Weiterhin könnten Schlussfolgerungen, die auf keiner belastbaren Datengrundlage fußen, für den Zweck der Evaluierung jedoch erforderlich sind, durch eine transparente Darstellung der verbleibenden Unsicherheit kenntlich gemacht werden.

Darüber hinaus wird der überwiegende Teil der Ergebnisse und Schlussfolgerungen zwar plausibel dargestellt (Q-15), aber selten durch Quellen hinterlegt (Q-14). Die Qualität der Evaluierung ließe sich durch zwei Maßnahmen erhöhen: durch verbesserte methodische Verfahren, die zu einem besseren Wirkungs- und Nachhaltigkeitsnachweis führen, sowie durch eine höhere Nachvollziehbarkeit der Evaluierungsergebnisse.

Im Sinne der Nachvollziehbarkeit zeigt sich auch, dass in nur 68 Prozent der Fälle die methodische Vorgehensweise beschrieben wurde (Q-05). In diesen Fällen wurde die Methodik auf Basis des Berichtes rekonstruiert. Der überwiegende Teil der Evaluierungen beinhaltete Feldmissionen und nutzte vor Ort verschiedene Verfahren der Befragung. In 83 Prozent der Evaluierungen kamen Leifadeninterviews zum Einsatz, bei 26 Prozent wurden Gruppendiskussionen und bei 13 Prozent standardisierte Befragungen zur Erhebung der Daten genutzt (siehe Abbildung 3). In 58 Prozent der Berichte wurden die befragten Gruppen vorgestellt. Allerdings wird nur in 15 Prozent der Evaluierungsberichte das zugrunde liegende Auswahlverfahren beschrieben, in den restlichen Fällen bleibt die Auswahl für die Leserin bzw. den Leser willkürlich.

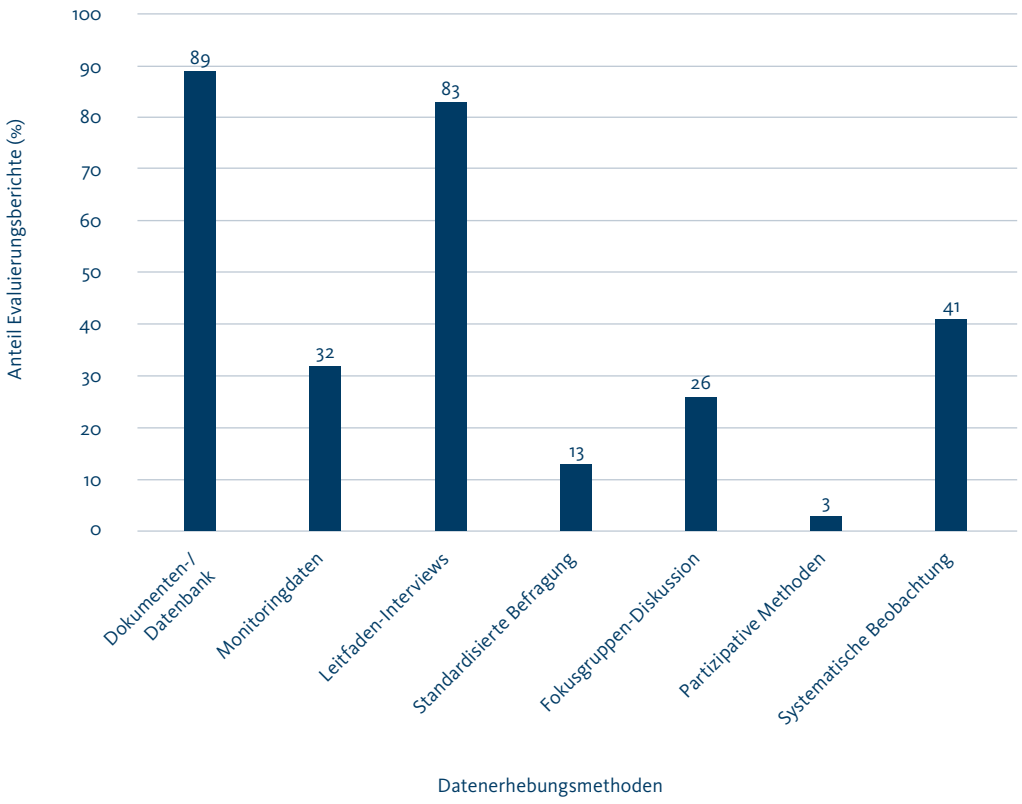
Aus der Gesamtbetrachtung aller Kriterien ergibt sich, dass keiner der 513 Evaluierungsberichte alle 16 Qualitätskriterien erfüllt (siehe Abbildung 1). Allerdings können auch Evaluierungen, die nicht allen Qualitätsansprüchen genügen, durchaus zu glaubwürdigen Ergebnissen kommen. Zum Beispiel ist es in kausalanalytischer Hinsicht nicht immer notwendig, Vorher-Nachher-Vergleiche mit Kontrollgruppenvergleichen und zusätzlichen theoriebasierten Designs zu kombinieren, auch wenn dies die Belastbarkeit des Wirkungsnachweises grundsätzlich erhöht. Generell ergibt sich das geeignete Evaluierungsdesign erst durch die Fragestellung einer Evaluierung und die Attribute des Evaluierungsgegenstandes. Da die Fragestellungen, wie oben gezeigt, jedoch selten evaluierungsspezifisch sind, sondern sich zumeist aus den Vorgaben ergeben, richtet sich die Wahl des geeigneten Evaluierungsdesigns in den Modulevaluierungen von GIZ und KfW vornehmlich nach den Charakteristika des Evaluierungsgegenstandes und den verfügbaren beziehungsweise umsetzbaren Designs. Zusätzlich beinhalten die Evaluierungen jedoch auch DO-spezifische Auswertungsverfahren, die aber wegen mangelnder Vergleichbarkeit kein Bestandteil der vorliegenden Meta-Evaluierung sind. So finden sich in Evaluierungen von FZ-Vorhaben im Bereich wirtschaftlicher Infrastruktur einzel- und gesamtwirtschaftliche Rechnungen, die in der Evaluierung von TZ-Maßnahmen keine Rolle spielen.

Für die vergleichende Betrachtung der Qualität wird der in Kapitel 3.2 vorgestellte Qualitätsindex herangezogen. Ein Evaluierungsbericht, der alle sechzehn Qualitätskriterien erfüllt, erhält demnach einen Index-Wert von 1. Wie in Kapitel 3.2 dargestellt, werden Kriterien, die für die Überprüfung der Belastbarkeit der Ergebnisse von besonderem Wert sind, doppelt gewichtet. Erfüllt ein Bericht kein einziges Kriterium, erhält er einen Wert von 0. Über alle 513 untersuchten Evaluierungsberichte hinweg wird ein durchschnittlicher Qualitätsindex-Wert von 0,34 erreicht. Dieses Ergebnis zeigt, dass eine hohe Zahl der Berichte nicht alle angelegten Qualitätsmerkmale auf nachvollziehbare Weise erfüllt. Insbesondere bei Kriterien zur Belastbarkeit des Wirkungs- bzw. Nachhaltigkeitsnachweises besteht Potenzial zur Erhöhung der methodischen Güte.

Die differenzierte Betrachtung der Qualität nach Evaluierungstypen (siehe Abbildung 5 unten) kommt zu folgendem Ergebnis: Die Ex-post- und Schluss-Evaluierungen der GIZ weisen mit einem mittleren Index-Wert von 0,6 die höchste Qualität auf. Es folgen die Ex-post-Evaluierungen der KfW und die PEV der GIZ mit einem Wert von ca. 0,3. Die geringste Qualität weisen mit einem durchschnittlichen Index-Wert von 0,2 die PFK auf. Die Unterschiede zwischen diesen drei Gruppen sind statistisch signifikant.[12] Diese Ergebnisse zeigen, dass sich höherer Aufwand lohnt: Die früheren Ex-post- und Schluss-Evaluierungen der GIZ waren in der Regel vergleichsweise umfangreicher und aufwändiger. Allerdings wurden zwischen 2006 und dem Ende des UE-Programms im Jahre 2014 nur 100 solcher Evaluierungen durchgeführt. Demgegenüber werden die Ex-post-Evaluierungen der KfW und die dezentralen Evaluierungen mit vergleichsweise geringem Aufwand durchgeführt, decken jedoch weite Teile des Portfolios von GIZ- und KfW-Vorhaben ab. Die Qualität steht somit im Spannungsverhältnis zwischen dem Umfang einer Evaluierung und dem Deckungsgrad der Evaluierungen insgesamt. Insgesamt lässt sich feststellen, dass die Qualität der Evaluierungsberichte über die Zeit zunimmt: Während Evaluierungen, die im Jahr 2006 durchgeführt wurden, einen Index-Wert von ca. 0,3 erreichten, lag dieser Wert zehn Jahre später bei knapp 0,4. Eine differenzierte Betrachtung der einzelnen Evaluierungstypen bestätigt dies: Insbesondere Ex-post-Evaluierungen und Schluss-Evaluierungen der GIZ weisen im betrachteten Zeitraum eine starke

[12] *Der Welch-Test zeigt, dass die Gruppen sich signifikant unterscheiden ($p < 0{,}01$) und die Unterschiede zwischen den Evaluierungstypen somit mit hoher Wahrscheinlichkeit nicht dem Zufall geschuldet sind. Ein Games-Howell-Test zum direkten Vergleich der Gruppen bestätigt dieses Ergebnis.*

Abbildung 3: Anteil Evaluierungsberichte nach verwendeten Datenerhebungsmethoden



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungen, bei denen die jeweilige Datenerhebungsmethode zur Anwendung kam. N = 513

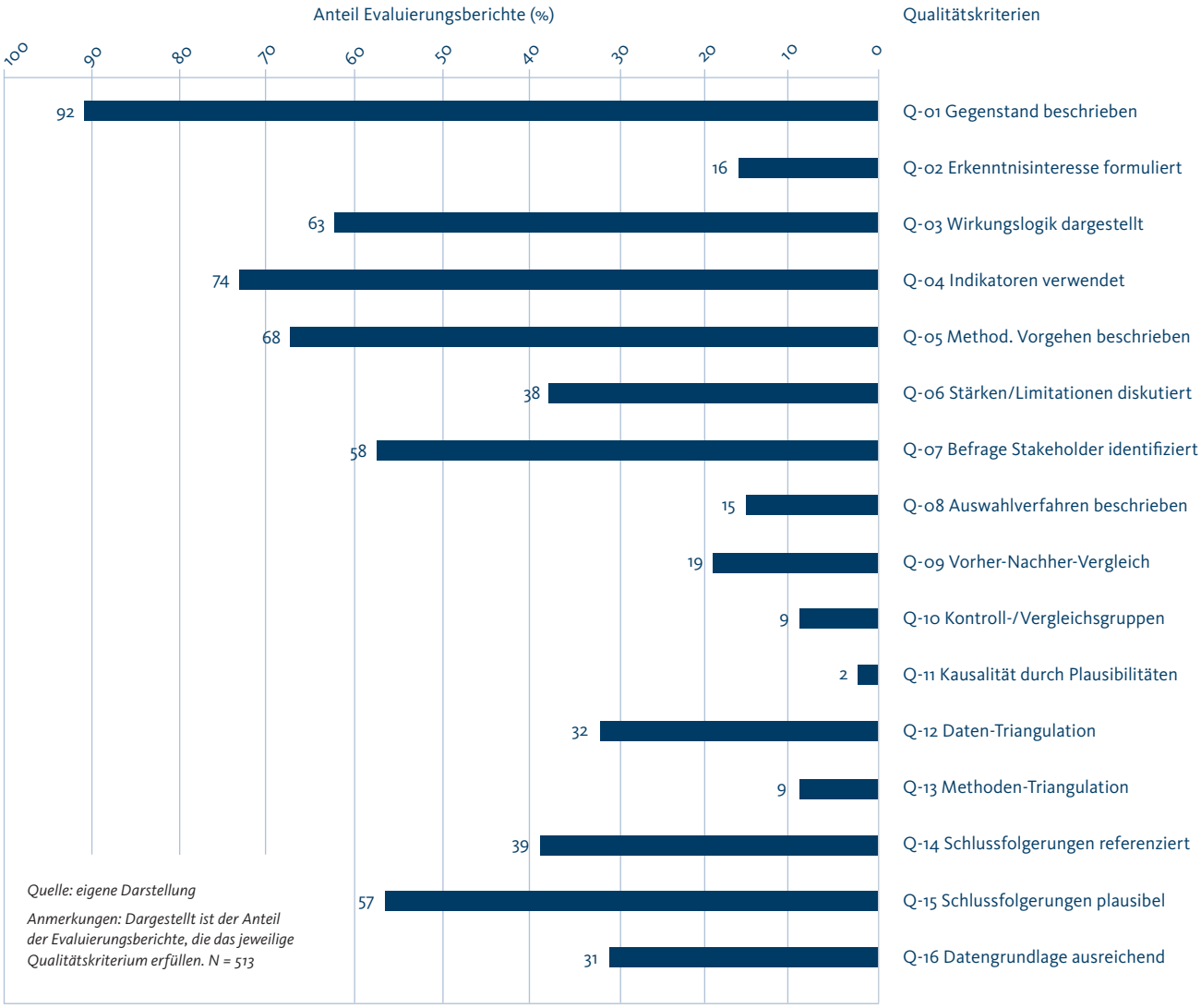
Qualitätssteigerung auf, welche bei GIZ-PFK und KfW-Ex-post-Evaluierungen schwächer ausfällt. Bei PEV ist aufgrund des kurzen Zeitraums nur eine geringe Veränderung festzustellen.

4.2 Die Bewertung von Nachhaltigkeit in Evaluierungen der GIZ und KfW

Das folgende Kapitel befasst sich mit dem Nachhaltigkeitsverständnis der deutschen EZ. Grundlage der Analyse bilden die Ergebnisse der quantitativen Inhaltsanalyse entlang der Bewertungskriterien der Nachhaltigkeit. Die Diskussion der Bewertung ist dabei dem konzeptionellen Rahmen der Nachhaltigkeit (siehe Kapitel 2.2) entsprechend strukturiert: Nach einigen übergreifenden Erkenntnissen (Kapitel 4.2.1) werden die Ergebnisse zu den Bereichen Kontext (4.2.2), Implementierung (4.2.3), Outcome (4.2.4), Kapazitäten vor Ort (4.2.5), Impact (4.2.6), Absehbarkeit der Wirkungen (4.2.7) und Zusammenspiel der Dimensionen der Nachhaltigkeit (4.2.8) diskutiert. Das zugrunde gelegte Verständnis der einzelnen Nachhaltigkeitskriterien findet sich zur Übersicht in Tabelle 5 im Anhang. In der Diskussion der Bewertungspraxis werden die Ergebnisse der Qualitätsuntersuchung (aus Kapitel 4.1) hinzugezogen, um die Belastbarkeit der Ergebnisse einzuordnen.

Wie in Kapitel 3.3 dargestellt, wurden die Bewertungskriterien auf der Grundlage eines integrierten Ansatzes entwickelt, der sowohl eine deduktive als auch eine induktive Herangehensweise umfasst. Bei der Definition der einzelnen Bewertungskriterien wurde dabei auf eine möglichst eindeutige Trennschärfe zwischen den Kriterien geachtet (siehe Tabelle 5). Aufgrund von Unterschieden in der subjektiven Betrachtung der Gutachter kann eine konzeptionelle Unschärfe zwischen den Kriterien allerdings nie vollständig ausgeschlossen werden. Aus empirischer Sicht scheint dieses Risiko jedoch überschaubar: Anhand

Abbildung 4: Anteil Evaluierungsberichte nach erfüllten Qualitätskriterien



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte, die das jeweilige Qualitätskriterium erfüllen. N = 513

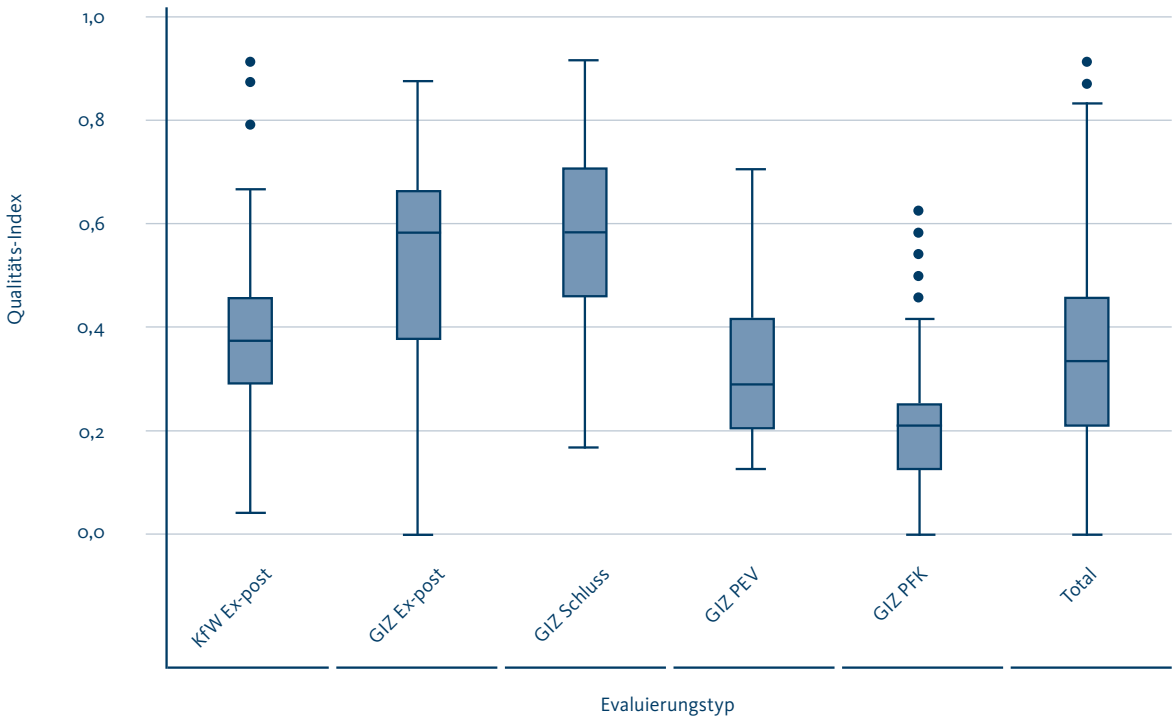
der Korrelationen zwischen den einzelnen Kriterien finden sich nur wenige Zusammenhänge von statistischer Bedeutung. Allein die Kriterien „Akzeptanz und Ownership" der Zielgruppe, „Nutzung des Outputs" durch die Partner und Träger sowie „Synergien zwischen den Dimensionen der Nachhaltigkeit" weisen starke Zusammenhänge mit anderen Nachhaltigkeitskriterien auf.[13] Insgesamt scheint eine getrennte Diskussion der einzelnen Bewertungskriterien jedoch zulässig.

4.2.1 Übergreifende Erkenntnisse

In der Gesamtschau der Bewertungspraxis von Nachhaltigkeit wird deutlich, dass diese auf einem umfassenden Verständnis von Nachhaltigkeit basiert. Dies zeigt sich an der Vielzahl der in die Bewertung einfließenden Bereiche. Abbildung 6 zeigt die prozentuale Häufigkeit der übergeordneten Kriterien und Bereiche, wenn mindestens ein Kriterium aus dem Bereich in die Bewertung eingeht. Hinter den übergeordneten Kriterien

[13] *Mit Hilfe des Produkt-Moment-Korrelationskoeffizienten nach Pearson wurden positive Korrelationen ab einer Stärke von 0,7 sowie negative Korrelationen ab einer Stärke von 0,5 betrachtet. Ferner wurden nur Variablenpaare analysiert, zu denen mindestens 10 Evaluierungsberichte vorlagen und welche auf dem 5-Prozent-Niveau signifikant waren.*

Abbildung 5: Qualitäts-Index nach Evaluierungstyp



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Qualitätsindex nach Evaluierungstyp. Der Qualitätsindex wird aus den in Abbildung 4 vorgestellten Qualitätskriterien Q-01 bis Q-16 gebildet. Die Kriterien Q-9 bis Q-16 werden doppelt gewichtet, wobei ein Qualitätsindex von 1 die höchste und ein Qualitätsindex von 0 die niedrigste methodische Qualität eines Berichtes anzeigt. Die Grafik zeigt die Verteilung der Daten pro Evaluierungstyp als Boxplots. Die Boxen in der Mitte repräsentieren die mittleren 50 Prozent einer Verteilung, die durch den Median geteilt werden. Oberhalb der Boxen kennzeichnen die Striche die Werte, die größer als das dritte Quartil sind; die Striche unterhalb der Boxen kennzeichnen die Werte, die kleiner als das zweite Quartil sind. Die Punkte bilden die Ausreißer ab. N = 513

stehen wiederum einzelne Bewertungskriterien.[14] Trotz der breiten Basis gibt es Bereiche, die deutlich seltener in die Bewertung eingehen als andere. Dies gibt einen ersten Hinweis darauf, dass die Bewertung über einzelne Evaluierungen hinweg nicht einheitlich erfolgt. Eine Ursache für die geringe Standardisierung liegt vermutlich in nicht hinreichenden Vorgaben dazu, wie Nachhaltigkeit im Sinne nachhaltiger Entwicklung über die fünf DAC-Kriterien hinweg konzeptionell verstanden wird.

Bereiche, die in den Evaluierungen vergleichsweise häufig in die Diskussion von Nachhaltigkeit eingehen, sind „Outcome" (in 87 % aller Evaluierungen genannt) und die „Kapazitäten vor Ort" (in 86 % aller Evaluierungen, siehe Abbildung 6).[15] Dies zeigt, dass direkte Wirkungen und Kapazitäten vor Ort in der Bewertungspraxis eine bedeutende Rolle spielen und somit auch ein integraler Bestandteil des zugrunde liegenden Nachhaltigkeitsverständnisses sind.

Die Ergebnisse lassen weiterhin erkennen, dass die durch den BMZ-Leitfaden vorgegebenen Prüffragen durchaus Berücksichtigung finden, wenn auch nicht so häufig und systematisch wie erwartet. Mit Blick auf das Evaluierungskriterium „Nachhaltigkeit" fordern die ersten beiden Prüffragen eine Auseinandersetzung mit der Absehbarkeit der Wirkungen sowie dem Kontext einer Maßnahme. Die empirischen Ergebnisse dieser

[14] *Ein übergeordnetes Kriterium gilt als „berichtet", wenn ein Evaluierungsbericht eine entsprechende positive oder negative Aussage über mindestens ein zugehöriges einzelnes Kriterium mit Bezug auf die Nachhaltigkeit trifft. Unklare (d. h. weder positive noch negative) Aussagen wurden in die Datenerhebung mit aufgenommen, in der nun folgenden Analyse jedoch nicht berücksichtigt.*

[15] *Ein Bereich gilt als „berichtet", wenn mindestens eines der ihm zugeordneten Kriterien in die Nachhaltigkeitsbewertung eingeht. Bei dem Bereich „Kapazitäten vor Ort" kommen hierfür drei Kriterien in Frage, im Bereich „Outcome" fünf. Dies führt dazu, dass der Bereich „Outcome" schneller als „berichtet" gilt.*

Meta-Evaluierung bestätigen eine vergleichsweise hohe Relevanz dieser beiden Bereiche. Allerdings werden Kontextfaktoren und die Absehbarkeit der Wirkungen auch nur in rund jedem zweiten Bericht in die Nachhaltigkeitsbewertung einbezogen. Die dritte Prüffrage des Evaluierungskriteriums der Nachhaltigkeit betrifft die Risiken und Potenziale im Umfeld der Maßnahme. Die Beantwortung dieser Frage findet sich empirisch in verschiedenen Kriterien der Bereiche „Implementierung" und „Outcome".

Es hat sich aber auch gezeigt, dass eine Diskussion möglicher „nicht intendierter Wirkungen" deutlich seltener in die Nachhaltigkeitsbewertung eingeht. Da die Identifizierung von nicht intendierten Wirkungen zu den grundsätzlichen Schwierigkeiten in Evaluierungen gehört, ist dieses Ergebnis wenig erstaunlich. Allerdings zeigt es auch, dass die Evaluierungen hier hinter den eigenen Ansprüchen zurückbleiben, ist doch die Diskussion nicht intendierter Wirkungen aus konzeptioneller Sicht elementarer Bestandteil des Impact-Kriteriums. Zudem legt die BMZ-Definition des Impact-Kriteriums nahe, verschiedene Dimensionen von Wirkungen zu betrachten und diese nach Möglichkeit miteinander in Bezug zu setzen. Auch dies erfolgt bislang äußerst selten. Auch hier kommt sicherlich die methodische Schwierigkeit, Wechselwirkungen zwischen einzelnen Zieldimensionen systematisch zu erfassen, zum Tragen. Bei der Analyse der Evaluierungsqualität schnitten die untersuchten Evaluierungen insbesondere beim Wirkungsnachweis vergleichsweise schlecht ab. Eine Ursache hierfür sind vermutlich die fehlenden Voraussetzungen für die Evaluierbarkeit nicht intendierter Wirkungen und Wechselwirkungen zwischen den Dimensionen. Beide Aspekte sind bislang selten expliziter Bestandteil der Wirkungslogiken von GIZ- und KfW-Vorhaben. Die frühe Identifikation potenzieller Neben- und Wechselwirkungen von Vorhaben ist jedoch Voraussetzung für deren spätere Überprüfung.

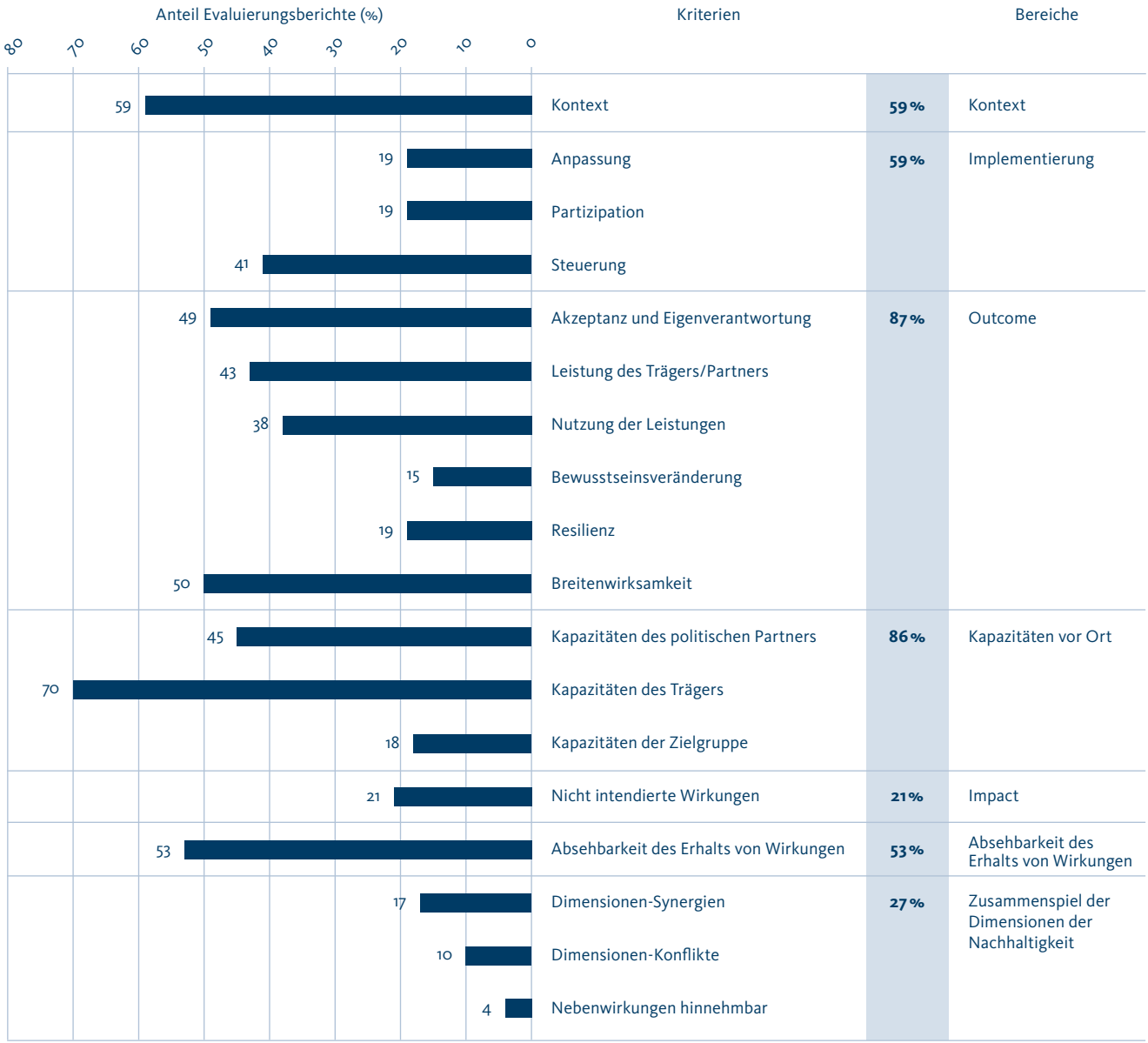
Die Ergebnisse zeigen, dass das bisherige Verständnis von Nachhaltigkeit zwar erkennbar über den Aspekt der Dauerhaftigkeit hinausgeht, jedoch noch nicht im Einklang mit dem Verständnis der Nachhaltigkeit im Sinne der Agenda 2030 steht.

Für die Beurteilung des Nachhaltigkeitsverständnisses entlang der angelegten Bewertungskriterien interessiert jedoch nicht nur, ob bestimmte Kriterien für die Einschätzung von Nachhaltigkeit herangezogen werden, sondern auch, ob deren Vorhandensein bzw. Nicht-Vorhandensein nach Einschätzung der Evaluatorinnen und Evaluatoren als förderlich bzw. hinderlich für die Nachhaltigkeit von Vorhaben angesehen wird. Abbildung 7 zeigt, dass laut der Berichte alle übergeordneten Bewertungskriterien die Bewertung sowohl positiv als auch negativ beeinflussen können. Hierzu ein Beispiel: Stellt eine Evaluierung Akzeptanz und Ownership auf Seiten der Partner fest, geht dies in aller Regel als Erfolgsfaktor in die Bewertung von Nachhaltigkeit ein. Stellt eine Evaluierung hingegen fest, dass der Partner keine Ownership zeigt, geht dieses Ergebnis als Herausforderung in die Nachhaltigkeitsbewertung ein und korrigiert die Nachhaltigkeitsnote schließlich nach unten. Da der theoretisch mögliche Fall, dass ein Kriterium sich, obwohl vorhanden, negativ auf die Nachhaltigkeitsbewertung auswirkt, nur sehr selten vorkam, werden diese Fälle in der Analyse nicht differenziert betrachtet. Ein Beispiel wäre das Kriterium „Nutzung des Outputs": Während die Nutzung des Outputs in der Regel positiv in die Nachhaltigkeitsbewertung eingeht, würde die Übernutzung des Outputs im negativen Sinne eingehen.

Aus Sicht der Steuerung und der Evaluierung von Vorhaben stellt sich die Frage, welche Kriterien überwiegend in einen positiven oder einen negativen Zusammenhang mit Nachhaltigkeit gesetzt werden. Hierbei zeigt sich, dass den meisten Kriterien in den Evaluierungen ein überwiegend positiver Einfluss zugesprochen wird. Lediglich die Kriterien „Kontext" sowie die „Partner- und „Trägerkapazitäten" werden überwiegend als hemmende Nachhaltigkeitsfaktoren angesehen. Dieses Ergebnis überträgt sich auch auf die zugehörigen Bereiche: Während „Kontext" und „Kapazitäten vor Ort" scheinbar als hemmend für die Nachhaltigkeit betrachtet werden, schneiden die Bereiche „Implementierung", „Outcome" und „Absehbarkeit der Wirkungen" vergleichsweise gut ab. Auch die Bereiche „nicht intendierte Wirkungen" und „Zusammenspiel der Dimensionen" werden überwiegend positiv in Bezug zur Nachhaltigkeit gesetzt; allerdings wurde zu diesen Bereichen auch deutlich seltener berichtet.[16]

[16] *Kriterien, die von nur sehr wenigen Vorhaben thematisiert wurden (weniger als 5 % der Stichprobengröße), wurden aufgrund der geringen Aussagekraft der Ergebnisse nicht mit in die weitere Analyse einbezogen.*

Abbildung 6: Anteil Evaluierungsberichte mit Bezug zu Bewertungskriterien und -bereichen



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte, die bei der Bewertung von Nachhaltigkeit zu mindestens einem differenzierten Kriterium des jeweiligen Nachhaltigkeitskriteriums Bezug nehmen. Die blau hinterlegten Prozentzahlen zeigen den Anteil aller Berichte (in %), die mindestens ein Kriterium des Bereichs berichtet haben. N = 513

Diese Ergebnisse zeigen eine Tendenz: Während externe Faktoren im Umfeld der Entwicklungsmaßnahme eher als hinderliche Faktoren in die Nachhaltigkeitsbewertung eingehen, werden Kriterien, die im Gestaltungsbereich der Vorhaben liegen, als eher förderlich für die Nachhaltigkeit eingeschätzt. Auf diese Tatsache wird in der nun folgenden Darstellung der Ergebnisse zu den einzelnen Nachhaltigkeitsbereichen näher eingegangen. Darüber hinaus ist bemerkenswert, dass die GIZ die Nachhaltigkeit ihrer Projekte in den Evaluierungen signifikant positiver einschätzt als die KfW. Diese Feststellung wird auch durch die begleitende Evaluierungssynthese bestätigt (Noltze et al., 2018). Im regionalen Vergleich fällt auf, dass die Nachhaltigkeitsbewertung für die Region Subsahara-Afrika in allen Bereichen außer „Kontext" deutlich schlechter ausfällt als für andere Regionen. Die Evaluierungssynthese hat allerdings gezeigt, dass dieses Ergebnis nicht robust ist, wenn für weitere Variablen kontrolliert wird (Noltze et al., 2018). Ferner schneiden überregionale Vorhaben in allen Bereichen deutlich positiver ab als Vorhaben in jeweils einzelnen Regionen; einzige Ausnahme bildet wiederum der Kontext. Dies kann zum einen durch Synergieeffekte zwischen den verschiedenen Programmen begründet sein, zum anderen durch die Möglichkeit eines im Vergleich zu einzelnen Programmen „holistischeren" Ansatzes. Beispielsweise können sich hier Effekte, die sich durch die Situation in Nachbarländern ergeben, auswirken oder auch eine größere Bandbreite an Stakeholdern in den Prozess einbezogen werden.

4.2.2 Kontext

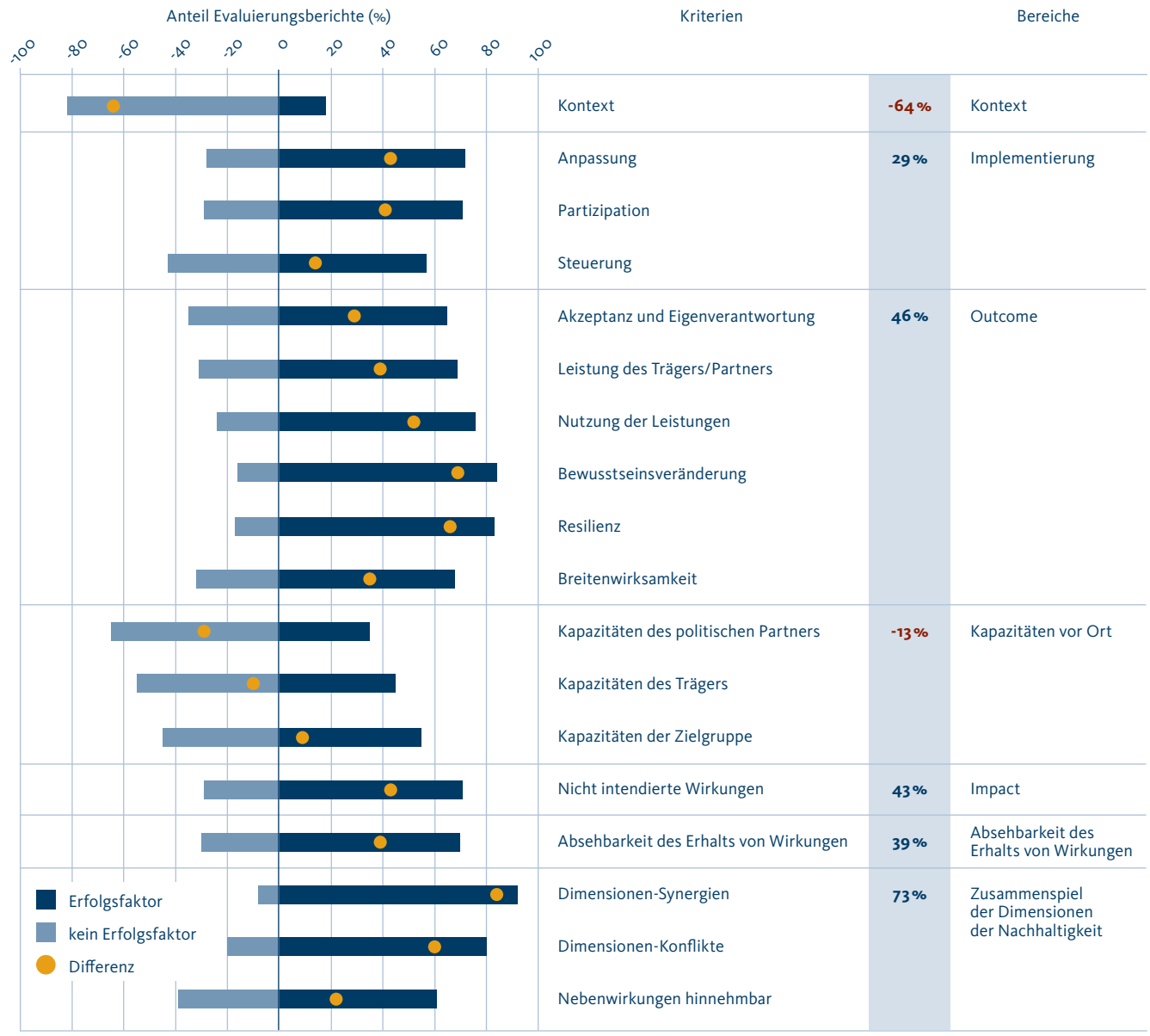
Ein erster und erwartungsgemäß wichtiger Bereich in der Nachhaltigkeits-Bewertung von Vorhaben ist die Auseinandersetzung mit dem jeweiligen Kontext. Nach dem Bewertungsraster dieser Meta-Evaluierung wurden diejenigen Kontextfaktoren in der Analyse berücksichtigt, die laut der Berichte einen direkten Einfluss auf die Wirkungen oder auf die Absehbarkeit des Erhalts der Wirkungen haben. Dabei wurden die Kontextfaktoren nach der sozialen, ökonomischen, ökologischen und politischen Dimension differenziert betrachtet. Der Großteil der Evaluierungen (59 %) bezieht den Kontext in die Bewertung der Nachhaltigkeit von Vorhaben ein (siehe Abbildung 6 in Kapitel 4.2.1). Der Schwerpunkt der Berichterstattung liegt hierbei auf politischen Aspekten; knapp die Hälfte aller Evaluierungsberichte bezieht dieses Kriterium ein (siehe Abbildung 11 im Anhang). Während wirtschaftliche Kontextfaktoren noch in rund einem Viertel aller Berichte thematisiert werden, spielen soziale und ökologische Kontextfaktoren in der Nachhaltigkeitsbewertung nur selten eine Rolle. Hierbei berichtet die KfW häufiger zu Kontextfaktoren als die GIZ (siehe Abbildung 13 im Anhang), insbesondere hinsichtlich wirtschaftlicher Aspekte (siehe Abbildung 14 im Anhang)[17]. Die Ursache hierfür könnte in strukturellen Unterschieden zwischen TZ- und FZ-Vorhaben liegen. FZ-Vorhaben kommen in der Regel ohne Vor-Ort-Präsenz aus. Gleichzeitig werden teilweise erhebliche Mittel in die Verantwortung der Partner und Träger übergeben. Demgemäß ist der Kontext von besonderer Bedeutung und wird später auch in den Ex-post-Evaluierungen entsprechend berücksichtigt.

Auch die Richtung von Kontexteinflüssen auf die Nachhaltigkeitsbewertung ist von Bedeutung: Kontextfaktoren werden im Vergleich zu anderen Faktoren auf überwiegend negative Weise mit der Nachhaltigkeit von Vorhaben in Verbindung gebracht. Dies bedeutet zum Beispiel, dass eine bestimmte politische Entwicklung, etwa vor wegweisenden Wahlen, als Unsicherheitsfaktor in die Nachhaltigkeitsbewertung eingeht und die Bewertung schließlich nach unten korrigiert wird. Bei der Gesamtbetrachtung ergibt sich ein klares Bild: Kontextfaktoren bilden in der Nachhaltigkeitsbewertung insgesamt einen kritischen Bereich. Die hohe negative Differenz des Gesamtbereichs „Kontext" (siehe Abbildung 7 in Kapitel 4.2.1) ist insbesondere auf die negative Bilanz zu sozialen (bzw. wirtschaftlichen) Aspekten zurückzuführen (siehe Abbildung 8 in Kapitel 4.2.1): In der Differenz kommen hier knapp 90 Prozent (bzw. 70 Prozent) der Evaluierungen, die zu diesem Kriterium berichten, zu einer negativen Einschätzung. Dies ist über alle Sektoren konstant, lediglich Evaluierungen zu Vorhaben im Gesundheitsbereich kommen zu einer vergleichsweis ausgewogenen Einschätzung des Einflusses von Kontextfaktoren auf die Nachhaltigkeit (siehe Abbildung 21 und Abbildung 22 im Anhang).

Für die Evaluierung von Wirkungen und ihrer Nachhaltigkeit ist der Zeitpunkt der Messung von zentraler Bedeutung. Ex-post-Evaluierungen, deren Beobachtungen in einem zeitlichen Abstand zum Ende eines Vorhabens erfolgen, kommt daher eine wichtige Rolle zu. Bei einer vergleichenden Betrachtung von

[17] *Da GIZ und KfW mit einer unterschiedlichen Anzahl von Evaluierungsberichten in die Analyse einflossen (GIZ: n = 341, KfW: n = 172), wurden die Häufigkeiten in diesen beiden Abbildungen um die unterschiedliche Anzahl korrigiert. 100 Prozent der Skala bedeuten hier also 100 Prozent der GIZ-Berichte bzw. 100 Prozent der KfW-Berichte.*

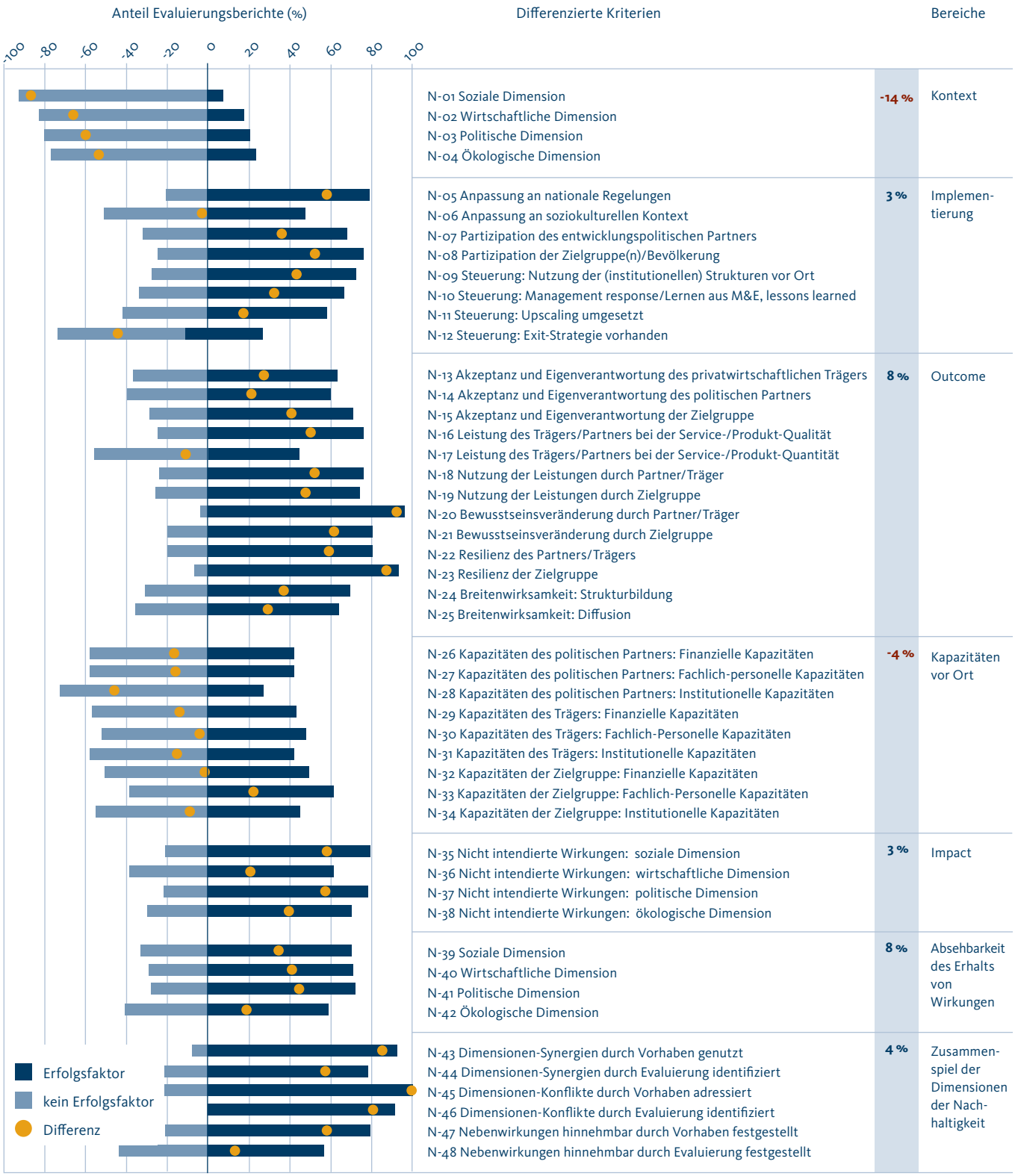
Abbildung 7: Einfluss der Nachhaltigkeitskriterien und -bereiche auf die Nachhaltigkeitsbewertung



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die gesamte Balkenlänge stellt hierbei jeweils 100 Prozent der zu diesem Kriterium berichtenden Evaluierungen dar. Die Balken oberhalb (bzw. unterhalb) der Achse repräsentieren die Anzahl der Evaluierungsberichte, die dem Kriterium einen positiven (bzw. negativen) Einfluss auf die Nachhaltigkeit zuschreiben. Die Punkte stellen die Differenz zwischen dem Anteil der positiven und dem Anteil der negativen Bewertungen eines Kriteriums dar. Die blau hinterlegten Prozentzahlen zeigen die durchschnittlichen Werte pro Bereich. N = 513

Abbildung 8: Relativer Anteil Evaluierungsberichte nach differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung



Differenzierte Kriterien		Bereiche
N-01 Soziale Dimension N-02 Wirtschaftliche Dimension N-03 Politische Dimension N-04 Ökologische Dimension	-14 %	Kontext
N-05 Anpassung an nationale Regelungen N-06 Anpassung an soziokulturellen Kontext N-07 Partizipation des entwicklungspolitischen Partners N-08 Partizipation der Zielgruppe(n)/Bevölkerung N-09 Steuerung: Nutzung der (institutionellen) Strukturen vor Ort N-10 Steuerung: Management response/Lernen aus M&E, lessons learned N-11 Steuerung: Upscaling umgesetzt N-12 Steuerung: Exit-Strategie vorhanden	3 %	Implementierung
N-13 Akzeptanz und Eigenverantwortung des privatwirtschaftlichen Trägers N-14 Akzeptanz und Eigenverantwortung des politischen Partners N-15 Akzeptanz und Eigenverantwortung der Zielgruppe N-16 Leistung des Trägers/Partners bei der Service-/Produkt-Qualität N-17 Leistung des Trägers/Partners bei der Service-/Produkt-Quantität N-18 Nutzung der Leistungen durch Partner/Träger N-19 Nutzung der Leistungen durch Zielgruppe N-20 Bewusstseinsveränderung durch Partner/Träger N-21 Bewusstseinsveränderung durch Zielgruppe N-22 Resilienz des Partners/Trägers N-23 Resilienz der Zielgruppe N-24 Breitenwirksamkeit: Strukturbildung N-25 Breitenwirksamkeit: Diffusion	8 %	Outcome
N-26 Kapazitäten des politischen Partners: Finanzielle Kapazitäten N-27 Kapazitäten des politischen Partners: Fachlich-personelle Kapazitäten N-28 Kapazitäten des politischen Partners: Institutionelle Kapazitäten N-29 Kapazitäten des Trägers: Finanzielle Kapazitäten N-30 Kapazitäten des Trägers: Fachlich-Personelle Kapazitäten N-31 Kapazitäten des Trägers: Institutionelle Kapazitäten N-32 Kapazitäten der Zielgruppe: Finanzielle Kapazitäten N-33 Kapazitäten der Zielgruppe: Fachlich-Personelle Kapazitäten N-34 Kapazitäten der Zielgruppe: Institutionelle Kapazitäten	-4 %	Kapazitäten vor Ort
N-35 Nicht intendierte Wirkungen: soziale Dimension N-36 Nicht intendierte Wirkungen: wirtschaftliche Dimension N-37 Nicht intendierte Wirkungen: politische Dimension N-38 Nicht intendierte Wirkungen: ökologische Dimension	3 %	Impact
N-39 Soziale Dimension N-40 Wirtschaftliche Dimension N-41 Politische Dimension N-42 Ökologische Dimension	8 %	Absehbarkeit des Erhalts von Wirkungen
N-43 Dimensionen-Synergien durch Vorhaben genutzt N-44 Dimensionen-Synergien durch Evaluierung identifiziert N-45 Dimensionen-Konflikte durch Vorhaben adressiert N-46 Dimensionen-Konflikte durch Evaluierung identifiziert N-47 Nebenwirkungen hinnehmbar durch Vorhaben festgestellt N-48 Nebenwirkungen hinnehmbar durch Evaluierung festgestellt	4 %	Zusammenspiel der Dimensionen der Nachhaltigkeit

Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen differenzierten Kriterium entweder einen positiven (dunkelblau) oder einen negativen (hellblau) Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Einzelne differenzierte Kriterien enthalten nur die Berichte, die das jeweilige differenzierte Kriterium zur Bewertung von Nachhaltigkeit heranziehen. Die Punkte stellen die Differenz zwischen dem Anteil der positiven und negativen Bewertungen eines differenzierten Kriteriums dar. Die blau hinterlegten Prozentzahlen zeigen die durchschnittlichen Werte pro Bereich. N = 513

Ex-post-Evaluierungen (von GIZ und KfW) einerseits und den übrigen eingesetzten Evaluierungsformaten andererseits (PFK, PEV, Schluss-Evaluierungen der GIZ) zeigt sich, dass die Evaluierungstypen, die zu einem vergleichsweise frühen Zeitpunkt zum Einsatz kommen, ökologische Kontextfaktoren als deutlich kritischer einschätzen (siehe Abbildung 19 im Anhang). Bei den Ex-post-Evaluierungen gibt es hingegen vergleichsweise mehr Fälle, die den Einfluss des ökologischen Kontextes positiv beurteilen, wenn die Beurteilung insgesamt auch immer noch überwiegend negativ ist. Eine mögliche Erklärung ist, dass Erfolge in der ökologischen Dimension erst nach einem längeren Zeitraum eintreten bzw. dass sich die entsprechenden Wirkungen erst vergleichsweise spät messen lassen. Weiterhin finden sich auch Unterschiede zwischen den DO: Ex-post-Evaluierungen der GIZ beurteilen ökologische Kontextfaktoren kritischer als KfW-Ex-post-Evaluierungen (siehe Abbildung 19 im Anhang). Dieses Ergebnis deutet auf systematische Unterschiede in der Bewertung von Nachhaltigkeit je nach Evaluierungstyp hin. In Kapitel 4.1 wurde gezeigt, dass sich die Evaluierungstypen nicht nur hinsichtlich des Evaluierungszeitpunktes, sondern auch in der Qualität unterscheiden. Der mögliche Einfluss der Qualität auf die Bewertung von Nachhaltigkeit bedarf somit besonderer Aufmerksamkeit und wird abschließend in Kapitel 4.3 diskutiert.

Insgesamt haben Kontextfaktoren laut der Evaluierungsberichte einen großen Einfluss auf die Bewertung der Nachhaltigkeit von Vorhaben. Zudem zeigt sich, dass sie vor allem als negative Beeinträchtigung wahrgenommen werden. Dies birgt allerdings die Gefahr einer systematischen Verzerrung in der Berichterstattung: Möglicherweise kommen negative Kontexteinflüsse in den Erhebungen der Evaluierungen eher zur Sprache und nehmen anschließend auch mehr Raum in der Darstellung der Ergebnisse ein, während ein neutraler oder positiver Kontext seltener Erwähnung findet. In der begleitenden Evaluierungssynthese wurden aus diesem Grund weitere, nicht von den Evaluierungsberichten herausgestellte Kontextfaktoren in die Kausalanalyse integriert. Dabei hat sich gezeigt, dass insbesondere in Ex-post-Evaluierungen ein signifikant negativer Zusammenhang zwischen der allgemeinen Einkommenssituation eines Landes und der vergebenen Nachhaltigkeitsnote eines Vorhabens besteht.[18]

4.2.3 Implementierung

Der zweite analysierte Bereich der Nachhaltigkeitsbewertung beinhaltet Aspekte der Implementierung. Darunter fallen nach dem Bewertungsraster die Kriterien „Alignment", „Partizipation" und „Steuerung". Insgesamt zeigt sich, dass Kriterien aus dem Bereich der Implementierung eine mäßige Rolle in der Nachhaltigkeitsbewertung von Vorhaben spielen. Zwar beziehen 59 Prozent der Evaluierungen mindestens ein Kriterium aus diesem Bereich in die Bewertung ein (siehe Abbildung 6 in Kapitel 4.2.1), doch werden die einzelnen Kriterien für sich betrachtet nur relativ selten mit Nachhaltigkeit in Verbindung gebracht (siehe Abbildung 13 im Anhang).

In der Wirksamkeitsdebatte nimmt das Kriterium „Alignment" seit geraumer Zeit eine zentrale Bedeutung ein. Es wird meist als Voraussetzung für die Akzeptanz und die Eigenverantwortung der Partner, Träger und Zielgruppen und somit als elementar für die Effektivität der EZ angeführt (Hartmuth, 2004; Klingebiel, 2013; OECD, 2017). In dieser Analyse bzw. in den Evaluierungsberichten wird Alignment – im Sinne der Anpassung einer Maßnahme an lokale Strukturen – entweder als Anpassung an nationale Entwicklungsstrategien oder als Anpassung an den soziokulturellen Kontext der Zielgruppen verstanden. Nach den Ergebnissen dieser Meta-Evaluierung scheint die Wirkungskette von Akzeptanz und Ownership über Nutzung des Outputs bis hin zu den Wirkungen und letztlich der Nachhaltigkeit im Sinne von Dauerhaftigkeit der Wirkungen jedoch relativ lang zu sein. Zumindest bringt nur etwa jeder zehnte Evaluierungsbericht „Alignment" in einen direkten Zusammenhang mit der Nachhaltigkeit von Vorhaben. Wenn dies erfolgt, dann allerdings überwiegend im positiven Sinne: 70 Prozent der Evaluierungsberichte, die zum Kriterium Alignment berichten, sehen hierin einen Erfolgsfaktor.

Auch „Partizipation" steht seit geraumer Zeit im Zentrum der Wirksamkeitsdebatte und gehört somit zu den Erfolgskriterien der Evaluierung der EZ (OECD, 2010b). Neben dem Grad der Partizipation lassen sich auch die jeweiligen Akteursgruppen unterscheiden. Im Bewertungsraster dieser Meta-Evaluierung wurde Partizipation dann berücksichtigt, wenn laut der Berichte die Partner oder Zielgruppen wenigstens konsultiert wurden und dies für die Nachhaltigkeit von Vorhaben von Bedeutung

[18] *Hierfür wurden zur breiteren Überprüfung des Kontextes das Bruttoinlandsprodukt (BIP) pro Kopf in aktuellen US-Dollar sowie empfangene Gelder der EZ (ODA-Nettomittel) in Prozent des BIP und der sogenannte Freedom-House-Index herangezogen. Der Freedom-House-Index gibt Auskunft über das Maß an politischen Rechten und zivilen Freiheiten einer Gesellschaft (Freedom House, 2016).*

war. Dabei hat sich gezeigt, dass Partizipation eher selten in die Nachhaltigkeitsbewertung eingeht, dann allerdings überwiegend als Erfolgsfaktor beschrieben wird. Die GIZ kommt insbesondere hinsichtlich der Partizipation auf Ebene der Zielgruppen zu deutlich positiveren Einschätzungen als die KfW (siehe Abbildung 15 und Abbildung 16 im Anhang). Zudem variieren die Ergebnisse je nach Evaluierungstyp: Ex-post-Evaluierungen von GIZ und KfW schätzen den Einfluss von Partizipation auf die Nachhaltigkeit von Vorhaben weder positiv noch negativ ein, dezentrale und Schluss-Evaluierungen der GIZ kommen zu einem deutlich positiven Ergebnis (siehe Abbildung 17 und Abbildung 18 im Anhang). Dies deutet darauf hin, dass Evaluierungen, die im Verlauf oder kurz nach Beendigung eines Vorhabens stattfinden, die Bedeutung der Partizipation systematisch positiver einschätzen als Evaluierungen, die in zeitlichem Abstand zum Ende der Vorhaben durchgeführt werden.

Als letzter Teil im Bereich der Implementierung wurde der Stellenwert von steuerungsrelevanten Bewertungskriterien untersucht. Dabei wurde überprüft, inwieweit die Nutzung lokaler Strukturen in der Steuerung, beim Einbeziehen von Empfehlungen aus Monitoring und Evaluierung („Management Response") sowie in der Formulierung von Upscaling- und Exit-Strategien als relevant für die Nachhaltigkeitsbewertung angesehen wurde. Hierbei hat sich gezeigt, dass knapp die Hälfte aller Evaluierungsberichte mindestens eines dieser vier Kriterien in die Nachhaltigkeitsbewertung einbezog (siehe Abbildung 6 in Kapitel 4.2.1). Für den Steuerungsaspekt gilt insgesamt, dass diese Kriterien sowohl als Erfolgs- als auch als Misserfolgsfaktoren in die Bewertung eingingen, allerdings in der Summe zumeist positiv dargestellt wurden. Eine Ausnahme bildete das Bewertungskriterium „Exit-Strategie"; dieses wurde in den Evaluierungsberichten überwiegend als problematisch für die Nachhaltigkeit gesehen. Dieser Zusammenhang zeigt sich jedoch nahezu ausschließlich in Evaluierungen der GIZ, während KfW- Evaluierungen hier relativ ausgewogen berichten (siehe Abbildung 15 im Anhang). Dies erscheint vor dem Hintergrund der verschiedenen Implementierungsansätze plausibel. Während die GIZ in Vorhaben mit starker Vor-Ort-Präsenz eher auf funktionierende Phasing-Out-Konzepte angewiesen ist, sind die Modalitäten einer Übergabe in FZ-Vorhaben zumeist bereits im Modulvorschlag spezifiziert. Anhand der Daten dieser Meta-Evaluierung kann jedoch nicht abschließend erklärt werden, ob der überwiegend negative Nachhaltigkeitsbezug an dem Fehlen einer Exit-Strategie liegt oder vielmehr daran, dass vorhandene Exit-Strategien schlecht aufgesetzt waren. Auch van Tulder und Pfisterer (2008) betonen den hohen Stellenwert gut umgesetzter Exit- oder Phasing-Out-Strategien für die Nachhaltigkeit von Vorhaben.

Zusammenfassend betrachtet werden Implementierungsaspekte überwiegend positiv mit der Nachhaltigkeit von Vorhaben in Verbindung gebracht. Unterschiede zeigen sich in der Bewertungspraxis der beiden DO und lassen sich durch unterschiedliche Implementierungsstrukturen der TZ und FZ erklären. Darüber hinaus sind im Bereich der Implementierung keine nennenswerten Unterschiede zwischen Sektoren oder Regionen erkennbar.

4.2.4 Outcome

Die direkten und indirekten, kurz- und mittelfristigen Wirkungen einer Entwicklungsmaßnahme bilden einen weiteren Aspekt der Nachhaltigkeits-Bewertung. Sie wurden im Bewertungsraster mit dem Kurztitel „Outcome" zusammengefasst und werden durch eine Vielzahl unterschiedlicher Kriterien beschrieben. Zu den häufig herangezogenen Kriterien in diesem Bereich gehören „Akzeptanz und Ownership", „Output des Trägers/Partners", „Nutzung des Outputs" und „Reichweite/Breitenwirksamkeit" (siehe Abbildung 6 in Kapitel 4.2.1 und Abbildung 11 im Anhang). Knapp die Hälfte der untersuchten Evaluierungsberichte erwähnt diese Kriterien. Andere Aspekte in diesem Bereich, wie „Bewusstseinsveränderung" und „Resilienz/Anpassungsfähigkeit", wurden hingegen eher selten in die Nachhaltigkeitsbewertung einbezogen.

Im Bereich „Outcome" wurde zunächst die Rolle von „Akzeptanz und Ownership" untersucht. Beide Konzepte gehören seit jeher zur Debatte um die Wirksamkeit der EZ und werden entsprechend auch mit Nachhaltigkeit in Verbindung gebracht (OECD, 2008). Die Annahme ist, dass Akzeptanz und Ownership Voraussetzung für eine erfolgreiche EZ und die Fortführung der erreichten Erfolge über die Zeit sind (Russ-Eft, 2014; Stockmann und Silvestrini, 2011). Da beide Konzepte in den Berichten stets in einem engen Zusammenhang stehen, wurden sie in dieser Meta-Evaluierung gemeinsam erfasst.

Dabei wurde untersucht, inwieweit die Evaluierungen die Eigeninitiative von lokalen Akteuren in die Bewertung von Nachhaltigkeit einbezogen haben. Die Konzepte wurden getrennt nach Akteursgruppen, also „politischen Partnern", „durchführenden Trägern" sowie „Zielgruppen", analysiert. Dabei hat sich gezeigt, dass „Akzeptanz" und „Ownership" in etwa jeder zweiten Evaluierung mit Nachhaltigkeit in Verbindung gebracht wurden (siehe Abbildung 6 in Kapitel 4.2.1) und dass sie überwiegend als Erfolgskriterien in die Nachhaltigkeits-Bewertung eingingen. Über alle Evaluierungstypen hinweg schätzt die GIZ Akzeptanz und Ownership in ihren Vorhaben deutlich positiver ein als die KfW.

Einen weiteren Analyseaspekt bilden die direkten Leistungen der Vorhaben. Unter dem Begriff „Qualität und Quantität des Outputs" wurde untersucht, inwieweit die Qualität und Quantität der Leistungen als hinreichend eingeschätzt wurden, um die Ziele der Vorhaben zu erreichen. Dabei zeigte sich, dass die Qualität der Leistungen deutlich häufiger in die Nachhaltigkeits-Bewertung einbezogen wurde als die Quantität (siehe Abbildung 11 im Anhang). Zudem weist die Quantität der Leistungen eine – wenn auch geringe – negative Differenz auf (siehe Abbildung 8 im Kapitel 4.2.1). Dies kann bedeuten, dass sich in den Evaluierungsberichten Quantität im Vergleich zur Qualität zwar unbedeutender für die Bewertung von Nachhaltigkeit darstellt, das Fehlen einer gewissen Quantität sich jedoch auch negativ auf die Nachhaltigkeit von Vorhaben auswirkt.

Aufbauend auf diesen Ergebnissen wurde durch die Meta-Evaluierung überprüft, inwieweit die Bereitstellung von Leistungen auch deren Nutzung nach sich zieht. Bei der „Nutzung des Outputs" wurde zwischen der Nutzung durch die Partner und/oder Träger und der Nutzung durch die Zielgruppen unterschieden. Dabei hat sich gezeigt, dass knapp jeder vierte Bericht die Nutzung des Outputs in die Nachhaltigkeitsbewertung einbezieht (siehe Abbildung 6 in Kapitel 4.2.1). Während Evaluierungen der GIZ vornehmlich über die Nutzung des Outputs durch die Partner bzw. Träger berichten, thematisieren KfW-Evaluierungen häufiger die Output-Nutzung durch die Zielgruppe (siehe Abbildung 14 im Anhang). Auch hier liegt eine mögliche Erklärung in den unterschiedlichen Implementierungsstrukturen: Viele TZ-Leistungen werden durch lokales Personal vor Ort erbracht und richten sich anschließend zunächst an die Partner oder die lokalen Trägerstrukturen, die wiederum Leistungen für die Zielgruppen bereitstellen. FZ-Leistungen werden hingegen in der Regel von den lokalen Trägern erbracht und kommen direkt bei den Zielgruppen an. Die Nutzung der direkten Leistungen geht anschließend überwiegend als Erfolgsfaktor in die Bewertung der Nachhaltigkeit ein. Auch hier zeigen sich Unterschiede zwischen den DO: Evaluierungen der GIZ stellen den Zusammenhang zwischen der Nutzung des Outputs und der Nachhaltigkeit positiver dar als diejenigen der KfW (siehe Abbildung 15 und Abbildung 16 im Anhang).

Weiterhin wurde im Bereich „Outcome" untersucht, ob die Vorhaben über „Bewusstseinsveränderung" auf der Ebene der Partner/Träger bzw. Zielgruppen zur Nachhaltigkeit beigetragen haben. In diesem Zusammenhang wurde analysiert, inwieweit die Evaluierungen auf Langfristigkeit angelegte Verhaltensänderungen der Akteure einbeziehen. Die Annahme ist, dass Veränderungen im Bewusstsein der Akteure einen besonders starken Einfluss auf die Nachhaltigkeit von Vorhaben haben (Stadtler, 2016; Von Raggamby und Rubik, 2012). Die Ergebnisse entsprechen dieser Annahme auf den ersten Blick nur bedingt: Nur 15 Prozent aller Evaluierungen ziehen Bewusstseins- und Verhaltensänderungen der Zielgruppen bei der Betrachtung von Nachhaltigkeit von Vorhaben heran (siehe Abbildung 6 in Kapitel 4.2.1). Doch wenn dazu berichtet wurde, erfolgte dies sehr positiv: „Bewusstseinsveränderung" weist mit 70 Prozent aller hierzu berichtenden Evaluierungen die höchste Positiv-Negativ-Differenz aller Outcome-Kriterien auf. Insofern bestätigt sich die Annahme, dass Bewusstseinsveränderungen einen vergleichsweise starken positiven Einfluss auf die Bewertung von Nachhaltigkeit haben. Hinsichtlich des Trägers/Partners schätzte die GIZ die Bewusstseinsveränderung über alle Evaluierungsberichte hinweg sehr positiv ein (siehe Abbildung 15 im Anhang). Auch die Evaluierungen der KfW kommen hier zu positiven Ergebnissen, wenn auch in der Positiv-Negativ-Differenz deutlich ausgewogener: Sie beträgt etwa 30 Prozent der zu Bewusstseinsveränderung berichtenden KfW-Evaluierungen.

Zusätzlich untersuchte diese Meta-Evaluierung den Stellenwert des Analyseaspekts „Resilienz/Anpassungsfähigkeit". Dabei wurde anhand der Evaluierungsberichte überprüft,

inwieweit die Vorhaben die Partner/Träger bzw. Zielgruppen in die Lage versetzt haben, Entwicklungspotenziale und -risiken eigenständig zu erkennen und in Handlungen umzusetzen. Es zeigte sich, dass die Resilienz bzw. Anpassungsfähigkeit der Akteure nur selten in die Nachhaltigkeitsbewertung einbezogen wurde. Nur etwa jeder fünfte Evaluierungsbericht geht auf dieses Kriterium ein (siehe Abbildung 6 in Kapitel 4.2.1). Allerdings wird Resilienz überwiegend als Erfolgsfaktor für die Nachhaltigkeit gesehen. Insbesondere die Resilienz der Zielgruppen wird dabei als Erfolg versprechend dargestellt (siehe Abbildung 7 und Abbildung 8 in Kapitel 4.2.1). Ein überraschendes Ergebnis ist, dass allein im Bildungssektor die Resilienz bzw. Anpassungsfähigkeit als überwiegend hemmender Faktor für Nachhaltigkeit eingeschätzt wird (siehe Abbildung 21 im Anhang). Hier bestand die Annahme, dass Bildungsmaßnahmen der Resilienz der Zielgruppe zuträglich sind. Dies lässt sich im Rahmen der vorliegenden Untersuchung empirisch nicht bestätigen.

Schließlich wurde auch der Analyseaspekt „Reichweite und Breitenwirksamkeit" auf seine Bedeutung im Rahmen der Nachhaltigkeitsbewertung von Vorhaben untersucht. Hier wurden vor allem die Aussagen zu den Kriterien „Strukturbildung" und „Diffusion" betrachtet. Mit dem Kriterium der Strukturbildung wurde überprüft, inwieweit Veränderungen auf Systemebene eingetreten sind, die auch für die Bewertung der Nachhaltigkeit herangezogen wurden. Unter dem Aspekt der Diffusion wurde analysiert, inwieweit sich die Leistungen und Innovationen über die ursprüngliche Zielgruppe hinaus verbreitet haben. Auf Basis der Literatur wurde angenommen, dass Breitenwirksamkeit einen relevanten Erfolgsfaktor für die Nachhaltigkeit von Vorhaben darstellt (Stadtler, 2016; Vahlhaus, 2014; Von Raggamby und Rubik, 2012). Die Ergebnisse bestätigen diese Annahme vollumfänglich: Breitenwirksamkeit wird in über der Hälfte aller Evaluierungen mit Nachhaltigkeit in Verbindung gebracht und gehört damit zu den am häufigsten behandelten Kriterien im Bereich „Outcome" (siehe Abbildung 6 in Kapitel 4.2.1). Bei näherer Betrachtung zeigt sich, dass dabei das Kriterium „Strukturbildung" einen deutlich höheren Stellenwert einnimmt als das Kriterium „Diffusion". Insbesondere die Evaluierungen der GIZ bringen Strukturbildung häufig in die Nachhaltigkeitsdiskussion ein (siehe Abbildung 11 und Abbildung 14 im Anhang). Eine mögliche Erklärung liegt hier erneut in strukturellen Unterschieden zwischen TZ- und FZ-Vorhaben: Während FZ-Vorhaben zum großen Teil Ressourcen für den Aufbau von Infrastruktur einsetzen und Kapazitätsbildung mit Zielgruppen und Multiplikatoren nur als „Begleitmaßnahme" stattfindet, ist Letzteres in TZ-Vorhaben ein zentrales Anliegen.

In der Gesamtbetrachtung zeigt sich, dass laut der Evaluierungsberichte eine Reihe von Bewertungskriterien im Bereich „Outcome" einen deutlich positiven Einfluss auf die Nachhaltigkeit von Vorhaben zu haben scheint. Dies zeigen unter anderem die Kriterien „Bewusstseinsveränderung" und „Resilienz". Insbesondere die Evaluierungen der GIZ stellen hier deutlich positive Bezüge her. Eine sektorale Besonderheit zeigt sich dabei in Evaluierungen von Vorhaben im Transportsektor, dort wird der Einfluss von Outcome-Kriterien auf die Nachhaltigkeit deutlich kritischer beurteilt. Dieses Ergebnis könnte damit zusammenhängen, dass der Erfolg der erbrachten Leistungen im Transportsektor in hohem Maße von der Nachfrage abhängt. Ein nachfragebezogener Indikator wäre hier die wirtschaftliche Aktivität des Umfeldes der Maßnahme.

4.2.5 Kapazitäten vor Ort

Einen weiteren Bereich in der Nachhaltigkeits-Bewertung von Vorhaben bilden die Kapazitäten vor Ort. Der Begriff der Kapazitäten beinhaltet hier die finanziellen, personellen und institutionellen Beiträge der entwicklungspolitischen Partner, Träger und letztlich auch der Zielgruppen. Die Kapazitäten vor Ort stehen dabei insgesamt für die Fähigkeit der lokalen Akteure, die Leistungen weiterzuführen und die Wirkungen über die Zeit zu erhalten. In der Literatur wird den Kapazitäten vor Ort eine hohe Bedeutung für die Nachhaltigkeit beigemessen (Caspari, 2004; KfW Entwicklungsbank, 2003; Russ-Eft, 2014; Stockmann und Silvestrini, 2011). Dieser hohe Stellenwert lässt sich anhand der Ergebnisse auch empirisch belegen: 86 Prozent aller Evaluierungen beziehen die Kapazitäten vor Ort in die Bewertung ein (siehe Abbildung 6 in Kapitel 4.2.1). Am häufigsten werden dabei die Kapazitäten der Träger mit Nachhaltigkeit in Verbindung gebracht. Dieser hohe Wert ergibt sich vor allem durch die Ex-post-Evaluierungen der KfW; deren Vorhaben arbeiten vornehmlich über lokale Trägerstrukturen und messen diesen somit auch in Evaluierungen eine hohe Bedeutung bei.

Laut der Evaluierungsberichte gehen die Kapazitäten vor Ort überwiegend als negativ in die Nachhaltigkeitsbewertung ein. Dies lässt sich vermutlich durch die unzureichenden Kapazitäten der Partner, Träger und Zielgruppen in den Partnerländern der deutschen EZ erklären. Das Ergebnis ist dennoch erstaunlich, da unzureichende Kapazitäten eigentlich bereits in der Planung von Vorhaben berücksichtigt werden müssen und nicht erst mit einer Evaluierung in das Bewusstsein rücken sollten. Verwunderlich ist vor allem, dass insbesondere die Beiträge der politischen Partner als deutliche Herausforderung in die Nachhaltigkeitsbewertung eingehen. Diese sollten aber als Teil der Vertragsverhandlungen von Vorhaben eigentlich deutlich planbarer sein als etwa die Beiträge der Zielgruppen, die wiederum wesentlich häufiger als Erfolgsfaktor angeführt werden (siehe Abbildung 7). Aus Sicht der Vorhaben müsste geklärt werden, inwieweit sich die Einschätzung der Kapazitäten der Partner bereits bei der Planung von Vorhaben verbessern lässt, um später negative Auswirkungen auf die Nachhaltigkeit zu vermeiden. Eine fundierte Analyse der Kapazitäten vor Ort liegt letztlich auch im Interesse von Evaluierungen, die sich aus Sicht der Partner schnell dem Vorwurf ausgesetzt sehen, die Verantwortung für das Gelingen von Vorhaben auf externe Faktoren zu übertragen.

Interessant ist hier auch ein Vergleich zwischen GIZ und KfW. Beide Organisationen kommen in ihren Evaluierungsberichten insbesondere hinsichtlich der Trägerkapazitäten zu deutlich abweichenden Schlussfolgerungen: Während die GIZ die Trägerkapazitäten als problematisch für die Nachhaltigkeit von Vorhaben einschätzt, zieht die KfW hier eine positive Bilanz. Ein Grund liegt hier möglicherweise in den aufwändigeren Ex-ante-Prüfungen für FZ-Vorhaben, durch die schließlich verlässliche Partner identifiziert und gefunden werden. Demgegenüber sind möglicherweise in TZ-Projekten aufgrund des thematischen Schwerpunktes, zum Beispiel im Bereich guter Regierungsführung, nicht selten Träger mit hohem Beratungsbedarf involviert. Diese Annahme wird durch die sektorale Betrachtung bestätigt, in welcher der Sektor „Frieden“ unter dem Strich die mit Abstand negativste Nachhaltigkeitseinschätzung des Kriteriums „Trägerkapazitäten“ erfährt.

Der Gesamteindruck, dass die Kapazitäten vor Ort in der Bewertung nur eine geringe Differenz aufweisen, wird also in der differenzierten Betrachtung bestätigt, auch wenn moderate Unterschiede zwischen den verschiedenen Gruppen und Kriterien zu erkennen sind. Insgesamt kommt die GIZ hier zu kritischeren Bewertungen als die KfW, und die Ex-post-Evaluierungen liefern positivere Einschätzungen als die dezentralen Evaluierungsformate der GIZ. Wie schon in den zwei vorangegangenen Nachhaltigkeitsbereichen – „Implementierung“ und „Outcome“ – zeigt sich auch hier Subsahara-Afrika als Region, in der die Kapazitäten vor Ort deutlich negativer eingeschätzt werden als in anderen Regionen (siehe Abbildung 24 im Anhang). Auch Vorhaben in Lateinamerika und Europa/Kaukasus schätzen die Kapazitäten vor Ort negativ ein, jedoch deutlich ausgewogener.

4.2.6 Impact

Ein weiterer zentraler Bereich der Nachhaltigkeitsbewertung liegt in der Auseinandersetzung mit den übergeordneten entwicklungspolitischen Wirkungen (Impact). Dabei wird angenommen, dass Vorhaben, die zu entwicklungspolitischen Zielen beitragen, erfolgreicher und nachhaltiger sind als Vorhaben, die nur direkte Leistungen erbringen (Boone, 1995; Faust, 2007). Um den Stellenwert des Impacts für die Nachhaltigkeit von Vorhaben anhand der Evaluierungsberichte angemessen nachvollziehen zu können, wurden die Ergebnisse zur entwicklungspolitischen Wirksamkeit im Rahmen der vorliegenden Meta-Evaluierung umfassend analysiert. Hierzu gehörten 1) der Abgleich von beabsichtigten und nachgewiesenen Wirkungen auf Ebene der Oberziele nach Dimensionen und 2) die Aufnahme der Ergebnisse zu nicht intendierten Wirkungen entlang der Dimensionen.

Die Analyse der Stichprobe zeigt, dass soziale und wirtschaftliche Oberziele in rund 60 bzw. 50 Prozent der Vorhaben und damit am häufigsten genannt wurden. Politische und ökologische Oberziele wurden seltener angestrebt. Dabei haben viele Vorhaben Oberziele in mehr als einer Nachhaltigkeitsdimension.[19] In der Gegenüberstellung der Oberziele zeigt sich, dass die Evaluierungsberichte die Zielerreichung der Vorhaben sehr positiv einschätzen (siehe Abbildung 9). Dieser bemerkenswerte Zielerreichungsgrad hält sich über beide DOs und alle Evaluierungstypen hinweg. Lediglich die sektorale Betrachtung lässt leichte Unterschiede erkennen. So zeigt der Sektor „Frieden“ einen vergleichsweise geringen Zielerreichungsgrad. Dies

[19] *Hierauf wird in Kapitel 4.2.8 zum „Zusammenspiel der Nachhaltigkeitsdimensionen“ eingegangen.*

deckt sich mit der Beobachtung in dieser Untersuchung, dass Ergebnisse in diesem Sektor stark von Kontextfaktoren abhängig sind. Andere Sektoren hingegen weisen eine relativ hohe Zielerreichungsquote auf, so etwa die Sektoren „Demokratie", „Wirtschaft" oder „Energie" (siehe Abbildung 27 im Anhang).

Während die intendierten Wirkungen in der Mehrheit der Evaluierungen anhand der Oberziele Berücksichtigung finden, werden nicht intendierte Wirkungen kaum diskutiert: Lediglich jeder fünfte Bericht geht auf eine positive oder negative nicht intendierte Wirkung ein (siehe Abbildung 6 in Kapitel 4.2.1). Die Gegenüberstellung der positiven und negativen Bedeutung nicht intendierter Wirkungen für die Nachhaltigkeitseinschätzung ergibt ein deutlich positives Bild: 70 Prozent der Evaluierungen, die nicht intendierte Wirkungen anführen, bringen diese in einen positiven Zusammenhang mit der Nachhaltigkeit von Vorhaben (siehe Abbildung 7 in Kapitel 4.2.1).

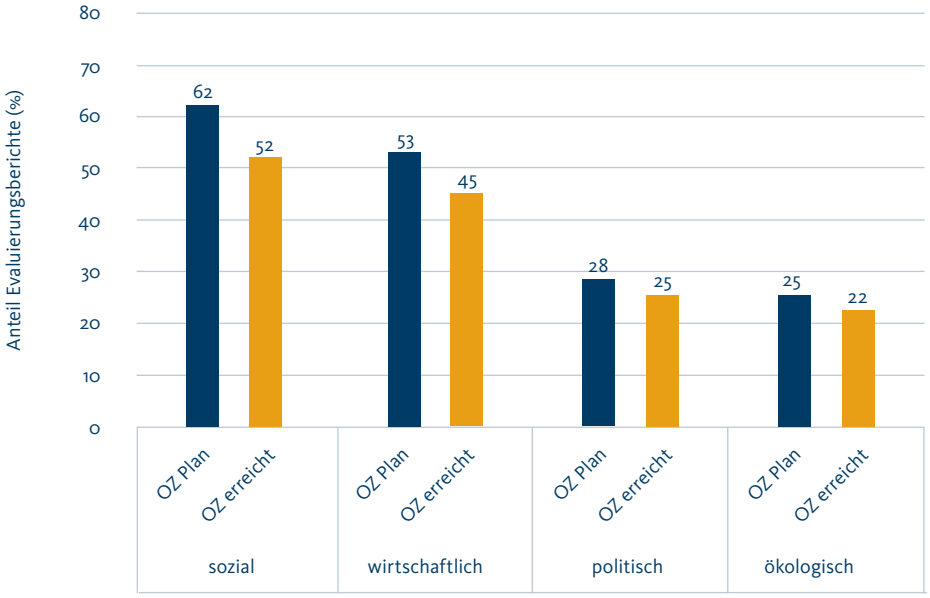
Auffällig ist die Einschätzung von wirtschaftlichen im Vergleich zu sozialen, politischen und ökologischen Aspekten. 70 bis 80 Prozent der Evaluierungen schätzen die nicht intendierten sozialen Wirkungen als positiv ein, die Differenz ist also positiv und liegt bei 50 bis 60 Prozent der Evaluierungen. Dies wird unter anderem von der GIZ getrieben, welche das Kriterium der „sozialen Nebenwirkungen" deutlich positiver einschätzt als die KfW (siehe Abbildung 15 im Anhang). Wirtschaftliche Aspekte hingegen schneiden schlechter ab: 40 Prozent der hierzu berichtenden Evaluierungen, also ein nicht geringer Anteil, kommen zu der Einschätzung, dass sich nicht intendierte Wirkungen negativ auf Nachhaltigkeit auswirken (siehe Abbildung 6 in Kapitel 4.2.1).

4.2.7 Absehbarkeit von Wirkungen

Die Absehbarkeit, dass Wirkungen über die Zeit erhalten bleiben, ist ein zentraler Aspekt in der Nachhaltigkeitsbewertung von Vorhaben. Für die Bewertung der „Dauerhaftigkeit" ist die Absehbarkeit in rein konzeptioneller Hinsicht sogar der zentrale Aspekt. Dieser wurde im Bewertungsraster der Meta-Evaluierung mit Blick auf das Erreichen der Oberziele über die Zeit untersucht. Daher ist es überraschend, dass nur etwa jeder zweite Evaluierungsbericht die Absehbarkeit des Erhalts von Wirkungen über die Zeit explizit diskutiert. Eine offensichtliche Erklärung liegt hier in der Tatsache, dass aufgrund von Unzulänglichkeiten im Wirkungsnachweis nicht alle Evaluierungen abschließende Aussagen zum Erreichen der beabsichtigten Oberziele treffen. Folgerichtig findet sich in diesen Fällen anschließend auch keine Einschätzung zur Absehbarkeit der Wirkungen. Nach dem Bewertungsraster wurde die Absehbarkeit des Erhalts der Wirkungen weiterhin nach den Dimensionen der Nachhaltigkeit differenziert untersucht. Da die Oberziele der Vorhaben von GIZ und KfW vornehmlich der sozialen und ökonomischen Dimension zugeordnet werden können (siehe Kapitel 4.2.6), wird auch die Absehbarkeit des Erhalts der Wirkungen über die Zeit vornehmlich innerhalb dieser beiden Dimensionen diskutiert (siehe Abbildung 11 im Anhang).

In der Regel geht die Absehbarkeit des Erhalts von Wirkungen bei den Evaluierungen als Erfolgsfaktor in die Bewertung der Nachhaltigkeit ein (siehe Abbildung 7 und Abbildung 8 in Kapitel 4.2.1). Dieses Ergebnis ist über beide DO und über die verschiedenen Evaluierungsformate hinweg konstant (siehe Abbildung 15 bis Abbildung 18 im Anhang). Eine Ausnahme bilden ökologische Wirkungen, die nach den Ergebnissen der Ex-post-Evaluierungen eine Gefährdung für die Nachhaltigkeit von Vorhaben darstellen. Eine mögliche Erklärung liegt darin, dass sich Wirkungen in der ökologischen Dimension erstens schwieriger erreichen und zweitens auch schwer erhalten lassen. Darüber hinaus zeigt sich eine sektorale Besonderheit: Die Absehbarkeit von Wirkungen wird im Energiesektor deutlich kritischer eingeschätzt. Dies ist auch der einzige Sektor, in dem der Faktor unter dem Strich als negativ für die Nachhaltigkeit eingeschätzt wird (siehe Abbildung 21 und Abbildung 22 im Anhang). Dies könnte darin begründet liegen, dass in diesem Sektor insbesondere Infrastrukturprojekte eine wichtige Rolle spielen, bei denen oftmals schwer einschätzbar ist, inwieweit notwendige Wartungsarbeiten über die Zeit sichergestellt sind.

Abbildung 9: Anteil Evaluierungsberichte nach angestrebten und erreichten Oberzielen und Dimensionen der Nachhaltigkeit



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil an Evaluierungsberichten nach angestrebten (blau) und erreichten (orange) Oberzielen. N = 513

4.2.8 Zusammenspiel der Dimensionen der Nachhaltigkeit

Laut dem BMZ-Leitfaden von 2006 sollen die Evaluierungen sowohl die Wirkungen als auch die Absehbarkeit der Wirkungen nach sozialen, ökonomischen, ökologischen und politischen Dimensionen untersuchen und darstellen. Da der Leitfaden das Zusammenspiel dieser Dimensionen nicht explizit fordert, stellte sich bei der vorliegenden Meta-Evaluierung die Frage, ob es dennoch in der Vergangenheit bereits eine Diskussion zu möglichen Wechselwirkungen gegeben hat. Die Motivation, dieser Frage gezielt nachzugehen, ergibt sich durch die prominente Herausstellung des Zusammenspiels der Nachhaltigkeitsdimensionen durch die Agenda 2030.

Im Rahmen der Analyse der „Wirkungspotenziale und Herausforderungen zwischen den Dimensionen“ wurde untersucht, inwieweit die Evaluierungsberichte herausgestellt haben, dass die Vorhaben Synergien zwischen den Dimensionen geschaffen, Zielkonflikte angesprochen oder mögliche Nebenwirkungen in den einzelnen Dimensionen als hinnehmbar eingeschätzt haben. Zusätzlich wurde geprüft, ob die Evaluierungen diese Einschätzungen bestätigten oder widerlegten. Im Vergleich zu den anderen Bereichen ist das Zusammenspiel der Dimensionen bislang deutlich seltener Bestandteil der Nachhaltigkeitsbewertung. In nur rund einem Viertel der Evaluierungen wurden solche Aspekte berücksichtigt. Am häufigsten wurden Synergien zwischen den Dimensionen herausgestellt (siehe Abbildung 6 in Kapitel 4.2.1 und Abbildung 11 im Anhang). Zudem unterscheidet sich die Bewertungspraxis der beiden DO voneinander: Die Evaluierungen der GIZ gehen häufiger auf Synergien ein, die Gutachter der KfW nehmen häufiger auf Konflikte zwischen den Dimensionen Bezug. Dieses Ergebnis scheint auf den ersten Blick plausibel, da z. B. durch den Bau von Anlagen, wie oft im Falle von FZ-Vorhaben, Schäden für Umwelt oder Anrainer-Gemeinschaften entstehen können. Ein solcher Zusammenhang scheint bei TZ-Vorhaben zunächst weniger offensichtlich, da hier vor allem kapazitätsbildende Maßnahmen im Mittelpunkt stehen. Nicht zuletzt aus diesem Grund schreibt die KfW in ihrer Nachhaltigkeitsrichtlinie eine entsprechende Prüfung vor (KfW Entwicklungsbank, 2016).

Jedoch können auch in TZ-Vorhaben Konflikte zwischen Dimensionen auftreten. Beispielsweise kann die Förderung von Entrepreneurship ohne entsprechende politische Begleitung – etwa durch begrenzende Vorgaben, deren Umsetzung über Inspektionen gesichert wird – zu erhöhten Umweltbelastungen führen. Eine umfassendere Berichterstattung kann daher zu einer kohärenteren Vorhabenplanung und -umsetzung beitragen.

Wenn Synergien zwischen den Dimensionen durch die Vorhaben gefördert und anschließend auch durch die Evaluierung bestätigt wurden, gingen diese in fast allen Fällen auch als Erfolgsfaktoren in die Nachhaltigkeitsbewertung ein (siehe Abbildung 7 und Abbildung 8 in Kapitel 4.2.1). Insgesamt wird dieses Kriterium jedoch in der Bewertungspraxis kaum untersucht. Diese fehlende Berücksichtigung des Zusammenspiels der Dimensionen bildet neben der unzureichenden Untersuchung nicht intendierter Wirkungen das zweite zentrale Defizit der aktuellen Evaluierungs- und Bewertungspraxis mit Blick auf die Anforderungen, die sich aus der Agenda 2030 ergeben.

4.3 Zusammenhang zwischen Qualität und der Bewertung von Nachhaltigkeit

Die vorliegende Meta-Evaluierung hat gezeigt, dass die deutsche EZ die Nachhaltigkeit ihrer Vorhaben umfassend evaluiert. Im Hinblick auf den konzeptionellen Rahmen der DAC-Kriterien ist Nachhaltigkeit somit ein umfassendes und übergreifendes Konstrukt, welches weit über das „Evaluierungskriterium Nachhaltigkeit“ hinausgeht. Die Ergebnisse der vorangegangenen Kapitel lassen jedoch vermuten, dass sowohl die Anzahl der eingesetzten Kriterien als auch die Einschätzung, ob sich einzelne Kriterien negativ bzw. positiv auf die Nachhaltigkeit von Vorhaben auswirken, auch von einzelnen Evaluierungstypen abhängen und somit nicht allein durch das zugrunde liegende Nachhaltigkeitsverständnis bestimmt werden. In dem Ergebniskapitel zur Qualität der Evaluierungsberichte wurde gezeigt, dass die Evaluierungstypen unterschiedliche Qualitäten aufweisen. Im Folgenden wird nun der Frage nachgegangen, inwieweit die methodische Qualität der Evaluierungen die Nachhaltigkeitsbewertung tatsächlich beeinflusst.

Im Hinblick auf die Bewertungsgrundlage der Nachhaltigkeit von Vorhaben sind zwei unterschiedliche Zusammenhänge von Relevanz: 1) der Zusammenhang zwischen der Qualität eines Evaluierungsberichtes und der Anzahl Kriterien, die dieser thematisiert, und 2) der Zusammenhang zwischen der Qualität eines Evaluierungsberichtes und der Neigung, Kriterien eher als positiv oder negativ für die Nachhaltigkeit einzuschätzen.

Dabei zeigt sich, dass es einen positiven Zusammenhang zwischen Qualität und Breite der Bewertungsgrundlage gibt (siehe linkes Schaubild in Abbildung 10). Dies bedeutet, dass Evaluierungen mit höherer Qualität tendenziell mehr Nachhaltigkeitskriterien in die Bewertung einbeziehen, die Einschätzung von Nachhaltigkeit also auf eine breitere Bewertungsbasis stellen. Im Hinblick auf den angemessenen Umgang mit einem umfassenden Nachhaltigkeitsverständnis könnte die Einbeziehung vielfältiger Bewertungskriterien der Belastbarkeit der Gesamtbewertung der Nachhaltigkeit durchaus zuträglich sein. Allerdings stellt sich die Frage, ob sich mit der Qualität nicht nur die Belastbarkeit, sondern auch die Bewertung der einzelnen Kriterien und schließlich auch die Nachhaltigkeitsnote verändert. Eine solche Erwartung lässt sich nicht bestätigen (siehe rechtes Schaubild). Die Meta-Evaluierung findet keinen statistischen Zusammenhang zwischen der Qualität einer Evaluierung und der Wirkungsrichtung der Bewertungskriterien. Die methodische Qualität einer Evaluierung hat demnach keinen positiven oder negativen Einfluss auf die Bewertung der Nachhaltigkeit. Zudem hat sich auch in der begleitenden Evaluierungssynthese kein Zusammenhang zwischen der Qualität einer Evaluierung und der Nachhaltigkeitsnote gezeigt (Noltze et al., 2018). Eine Vielfalt an Bewertungskriterien erhöht somit nur die empirische Basis der Untersuchung und steigert das Lernpotenzial, hat aber keinen Einfluss auf die abschließende Bewertung.

4.4 Evaluierung und Bewertung von Nachhaltigkeit im internationalen Vergleich

Als Teil der vorliegenden Meta-Evaluierung widmete sich die Kontextstudie der Frage nach der Angemessenheit der deutschen Evaluierungs- und Bewertungspraxis, wobei eine international vergleichende Perspektive eingenommen wurde.

Abbildung 10: Qualitätsindex nach Anzahl differenzierter Nachhaltigkeitskriterien und nach aggregiertem Einfluss auf die Nachhaltigkeitsbewertung

Quelle: eigene Darstellung

Anmerkungen: Im linken Schaubild ist der Zusammenhang zwischen dem Qualitätsindex eines Berichtes und der Anzahl an differenzierten Kriterien, die zur Bewertung von Nachhaltigkeit herangezogen wurden, dargestellt. Das rechte Schaubild zeigt den Zusammenhang zwischen dem Qualitätsindex eines Berichtes und dem aggregierten Einfluss („Positiv-Negativ-Einschätzung") aller zur Bewertung von Nachhaltigkeit herangezogenen differenzierten Kriterien. Der aggregierte Einfluss ergibt sich aus der Summe aller im Bericht als positiv (+1), neutral (0) oder negativ (-1) eingeschätzten differenzierten Kriterien. N = 513

Dabei wurden 40 Evaluierungseinheiten des OECD-DAC-EvalNets sowie neun multilaterale Organisationen auf ihren Umgang mit Nachhaltigkeit als Evaluierungskriterium hin überprüft (siehe Kapitel 3.4). Neben Informationen auf den offiziellen Webseiten der Einheiten gingen auch online verfügbare Standards, Leitlinien und Orientierungshilfen in die Auswertung ein.

Im Rahmen der Kontextstudie wurde insgesamt eine geringe Transparenz in der Evaluierungs- und Bewertungspraxis von Nachhaltigkeit festgestellt: Lediglich 18 der 40 Evaluierungseinheiten des DAC- EvalNet sowie sechs der neun multilateralen Organisationen stellen auf ihren Webseiten den Umgang mit Nachhaltigkeit in Projekt- und Programmevaluierungen transparent dar. Bei den verbleibenden Organisationen kann lediglich vermutet werden, dass Nachhaltigkeit als eines der fünf DAC-Kriterien für die Erfolgsbewertung von Vorhaben hinzugezogen wird.

Bei den vorhandenen Informationen der verbleibenden 24 Evaluierungseinheiten handelt es sich in erster Linie um die Darstellung des Nachhaltigkeitsverständnisses für Projekt- und Programmevaluierungen. Im Zentrum der Nachhaltigkeitsdefinitionen steht dabei die Dauerhaftigkeit der Wirkungen über die Zeit. Allerdings wird das Nachhaltigkeitsverständnis in der Regel anhand von Dimensionen der Nachhaltigkeit unterschieden. Nach der Häufigkeit der Nennungen ergibt sich dabei über alle 24 Evaluierungseinheiten hinweg die folgende Rangordnung: 1) finanzielle, 2) institutionelle, 3) politische, 4) soziale, 5) technische und 6) ökologische Nachhaltigkeit.

Neben der Definition des Nachhaltigkeitsverständnisses stand die Operationalisierung des Nachhaltigkeitskriteriums für den Zweck von Projektevaluierungen im Zentrum des Interesses der Kontextstudie. Dabei wurde untersucht, inwieweit das Nachhaltigkeitsverständnis von den Evaluierungseinheiten durch überprüfbare Bewertungskriterien konkretisiert wird. Die vergleichende Analyse ergab, dass nur wenige Evaluierungseinheiten die Operationalisierung von Nachhaltigkeit als Evaluierungskriterium transparent darstellen. Wenn dies erfolgt, dann gehört eine Risikobewertung im Umfeld der Maßnahme zu den gängigen Kriterien. Vereinzelt werden darüber hinaus Prinzipien aus der Wirksamkeitsdebatte der EZ, beispielsweise das Prinzip der Eigenverantwortung („Ownership"), als weitere Bewertungskriterien aufgeführt. Eine kondensierte Darstellung der Bewertungskriterien in einer einzelnen Handreichung, wie es der deutschen EZ durch den BMZ-Leitfaden von 2006 zum Umgang mit den DAC-Kriterien gelungen ist, sucht jedoch im internationalen Vergleich ihresgleichen.

Abschließend wurde untersucht, inwieweit die betrachteten Evaluierungseinheiten neben konkreten Bewertungskriterien auch systematische Bewertungsmaßstäbe anlegen, also beispielsweise Punkte- oder Notensysteme. Dabei hat sich gezeigt, dass neben Deutschland nur sehr wenige Länder auch Benotungssysteme vorgeben. Dazu gehören Japan, die Schweiz und Frankreich. Bei den multilateralen Organisationen sind solche Vorgaben gängiger: Neben dem Entwicklungsprogramm der Vereinten Nationen legen die multilaterale Banken – die Europäische Investitionsbank, die Weltbank sowie die Afrikanische und die Asiatische Entwicklungsbank – sowohl Kriterien als auch Punkte- und Notensysteme an.

Die Ergebnisse zeigen: Trotz der grundsätzlich einheitlichen Nutzung des engen Verständnisses von Nachhaltigkeit im Sinne der Dauerhaftigkeit wird Nachhaltigkeit als Evaluierungskriterium im OECD-Vergleich oft deutlich breiter, allerdings konzeptionell auch uneinheitlich, angelegt. Ein möglicher Grund hierfür liegt im Querschnittscharakter des Nachhaltigkeitskriteriums für sämtliche DAC-Kriterien – die Voraussetzungen für Nachhaltigkeit hängen konzeptionell mit denen der anderen DAC-Kriterien zusammen. Die geringe Harmonisierung des Nachhaltigkeitsverständnisses geht dabei Hand in Hand mit einer geringen Standardisierung der Bewertungspraxis. Nur wenige untersuchte Evaluierungseinheiten operationalisieren das Nachhaltigkeitsverständnis auf nachvollziehbare Weise durch konkrete Bewertungskriterien und Benotungssysteme. Aufgrund der geringen Standardisierung finden sich auch keine quantitativen Querschnittsauswertungen zum Thema Nachhaltigkeit in der EZ. Neben einigen wenigen nationalen Evaluierungseinheiten schaffen vor allem die internationalen Entwicklungsbanken die Voraussetzungen für die Aggregation von Wissen durch übergreifende Analysen. Auch in Deutschland sind die Voraussetzungen für quantitative Querschnittsauswertungen gegeben. Allerdings hat die GIZ das vierstufige Notensystem 2014 in eine Punkteskala übertragen, die im Rahmen der Gesamterfolgsbewertung wiederum in eine sechsstufige Notenskala überführt wird, während die KfW Nachhaltigkeit weiterhin anhand von vier Noten bewertet.

5.

SCHLUSSFOLGERUNGEN UND EMPFEHLUNGEN

Die vorliegende Meta-Evaluierung bildet die erste systematische und umfassende empirische Auseinandersetzung mit der Evaluierungs- und Bewertungspraxis zur Nachhaltigkeit von Vorhaben der deutschen EZ. Die zunehmende Bedeutung von Nachhaltigkeit in der internationalen Zusammenarbeit – die in der deutschen EZ bereits ein langjähriges Leitprinzip darstellt – verleiht ihr besondere Relevanz: Mit der Agenda 2030 bildet das Prinzip der Nachhaltigkeit heute mehr denn je den Orientierungsrahmen für die strategische und operative Ausrichtung der EZ.

Zentraler Zweck der Meta-Evaluierung ist es, die EZ bei der Entwicklung einer modernen, nachhaltigkeitsorientierten Evaluierungs- und Bewertungspraxis zu unterstützen. Hierfür wurde methodisches Neuland betreten: Erstmals wurde ein klassisches Meta-Evaluierungsdesign über die reine Qualitätsbewertung hinaus um die systematische Auseinandersetzung mit den inhaltlichen Bewertungskriterien eines Untersuchungsgegenstandes erweitert. Nur so war es möglich, das zugrunde liegende Nachhaltigkeitsverständnis sowie die daraus folgende Evaluierungs- und Bewertungspraxis systematisch zu untersuchen. Die Ergebnisse der Meta-Evaluierung wurden anschließend im Rahmen eines integrierten Forschungsdesigns auch herangezogen, um die Einflussfaktoren der Nachhaltigkeit im Rahmen der begleitenden Evaluierungssynthese zu untersuchen (Noltze et al., 2018).

5.1 Die Qualität der deutschen Evaluierungspraxis

Der erste Teil der vorliegenden Meta-Evaluierung widmet sich der Qualität der Evaluierungspraxis. Neben der Qualität einzelner Evaluierungen galt es dabei, die Strukturen des zugrunde liegenden Evaluierungssystems zu untersuchen. Die nachstehenden Schlussfolgerungen betreffen damit sowohl die Diskussion zur Qualität einzelner Evaluierungen bzw. Evaluierungstypen als auch die Rahmenbedingungen der deutschen EZ-Evaluierungspraxis. Dementsprechend beziehen sich die anknüpfenden Empfehlungen zunächst auf die Weiterentwicklung der Evaluierungspraxis und dann auf die Weiterentwicklung des Evaluierungssystems.

Die Ergebnisse zeigen, dass GIZ und KfW die aus Modulevaluierungen hervorgehenden Ergebnisse und Schlussfolgerungen auf eine dem Umfang dieser Evaluierungen angemessene evaluatorische Grundlage stellen. Neben der Gegenstandsbeschreibung enthält die überwiegende Mehrheit der Evaluierungsberichte eine nachvollziehbare Darstellung der zu überprüfenden Wirkungszusammenhänge und der methodischen Vorgehensweise. Die deutsche EZ zeichnet sich zudem durch einen hohen Deckungsgrad aus: Die GIZ unterzieht nahezu alle Module einer systematischen Erfolgsbewertung. Die KfW arbeitet mit einer repräsentativen Stichprobe; jährlich wird die Hälfte aller evaluierungsreifen Vorhaben pro Sektor einer Ex-post-Evaluierung unterzogen.

Die Meta-Evaluierung hat gezeigt, dass mit Blick auf die Evaluierungsqualität Verbesserungspotenzial besteht. Zum einen sollten mehr Anstrengungen unternommen werden, Ursache-Wirkungs-Beziehungen durch systematische Analyse- und Triangulationsverfahren aufzudecken; zum anderen sollte die Nachvollziehbarkeit von Ergebnissen und Schlussfolgerungen verbessert werden. Eine höhere Belastbarkeit und Nachvollziehbarkeit des Wirkungsnachweises ist ein Schlüssel für einen vertrauenswürdigen Nachhaltigkeitsnachweis nach den Prinzipien der Agenda 2030. Neben den rein methodischen Möglichkeiten liegt weiteres Potenzial in der Wahl des geeigneten Erhebungszeitpunktes: Insbesondere in den vielen dezentralen Evaluierungen der GIZ, die während des Verlaufs der Vorhaben durchgeführt werden, erfolgt der Wirkungsnachweis allein über Zukunftseinschätzungen und geht somit unweigerlich mit Unsicherheit einher. Im Rahmen von Ex-post-Evaluierungen besteht hingegen die Möglichkeit, Wirkungen und die Nachhaltigkeit von Wirkungen in gewissem zeitlichem Abstand zum Ende der Vorhaben tatsächlich zu beobachten.

Bei den Modulevaluierungen der KfW und GIZ handelt es sich in aller Regel um Soll-Ist-Vergleiche entlang ausgewählter Indikatoren der Wirkungslogik. Auch wenn solche Vergleiche die Zuordnungslücke nicht vollständig schließen können, ermöglichen sie doch eine Annäherung an das Aufdecken von Ursache-Wirkungsbeziehungen. Vor diesem Hintergrund ist es erstaunlich, dass nur wenige Evaluierungen nachweislich auf Monitoringdaten der Vorhaben oder der Träger zurückgreifen. Dies steht der Umsetzung von belastbaren Soll-Ist-Vergleichen entgegen.

Es ist nachvollziehbar, dass sich die Evaluierungsqualität in den dezentralen Evaluierungen aufgrund der Überfrachtung der Evaluierungsmissionen nur schwer steigern lässt. Bei der Mehrzahl der dezentralen GIZ-Evaluierungen liegt das Erkenntnisinteresse – auch motiviert durch die Vorbereitung des Antrags für eine mögliche Folgephase – neben Fragen entlang der DAC-Kriterien auf steuerungsrelevanten Aspekten. Dezentrale Evaluierungen haben folglich einen starken Prüfungs- und einen geringen Evaluierungscharakter. Abgesehen von möglichen inhärenten Interessenkonflikten, die sich zwischen Prüfungs- und Evaluierungsabsichten ergeben können, müssen auch die notwendigen Ressourcen berücksichtigt werden – unabhängig davon, ob beide Anliegen getrennt oder gemeinsam durchgeführt werden.

Empfehlungen zur Weiterentwicklung der Evaluierungspraxis:

1. Vor dem Hintergrund zunehmender Anforderungen an die Evaluierung als Instrument für Lernen und Rechenschaftslegung sollten GIZ und KfW Maßnahmen entwickeln, die sicherstellen, dass bestehende Potenziale zur Erhöhung der Evaluierungsqualität, insbesondere im Bereich des Wirkungs- und Nachhaltigkeitsnachweises, ausgeschöpft werden.
2. Aufgrund der anhaltend geringen Bedeutung, die Monitoringdaten in Modulevaluierungen beigemessen wird, sollten die DO systematisch untersuchen, welche Hindernisse hier bestehen und wie diese überwunden werden können. Dabei sollten sie prüfen, inwieweit sich die Monitoringsysteme der Vorhaben über die Zielsysteme der Vorhaben mit dem Zielsystem der nachhaltigen Entwicklungsziele (SDGs) verknüpfen lassen.
3. Im Sinne der Transparenz und als Anreiz für eine nachvollziehbare Berichtslegung sollten GIZ und KfW unter Abwägung der Chancen und Risiken die Möglichkeit prüfen, die Evaluierungsberichte – gegebenenfalls zunächst in einer Pilotphase – vollständig zu veröffentlichen und das BMZ über die Erfahrungen hierzu in Kenntnis setzen.
4. Um die Evaluierungsqualität zu steigern, wird der GIZ empfohlen, die Qualitätssicherung langfristig in der Stabsstelle Evaluierung zu verankern. Alle Modulevaluierungen sollten künftig durch die Stabsstelle gesteuert werden.
5. Für eine höhere Qualität von Evaluierungen sollte in der GIZ eine Trennung zwischen Prüfung und Evaluierung erfolgen.
6. Im Hinblick auf den geeigneten Zeitpunkt für einen aussagekräftigen Wirkungs- und Nachhaltigkeitsnachweis sollte das Format von Ex-post-Evaluierungen in der GIZ erneut an Bedeutung gewinnen. Bei der Durchführung von Ex-post-Evaluierungen sollten sowohl GIZ als auch KfW darauf achten, die Steuerungsrelevanz sicherzustellen. Dies kann beispielsweise durch thematische Fokussierung oder durch die geeignete Wahl des Evaluierungszeitpunktes erfolgen.

Empfehlungen zur Weiterentwicklung des Evaluierungssystems:

7. Im Sinne des gemeinsamen Lernens und der Rechenschaftslegung wird dem BMZ empfohlen, die Evaluierungspraxis von GIZ und KfW auf der Grundlage der Gemeinsamen Verfahrensreform (GVR) und der Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit zu harmonisieren. Dabei sollte das BMZ verbindliche Vorgaben zu Zeitpunkt, Umfang und Benotungssystem schaffen, um die Evaluierungstypen für Modulevaluierungen zu vereinheitlichen.
8. Dem BMZ wird empfohlen, durch die Festlegung einheitlicher Mindeststandards das Ausschöpfen von Potenzialen zur Erhöhung der Evaluierungsqualität in Modulevaluierungen zu unterstützen. Die Anforderungen an eine Evaluierung können beispielsweise durch die Entwicklung von Muster-Terms of Reference konkretisiert werden. Angesetzt werden kann aber schon auf früherer Ebene, etwa über die Anforderungen in den Ausschreibungen zu einer Evaluierungsmission (beispielsweise regelmäßige Fortbildungen der Evaluierenden).
9. Das BMZ sollte die DO dazu anhalten, die Evaluierungsberichte in sich nachvollziehbar zu gestalten, sodass sie für sich stehen können. Je nach Ausgang einer entsprechenden Prüfung sollte das BMZ die DO zu einer vollständigen Veröffentlichung der Evaluierungsberichte anhalten.
10. Das BMZ sollte dafür sorgen, dass neben der Qualitätssicherung der Modulevaluierungen durch die Evaluierungseinheiten von GIZ und KfW regelmäßig eine externe, organisationsübergreifende Meta-Evaluierung zu einer Stichprobe von Evaluierungen stattfindet.

5.2
Die Bewertung von Nachhaltigkeit in der deutschen EZ

Die Ergebnisse der vorliegenden Meta-Evaluierung belegen erstmalig empirisch, dass Nachhaltigkeit in der deutschen EZ-Evaluierungspraxis bereits als ein umfassendes Konzept verstanden, evaluiert und bewertet wird. Im Hinblick auf die breit geführte Diskussion zum Konzept der Nachhaltigkeit in der EZ ist dieses Ergebnis möglicherweise wenig überraschend, mit Blick auf die diesbezüglich deutlich enger gefassten Vorgaben der BMZ-Orientierungshilfe zum Umgang mit den DAC-Kriterien in Evaluierungen allerdings durchaus bemerkenswert: Die Ergebnisse zeigen, dass das Nachhaltigkeitsverständnis in der Evaluierungspraxis bereits deutlich über den Aspekt der Dauerhaftigkeit hinausgeht. Gleichzeitig zeigt sich aber auch, dass wesentliche Elemente der Agenda 2030, etwa die Diskussion um das Zusammenspiel der Dimensionen der Nachhaltigkeit, noch kein systematischer Bestandteil der Bewertungspraxis sind und insofern der Aspekt der nachhaltigen Entwicklung noch nicht vollständig abgedeckt wird. Somit widerlegen die Ergebnisse zwar die in EZ-Kreisen verbreitete Annahme, in den DAC-Kriterien sei ausschließlich ein enges Nachhaltigkeitsverständnis im Sinne der Dauerhaftigkeit angelegt. Sie weisen aber auch auf eine deutliche Diskrepanz zum modernen Verständnis von Nachhaltigkeit nach der Agenda 2030 hin.

Zudem zeigen die Ergebnisse, dass die Evaluierung und Bewertung von Nachhaltigkeit in der Praxis bislang unsystematisch und uneinheitlich erfolgt. Grund dafür ist ein fehlender konzeptioneller Rahmen für ein umfassendes Nachhaltigkeitsverständnis. Die Auswahl der konkreten Bewertungskriterien erfolgt bislang weitestgehend nach Willkür der Gutachter. Selbst die vorgeschlagenen Prüffragen der Orientierungshilfe des BMZ werden nicht systematisch berücksichtigt. In der Gesamtschau zeigt sich, dass die derzeitige Konzeption der DAC-Kriterien eine Evaluierung von Nachhaltigkeit im umfassenden Sinne zwar zulässt, jedoch keinesfalls systematisch und verbindlich vorgibt. Aufgrund der fehlenden Systematik ist ein einfacher Vergleich der Nachhaltigkeitsnote über verschiedene Vorhaben hinweg nur eingeschränkt möglich. Dies steht dem strategischen Lernen aus Evaluierungen entgegen. Ein rigoroser Vergleich zur Beurteilung von Nachhaltigkeit von Vorhaben ist derzeit nur unter erheblichem Aufwand möglich, beispielsweise im Rahmen der vorliegenden erweiterten Meta-Evaluierung und der begleitenden Evaluierungssynthese.

Die Meta-Evaluierung hat gezeigt, dass die Nachhaltigkeit in der Praxis über eine Vielzahl unterschiedlicher Kriterien bewertet wird. Neben dem Kontext von Maßnahmen und den Kapazitäten vor Ort gehen auch Ergebnisse zu den Leistungen und Wirkungen der Vorhaben in die Bewertung ein. Allerdings finden sich auch Kriterien, die entgegen den Vorabannahmen der Meta-Evaluierung vergleichsweise selten in die Bewertung einfließen. Dazu gehören vor allem Kriterien zu nicht intendierten Wirkungen sowie zum Zusammenspiel der Nachhaltigkeitsdimensionen. Letzteres ist überraschend, da das Zusammenspiel der Wirkungen nach Dimensionen der Nachhaltigkeit zwar bislang kein expliziter Bestandteil der Vorgaben darstellt, jedoch bereits seit geraumer Zeit Teil des entwicklungspolitischen Diskurses ist. Demgegenüber sollten nicht intendierte Wirkungen bereits in der Vergangenheit einen wichtigen Teil der Impact-Bewertung ausmachen. Im Hinblick auf die Agenda 2030, die das Zusammenspiel der Dimensionen prominent hervorhebt, besteht hier Potenzial zur Weiterentwicklung der Evaluierungs- und Bewertungspraxis. Ein Grund dafür, dass sowohl nicht intendierte Wirkungen als auch das Zusammenspiel der Dimensionen bisher kaum berücksichtigt werden, sind fehlende methodische Voraussetzungen. Hier müsste die spätere Evaluierbarkeit bereits bei der Planung der Module angelegt werden. Jedoch werden weder nicht intendierte Wirkungen noch das Zusammenspiel der Dimensionen bisher in der Formulierung der Wirkungslogiken systematisch berücksichtigt. In der Folge lassen sich diese Wirkungen auch in den Evaluierungen später kaum oder nur durch erheblichen Aufwand systematisch bearbeiten. Letztlich stellt sich aber auch die Frage nach der Angemessenheit der Betrachtungsebene. Vor dem Hintergrund zunehmend komplexer werdender EZ-Programme und der dazugehörigen TZ- und FZ-Module lassen sich viele Wechselwirkungen und nicht intendierte Wirkungen, insbesondere auf Impact-Ebene, erst auf der Ebene der Programme abschließend feststellen. Auf der Ebene einzelner Module können Evaluierungen diesbezüglich nur ein unvollständiges Bild liefern. Die Diskussion zum Zusammenspiel der Dimensionen steht letztlich stellvertretend für eine notwendige Diskussion um eine Vielzahl evaluatorischer Herausforderungen rund um die Prinzipien der Agenda 2030.

Die Ergebnisse der Meta-Evaluierung haben darüber hinaus auch einen interessanten Zusammenhang zwischen evaluatorischer Qualität und inhaltlichem Erkenntnisgewinn aufgedeckt: Mit zunehmender methodischer Qualität der Evaluierungen werden mehr Kriterien für die Bewertung hinzugezogen. Insofern stellen anspruchsvollere Evaluierungen die Bewertung von Nachhaltigkeit auf eine breitere Basis und sorgen somit auch für eine höhere Belastbarkeit der Aussagen. Ein Zusammenhang zwischen der Qualität und der Einzelbewertung eines Kriteriums oder der Gesamtbewertung der Nachhaltigkeit lässt sich aber nicht feststellen.

Die Auswertung der Bewertungskriterien hat zudem gezeigt, dass Kriterien im Bereich der Leistungen und Wirkungen der Vorhaben überwiegend als förderlich, Kriterien aus den Bereichen der Kapazitäten vor Ort und dem Kontext der Maßnahmen hingegen überwiegend als hemmend für die Nachhaltigkeit bewertet werden. Einerseits betont dieses Ergebnis die herausfordernden Rahmenbedingungen, in denen die deutsche EZ arbeitet. Andererseits birgt dieses Ergebnis auch das Risiko der Externalisierung von Verantwortung: Laut den Evaluierungsberichten liegen die Gründe für eine geringe Nachhaltigkeit vor allem außerhalb des Einflussbereichs der Vorhaben. Allerdings sollten schwierige Rahmenbedingungen nach Möglichkeit a priori bekannt sein und sich somit später auch nicht einseitig auf die Bewertung der Nachhaltigkeit auswirken. Diesbezüglich stellt sich die Frage, inwieweit potenzielle externe Risiken für die Nachhaltigkeit deutscher EZ-Vorhaben durch verbesserte Ex-ante-Prüfung und Planung weiter minimiert werden können.

Diese Schlussfolgerungen belegen letztlich den Mehrwert der vorliegenden erweiterten Meta-Evaluierung, in der die Bewertung der Qualität von Evaluierungen um eine Diskussion zu möglichen Nachhaltigkeitskriterien ergänzt wurde. Durch die Konzeption konnten strukturelle Erfolgsfaktoren und Hindernisse für die Bewertungspraxis von Nachhaltigkeit herausgestellt werden. Die übergreifende Betrachtung des Evaluierungsgegenstandes hat es zudem ermöglicht, Wissen auf globaler Ebene zu aggregieren. Durch die thematische Meta-Evaluierung konnte auch die Evaluierungssynthese auf eine breitere Datenbasis gestellt werden (Noltze et al., 2018).

Der zukünftige Umgang mit der Agenda 2030 und der Nachhaltigkeit von EZ-Vorhaben in Evaluierungen ist eine globale Aufgabe. Mit Blick auf die deutsche EZ hat die vorliegende Meta-Evaluierung konkreten Handlungsbedarf identifiziert. Die vorgestellten Schlussfolgerungen rufen nach einer Reform der bisherigen Evaluierungs- und Bewertungspraxis. Neben dem Harmonisierungs- und Koordinierungsgedanken der Erklärung von Paris zur Effektivität der EZ und dem Aktionsplan von Accra verlangt der universelle Charakter der Agenda 2030 dabei auch nach Austausch und Abstimmung auf internationaler Ebene (OECD, 2008; UN, 2015). Die nun folgenden Empfehlungen zielen somit darauf ab, die laufenden Reformprozesse auf Ebene der deutschen EZ, insbesondere im Rahmen der Gemeinsamen Verfahrensreform (GVR)[20], zu unterstützen. Darüber hinaus sollen sie die Diskussion auf internationaler Ebene, insbesondere im Rahmen des OECD-DAC, bereichern. Vor dem Hintergrund der laufenden Reformprozesse werden die Empfehlungen durch eine Reihe konzeptioneller Vorschläge ergänzt, die als Denkanstöße zu verstehen sind – in dem Bewusstsein, sich damit an ein System zu richten, dem auch das DEval angehört.

Entsprechend der differenzierten Darstellung der Empfehlungen zur Qualität der Evaluierungs- und Bewertungspraxis (Kapitel 5.1) werden im Folgenden auch die Empfehlungen zur Bewertung von Nachhaltigkeit in zwei Bereiche unterteilt: Empfehlungen zur Weiterentwicklung der Evaluierungspraxis und solche zur Weiterentwicklung des Evaluierungssystems.

Empfehlungen zur Weiterentwicklung der Evaluierungspraxis:

Die folgenden Empfehlungen richten sich sowohl an das BMZ als auch an die Durchführungsorganisationen. Die Umsetzung sollte auf der Grundlage eines gemeinsamen Prozesses unter Federführung des BMZ und unter Beteiligung der DO und des DEval erfolgen. Es wird empfohlen, diesen Prozess inklusive einer Pilotierungsphase bis Ende 2018 abzuschließen, um eine Agenda-2030-konforme Evaluierungs- und Bewertungspraxis der deutschen EZ ab 2019 zu gewährleisten.

[20] *Die im Juni 2017 in Kraft getretene Gemeinsame Verfahrensreform (GVR) des BMZ bildet die Grundlage für die zukünftige Gestaltung, Umsetzung und Evaluierung von Länderstrategien, EZ-Programmen und -Modulen mit dem Ziel, die Wirksamkeit der EZ zu verbessern. An verschiedenen Stellen nimmt die GVR dabei Bezug auf die Prinzipien und Entwicklungsziele der Agenda 2030. Auf der Basis der GVR arbeiten die DO an organisationsspezifischen Leitlinien für die Entwicklung Agenda-2030-konformer Vorhaben und der begleitenden Prüf- und Evaluierungssysteme. In der GIZ erfolgt die Reformierung des organisationsinternen Evaluierungssystems im Rahmen einer 2017 vorgelegten Evaluierungspolitik.*

11. Dem BMZ und den DO wird empfohlen, die Nachhaltigkeit von Vorhaben im Sinne der Prinzipien der Agenda 2030 für nachhaltige Entwicklung zukünftig im Rahmen eines zusätzlichen Bewertungskriteriums zu evaluieren.
    - Solch ein zusätzliches Erfolgskriterium könnte konzeptionell so angelegt werden, dass es die fünf Erfolgskriterien nach OECD-DAC um eine Bewertung des Beitrags zur nachhaltigen Entwicklung nach dem Verständnis der Agenda 2030 ergänzt. Die Operationalisierung des zusätzlichen Erfolgskriteriums durch geeignete Prüffragen könnte gemäß der Struktur der Agenda-2030-Prinzipien[21] erfolgen. Der Mehrwert des zusätzlichen Kriteriums läge im Sinne des Lernens und der Rechenschaftslegung darin, die spezifischen Beiträge deutscher Entwicklungsmaßnahmen zur Umsetzung der Agenda 2030 kondensiert darzustellen. Ein solches zusätzliches Kriterium böte zudem die Grundlage für eine zukünftige aggregierte Wirkungsberichterstattung im Hinblick auf die Agenda 2030. Gleichzeitig könnten die Systematik der DAC-Kriterien und damit die Vergleichbarkeit zu früheren Erfolgsbewertungen sowie die internationale Harmonisierung beibehalten werden.
    - Alternativ könnten Prüffragen nach den Prinzipien der Agenda 2030 in die DAC-Kriterien integriert werden. Dies hätte den Vorteil, dass die DAC-Kriterien weiterhin die alleinige Grundlage der Erfolgsbewertung bilden. Allerdings würden sie inhaltlich verändert werden und wären damit historisch wie international nicht mehr vergleichbar. Zudem kann erwartet werden, dass die Einbeziehung einer aktuellen Entwicklungsagenda den zeitlosen Prüfcharakter der DAC-Kriterien ein Stück weit auflösen würde.
    - Unabhängig davon, ob der Umgang mit der Agenda 2030 in Evaluierungen zukünftig als gesondertes Kriterium oder integriert in die DAC-Kriterien erfolgen sollte, wäre eine Abstimmung auf internationaler Ebene im Rahmen des OECD-DAC sinnvoll.
12. Einhergehend mit dem Einbeziehen von Nachhaltigkeit im Sinne der Agenda 2030 als zusätzliches Bewertungskriterium wird dem BMZ die konzeptionelle Schärfung der DAC-Kriterien und eine höhere Verbindlichkeit der BMZ-Orientierungshilfe zum Umgang mit den DAC-Kriterien empfohlen.
    - Dabei könnten, unter Berücksichtigung eines geeigneten Maßes an fallspezifischer Offenheit und übergeordneter Verbindlichkeit, die bereits vorhandenen Prüffragen auf ihren genuinen Charakter und die Trennschärfe zu den Prüffragen der anderen DAC-Kriterien hin überprüft und gegebenenfalls konkretisiert werden.
    - Die Orientierungshilfe könnte zudem durch eine Evaluierungsmatrix übersichtlich dargestellt werden. Soweit möglich, könnte diese auch Vorschläge zum Gewichten der einzelnen Prüffragen sowie Definitionen für eine intersubjektiv vergleichbare Notengebung enthalten.
13. Im Rahmen der Reform der Evaluierungskriterien für die Erfolgsbewertung von EZ-Vorhaben wird dem BMZ empfohlen, das bisherige Evaluierungskriterium der Nachhaltigkeit nach OECD-DAC im Sinne der Dauerhaftigkeit von Wirkungen zu erhalten und die entsprechenden Prüffragen auf diesen Aspekt auszurichten.
    - Um die konzeptionelle Abgrenzung zwischen dem Aspekt der Dauerhaftigkeit und dem Aspekt der nachhaltigen Entwicklung nach der Agenda 2030 zu erreichen, sollte auch über eine sprachliche Differenzierung nachgedacht werden. Im deutschen Sprachgebrauch wäre eine Trennung zwischen den Begriffen „Dauerhaftigkeit“ und „Nachhaltigkeit“ eine mögliche Option. Hierbei sollte auf die internationale Anschlussfähigkeit geachtet werden.
14. Mit Blick auf die Prinzipien der Agenda 2030 sollten GIZ und KfW untersuchen, wie in Evaluierungen künftig die nicht intendierten Wirkungen eines Vorhabens und die Wechselwirkungen zwischen den Dimensionen der Nachhaltigkeit identifiziert und geprüft werden können.
    - Dabei könnten erwartete und tatsächliche Synergien und Konflikte zwischen entwicklungspolitischen Zielen dargestellt und überprüft werden. Die Verantwortung würde bei der Planung von Vorhaben beginnen: Bereits in den Modulvorschlägen könnten nicht intendierte Wirkungen und Wechselwirkungen als elementarer Bestandteil integrierter Ansätze diskutiert werden. Eine solche Verankerung würde einer entsprechenden Vorgabe des BMZ bedürfen.
    - Die nicht intendierten Wirkungen sowie die möglichen Potenziale und Risiken der Wechselwirkungen zwischen den Dimensionen sollten dabei nach Möglichkeit in

[21] *Zu den Prinzipien gehören: gemeinsame Verantwortung; Zusammenspiel der Dimensionen; niemanden zurücklassen; Universalität und Rechenschaftspflicht.*

multidisziplinären Teams unter Einbeziehung verschiedener sektoraler Perspektiven herausgearbeitet werden. Kommt es hierbei zu widersprüchlichen Erwartungen, könnten diese dokumentiert werden, um die Theoriebildung langfristig auf nachvollziehbare Weise zu verbessern. Die Herausstellung solcher (möglichen) Wirkungen würde die Evaluierbarkeit erleichtern und ist somit auch der Effizienz einer Evaluierung zuträglich.

- Die Suche nach möglichen nicht intendierten Wirkungen könnte durch die Nutzung vorhandener Rahmenwerke unterstützt werden, etwa die Standards zur Umwelt- und Sozialverträglichkeitsprüfung, die IFC Performance Standards oder die Environmental and Social Safeguards der Weltbank sowie die Environmental, Health and Safety Guidelines und die Kernarbeitsnormen der Internationalen Arbeitsorganisation.
- Zudem sollte auch der Umgang mit den weiteren Prinzipien der Agenda 2030, z. B. dem Mandat, „niemanden zurückzulassen", geklärt werden.

Empfehlungen zur Weiterentwicklung des Evaluierungssystems:

15. Dem BMZ wird empfohlen, eine übergeordnete Evaluierungsstrategie zu entwickeln, die sich über die Zeit thematische Schwerpunkte setzt.
    - Eine übergeordnete Evaluierungsstrategie des BMZ könnte durch die DO in strategische Evaluierungsprogramme überführt und zusätzlich durch thematische Querschnittsauswertungen durch das DEval begleitet werden.
    - Die inhaltliche Ausgestaltung der Evaluierungsstrategien und -programme bis 2030 könnte sich dabei sowohl an den Prinzipien der Agenda 2030 als auch an dem begleitenden Zielsystem der SDGs orientieren. Mit der Evaluierungsstrategie könnten auch die Angemessenheit des Deckungsgrades von Evaluierungen und das Erstellen geeigneter Stichprobenpläne geprüft werden.
16. In der Evaluierungsstrategie sollte das BMZ definieren, welche Anforderungen sich aus den Fragestellungen um die Agenda 2030 für die jeweiligen Evaluierungen – also auf Ebene der Module, der Programme und der Länderstrategien – ergeben.
    - Die abschließende Bewertung der Beiträge der deutschen EZ zu den SDGs und den Prinzipien der Agenda 2030 könnte zukünftig vor allem auf der Programm-Ebene erfolgen. Dabei ist festzuhalten, dass sich auch in Evaluierungen auf Modul-Ebene weiterhin vielfältige Möglichkeiten bieten, den Beitrag zum Leitbild einer nachhaltigen Entwicklung nach der Agenda 2030 zu bewerten.
    - Da auf Programm-Ebene oftmals mehrere Akteure involviert sind, könnten die einzelnen Beiträge in Form von Gemeinschaftsevaluierungen der DO erfasst und zusammengeführt werden. Dabei sollte auch geklärt werden, wie entsprechend mit der Evaluierung von Sektor- und Globalvorhaben umgegangen werden kann.
    - Die Auswahl von Programm- und Modulevaluierungen im Rahmen der Evaluierungsstrategie könnte einem zweistufigen Selektionsprozess folgen: In einem ersten Schritt könnten die Programme, die für eine Evaluierung in Frage kommen, ausgewählt werden. Dabei könnten auch politische Fristen (Zeitpunkt von Regierungsverhandlungen, Bericht eines Partnerlandes vor den Vereinten Nationen etc.) in die Entscheidungen einbezogen werden. In einem zweiten Schritt würde die Auswahl der Modulevaluierungen erfolgen.

6.

LITERATUR

Ashoff, G. (2015), „Die Global Governance-Qualität der internationalen Aid Effectiveness Agenda: eine theoretische Analyse und Bewertung der Systemreform der internationalen Entwicklungszusammenarbeit", Deutsches Institut für Entwicklungspolitik, 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 (2008), *„Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit"*, Nr. 165, BMZ Konzepte, Bonn/Berlin.

Boone, P. (1996), *„Politics and the effectiveness of foreign aid"*, NBER Working Paper Series, Cambridge; Massachusetts, S. 289–329.

Carlsson, J. und L. Wohlgemuth (1996), *„Capacity Building and Networking - A meta-evaluation of African regional research networks"*, Sida Evaluation, Department for Evaluation and Internal Audit, Stockholm.

Caspari, A. (2004), „Evaluation der Nachhaltigkeit von Entwicklungszusammenarbeit. Zur Notwendigkeit angemessener Konzepte und Methoden", Sozialwissenschaftliche Evaluationsforschung, VS Verlag für Sozialwissenschaften, Wiesbaden.

Cutter, A. (2014), *„Sustainable Development Goals (SDG) and Integration: Achieving a better balance between the economic, social and environmental dimensions"*, Rat für Nachhaltige Entwicklung, Berlin.

Dietz, F. und A. Hanemaaijer (2012), „How to select policy-relevant indicators for sustainable development", in von Raggamby, A. und F. Rubik (Hrsg.), *Sustainable development, evaluation and policy-making: theory, practise and quality assurance*, Edward Elgar, Cheltenham, S. 21–35.

Faust, J. (2007), „Assessing Aid: die makroquantitative Forschung zur Effektivität der Entwicklungszusammenarbeit", in Hemmer, H.-R. (Hrsg.), *Zur Wirksamkeitsdebatte in der Entwicklungszusammenarbeit (EZ)*, Erfurt.

Freedom House (2016), *„Freedom in the World"*, New York.

GIZ (2016), *„Meta-Evaluierung der Projektevaluierungen (PEV)"*, Bonn.

Grunwald, A. und J. Kopfmüller (2006), „Nachhaltigkeit", Campus Einführungen, Campus-Verlag, Frankfurt am Main.

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

Hartmuth, G. (2004), *„Nachhaltige Entwicklung im lokalen Kontext - Schritte zur Entwicklung eines kommunalen Nachhaltigkeits-Indikatorensystems"*, Nr. 6, UFZ Diskussionspapiere, Umweltforschungszentrum, Leipzig.

Islam, S.M.N. und M.F. Clarke (2005), „The welfare economics of measuring sustainability: a new approach based on social choice theory and systems analysis", *Sustainable Development*, Vol. 13, Nr. 5, S. 282–296.

KfW Entwicklungsbank (2003), *„FZ-Projekte und Nachhaltigkeit. Zur Berücksichtigung der Nachhaltigkeit durch die KfW in Schlussprüfungen von FZ-Vorhaben: Grundsätzliche Überlegungen"*, Nr. 33, Diskussionsbeiträge, KfW Entwicklungsbank, Frankfurt am Main.

KfW Entwicklungsbank (2016), *„KfW Nachhaltigkeitsrichtlinie"*, KfW Entwicklungsbank, Frankfurt am Main.

Klasen, S. (2015), *„SDG - Den Ärmsten der Welt einen Bärendienst erwiesen"*, Nr. 3, Meinungsforum Entwicklungspolitik, KfW Entwicklungsbank, Frankfurt am Main.

Klingebiel, S. (2013), „Entwicklungszusammenarbeit - eine Einführung", Deutsches Institut für Entwicklungspolitik, Bonn.

König, J. und J. Thema (Hrsg.) (2011), „Nachhaltigkeit in der Entwicklungszusammenarbeit: theoretische Konzepte, strukturelle Herausforderungen und praktische Umsetzung", Globale Gesellschaft und internationale Beziehungen, Verlag für Sozialwissenschaft, Wiesbaden, 1. Auflage.

Landis, J.R. und G.G. Koch (1977), „The Measurement of Observer Agreement for Categorical Data", *Biometrics*, Vol. 33, Nr. 1, S. 159.

Leeuw, F.L. und L.J. Cooksy (2005), „Evaluating the performance of development agencies: The role of metaevaluations.", in Pitman, G.K., O.N. Feinstein und G.K. Ingram (Hrsg.), *Evaluating development effectiveness*, World Bank series on Evaluation and Development, Transaction, New Brunswick, S. 95–108.

Meadows, D.H. et al. (1972), „The limits to growth: a report for the Club of Rome's Project on the Predicament of Mankind", Universe Books, New York.

Noltze, M. et al. (2018), *„Evaluationssynthese von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit"*, DEval, Bonn.

Nuscheler, F. (2007), „Wie geht es weiter mit der Entwicklungspolitik?", *Aus Politik und Zeitgeschichte*, Vol. 48, S. 3–10.

OECD (1991), *„DAC Criteria for Evaluating Development Assistance Factsheet"*, Paris.

OECD (2008), *„The Paris Declaration on Aid Effectiveness and the Accra Agenda for Action"*, Paris.

OECD (2010a), „Evaluation in Development Agencies", Better Aid, OECD Publishing, Paris.

OECD (2010b), „Quality Standards for Development Evaluation", DAC Guidelines and Reference Series, OECD Publishing, Paris.

OECD (2016a), „Better Policies for Sustainable Development 2016. A New Framework for Policy Coherence", OECD Publishing, Paris.

OECD (2016b), „Evaluation Systems in Development Cooperation: 2016 Review", OECD Publishing, Paris.

OECD (2017), „DAC Criteria for Evaluating Development Assistance", *Organisation for Economic Co-operation and Development, http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm*, zugegriffen 23.03.2017.

Patton, M.Q. (2008), „Utilization-focused evaluation", SAGE Publications, Thousand Oks, Calif., 4. Aufl.

Preiß, J. (2017), *„Evaluierung von Nachhaltigkeit und ihre Determinanten in der Entwicklungszusammenarbeit. Eine empirische Analyse anhand von Projekten der Weltbank"*, unveröffentlichte Masterarbeit, Freie Universität Berlin, Berlin.

Raetzell, L. und M. Krämer (2013), *„Meta-Evaluation Gesundheit"*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), Bonn.

Russ-Eft, D.F. (2014), „Human resource development, evaluation, and sustainability: what are the relationships?", *Human Resource Development International*, Vol. 17, Nr. 5, S. 545–559.

Scriven, M. (1991), „Evaluation Thesaurus", Sage Publications, Newbury Park; London; New Delhi, 4. Aufl.

Scriven, M. (2009), „Meta-Evaluation revisited", *Journal of MultiDisciplinary Evaluation*, Vol. 6, Nr. 11, S. iii–viii.

Stadtler, L. (2016), „Scrutinizing Public–Private Partnerships for Development: Towards a Broad Evaluation Conception", *Journal of Business Ethics*, Vol. 135, Nr. 1, S. 71–86.

Stockmann, R. und W. Gaebe (Hrsg.) (1993), „Hilft die Entwicklungshilfe langfristig? Bestandsaufnahme zur Nachhaltigkeit von Entwicklungsprojekten", Westdeutscher Verlag GmbH, Opladen.

Stockmann, R. und S. Silvestrini (2011), *„Synthese und Metaevaluierung Berufliche Bildung"*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

Stockmann, R. und S. Silvestrini (2012), „Ergebnispräsentation: Synthese und Meta-Evaluierung Berufliche Bildung", gehalten auf der GIZ Dialogtag, Bonn.

Stufflebeam, D.L. (2001), „The Metaevaluation Imperative", *American Journal of Evaluation*, Vol. 22, Nr. 2, S. 183–209.

van Tulder, R. und S. Pfisterer (2008), *„From Idea to Partnership: Reviewing the Effectiveness of Development Partnerships in Zambia, Columbia and Ghana"*, Expert Centre for Sustainable Business & Development Cooperation, Maastricht.

UN (2015), „Transforming our world. The 2030 Agenda for Sustainable Development", New York.

Vahlhaus, M. (2014), *„Der Weg: Scaling-Up. Das Ziel: Breitenwirksamkeit"*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn, Eschborn.

Von Raggamby, A. und F. Rubik (Hrsg.) (2012), „Sustainable development, evaluation and policy-making: theory, practise and quality assurance", Evaluating sustainable development, Edward Elgar, Cheltenham.

Widmer, T. (2006), „Meta-Evaluation. Kriterien zur Bewertung von Evaluationen", Verlag Paul Haupt, Zürich.

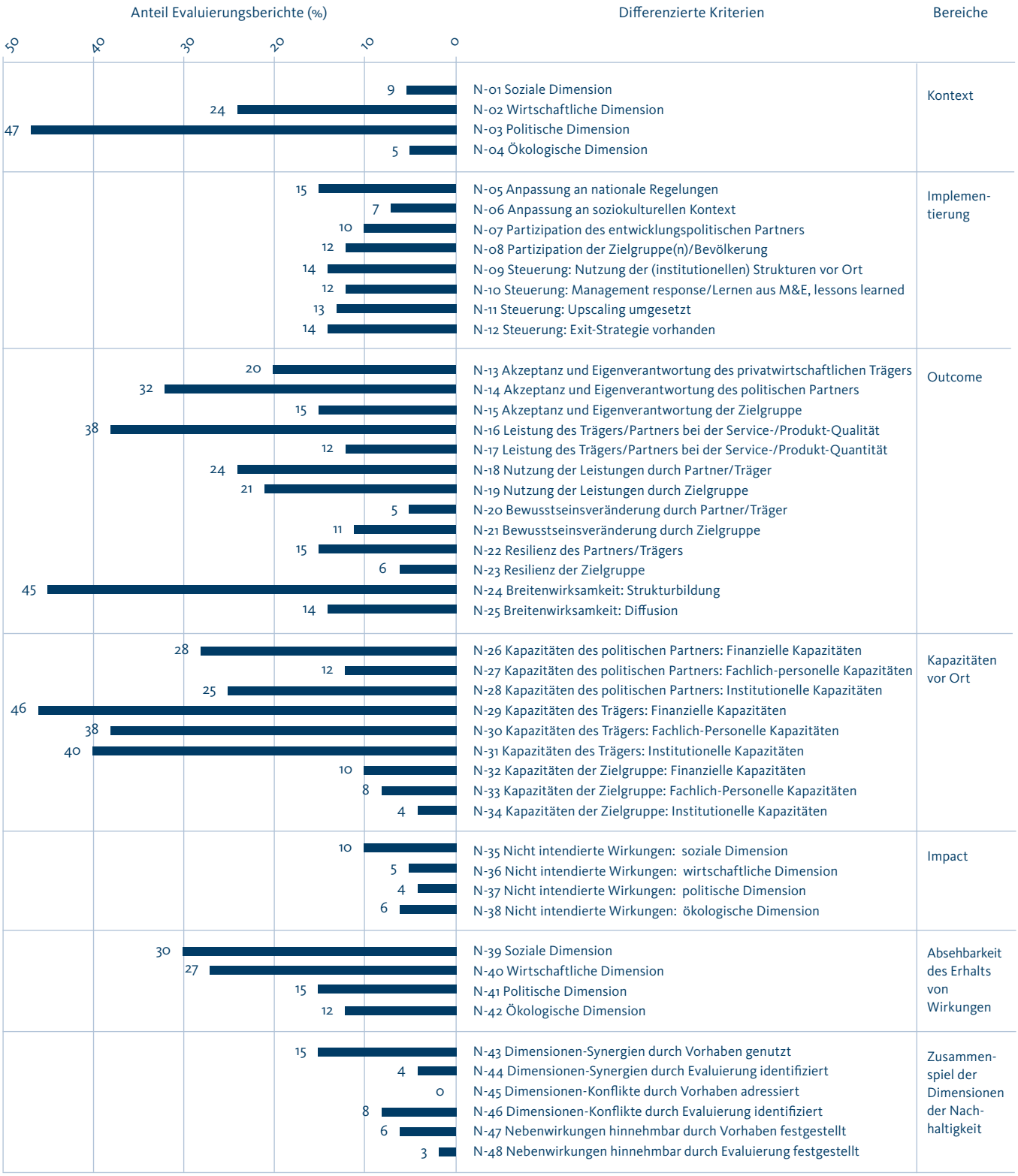
World Commission on Environment and Development (1987), „Our Common Future", Oxford University Press, Oxford.

7.

ANHANG

7.1 Abbildungen

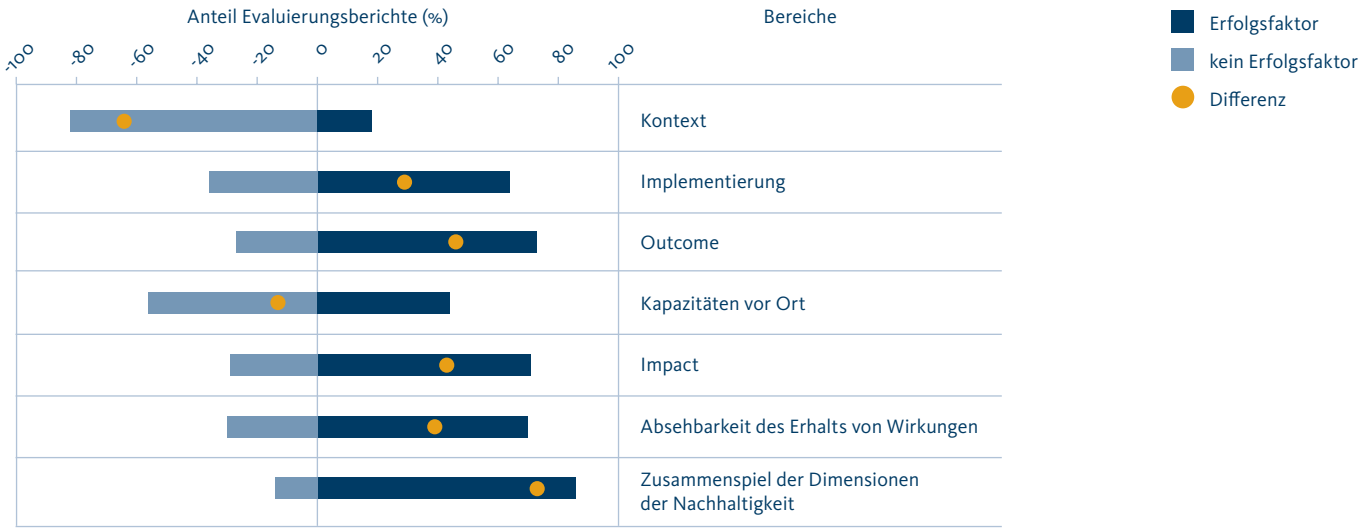
Abbildung 11: Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil an Evaluierungsberichten, der zu dem jeweiligen differenzierten Kriterium bei der Bewertung von Nachhaltigkeit Bezug nimmt. N = 513

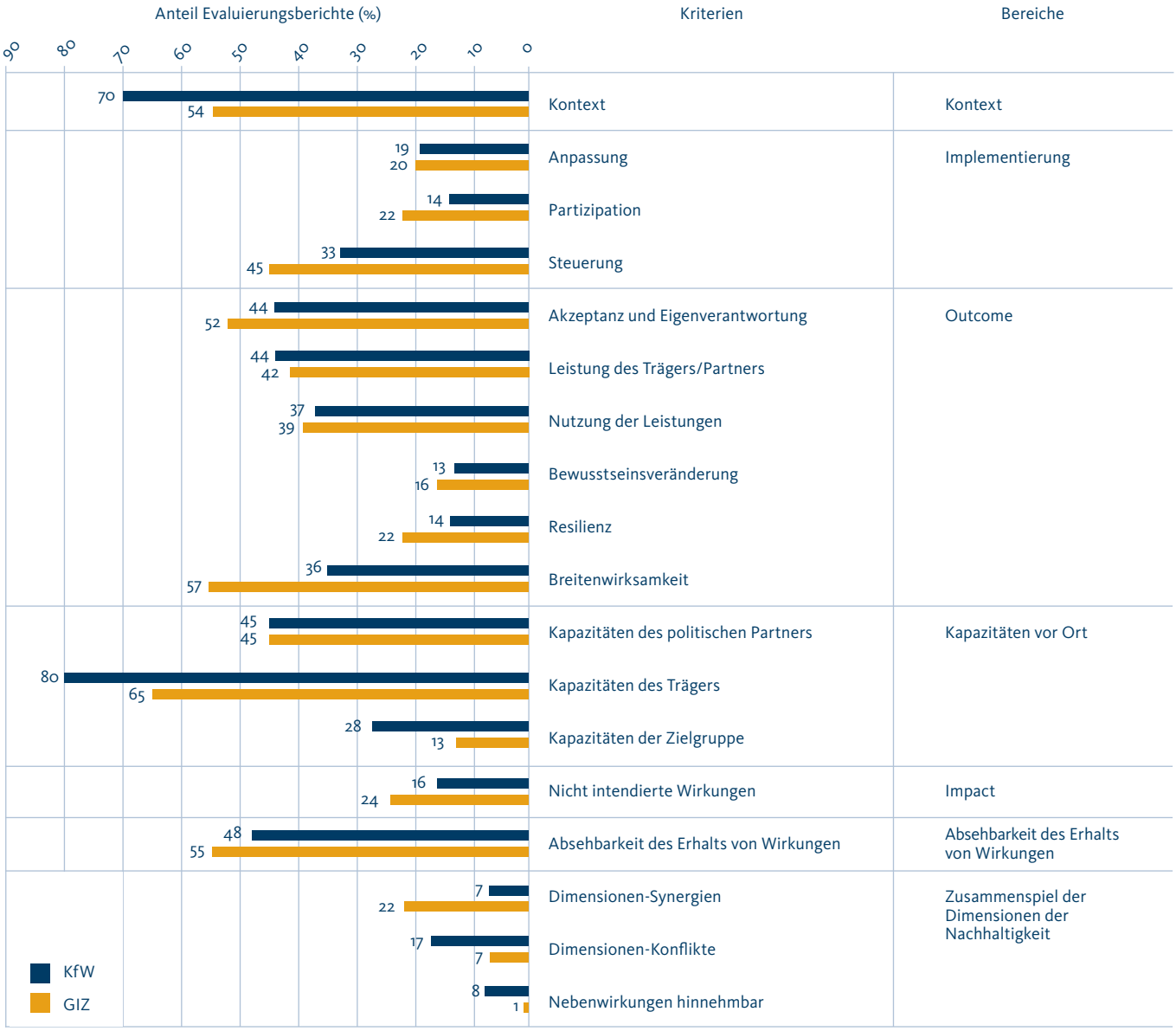
Abbildung 12: Relativer Anteil Evaluierungsberichte nach Nachhaltigkeitsbereich und Einfluss auf Nachhaltigkeitsbewertung



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Bereich entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Einzelne Bereiche enthalten nur die Berichte, die zu mindestens einem differenzierten Kriterium des jeweiligen Bereiches bei der Bewertung von Nachhaltigkeit Bezug nehmen. Die Punkte stellen die Differenz zwischen dem Anteil der positiven und negativen Bewertungen eines Bereiches dar. N = 513

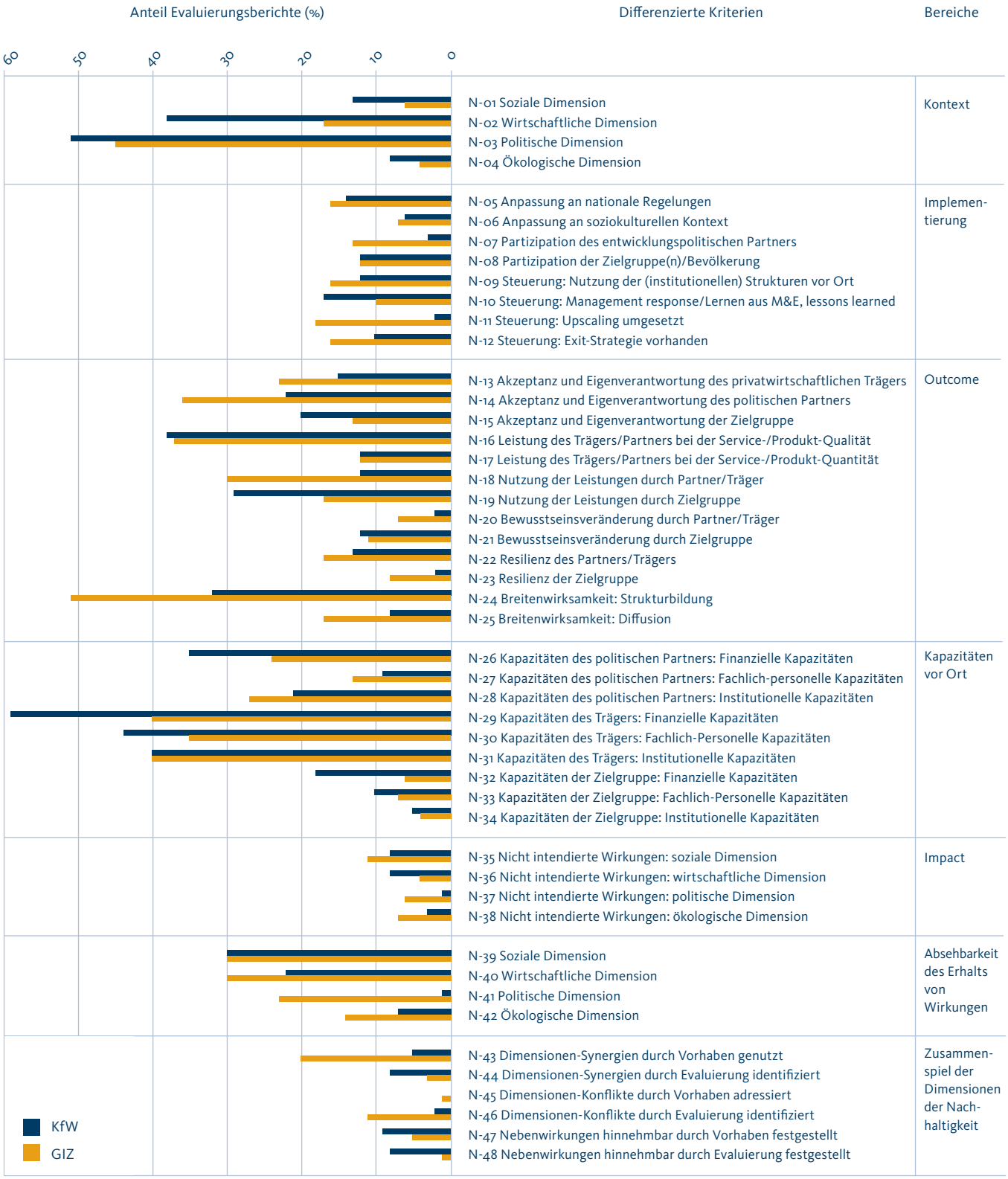
Abbildung 13: Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte, die bei der Bewertung von Nachhaltigkeit zu mindestens einem differenzierten Kriterium aus dem jeweiligen Nachhaltigkeitskriterium Bezug nehmen. Die Evaluierungsberichte sind unterteilt nach KfW (n = 172) und GIZ (n = 341). N = 513

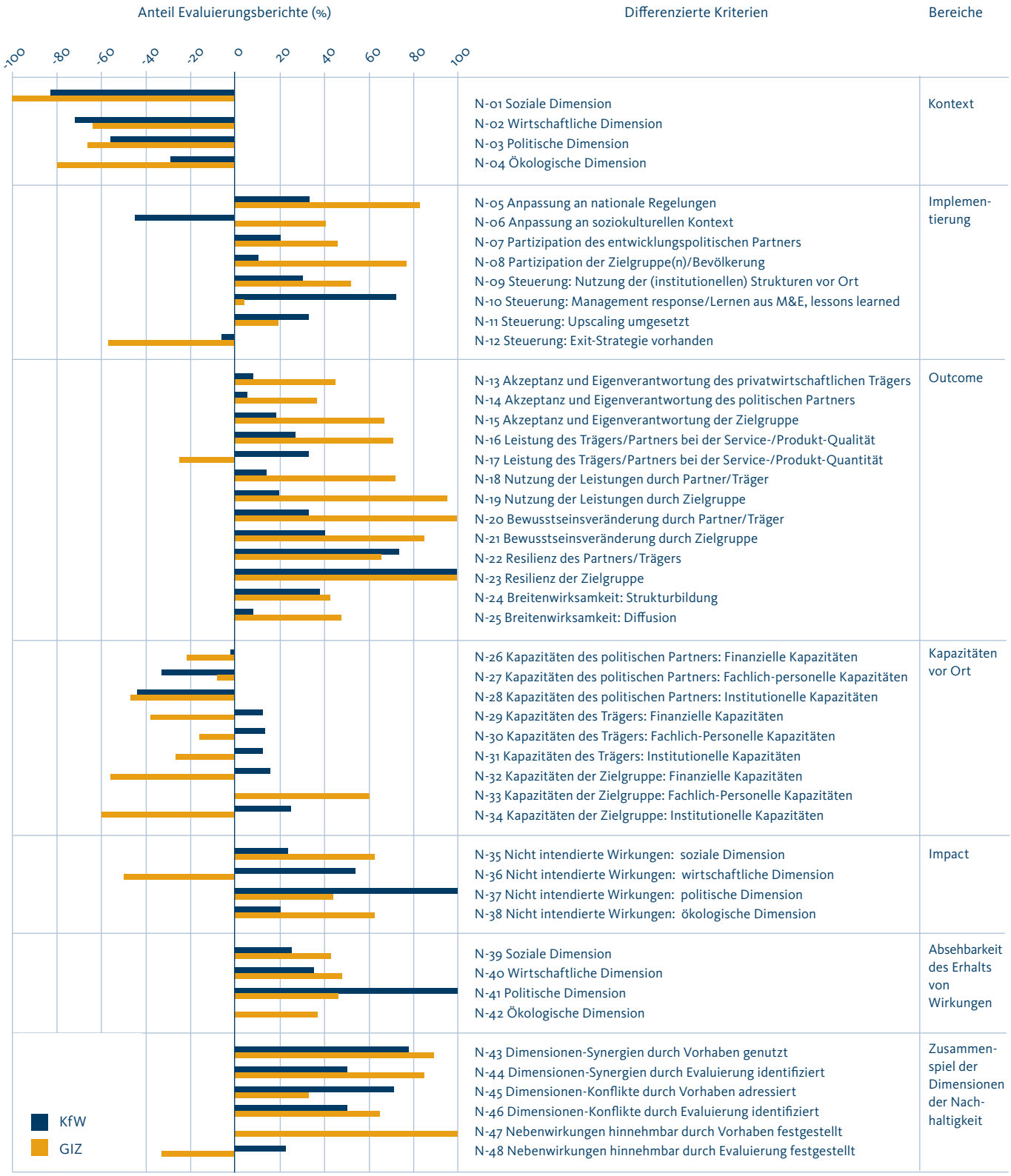
Abbildung 14: Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte, die bei der Bewertung von Nachhaltigkeit zu den jeweiligen differenzierten Nachhaltigkeitskriterien Bezug nehmen. Die Evaluierungsberichte sind unterteilt nach KfW (blau, n = 172) und GIZ (orange, n = 341). N = 513

Abbildung 15: Relativer Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisationen



Quelle: eigene Darstellung

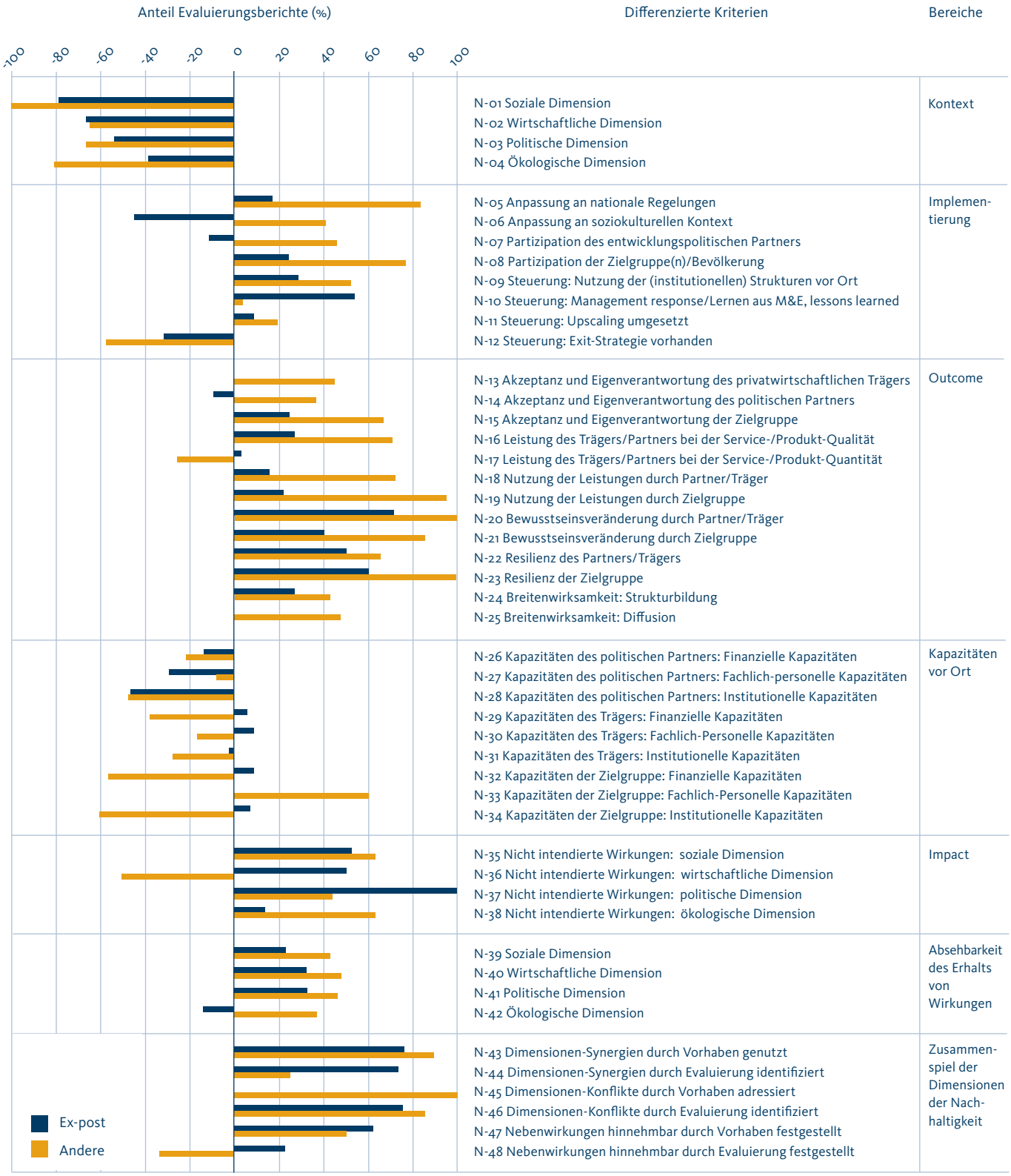
Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen differenzierten Kriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach KfW (n = 172) und GIZ (n = 341). N = 513

Abbildung 16: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation

Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Kriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach KfW (n = 172) und GIZ (n = 341). N = 513

Abbildung 17: Relativer Anteil Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Evaluierungstyp



Quelle: eigene Darstellung

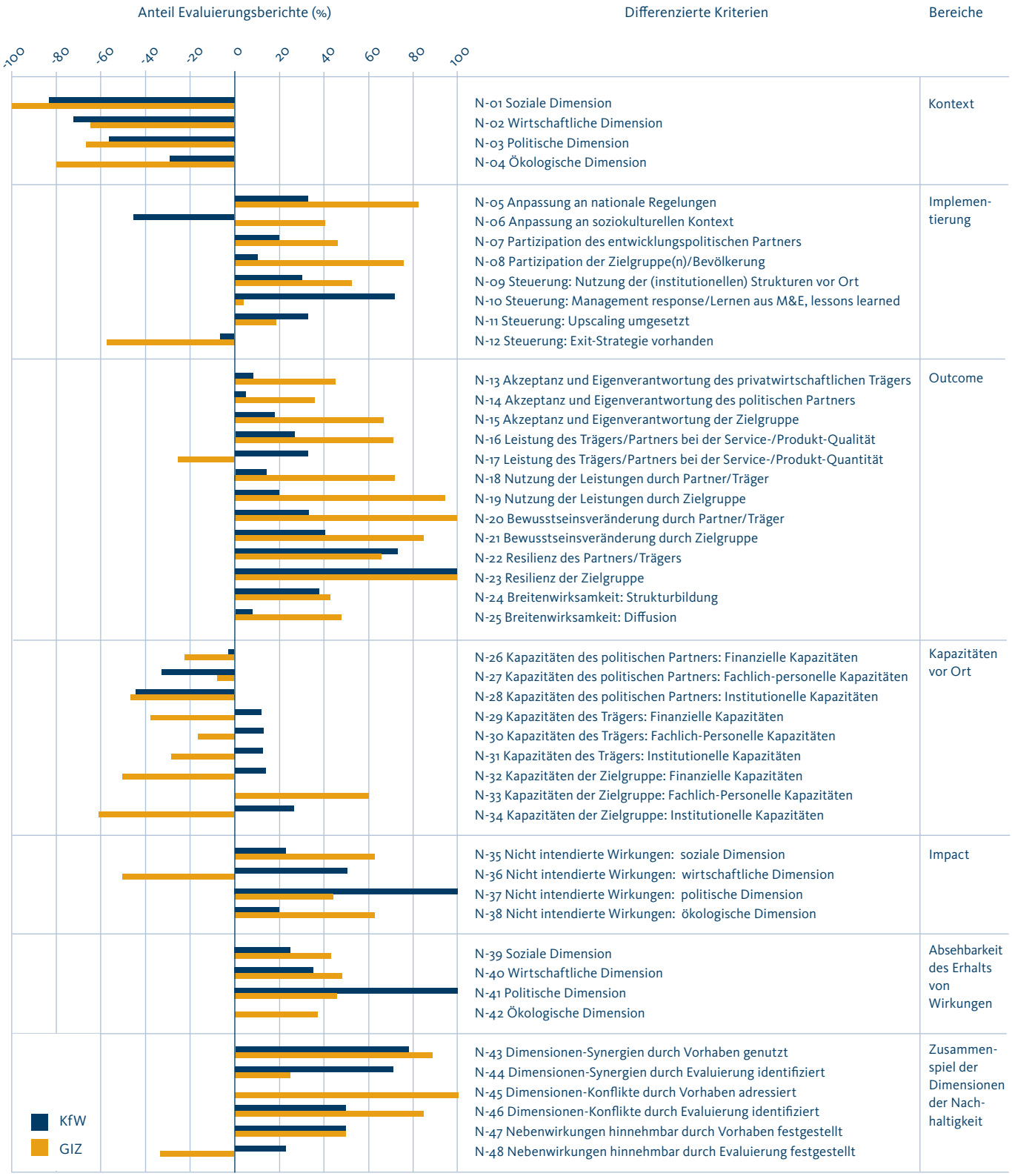
Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen differenzierten Kriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach Ex-post-Evaluierungen (blau, n = 219,) und PFK, PEV und Schluss-Evaluierungen (orange, n = 294). N = 513

Abbildung 18: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Evaluierungstyp

Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach Ex-post-Evaluierungen (blau, n = 219) und PFK, PEV und Schluss-Evaluierungen (orange, n = 294). N = 513

Abbildung 19: Relativer Anteil Ex-post-Evaluierungsberichte mit Bezug zu differenzierten Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation



Quelle: eigene Darstellung

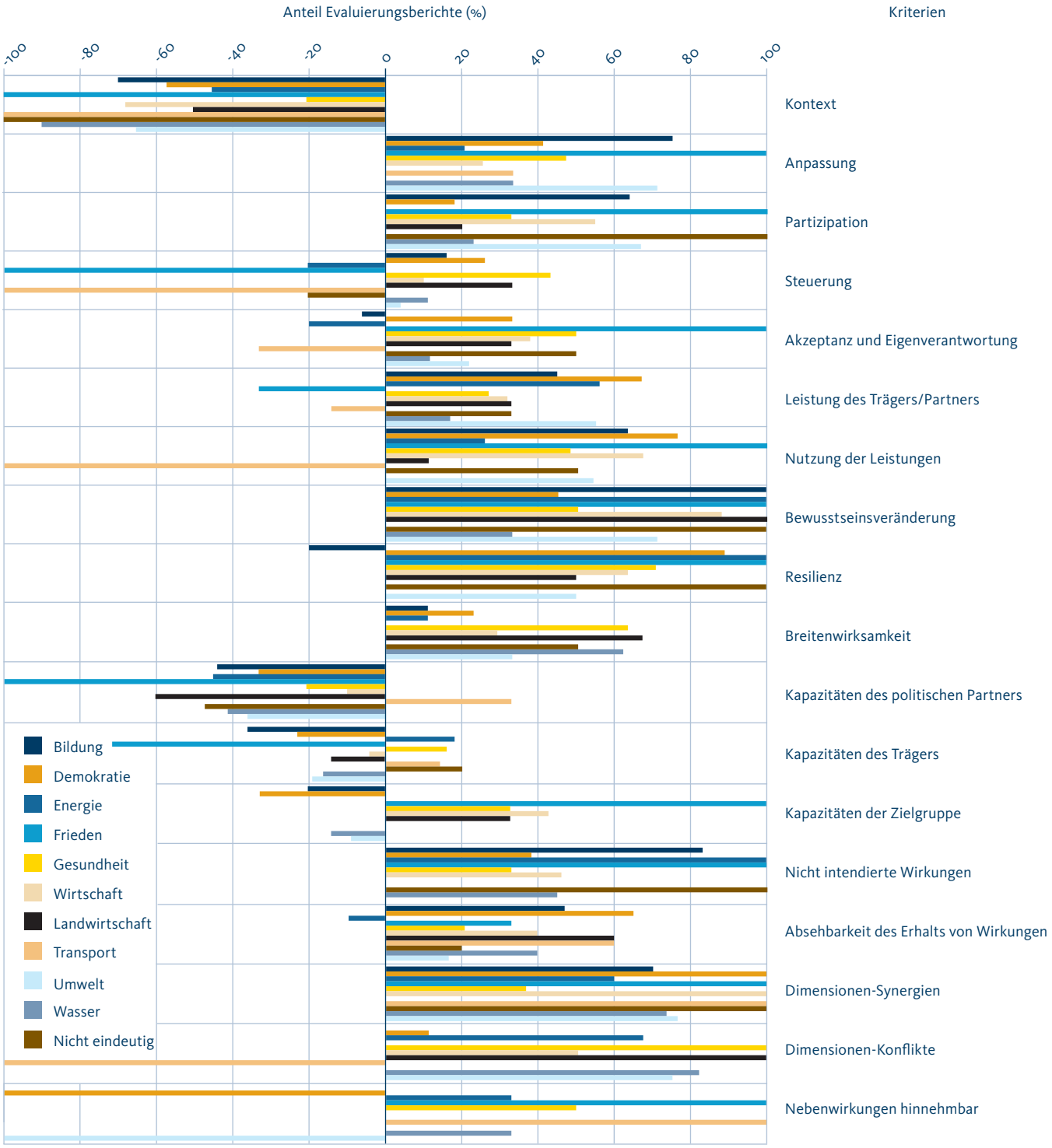
Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen differenzierten Kriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Abbildung enthält ausschließlich Ex-post-Evaluierungen. Diese sind unterteilt nach KfW (blau, n = 172) und GIZ (orange, n = 38). N = 210

Abbildung 20: Relativer Anteil Ex-post-Evaluierungen mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Durchführungsorganisation

Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Abbildung enthält ausschließlich Ex-post-Evaluierungen. Diese sind unterteilt nach KfW (blau, n = 172) und GIZ (orange, n = 38). N = 210

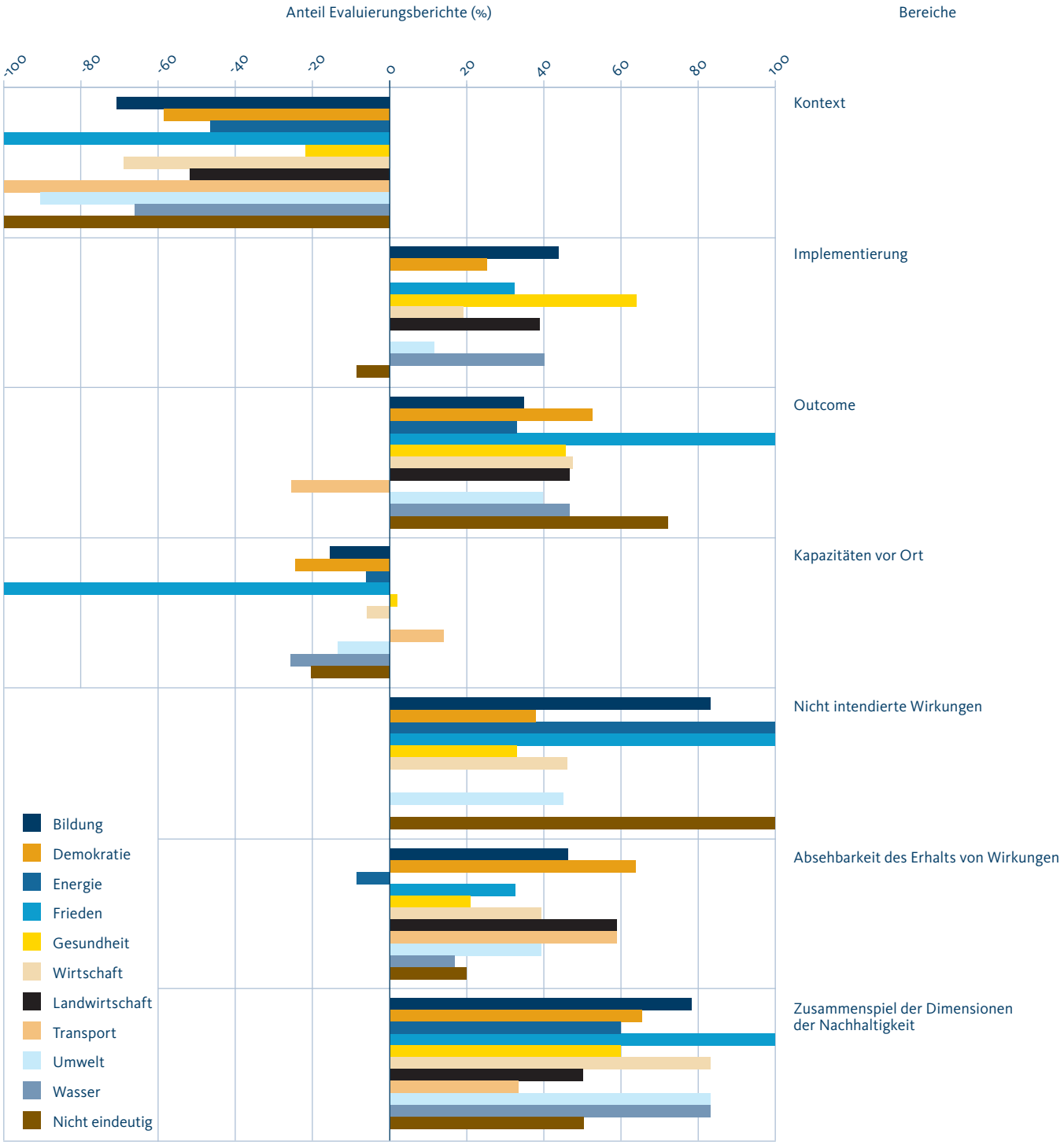
Abbildung 21: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Sektor



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach dem Sektor, in dem das Vorhaben umgesetzt wird. N = 513

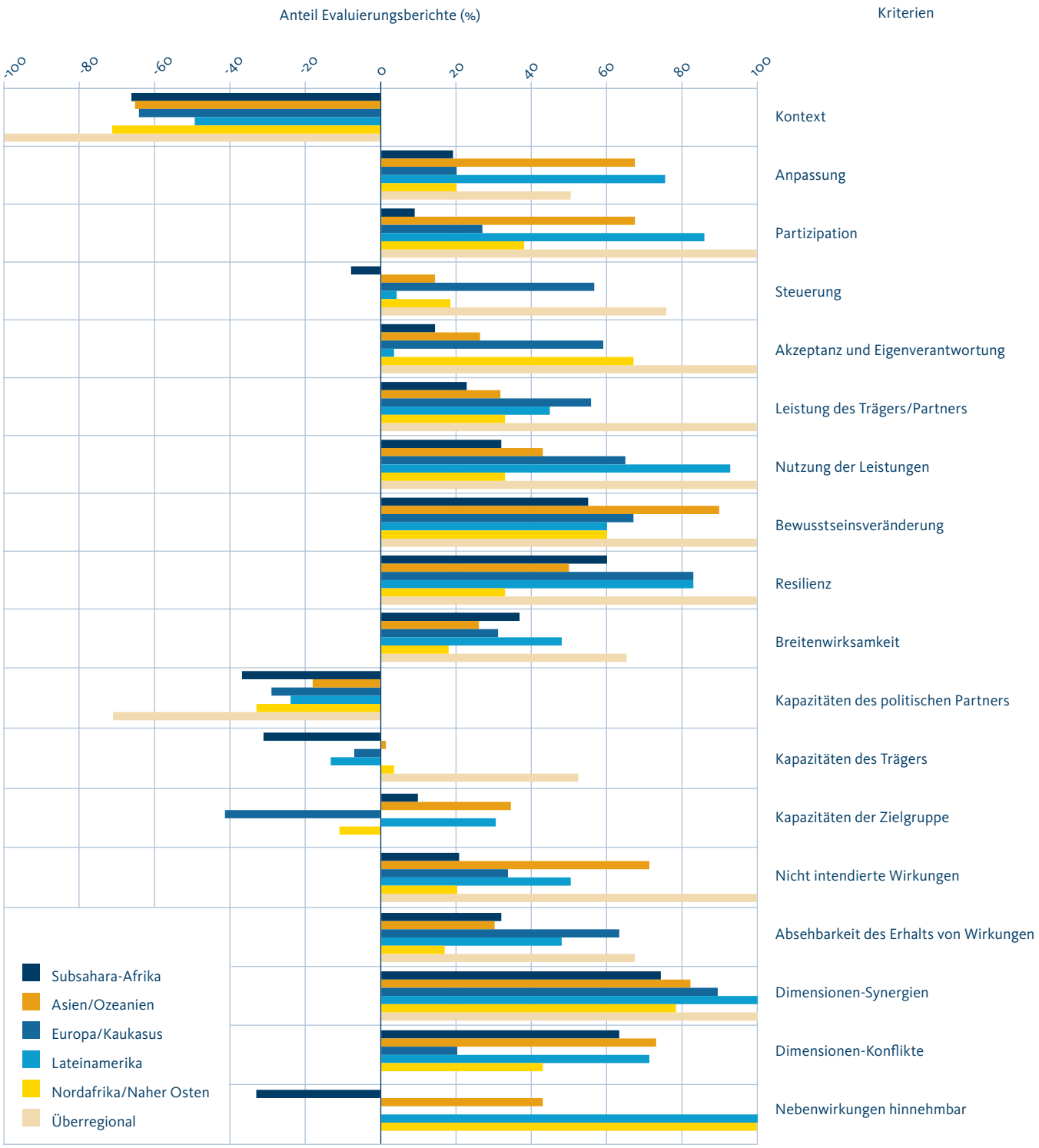
Abbildung 22: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitsbereichen und Einfluss auf Nachhaltigkeitsbewertung nach Sektor



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach dem Sektor, in dem das Vorhaben umgesetzt wird. N = 513

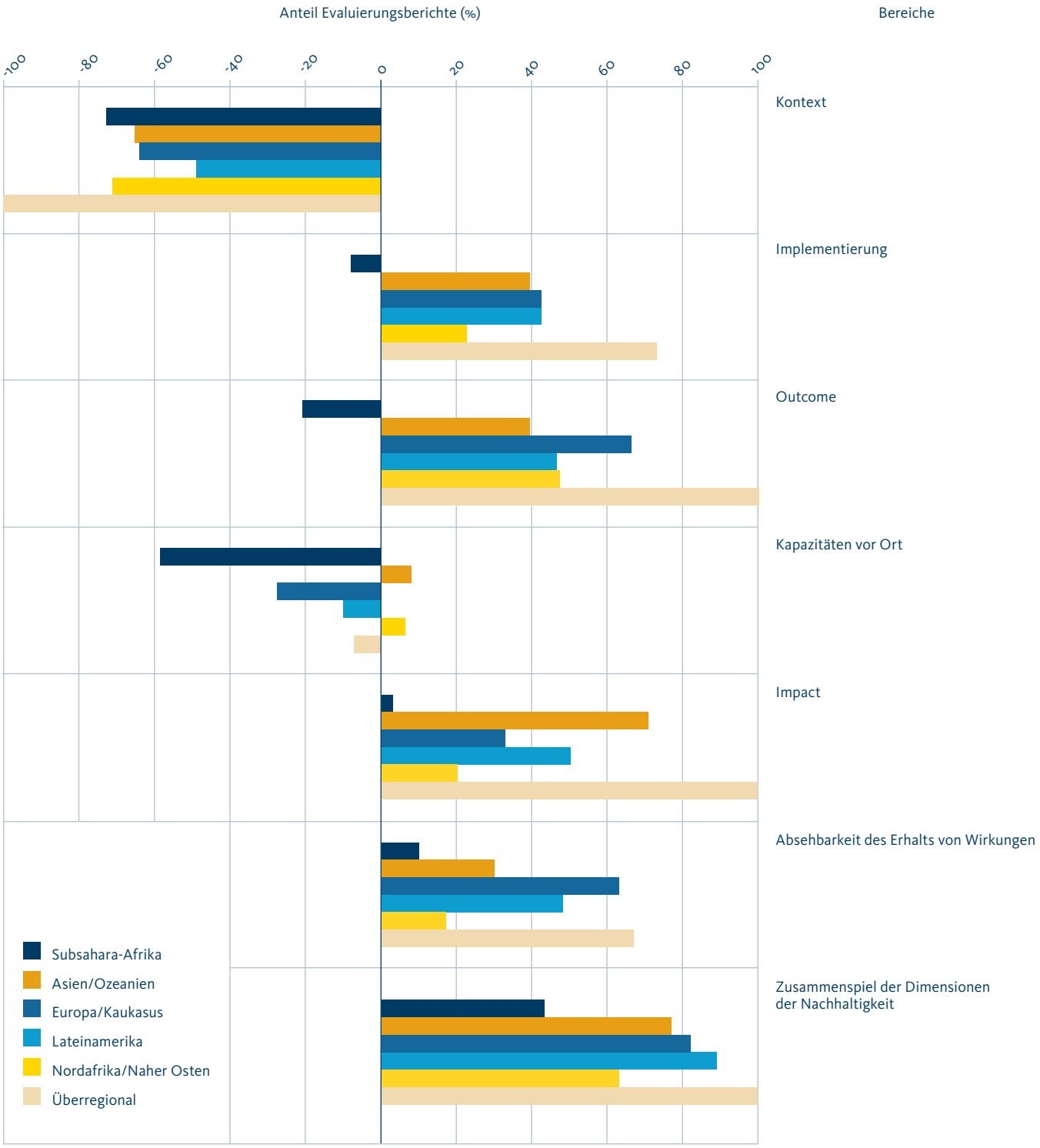
Abbildung 23: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitskriterien und Einfluss auf Nachhaltigkeitsbewertung nach Region



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die den jeweiligen Nachhaltigkeitskriterien entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach der Region, in der das Vorhaben umgesetzt wird. N = 513

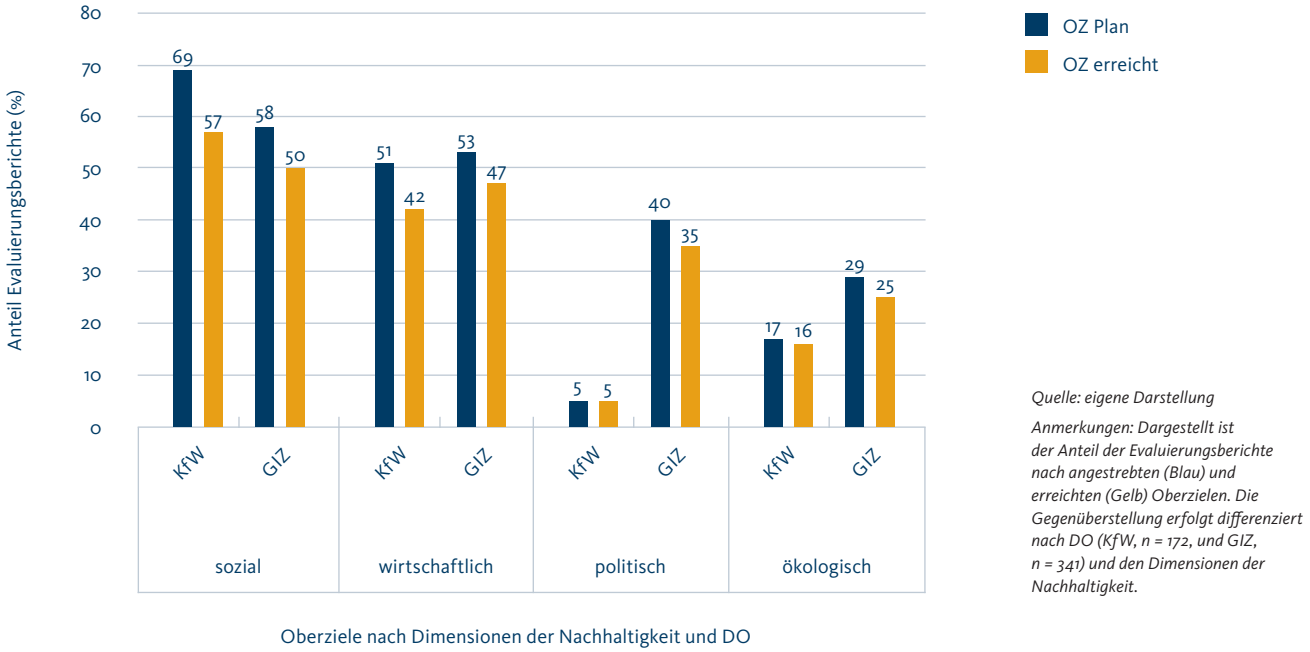
Abbildung 24: Relativer Anteil Evaluierungsberichte mit Bezug zu Nachhaltigkeitsbereichen und Einfluss auf Nachhaltigkeitsbewertung nach Region



Quelle: eigene Darstellung

Anmerkungen: Die Balken zeigen den relativen Anteil der Evaluierungsberichte, die dem jeweiligen Nachhaltigkeitskriterium entweder einen positiven oder einen negativen Einfluss auf die Nachhaltigkeit eines Vorhabens zuschreiben. Die Evaluierungsberichte sind unterteilt nach dem Sektor, in dem das Vorhaben umgesetzt wird. N = 513

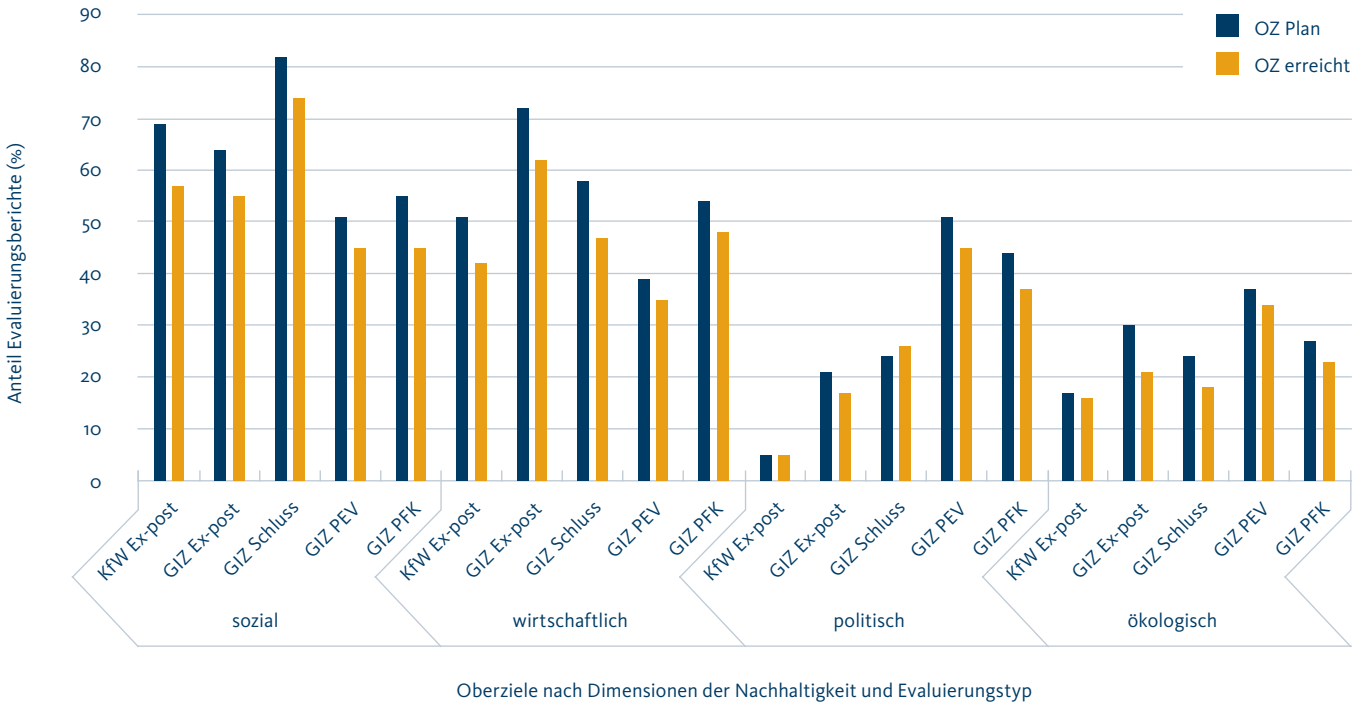
Abbildung 25: Anteil Evaluierungsberichte nach angestrebten und erreichten Oberzielen nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte nach angestrebten (Blau) und erreichten (Gelb) Oberzielen. Die Gegenüberstellung erfolgt differenziert nach DO (KfW, n = 172, und GIZ, n = 341) und den Dimensionen der Nachhaltigkeit.

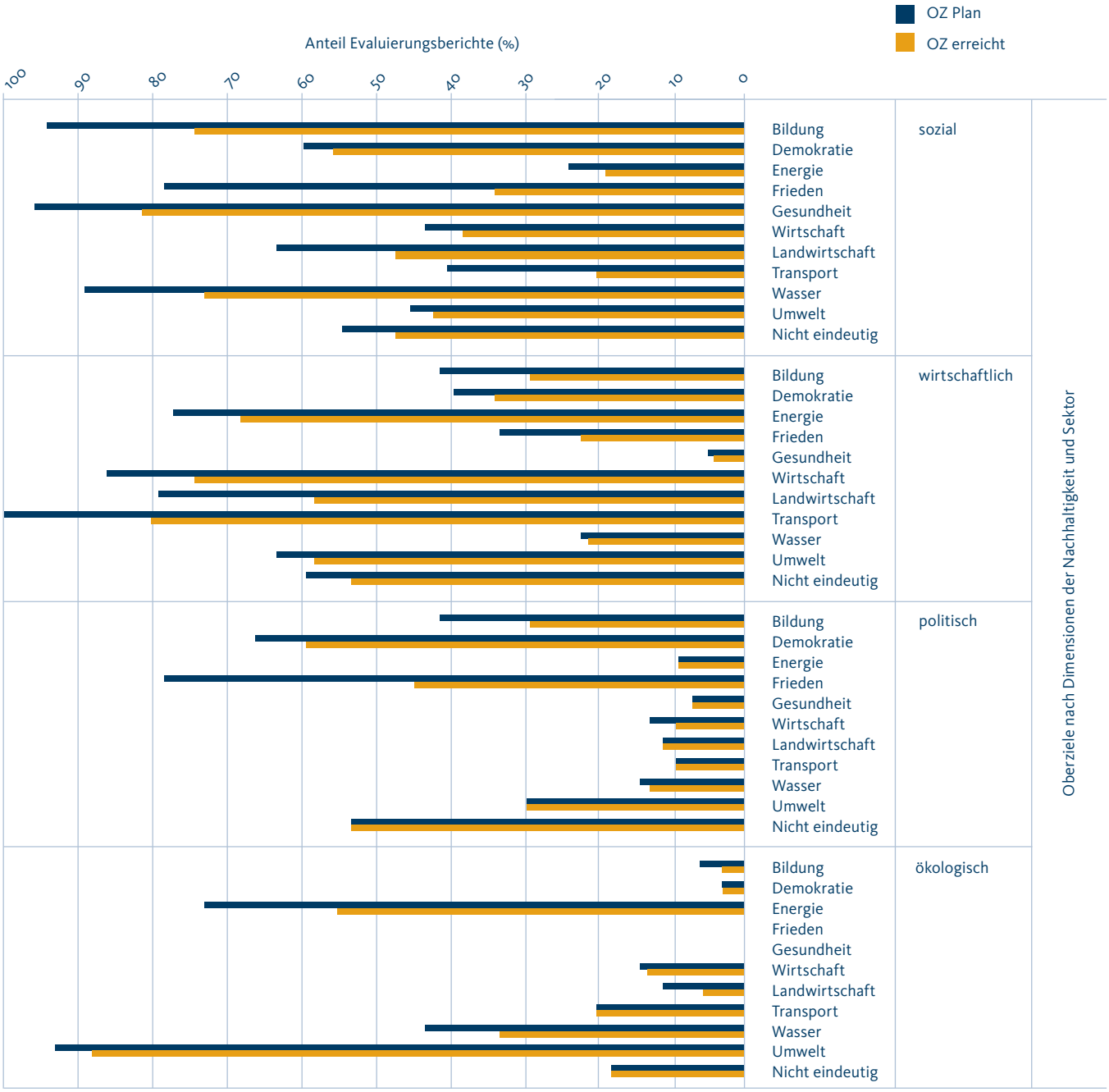
Abbildung 26: Anteil Evaluierungsberichte nach angestrebtem und erreichtem Oberziel, Evaluierungstyp und Nachhaltigkeitsdimension



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte nach Oberziel und nach Erreichen des Oberziels. Die Evaluierungsberichte sind unterteilt nach Evaluierungstyp. Diese sind: KfW-Ex-post-Evaluierungen (n = 172), GIZ-Ex-post-Evaluierungen (n = 47), GIZ-Schluss-Evaluierungen (n = 38), GIZ-PEV (n = 82), und GIZ-PFK (n = 174). Innerhalb eines Balkenpaares eines Evaluierungstyps sind die Oberziele eines Vorhabens differenziert nach geplanten und tatsächlich erreichten Oberzielen dargestellt.

Abbildung 27: Anteil Evaluierungsberichte nach angestrebtem und erreichtem Oberziel, Sektor und Nachhaltigkeitsdimension



Quelle: eigene Darstellung

Anmerkungen: Dargestellt ist der Anteil der Evaluierungsberichte nach Oberziel und nach Erreichen des Oberziels. Die Evaluierungsberichte sind unterteilt nach dem Sektor, in dem das Vorhaben umgesetzt wird. Diese sind: Bildung (n = 34), Demokratie (n = 95), Energie (n = 22), Frieden (n = 9), Gesundheit (n = 57), Wirtschaft (n = 127), Landwirtschaft (n = 19), Transport (n = 10), Wasser (n = 63), Umwelt (n = 60), und „Nicht eindeutig" (n = 17). Innerhalb eines Balkenpaares eines Sektors sind die Oberziele eines Vorhabens differenziert nach geplanten und tatsächlich erreichten Oberzielen dargestellt.

7.2
Tabellen

Tabelle 4: Analyseraster der Qualitätsbewertung

Bereiche	Nr. [22]	Kriterien	Definition des Kriteriums
1. Hintergrund	Q-01	Gegenstand (Vorhaben) beschrieben	Das Kriterium ist erfüllt, wenn 1) Ziele, 2) Zielgruppe, 3) Kontext sowie 4) relevante Akteure (politischer Partner und/oder Träger) der EZ-Maßnahme dargestellt sind und somit eine Eingrenzung des Gegenstandes vorgenommen wurde.
	Q-02	Erkenntnisinteresse formuliert/ operationalisiert	Das Kriterium ist erfüllt, wenn das Erkenntnisinteresse und/oder Evaluierungsfragen spezifiziert bzw. konkretisiert wurden.
2. Darstellung der Wirkungs-zusammenhänge	Q-03	Wirkungslogik/Wirkungskette dargestellt	Das Kriterium ist erfüllt, wenn bei der Darstellung der intendierten Wirkungen der EZ-Maßnahme zwischen verschiedenen Wirkungsebenen (Input-Output-Outcome-Impact) unterschieden wird und diese logisch aufeinander aufbauen (und/oder ggf. Wirkungs-hypothesen formuliert werden).
	Q-04	Wirkungslogik überwiegend durch Indikatoren operationalisiert	Das Kriterium ist erfüllt, wenn der Zielerreichungsgrad der Mehrheit der Programmziele messbar gemacht/anhand von Indikatoren abgeschätzt wird.
3. Methodisches Vorgehen	Q-05	Methodisches Vorgehen beschrieben	Das Kriterium ist erfüllt, wenn die in der Evaluierung zur Anwendung kommenden Arbeitsschritte zur Datenerhebung und Auswertung beschrieben und operationalisiert sind.
	Q-06	Stärken und/oder Limitationen des methodischen Vorgehens identifiziert	Das Kriterium ist erfüllt, wenn begründet wird, warum die angewandten Methoden dem Gegenstand der Evaluierung angemessen sind. Vorteile und Limitationen des methodischen Vorgehens werden diskutiert.
	Q-07	Befragte Gesprächspartner identifiziert	Das Kriterium ist erfüllt, wenn die zur Datenerhebung konsultierten/befragten Gesprächspartner identifiziert wurden.
	Q-08	Auswahlverfahren der Gesprächspartner beschrieben	Das Kriterium ist erfüllt, wenn die Auswahl der Gesprächspartner beschrieben wurde bzw. die Auswahlkriterien dargestellt sind.
4. Daten-erhebungs-methoden		Dokumenten-/ Datenbankanalyse	Das Kriterium ist erfüllt, wenn Dokumente und/oder Daten aus Sekundärdatenbanken analysiert werden.
		Monitoringdaten verwendet	Das Kriterium ist erfüllt, wenn Daten aus Monitoringdaten analysiert werden.
		Leitfaden-Interviews	Das Kriterium ist erfüllt, wenn Leitfadeninterviews zur Anwendung kommen.
		Standardisierte Interviews	Das Kriterium ist erfüllt, wenn standardisierte Interviews zur Anwendung kommen.
		Fokusgruppen-Diskussion	Das Kriterium ist erfüllt, wenn Fokusgruppen-Diskussionen zur Anwendung kommen.
		Partizipative Methoden	Das Kriterium ist erfüllt, wenn partizipative Datenerhebungsmethoden (Problem Tree, SWOT-Analyse, etc.) zur Anwendung kommen und/oder die Befragten die Gesprächsthemen mitentwickeln.
		Systematische Beobachtungen	Das Kriterium ist erfüllt, wenn systematische Beobachtungen (Begehungen, Probenprüfung etc.) gemacht werden.

[22] *Eine Nummer „Q-...“ erhalten diejenigen Kriterien, die aufgrund ihrer Aussagekraft hinsichtlich der Qualität der Evaluierungsberichte im Rahmen des Qualitätsindex Eingang in die Qualitätsbewertung gefunden haben.*

5. Evaluierungs-design	Q-09	Vorher-Nachher-Vergleich	Das Kriterium ist erfüllt, wenn die Wirkungen des EZ-Vorhabens anhand eines Vergleichs der Mehrzahl aller Indikatoren vor Vorhabenbeginn und nach Vorhabenende ermittelt werden.
	Q-10	Kontroll-/Vergleichsgruppe einbezogen	Das Kriterium ist erfüllt, wenn die Wirkungen des EZ-Vorhabens anhand eines Vergleichs zwischen Kontroll- (außerhalb des Einflussbereichs der EZ-Maßnahme) und Interventions-Gruppe (innerhalb des Einflussbereichs der EZ-Maßnahme) ermittelt werden.
	Q-11	Kausalität über Plausibilitäten hergeleitet	Das Kriterium ist erfüllt, wenn die Wirkungen des EZ-Vorhabens auf der Grundlage eines systematischen Verfahrens anhand von Plausibilitäten (insbesondere theoriebasierter Ansätze, z. B. durch Kontributionsanalysen) ermittelt werden.
6. Robustheit der Ergebnisse	Q-12	Daten-Triangulation angewandt	Das Kriterium ist erfüllt, wenn die der Analyse zugrunde liegenden Daten aus verschiedenen Quellen (im Sinne von Stakeholdergruppen und/oder Erhebungsinstrumenten) stammen (> 1 Quelle).
	Q-13	Methoden-Triangulation angewandt	Das Kriterium ist erfüllt, wenn die Auswertung der Daten derselben Quelle durch verschiedene Methoden erfolgt (> 1 Methode).
		Forscherinnen- und Forscher-Triangulation	Das Kriterium ist erfüllt, wenn an der Analyse mindestens zwei Forscherinnen/Forscher beteiligt sind und wenn in Aussagen transparent gemacht wird, durch welche/n Forscherin/Forscher diese gestützt bzw. nicht gestützt wird.[23]
7. Auswertung und Schluss-folgerungen	Q-14	Schlussfolgerungen durch Daten überwiegend referenziert	Das Kriterium ist erfüllt, wenn der überwiegende Anteil der Ergebnisse und Schlussfolgerungen mit der Datengrundlage /-analyse in der Mehrheit der Schlussfolgerungen in Bezug gesetzt wird.
	Q-15	Schlussfolgerungen aus Daten überwiegend plausibel begründet	Das Kriterium ist erfüllt, wenn der überwiegende Anteil der Ergebnisse und Schlussfolgerungen mit Blick auf die Wirkung auf der Grundlage der verwendeten Daten nachvollziehbar ist.
	Q-16	Datengrundlage ausreichend hinsichtlich Schlussfolgerungen	Das Kriterium ist erfüllt, wenn die Datengrundlage und die methodische Vorgehensweise qualitativ und quantitativ ausreichend sind, um die ausgesprochenen Schlussfolgerungen (im Sinne von erreichten Wirkungen) zu ziehen.

Quelle: eigene Darstellung

[23] *Aufgrund der schwierigen Umsetzung von Forscherinnen- und Forschertriangulation in der Praxis der Evaluierungsberichte findet dieses Kriterium in der Analyse keine weitere Berücksichtigung.*

Tabelle 5: Analyseraster der Nachhaltigkeitsbewertung

Bereiche	Kriterien	Nr.	Differenzierte Kriterien	Definition
1) Kontext	1. Kontext nach Dimensionen	N-01	Soziale Dimension	Das Kriterium gilt als vorhanden, wenn die berichteten Kontextfaktoren direkten Einfluss auf a) die Wirkungen der Maßnahme oder b) die Absehbarkeit des Erhalts ihrer Wirkungen haben.
		N-02	Wirtschaftliche Dimension	
		N-03	Politische Dimension	
		N-04	Ökologische Dimension	
2) Implementierung	2. Anpassung (Alignment)	N-05	Anpassung an nationale Regelungen	Das Kriterium gilt als vorhanden, wenn die Maßnahme im Einklang mit einer nationalen Strategie/einem nationalen Programm steht.
		N-06	Anpassung an soziokulturellen Kontext auf Ebene der Zielgruppen	Das Kriterium gilt als vorhanden, wenn die Maßnahme im Einklang mit gesellschaftlichen Konventionen steht.
	3. Partizipation	N-07	Partizipation des entwicklungspolitischen Partners	Das Kriterium gilt als vorhanden, wenn der Träger/Partner bei Entscheidungen in der Implementierung mindestens konsultiert wurde.
		N-08	Partizipation der Zielgruppe(n)/ Bevölkerung	Das Kriterium gilt als vorhanden, wenn die Zielgruppe(n) bei Entscheidungen in der Implementierung mindestens konsultiert wurde(n).
	4. Steuerung	N-09	Nutzung der (institutionellen) Strukturen vor Ort	Das Kriterium gilt als vorhanden, wenn bereits existierende Gremien, Arbeitsgruppen oder andere institutionelle Strukturen im Partnerland oder der Region für die Umsetzung des Vorhabens genutzt werden.
		N-10	Management response / Lernen aus M&E / Lessons learned	Das Kriterium gilt als vorhanden, wenn Monitoring-/Evaluierungsergebnisse bei Maßnahmenstrukturen und/oder Maßnahmenprozessen berücksichtigt wurden.
		N-11	Upscaling-Strategie	Das Kriterium gilt als vorhanden, wenn die Aktivitäten auf eine oder mehrere Provinzen und/oder Zielgruppen bzw. Stakeholdergruppen ausgedehnt wurden und/oder eine Systematisierung von Pilotvorhaben stattfand – z. B. wenn mehrere kleinere Programmstränge beendet und in ein größeres Programm/eine nationale Strategie überführt wurden.
		N-12	Exit-Strategie	Das Kriterium gilt als vorhanden, wenn gemeinsam mit dem Partner/ Träger ein Konzept zur Fortführung der Aktivitäten ohne die deutsche EZ entwickelt wurde und/oder Schritte einer allmählichen Leistungsreduzierung oder einer nach Projektende reduziert fortlaufenden Aktivität der deutschen EZ beschrieben sind.

Bereiche	Kriterien	Nr.	Differenzierte Kriterien	Definition
3) Outcome	5. Akzeptanz und Eigenverantwortung (Ownership)	N-13	Akzeptanz und Eigenverantwortung des privatwirtschaftlichen Trägers	Das Kriterium gilt als vorhanden, wenn dieser Initiative gezeigt hat und/oder sich überwiegend an Zusagen/eigene Verpflichtungen gehalten hat und/oder Verantwortung übernommen hat.
		N-14	Akzeptanz und Eigenverantwortung des politischen Partners	Das Kriterium gilt als vorhanden, wenn dieser Initiative gezeigt hat und/oder sich überwiegend an Zusagen/eigene Verpflichtungen gehalten hat und/oder Verantwortung übernommen hat.
		N-15	Akzeptanz und Eigenverantwortung der Zielgruppe	Das Kriterium gilt als vorhanden, wenn diese Initiative gezeigt hat und/oder sich überwiegend an Zusagen/eigene Verpflichtungen gehalten hat und/oder Verantwortung übernommen hat.
	6. Leistungen (Outputs) des Trägers/Partners	N-16	Service-/Produkt-Qualität	Das Kriterium gilt als vorhanden, wenn die Qualität des Outputs überwiegend als ausreichend eingeschätzt wird, um die Programmziele erreichen zu können.
		N-17	Service-/Produkt-Quantität	Das Kriterium gilt als vorhanden, wenn die Quantität des Outputs überwiegend als ausreichend eingeschätzt wird, um die Programmziele erreichen zu können.
	7. Nutzung der Leistungen (Outputs)	N-18	Nutzung der Leistungen durch Partner/Träger	Das Kriterium gilt als vorhanden, wenn Leistungen der Maßnahme (Konzepte, Materialien) vom Partner/Träger angewendet werden
		N-19	Nutzung der Leistungen durch Zielgruppe	Das Kriterium gilt als vorhanden, wenn Leistungen der Maßnahme (Konzepte, Materialien) von der Zielgruppe genutzt werden.
	8. Bewusstseinsveränderung	N-20	Bewusstseinsveränderung bei Partner/Träger	Das Kriterium gilt als vorhanden, wenn beim Partner/Träger eine Bewusstseinsveränderung über die Nutzung des Outputs hinaus (im Sinne von Verhaltensänderungen auch außerhalb des Vorhabens/ohne Anreize) zu beobachten ist.
		N-21	Bewusstseinsveränderung bei Zielgruppe	Das Kriterium gilt als vorhanden, wenn bei der Zielgruppe eine Bewusstseinsveränderung über die Nutzung des Outputs hinaus (im Sinne von Verhaltensänderungen auch außerhalb des Vorhabens/ohne Anreize) zu beobachten ist.
	9. Resilienz und Anpassungsfähigkeit	N-22	Resilienz und Anpassungsfähigkeit bei Partner/Träger	Das Kriterium gilt als vorhanden, wenn dieser in der Lage ist, Chancen und Herausforderungen selbst zu erkennen und entsprechend zu handeln.
		N-23	Resilienz und Anpassungsfähigkeit bei Zielgruppe	Das Kriterium gilt als vorhanden, wenn diese in der Lage ist, Chancen und Herausforderungen selbst zu erkennen und entsprechend zu handeln.
	10. Reichweite und Breitenwirksamkeit	N-24	Strukturbildung (direkt)	Das Kriterium gilt als vorhanden, wenn Veränderungen nicht nur auf individueller Ebene, sondern auf System-Ebene stattfinden.
		N-25	Diffusion (indirekt)	Das Kriterium gilt als vorhanden, wenn sich Konzepte oder Ideen auf Menschen, die nicht zur ursprünglichen Zielgruppe gehörten, übertragen.

Bereiche	Kriterien	Nr.	Differenzierte Kriterien	Definition
4) Kapazitäten vor Ort	11. Kapazitäten des politischen Partners	N-26	Finanzielle/wirtschaftliche Beiträge	Das Kriterium gilt als vorhanden, wenn durch politische Partner zu erbringende finanzielle/wirtschaftliche Beiträge gemäß Vereinbarung geleistet werden bzw. die Beiträge zur erfolgreichen Fortführung der Aktivitäten ausreichend sind.
		N-27	Personelle/fachlich-technische Kapazitäten	Das Kriterium gilt als vorhanden, wenn a) ausreichend Personal vorhanden ist und b) das Personal ausreichend qualifiziert ist, um die Aktivitäten der Maßnahme erfolgreich weiterzuführen.
		N-28	Institutionelle/organisationale Beiträge	Das Kriterium gilt als vorhanden, wenn ein ausreichendes Maß an institutioneller Unabhängigkeit und organisationaler Effektivität/ Effizienz gegeben ist, um Programmziele zu erreichen bzw. wenn institutionelle Beiträge gemäß Vereinbarung geleistet werden.
	12. Kapazitäten des Trägers	N-29	Finanzielle/wirtschaftliche Beiträge	Das Kriterium gilt als vorhanden, wenn durch den Träger zu erbringende finanzielle/wirtschaftliche Beiträge gemäß Vereinbarung geleistet werden bzw. die Beiträge zur erfolgreichen Fortführung der Aktivitäten ausreichend sind.
		N-30	Personelle/fachlich-technische Kapazitäten	Das Kriterium gilt als vorhanden, wenn a) ausreichend Personal vorhanden ist und b) das Personal ausreichend qualifiziert ist, um die Aktivitäten der Maßnahme erfolgreich weiterzuführen.
		N-31	Institutionelle/organisationale Kapazitäten	Das Kriterium gilt als vorhanden, wenn ein ausreichendes Maß an institutioneller Unabhängigkeit und organisationaler Effektivität/ Effizienz gegeben ist, um Programmziele zu erreichen.
	13. Kapazitäten der Zielgruppe	N-32	Finanzielle/wirtschaftliche Beiträge	Das Kriterium gilt als vorhanden, wenn durch die Zielgruppe zu erbringende finanzielle/wirtschaftliche Beiträge gemäß Vereinbarung geleistet werden bzw. die Beiträge zur erfolgreichen Fortführung der Aktivitäten ausreichend sind.
		N-33	Personelle/fachlich-technische Kapazitäten	Das Kriterium gilt als vorhanden, wenn die Zielgruppen ausreichend qualifiziert sind bzw. die Beschaffung des nötigen Know-hows gesichert ist, um die Aktivitäten der Maßnahme erfolgreich weiterzuführen.
		N-34	Institutionelle/organisationale Kapazitäten	Das Kriterium gilt als vorhanden, wenn ein ausreichendes Maß an institutioneller Unabhängigkeit und organisationaler Effektivität/ Effizienz des Nutzers gegeben ist, um Programmziele zu erreichen.
5) Impact	14. Nicht intendierte Wirkungen nach Dimensionen	N-35	Soziale Gerechtigkeit	Das Kriterium gilt als vorhanden, wenn die Maßnahme außerhalb des Oberziels/Programmziels zu Veränderungen in sozialen Aspekten beiträgt.
		N-36	Wirtschaftliche Aspekte	
		N-37	Politische Aspekte	
		N-38	Ökologische Aspekte	
6) Absehbarkeit des Erhalts von Wirkungen	15. Absehbarkeit des Erhalts von Wirkungen nach Dimensionen	N-39	Soziale Aspekte	Das Kriterium gilt als vorhanden, wenn die Faktoren, die eine Fortdauer der positiven Wirkungen sichern bzw. die Wirkungen steigern, überwiegen.
		N-40	Wirtschaftliche Aspekte	
		N-41	Politische Aspekte	
		N-42	Ökologische Aspekte	

Bereiche	Kriterien	Nr.	Differenzierte Kriterien	Definition
7) Zusammenspiel der Dimensionen der Nachhaltigkeit	16. Dimensionen-Synergien	N-43	Schaffung von Synergien durch Vorhaben	Das Kriterium gilt als vorhanden, wenn Maßnahmen Wirkungen in unterschiedlichen Nachhaltigkeitsdimensionen entfalten, die in einem synergetischen Zusammenspiel stehen.
		N-44	Identifizierung von Synergien durch Evaluierung	Das Kriterium gilt als vorhanden, wenn die Evaluierung Potenziale für Synergien identifiziert.
	17. Dimensionen-Konflikte	N-45	Identifizierung von Zielkonflikten durch Vorhaben	Das Kriterium gilt als vorhanden, wenn Zielkonflikte zwischen Dimensionen durch das Vorhaben identifiziert werden.
		N-46	Identifizierung von Zielkonflikten durch Evaluierung	Das Kriterium gilt als vorhanden, wenn die Evaluierung Zielkonflikte zwischen Dimensionen identifiziert.
	18. Nebenwirkungen hinnehmbar	N-47	Einstufung eventueller Kompensationsmaßnahmen durch Vorhaben als ausreichend und/oder von möglichen Neben-wirkungen als „hinnehmbar"	Das Kriterium gilt als vorhanden, wenn das Vorhaben feststellt, dass umgesetzte Kompensationsmaßnahmen (zur Minimierung von Zielkonflikten zwischen Dimensionen) ausreichend bzw. eventuelle vom Vorhaben ausgelöste Nebenwirkungen „hinnehmbar" sind.
		N-48	Einstufung von eventuellen Nebenwirkungen durch Evaluierung als „hinnehmbar"	Das Kriterium gilt als vorhanden, wenn die Evaluierung feststellt, dass vom Vorhaben umgesetzte Kompensationsmaßnahmen ausrei-chend bzw. eventuelle Nebenwirkungen (im Sinne von Zielkonflikten zwischen Dimensionen) „hinnehmbar" sind.

7.3
Evaluierungsteam und Mitwirkende

Kernteam	
Dr. Sven Harten	Abteilungsleiter
Dr. Martin Noltze	Senior-Evaluator und Teamleiter
Dr. Michael Euler	Evaluator
Ida Verspohl	Evaluatorin
Cornelia Michels-Lampo	Projektadministratorin

Mitwirkende	Funktion
Prof. Dr. Sebastian Vollmer	Externer Gutachter
Dr. Kerstin Guffler	DEval-interne Gutachterin
Solveig Gleser	DEval-interne Gutachterin
Thomas Wencker	DEval-interner Gutachter
Jana Preiß	Assoziierte Masterstudentin
Niklas Witzig	Praktikant
Grisel Orozco	Praktikantin
Helena Heberer	Studierende Beschäftigte
Sarah Stahlmann	Studierende Beschäftigte
Lea Smidt	Studierende Beschäftigte

7.4 Zeitplan

Konzeptionsphase		**Vorbereitende Phase und Festlegung des Evaluierungsgegenstandes**
	04/2016 – 05/2016	Klärungsgespräche mit BMZ und DO
	06/2016 – 07/2016	Erstellung des Konzeptpapiers
	08/2016	Referenzgruppentreffen zur Diskussion des Konzeptpapiers
	08/2016	Fertigstellung des Konzeptpapiers
Inception-Phase		**Entwicklung der methodischen Vorgehensweise**
	08/2016 – 10/2016	Erarbeitung des Inception-Reports
	10/2016	Referenzgruppensitzung zur Diskussion des Inception-Reports
	02/2017	Fertigstellung des Inception-Reports
Erhebungs- und Synthesephase		**Datenerhebung und Datenanalyse**
	10/2016 – 11/2016	Einholung von Daten und Dokumenten von den DO
	11/2016	Aufbau Datensatz und Stichprobenziehung
	12/2016 – 02/2017	Einholung von Sekundärdaten im Rahmen der Evaluierungssynthese
	12/2016 – 04/2017	Durchführung der quantitativen Inhaltsanalyse
	02/2017	Durchführung der Kontextstudie und Portfolioanalyse
	03/2017 – 04/2017	Analyse und Zusammenführung der Ergebnisse aus der Meta-Evaluierung und Evaluierungssynthese
	05/2017	Referenzgruppentreffen zu vorläufigen Ergebnissen und Schlussfolgerungen
Berichtslegung		**Erstellung der Evaluierungsberichte und Disseminierung**
	06/2017 – 07/2017	Verfassen der Evaluierungsberichte der Meta-Evaluierung und Evaluierungssynthese
	08/2017	Versand der Evaluierungsberichte an die Referenzgruppe
	09/2017	Referenzgruppentreffen zur Vorstellung der Evaluierungsberichte
	01/2018	Veröffentlichung der Evaluierungsberichte
	2018	Disseminierung

Deutsches Evaluierungsinstitut der
Entwicklungszusammenarbeit (DEval)

Fritz-Schäffer-Straße 26
53113 Bonn, Deutschland

Tel: +49 (0)228 33 69 07-0
Fax: +49 228 24 99 29-904

Mail: info@DEval.org
www.DEval.org

DEval
DEUTSCHES
EVALUIERUNGSINSTITUT
DER ENTWICKLUNGS-
ZUSAMMENARBEIT