NACHHALTIGKEIT IN DER DEUTSCHEN ENTWICKLUNGS-ZUSAMMENARBEIT

Evaluierungssynthese
2018

DEval
DEUTSCHES EVALUIERUNGSINSTITUT DER ENTWICKLUNGS-ZUSAMMENARBEIT

Die vorliegende Evaluierungssynthese „Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit“ ist Teil des DEval-Themenschwerpunktes Nachhaltigkeit. Die Evaluierungssynthese wird durch eine begleitende Meta-Evaluierung unterstützt. 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

Evaluierungssynthese
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), FO Travel/Alamy Stock Foto (Kap. 1), themacx/iStock.com (Kap. 2), Nikon'as/Fotolia.com (Kap. 3), epicurean/iStock.com (Kap. 4 + 6), andresr/iStock.com (Kap. 5), Steve Bloom Images/Alamy Stock Foto (Kap. 7)

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

Druck
Bonifatius,
Paderborn

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

Druck ISBN: 978-3-96126-051-5
PDF ISBN: 978-3-96126-052-2

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 Evaluierungssynthese und der begleitenden Meta-Evaluierung 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 erhebt Nachhaltigkeit zum Leitbild globalen menschlichen Handelns. Die in der Agenda 2030 definierten nachhaltigen Entwicklungsziele (SDGs) verbinden wirtschaftlichen Fortschritt mit sozialer Gerechtigkeit und der schonenden Nutzung ökologischer Ressourcen. Dabei liegt die Umsetzung der Agenda 2030 in der Verantwortung aller Staaten. Gleichzeitig erfordert sie neue Kooperationen zwischen Politik, Privatwirtschaft, Wissenschaft und Zivilgesellschaft.

Auch die internationale Entwicklungszusammenarbeit (EZ) hat sich zu einer entsprechenden Neuausrichtung verpflichtet: Konzeption und Umsetzung von EZ-Maßnahmen müssen künftig den Zielen und Handlungsprinzipien der Agenda 2030 gerecht werden. Dies ist eine zentrale Herausforderung für die internationale EZ. Auf Ebene einzelner Maßnahmen bedarf dies insbesondere der Reflexion über soziale, wirtschaftliche und ökologische Wechselwirkungen sowie über Auswirkungen auf benachteiligte Gruppen. Um diesen Prozess zu unterstützen, sind evidenzbasierte Handlungsempfehlungen nötig. Derzeit gibt es nur eine begrenzte Anzahl an Vorhaben, die explizit in Anlehnung an die Agenda 2030 und deren Prinzipien konzipiert wurden. Eine empirische Auseinandersetzung mit der Nachhaltigkeit von EZ-Maßnahmen ist dennoch möglich.

Bereits seit 2006 wird die Nachhaltigkeit in Evaluierungen von Vorhaben der deutschen EZ systematisch überprüft und benotet. In dem Jahr verabschiedete das Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) die „Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen". In Anlehnung an die vom Entwicklungsausschuss (DAC) der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) 1991 beschlossenen Prinzipien zur Evaluierung der Entwicklungszusammenarbeit enthält dieser Leitfaden Vorgaben zur Bewertung der Evaluierungskriterien Relevanz, Effektivität, Effizienz, übergeordnete entwicklungspolitische Wirkungen (Impact) und Nachhaltigkeit. Demnach wird die Nachhaltigkeit einzelner Maßnahmen anhand von verbindlichen Leitfragen bewertet. Als Ergebnis der Nachhaltigkeitsprüfung wird eine Note vergeben. Die Nachhaltigkeit von Vorhaben wird dabei konzeptionell in enger Verbindung zur entwicklungspolitischen Wirksamkeit beurteilt. Daher kann erwartet werden, dass die bisherige Bewertungspraxis – über die entwicklungspolitische Wirksamkeit – bereits einige der Prinzipien der Agenda 2030 abdeckt.

Die hier durchgeführte Evaluierungssynthese zielt darauf ab, das Zusammenspiel verschiedener Determinanten bei der Nachhaltigkeitsbewertung von Vorhaben besser zu verstehen. Zweck der Untersuchung ist es, dazu beizutragen, die strategische und operative Ausrichtung der deutschen EZ besser an den neuen Anforderungen des modernen Nachhaltigkeitsverständnisses der Agenda 2030 auszurichten. Dies trägt der gesteigerten Bedeutung von Nachhaltigkeit im Kontext der SDG-konformen Evaluierung von Vorhaben in der deutschen EZ Rechnung.

Die vorliegende Evaluierungssynthese beinhaltet eine erste umfassende und systematische Auseinandersetzung mit der aggregierten Nachhaltigkeitsbewertung in Evaluierungen der deutschen finanziellen und technischen EZ. Dabei beschränkt sich die Untersuchung auf die Evaluierungs- und Bewertungspraxis der beiden großen staatlichen Durchführungsorganisationen (DO) – der KfW Entwicklungsbank (KfW) und der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ). Da der Evaluierungsgegenstand möglichst umfassend bearbeitet werden soll, wird er weder auf bestimmte Sektoren noch auf bestimmte Regionen oder Typen von Vorhaben beschränkt. Neben rein bilateralen Vorhaben in bestimmten Ländern sind auch Regional-, Sektor- und Globalvorhaben Teil der Untersuchung.

Methodisches Vorgehen

Die Analyse der Einflussfaktoren auf die vergebene Nachhaltigkeitsnote erfolgt durch multivariate Regressionsmodelle. Diese Modelle erlauben es, den Einfluss verschiedener Faktoren auf die zu erklärende Variable – die Nachhaltigkeitsnote von Vorhaben – zu bestimmen. Durch die begrenzte Datenverfügbarkeit können dabei nur bestimmte Einflussfaktoren berücksichtigt werden. Die Untersuchung beschränkt sich daher auf Merkmale der Vorhaben, Faktoren ihrer Implementierung sowie verfügbare Kontextinformationen. Zu Letzteren gehören sowohl Merkmale des unmittelbaren Kontextes der

Entwicklungsmaßnahmen als auch makroquantitative Indikatoren auf der Ebene der Partnerländer. Ferner greift die Analyse auf Ergebnisse der begleitenden Meta-Evaluierung zum Thema Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit zurück. Die Ergebnisse der Meta-Evaluierung erlauben es einerseits, die zur Bewertung von Nachhaltigkeit herangezogenen Kriterien in den Analysen zu berücksichtigen. Andererseits wird die im Rahmen der Meta-Evaluierung durchgeführte Qualitätsbewertung der Evaluierungen als Gewichtungsfaktor für einzelne Beobachtungen in den Regressionsmodellen herangezogen. Dabei werden keine Beobachtungen von der Analyse ausgeschlossen, die Gewichtung einzelner Beobachtungen stellt jedoch sicher, dass die glaubwürdigsten Ergebnisse den größten Einfluss bei der Synthese erhalten.

Zentrale Ergebnisse, Schlussfolgerungen und Empfehlungen

Einflussfaktoren auf die Nachhaltigkeitsbewertung von Vorhaben

In der Evaluierungspraxis von KfW und GIZ variiert die vergebene Nachhaltigkeitsnote nur geringfügig. Über 84 Prozent der übermittelten Evaluierungen bewerten die Nachhaltigkeit mit der Notenstufe 2 oder 3. Dabei geht eine bessere Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz und übergeordnete entwicklungspolitische Wirkungen (Impact) auch mit einer besseren Nachhaltigkeitsbewertung einher. So ist die Durchschnittsnote aller DAC-Kriterien (ohne Nachhaltigkeit) – nach statistischem Signifikanzniveau und nach Effektstärke – in allen Regressionsmodellen der wichtigste Einflussfaktor der Nachhaltigkeitsnote.

Nachhaltigkeit ist demnach ein übergeordnetes Evaluierungskriterium. Dieses enthält kaum genuine Einflussfaktoren, die trennscharf von den übrigen DAC-Kriterien abzugrenzen sind. Dennoch zeigen die Regressionsmodelle, dass bestimmte Faktoren bei der Notenvergabe besonders bedeutend sind. Insbesondere die aus der begleitenden thematischen Meta-Evaluierung gewonnenen Informationen erlauben Rückschlüsse auf die Nachhaltigkeit einzelner Maßnahmen. Die Ergebnisse der begleitenden Meta-Evaluierung zeigen außerdem, dass Nachhaltigkeit in der Evaluierungspraxis zwar anhand umfassender Kriterien bewertet wird, diese Bewertung aber gleichzeitig unsystematisch und uneinheitlich geschieht. Die Bewertung der Nachhaltigkeit steht zudem über die Bewertung der entwicklungspolitischen Wirksamkeit auch immer im Zusammenhang mit der Bewertung der anderen DAC-Kriterien.

Auch die angewandten Evaluierungsformate bedingen Unterschiede in der Bewertung von Nachhaltigkeit. Während Ex-post-Evaluierungen ihre Bewertung auf Beobachtungen gründen, geschieht die Nachhaltigkeitsbewertung im Rahmen von Projektfortschrittskontrollen (PFK), Projektevaluierungen (PEV) und Schluss-Evaluierungen in Form einer Prognose. Im Vergleich zu den übrigen Evaluierungsformaten bewerten Ex-post-Evaluierungen die Nachhaltigkeit von Vorhaben tendenziell am schlechtesten. Doch nicht nur das Ergebnis der Bewertung, sondern auch die Kriterien, die der Bewertung zugrunde liegen, unterscheiden sich je nach Format. Ein Vergleich der Nachhaltigkeitsnoten zwischen verschiedenen Vorhaben ist somit nur eingeschränkt möglich. Generell kann aber festgehalten werden, dass in Ex-post-Evaluierungen die Rolle und die Beiträge der entwicklungspolitischen Partner und Zielgruppen von besonderer Bedeutung für die Nachhaltigkeit von Vorhaben sind. Demgegenüber werden bei der Nachhaltigkeitsbewertung in PFK, PEV und Schluss-Evaluierungen vor allem die direkten Leistungen und die Umsetzung einer Maßnahme sowie der Implementierungskontext berücksichtigt.

Neben diesen Unterschieden weisen die ermittelten Einflussfaktoren in den verschiedenen Evaluierungsformaten aber auch Gemeinsamkeiten auf. So wird der Absehbarkeit des Erhalts von Wirkungen sowohl in Ex-post-Evaluierungen als auch in PFK, PEV und Schluss-Evaluierungen ein signifikant positiver Einfluss auf die Nachhaltigkeit von Vorhaben zugesprochen. Dies zeigt, dass neben den Leistungen und Wirkungen von Vorhaben die Dauer von Wirkungen – ein konzeptionelles Kernelement der Nachhaltigkeitsbewertung – in allen Evaluierungstypen einen signifikanten Einfluss auf die Notenvergabe hat.

Empfehlungen im Sinne der Stärkung der Nachhaltigkeit von Vorhaben

Die im Folgenden genannten Empfehlungen ergeben sich aus den Ergebnissen und Schlussfolgerungen der Evaluierungssynthese. Aufgrund der Komplexität der Empfehlungen werden diese – in den jeweiligen Unterpunkten – durch Anregungen und Gedanken, die sich vornehmlich auf die Umsetzung beziehen, ergänzt.

Dem BMZ und den DO wird empfohlen, die Kapazitäten der Partner und Träger vor Ort bei der Planung und Durchführung von Vorhaben stärker zu berücksichtigen und systematisch zu fördern.

- In diesem Sinne könnte eine explizite Abschätzung der Kapazitäten aller relevanten Partner und Träger bereits bei der Planung von Vorhaben in das Votum zur Förderwürdigkeit eines Moduls einfließen. Dabei sollte sichergestellt werden, dass auf Seiten der Partner und Träger die technischen, finanziellen und institutionellen Voraussetzungen für die Fortführung der vormals durch die Maßnahme erbrachten Leistungen gegeben sind.
- Darüber hinaus könnte die Prüfung der Partner- und Trägerkapazitäten in regelmäßigen Abständen während eines laufenden Vorhabens wiederholt werden. Die Übertragung der Leistungen auf die Partner zum Ende eines Vorhabens könnte zudem durch das Entwickeln langfristiger Exit-Strategien sichergestellt werden.
- Durch die Stärkung des Partnersystems könnte die Eigenverantwortung der Partnerländer hinsichtlich der Umsetzung der Agenda 2030 sichergestellt werden.

GIZ und KfW wird empfohlen, steuerungsrelevante Faktoren eines Vorhabens zukünftig nicht nur im Hinblick auf die Wirksamkeit, sondern auch im direkten Bezug zur Nachhaltigkeit zu verstehen und zu berücksichtigen.

- Hierzu zählen insbesondere die Nutzung institutioneller Strukturen vor Ort, die Aufbereitung von Lernerfahrungen sowie das Entwickeln von Upscaling- und Exit-Strategien.

Systematisches Lernen aus Evaluierungen

Die Vergleichbarkeit von Evaluierungsergebnissen ist eine zentrale Voraussetzung für die Durchführung von Evaluierungssynthesen. Durch die Aggregation von Erkenntnissen aus einzelnen Evaluierungsberichten wird systematisches, strategisches und institutionenübergreifendes Lernen gefördert. Leider lassen sich die in den Evaluierungsberichten getätigten Aussagen zur Nachhaltigkeit von EZ-Maßnahmen nur bedingt vergleichen. Dies hat verschiedene Gründe:

Erstens bieten die Leitfragen zur Bewertung von Nachhaltigkeit zwar eine Orientierung bei der Notenvergabe, reichen für eine Operationalisierung aber nicht aus. Dies zeigt sich daran, dass die konkreten Bewertungskriterien, die sich hinter jeder einzelnen Note verbergen, vielfältig und nicht immer eindeutig zu benennen sind. Eine gewisse Flexibilität bei der Bewertung ist zwar aufgrund des diversen Portfolios an umgesetzten Maßnahmen notwendig; dennoch muss die Nachhaltigkeitsbewertung auch für Außenstehende nachvollziehbar und vergleichbar sein. Dieser Gedanke spiegelt sich mit dem Prinzip der gemeinsamen Rechenschaftslegung auch in der Agenda 2030 wider.

Zweitens weisen die hier betrachteten DO in der Bewertungspraxis und im Evaluierungsmanagement systematische Unterschiede auf. Die Ergebnisse belegen, dass Evaluierungen der GIZ – bei gleicher Anzahl positiv bewerteter Kriterien – signifikant bessere Nachhaltigkeitsnoten vergeben als Evaluierungen der KfW. Darüber hinaus führt die Anwendung verschiedener Evaluierungstypen sowohl innerhalb als auch zwischen den DO zu strukturellen Unterschieden in der Bewertung von Nachhaltigkeit. Auch im Evaluierungsmanagement der DO zeigen sich grundlegende Unterschiede. Bei der KfW werden alle Ex-post-Evaluierungen durch die Evaluierungsabteilung inhaltlich geprüft. Dabei wird die Bewertung einzelner Maßnahmen in den Kontext der Bewertung vergleichbarer Maßnahmen gesetzt. Demgegenüber liegt die Durchführung von PFK und PEV dezentral im Verantwortungsbereich des oder der Auftragsverantwortlichen einer Maßnahme. Während bei der KfW ein Kernteam an Mitarbeitenden alle Berichte kontrolliert und somit ein Mindestmaß an Vergleichbarkeit schafft, ist ein organisationsweiter Abgleich einzelner Berichte bei der GIZ im System dezentraler Evaluierungen nicht möglich. Es ist daher zu vermuten, dass Bewertungen von GIZ-Vorhaben insgesamt heterogener sind und stärker als bei der KfW von Eigenschaften der Autoren abhängen.

Drittens sind die von den DO aufbereiteten Meta-Daten von Evaluierungen und Vorhaben nur teilweise deckungsgleich. Für die vorliegende Analyse relevante Informationen wurden teilweise unvollständig oder aber nur von einer DO systematisch erfasst.

Die unzureichende Vergleichbarkeit der Nachhaltigkeitsbewertung erschwert es, nachhaltigkeitsförderliche Faktoren zu bestimmen. So ist beispielsweise anhand der vorliegenden Informationen nicht abschließend zu klären, ob die in den Modellen integrierten makro-ökonomischen und politischen Indikatoren tatsächlich keinen Einfluss auf die Nachhaltigkeit von Vorhaben ausüben oder ob aufgrund mangelnder Vergleichbarkeit und Transparenz der Bewertungsgrundlage kein Zusammenhang festgestellt werden kann. Das Potenzial zur Gewinnung strategischer und steuerungsrelevanter Erkenntnisse aus Evaluierungssynthesen ist somit stark eingeschränkt.

Empfehlungen hinsichtlich der Stärkung des systematischen, strategischen und institutionenübergreifenden Lernens
Auch die folgenden Empfehlungen werden durch Anregungen und Gedanken, die sich vornehmlich auf die Umsetzung beziehen, ergänzt.

Um die systematische Bewertung von Nachhaltigkeit zu gewährleisten, wird dem BMZ und den DO empfohlen, einheitliche und verbindliche Kriterien zu entwickeln. Diese sollten als Grundlage der Notenvergabe dienen und hierfür transparent gewichtet werden.

- Um dabei dem heterogenen Portfolio deutscher FZ und TZ gerecht zu werden, sollte auf angemessene sektor- oder regionalspezifische Flexibilität der Kriterien geachtet werden. Ein verbindlicher Umgang mit den Kriterien könnte gegebenenfalls auch sektoral oder für TZ-/FZ-Module getrennt festgelegt werden.

Dem BMZ und den DO wird empfohlen, Meta-Daten zu Vorhaben und deren Evaluierungen – soweit möglich – zwischen den DO zu harmonisieren und zentral zu erfassen.

- Eine systematische und zentrale Erfassung der Meta-Daten von Vorhaben und Evaluierungen würde institutionsübergreifende, aggregierte Analysen erheblich erleichtern und somit beschleunigen.
- Vor diesem Hintergrund könnten das BMZ und die DO prüfen, wie den Anforderungen der gemeinsamen Rechenschaftspflicht im Sinne der Agenda 2030 durch die Erfassung und Aufbereitung von Meta-Daten Rechnung getragen werden kann.

INHALT

Abbildungen

Tabellen

ABKÜRZUNGEN UND AKRONYME

BIP
Bruttoinlandsprodukt

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

KfW
KfW Entwicklungsbank

ODA
Öffentliche Entwicklungszusammenarbeit (Official Development Assistance)

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

PEV
Projektevaluierung

PFK
Projektfortschrittskontrolle

SDGs
Nachhaltige Entwicklungsziele (Sustainable Development Goals)

TZ
Technische Zusammenarbeit

UE
Unabhängige Evaluierungen der GIZ

USD
US-Dollar

1.

EINLEITUNG

Die vorliegende Evaluierungssynthese bildet eine erste umfassende empirische Auseinandersetzung mit der Nachhaltigkeit von Vorhaben der deutschen bilateralen Entwicklungszusammenarbeit (EZ) sowie deren Einflussfaktoren. Grundlage der Betrachtung bilden Evaluierungen der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ) und der KfW Entwicklungsbank (KfW) von Vorhaben, die durch öffentliche Entwicklungsgelder des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) finanziert werden.

1.1 Hintergrund

Der Erfolg der Entwicklungszusammenarbeit bemisst sich an der Nachhaltigkeit ihrer Wirkungen. Mit der Einführung der Agenda 2030 für nachhaltige Entwicklung wurde Nachhaltigkeit zum Leitprinzip menschlichen Handelns erhoben. Die Umsetzung der Agenda 2030 obliegt der Verantwortung aller Staaten. Auch die internationale EZ muss sich neu ausrichten. Auf übergeordneter Ebene geht es dabei vor allem um Fragen der Kohärenz der EZ mit anderen Politikfeldern, der Etablierung von Partnerschaften zwischen Politik, Wirtschaft, Zivilgesellschaft und Wissenschaft sowie der Bereitstellung von Finanzmitteln zum Erreichen der in der Agenda definierten nachhaltigen Entwicklungsziele (SDGs). Auf Ebene einzelner EZ-Maßnahmen hat die Agenda 2030 Auswirkungen auf deren Konzeption, Planung und Implementierung. Dies betrifft vor allem die Frage, wie die Nachhaltigkeit der durch einzelne Maßnahmen erzielten Wirkungen im Sinne der Agenda 2030 sichergestellt werden kann. Dabei sollen unter anderem soziale, wirtschaftliche und ökologische Wechselwirkungen sowie die Inklusion benachteiligter Gruppen berücksichtigt werden. Die Planung und Umsetzung von Agenda-2030-konformen Maßnahmen ist eine zentrale Herausforderung für die internationale EZ. Um diesen Prozess zu unterstützen, sind evidenzbasierte Handlungsempfehlungen nötig. Nach Wissen des Evaluierungsteams gibt es derzeit nur eine begrenzte Anzahl an Vorhaben, die explizit in Anlehnung an die Agenda 2030 und deren Prinzipien konzipiert wurden. Eine empirische Auseinandersetzung mit der Nachhaltigkeit von EZ-Maßnahmen ist dennoch möglich.

Bereits seit 2006 wird die Nachhaltigkeit von Vorhaben der deutschen EZ systematisch überprüft und benotet; in dem Jahr verabschiedete das Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ) eine Orientierungshilfe, die verbindliche Leitfragen für die Bewertung der Nachhaltigkeit einzelner EZ-Maßnahmen enthält (BMZ, 2006).[1] Nachhaltigkeit soll demnach anhand der Dauerhaftigkeit entwicklungspolitischer Wirkungen, der Stabilität des Umfelds im Hinblick auf soziale Gerechtigkeit, wirtschaftliche Leistungsfähigkeit, politische Stabilität und ökologisches Gleichgewicht sowie der Risiken und Potenziale für die (fortdauernde) Wirksamkeit überprüft werden (BMZ, 2006). Die Nachhaltigkeit von Vorhaben wird dabei konzeptionell in enger Verbindung zur entwicklungspolitischen Wirksamkeit beurteilt. Eine diese Evaluierungssynthese begleitende Meta-Evaluierung zeigt, dass die Bewertung von Nachhaltigkeit in der Praxis tatsächlich über mehrere Evaluierungskriterien hinweg erfolgt und dass Nachhaltigkeit dementsprechend umfassend verstanden, evaluiert und bewertet wird (Noltze et al., 2018). Daher kann erwartet werden, dass die bisherige Bewertungspraxis bereits einige der Prinzipien des Konzeptes nachhaltiger Entwicklung im Sinne der Agenda 2030 abdeckt. Eine systematische Analyse der Einflussfaktoren auf die vergebene Nachhaltigkeitsnote bietet somit die Möglichkeit, relevante Erkenntnisse für die Ausgestaltung von EZ-Maßnahmen im Agenda-2030-Zeitalter zu gewinnen.

1.2 Ziel der Evaluierungssynthese

Ziel der vorliegenden Evaluierungssynthese ist die umfassende und systematische Auseinandersetzung mit der Bewertung von Nachhaltigkeit deutscher EZ-Vorhaben. Durch die Identifizierung wichtiger Einflussfaktoren auf die vergebene Nachhaltigkeitsnote sollen Ansatzpunkte zur Steigerung der Nachhaltigkeit deutscher EZ-Vorhaben im Sinne der Agenda 2030 herausgearbeitet werden. Hierbei wird mit Hilfe statistischer Modelle ermittelt, inwieweit Faktoren auf Ebene der Evaluierungsberichte, der evaluierten Vorhaben und des Landes, in dem das Vorhaben umgesetzt wurde, Einfluss auf die vergebene Nachhaltigkeitsnote haben.

[1] *Daneben sind auch verbindliche Vorgaben für die Bewertung der Evaluierungskriterien Relevanz, Effektivität, Effizienz und Impact enthalten.*

Neben der Nutzung von Meta-Daten der Vorhaben und Evaluierungsberichte – etwa Dauer und Finanzvolumen von Vorhaben – greift die Analyse dabei auch auf die Ergebnisse der begleitenden Meta-Evaluierung zum Thema Nachhaltigkeit zurück (siehe Noltze et al., 2018). Auf Grundlage von Evaluierungsberichten der GIZ und KfW werden in der Meta-Evaluierung jene Kriterien erfasst, die in den Berichten zur Bewertung von Nachhaltigkeit herangezogen wurden.[2] Die hier durchgeführte Evaluierungssynthese gibt Gelegenheit, das Zusammenspiel verschiedener Determinanten besser zu verstehen. Dadurch kann sie dazu beitragen, die deutsche EZ im Kontext der Agenda 2030 sowohl strategisch als auch operativ stärker auf multidimensionale Nachhaltigkeit auszurichten. Dies trägt der gesteigerten Bedeutung von Nachhaltigkeit im Kontext der SDG-konformen Evaluierung von Vorhaben in der deutschen EZ Rechnung. Dabei wird davon ausgegangen, dass die Bewertung von Nachhaltigkeit durch Evaluierungen zwar kein objektiver Maßstab für die Nachhaltigkeit deutscher EZ-Vorhaben ist, jedoch eine bestmögliche Annährung an diese bietet. Den Anlass einer solchen Auseinandersetzung bildet die Einführung der Agenda 2030 für nachhaltige Entwicklung und die damit einhergehende Betonung der Nachhaltigkeit als zentrales Element der Wirksamkeitsdebatte.

1.3 Gegenstand

Den Gegenstand der Evaluierungssynthese bilden die Nachhaltigkeit von Vorhaben der deutschen EZ sowie deren Einflussfaktoren. Bei der vorliegenden Untersuchung geht es dabei ganz konkret um eine Auseinandersetzung mit der aggregierten Nachhaltigkeitsbewertung in Evaluierungen der deutschen finanziellen und technischen EZ. In der Evaluierungspraxis erfolgt die Bewertung anhand einer Note, deren Einflussfaktoren mittels statistischer Analysen untersucht werden. 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.

Grundsätzlich beschränkt sich diese erste systematische Untersuchung der Nachhaltigkeit von Vorhaben jedoch auf die Evaluierungs- und Bewertungspraxis der beiden großen staatlichen DO – der KfW und der GIZ.[3] Die beiden DO setzen jährlich einen wesentlichen Anteil der öffentlichen Entwicklungsfinanzierung um und verfügen über ein hoch diversifiziertes sektorales und regionales Portfolio. Gleichzeitig weisen beide DO einen hohen Deckungsgrad an Evaluierungen von Einzelvorhaben (den heutigen Modulen) auf. Nachhaltigkeit als Erfolgskriterium der deutschen EZ wird dabei seit 2006 in allen Evaluierungen bewertet. Die Bewertung basiert auf der BMZ-Orientierungshilfe zum Umgang mit den DAC-Kriterien. In der vorliegenden Evaluierungssynthese werden daher nur Evaluierungen berücksichtigt, die zwischen Juli 2006 und dem Zeitpunkt der Datenerhebung im Oktober 2017 durchgeführt und abgeschlossen wurden.

Bei der Ermittlung der Einflussfaktoren ergibt sich eine Eingrenzung des Gegenstandes auch durch die begrenzte Datenverfügbarkeit. Dabei beschränkt sich die Untersuchung auf Merkmale der Vorhaben, Faktoren aus dem Bereich der Implementierung der Vorhaben und verfügbare Kontextinformationen. Zu den Kontextfaktoren gehören sowohl Merkmale des unmittelbaren Kontextes der Entwicklungsmaßnahmen als auch makroquantitative Indikatoren auf der Ebene der Partnerländer. Die Erhöhung der Datenverfügbarkeit und die damit einhergehende Erweiterung des Evaluierungsgegenstandes wurde durch die begleitende Meta-Evaluierung zur Evaluierungs- und Bewertungspraxis der Nachhaltigkeit von Vorhaben unterstützt (Noltze et al., 2018).

[2] *Das methodische Vorgehen sowie die Ergebnisse der Meta-Evaluierung sind in Noltze et al. (2018) dargestellt.*
[3] *Andere staatliche DO, wie die Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) und die Physikalisch-Technische Bundesanstalt (PTB), sind nicht Teil der Betrachtung.*

1.4 Evaluierungsfragen

Die Evaluierungsziele werden durch fünf Evaluierungsfragen operationalisiert:

Evaluierungsfrage 1: Welche Besonderheiten zeigen sich bei der Gesamtbetrachtung von Nachhaltigkeit im Portfolio der Evaluierungen im Rahmen der deutschen EZ?

Evaluierungsfrage 2: Inwieweit beeinflussen programm- und projektspezifische Faktoren die Nachhaltigkeitsbewertung von Vorhaben?

Evaluierungsfrage 3: Inwieweit beeinflussen kontextspezifische Faktoren die Nachhaltigkeitsbewertung von Vorhaben?

Evaluierungsfrage 4: Inwieweit beeinflussen die zugrunde liegenden Bewertungskriterien die Nachhaltigkeitsbewertung von Vorhaben?

Evaluierungsfrage 5: Inwieweit beeinflusst die methodische Qualität der Evaluierungen die Nachhaltigkeitsbewertung?

1.5 Aufbau des Evaluierungsberichtes

Die Evaluierungssynthese gliedert sich wie folgt:

In Kapitel 2 wird die Evaluierungs- und Bewertungspraxis der Nachhaltigkeit von Vorhaben der deutschen EZ vorgestellt (Kapitel 2.1 und Kapitel 2.3). Dort werden auch mögliche Einflussfaktoren auf die Nachhaltigkeit identifiziert sowie ihr theoretischer Zusammenhang mit der Nachhaltigkeit von Vorhaben diskutiert (Kapitel 2.2). Das Kapitel schließt mit der Vorstellung der Datengrundlage (Kapitel 2.4) und der Vorstellung des Stichprobenplans der vorliegenden Untersuchung (Kapitel 2.5).

Kapitel 3 beschreibt die methodische Vorgehensweise der Evaluierung. Neben der empirischen Strategie (Kapitel 3.1) werden dort auch unterschiedliche Formen der statistischen Modellierung diskutiert (Kapitel 3.2) und Limitationen beziehungsweise Herausforderungen aufgezeigt (Kapitel 3.3).

Die Ergebnisse der Evaluierungssynthese werden in Kapitel 4 vorgestellt. Das Ergebniskapitel beginnt mit der Beschreibung der erklärenden Variablen (Kapitel 4.1) und der Diskussion der Nachhaltigkeitsnote als abhängige Variable der Untersuchung (Kapitel 4.2). Schließlich werden die Ergebnisse entlang der Evaluierungsfragen vorgestellt (Kapitel 4.3).

Die Schlussfolgerungen und Empfehlungen finden sich in Kapitel 5.

2.

NACHHALTIGKEIT IN DER DEUTSCHEN EZ

Im folgenden Kapitel wird zunächst erörtert, an welchen Leitfragen sich die Bewertung des Kriteriums Nachhaltigkeit orientiert. In diesem Zusammenhang wird auf die Grenzen der Nachhaltigkeitsbewertung hingewiesen. Auch werden mögliche Einflussfaktoren auf die vergebene Nachhaltigkeitsnote diskutiert. Im Anschluss wird die Evaluierungspraxis von GIZ und KfW erläutert. Abschließend wird die Datengrundlage der Evaluierung vorgestellt und die Verteilung der vergebenen Nachhaltigkeitsnote über das Berichtsportfolio abgebildet.

2.1
Bewertung von Nachhaltigkeit in Vorhaben der deutschen EZ

Die Nachhaltigkeit von Vorhaben der deutschen staatlichen EZ wird seit 2006 in allen Evaluierungen des BMZ und seiner DO systematisch bewertet. Die Überprüfung erfolgt auf Grundlage einer Orientierungshilfe für Evaluierungen des BMZ und der DO (BMZ, 2006). In Anlehnung an die vom OECD-DAC beschlossenen Prinzipien zur Evaluierung der Entwicklungszusammenarbeit (OECD, 1991) enthält die Orientierungshilfe Vorgaben zur Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz, übergeordnete entwicklungspolitische Wirkungen (Impact) und Nachhaltigkeit.

Die Nachhaltigkeit eines Vorhabens soll demnach entlang von drei zentralen Aspekten geprüft werden: erstens anhand der Dauerhaftigkeit entwicklungspolitischer Wirkungen der Entwicklungsmaßnahme; zweitens anhand der Stabilität des Umfelds der Entwicklungsmaßnahme bezüglich der Faktoren soziale Gerechtigkeit, wirtschaftliche Leistungsfähigkeit, politische Stabilität und ökologisches Gleichgewicht; und drittens anhand der Risiken und Potenziale für die nachhaltige Wirksamkeit der Entwicklungsmaßnahme (BMZ, 2006).[4]

Als Ergebnis dieser Prüfung wird eine Note von 1 bis 4 vergeben.[5] Ein Vorhaben erhält die Note 1, wenn die (bisher positive) entwicklungspolitische Wirksamkeit des Vorhabens mit hoher Wahrscheinlichkeit unverändert fortbestehen oder zunehmen wird. Die Note 2 erhält ein Vorhaben, wenn dessen 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 wird, sich aber mit hoher Wahrscheinlichkeit positiv entwickeln wird. Die Note 4 wird vergeben, wenn die entwicklungspolitische Wirksamkeit als nicht ausreichend eingeschätzt wird und sich mit hoher Wahrscheinlichkeit auch nicht verbessern wird. Ein Vorhaben wird als „nachhaltig" bewertet, wenn es eine Note zwischen 1 und 3 erhält. Als „nicht nachhaltig" gelten Vorhaben mit der Note 4.[6]

Bei genauerer Betrachtung der einzelnen Notenstufen zeigen sich zwei Auffälligkeiten: Erstens werden Vorhaben mit der Notenstufe 3 zwar formal als „nachhaltig" gewertet, doch ist die Note 3 inhaltlich gleichbedeutend mit einer nicht ausreichenden oder deutlich rückläufigen entwicklungspolitischen Wirksamkeit einer Maßnahme. Bei strengerer Auslegung dieser Definition könnten Vorhaben mit der Note 3 ebenso als „nicht nachhaltig" bewertet werden. Zweitens verdeutlichen die Definitionen aller Notenstufen, dass ein konzeptioneller Zusammenhang zwischen der entwicklungspolitischen Wirksamkeit und der Nachhaltigkeit von Vorhaben besteht. Ohne entwicklungspolitische Wirksamkeit eines Vorhabens kann es keine Nachhaltigkeit geben. Allerdings ist dieser Zusammenhang bislang eher impliziter Bestandteil der Konzeption der DAC-Kriterien. Aus ihm allein ergibt sich noch keine klare Handlungsanleitung für den Umgang mit Nachhaltigkeit in Evaluierungen. Vor diesem Hintergrund hat die begleitende Meta-Evaluierung empirisch überprüft, wie Nachhaltigkeit in der Praxis tatsächlich verstanden, evaluiert und bewertet wird (Noltze et al., 2018). Dabei hat sich gezeigt, dass die Nachhaltigkeit von Vorhaben in Evaluierungen einerseits umfassend und vielschichtig, andererseits jedoch auch unsystematisch und uneinheitlich untersucht, diskutiert und bewertet wird. Dieses Ergebnis zeigt, dass die in Evaluierungen vergebene Nachhaltigkeitsnote viel mehr Informationen enthält, als die

4 *Diese Prüffragen gehen über das Nachhaltigkeitsverständnis des OECD-DAC hinaus, in dem Nachhaltigkeit vor allem als Fortdauer von Wirkungen nach Vorhabenende definiert wird. Die vollständige Definition der OECD-DAC-Kriterien findet sich unter: http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm*

5 *In sogenannten Projektevaluierungen (PEV) der GIZ, die im April 2014 eingeführt wurden, wird Nachhaltigkeit auf einer sechsstufigen Skala bewertet.*

6 *In der Gesamtbewertung ist ein Vorhaben nur dann „erfolgreich", wenn die Bewertung der Nachhaltigkeit mindestens zufriedenstellend (Note 3) ist. Dies gilt ebenfalls für die Kriterien „Effektivität" und übergeordnete entwicklungspolitische Wirkungen (Impact).*

im BMZ-Leitfaden enthaltenen Prüffragen zunächst vermuten lassen. Für die vorliegende Evaluierungssynthese ist dieses Ergebnis von hoher Relevanz, denn die Untersuchung des mehrdimensionalen Konzepts der Nachhaltigkeit verlangt letztlich die Berücksichtigung vielfältiger Einflussfaktoren und geht somit insgesamt mit einem hohen Datenbedarf einher. Aus diesem Grund bezieht die Evaluierungssynthese auch zusätzliche Informationen aus der begleitenden Meta-Evaluierung in die Analyse ein.

2.2
Einflussfaktoren auf die Nachhaltigkeitsnote

Was macht EZ-Vorhaben nachhaltig? Die vorhandene Literatur zu Nachhaltigkeit in der EZ beantwortet diese Frage nur bedingt. Sie beinhaltet vor allem eine übergeordnete, konzeptionelle Diskussion und richtet ihren Blick weniger auf die Nachhaltigkeit einzelner Vorhaben. So werden darin vor allem die Bedeutung von Nachhaltigkeit für die EZ sowie die Herausforderungen nachhaltiger Entwicklung behandelt. Im Sammelband von König und Thema (2011) mit dem Titel „Nachhaltigkeit in der Entwicklungszusammenarbeit“ werden beispielsweise die Bedeutung von Nachhaltigkeit für die EZ verdeutlicht, die Konzepte „Nachhaltigkeit“ und „nachhaltige Entwicklung“ kritisch diskutiert, die Rolle der globalen Finanz- und Handelsordnung für nachhaltige Entwicklung hervorgehoben, Fragen der Kohärenz der Entwicklungspolitik erörtert sowie Erfahrungen bei der Evaluierung von Nachhaltigkeit in Maßnahmen der deutschen FZ beleuchtet. Caspari (2004) verdeutlicht die Komplexität des Evaluierungskriteriums Nachhaltigkeit und erarbeitet einen konzeptionellen Rahmen für eine einheitliche Nachhaltigkeitsbewertung. Im Sammelband von Von Raggamby und Rubik (2012) wird unter anderem diskutiert, wie die Evaluierung von Nachhaltigkeit zur Formulierung von Politiken beitragen kann. Außerdem werden politikrelevante Indikatoren sowie Methoden zur Evaluierung von Nachhaltigkeit dargestellt. Auch werden Qualitätsanforderungen von Evaluierungen für die Bewertung von Nachhaltigkeit deutlich gemacht. Eine Reihe neuerer Beiträge rückt die Bedeutung von Evaluierung für die Umsetzung der Agenda 2030 in den Fokus. Demnach sollen nationale Politiken zum Erreichen der SDGs durch nationale Evaluierungssysteme überprüft werden (Benoit et al., 2017; Ofir et al., 2016). Die Evaluierungsagenda einzelner Staaten soll dabei einen ganzheitlichen Ansatz verfolgen und Politiken und Projekte nicht isoliert, sondern im nationalen Kontext beurteilen (Ofir et al., 2016). Um der Komplexität der Agenda 2030 gerecht zu werden, soll die Überprüfung über ein reines Monitoring von Indikatoren hinausgehen. Insbesondere rigorose Wirkungsevaluierungen sollen aufdecken, warum, wie und unter welchen Bedingungen Politiken wirken und welche Gruppen von ihnen profitieren (Lucks et al., 2016; Schwandt et al., 2016). Außerdem soll durch die Einbeziehung eines breiten Spektrums von Akteuren ein länderspezifischer Fokus auf einzelne Indikatoren gelegt werden (Lucks et al., 2016).

Nach Wissen des Evaluierungsteams gibt es bislang keine empirischen Erkenntnisse zu den Faktoren, die die Nachhaltigkeitsbewertung einzelner Vorhaben beeinflussen. Allerdings gibt es eine Reihe von Studien, die die Einflussfaktoren auf die Bewertung des Gesamterfolges von Vorhaben der Weltbank sowie der Afrikanischen und Asiatischen Entwicklungsbank analysieren. In der Gesamtschau der Studien zeigt sich, dass der Erfolg einer Maßnahme vor allem von den Eigenschaften eines Vorhabens und dessen Umsetzungsmodalitäten, den Charakteristika der Evaluierung eines Vorhabens sowie von Kontextfaktoren auf Landesebene beeinflusst wird (Assefa et al., 2014; Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Kilby, 2013). Da der Gesamterfolg einer Maßnahme entscheidend von ihrer Nachhaltigkeit abhängt, wird geprüft, inwiefern die ermittelten Einflussfaktoren auch für die Bewertung von Nachhaltigkeit relevant sind.

Hinsichtlich der Einflussfaktoren auf Landesebene zeigen Denizer et al. (2013), dass der Erfolg einer Maßnahme positiv vom wirtschaftlichen Entwicklungsstand und der wirtschaftlichen Stabilität eines Landes beeinflusst wird. Daher wird geprüft, ob wirtschaftliche Entwicklung ebenfalls positive Auswirkungen auf die Nachhaltigkeit einer Maßnahme hat. So kann eine positive wirtschaftliche Entwicklung zu steigenden Staatseinnahmen führen, was wiederum den Spielraum für finanzielle oder personelle Beiträge des Partnerlandes bei der Umsetzung von Vorhaben erhöht (Bulman et al., 2015; Denizer et al., 2013; Hemmer und Lorenz, 2003). Auch die politischen Rechte und zivilen Freiheiten einer Gesellschaft korrelieren positiv mit dem Erfolg von Vorhaben (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995).

Darüber hinaus ist ein höheres Niveau von Rechtsstaatlichkeit und Demokratie innerhalb eines Landes förderlich für den Gesamterfolg der im Land umgesetzten Vorhaben (Chauvet et al., 2010; Denizer et al., 2013; Dollar und Levin, 2005). Rechtsstaatlichkeit begünstigt Investitionen, da ein erhöhtes Maß an Vertrauen zwischen verschiedenen Akteuren geschaffen und Transaktionskosten gesenkt werden (Dollar und Levin, 2005). Funktionierende demokratische Institutionen fördern die Rechenschaftspflicht von Regierungen gegenüber ihrer Bevölkerung. Diese stehen demnach unter dem Druck der Wählerinnen und Wähler, was ihr Interesse an der Umsetzung von wirksamen Vorhaben erhöht (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995). Es scheint plausibel, dass Rechtsstaatlichkeit und Demokratieniveau auch positiv auf die Nachhaltigkeit von Vorhaben wirken.

Die Ergebnisse mehrerer Studien zeigen, dass Faktoren auf Ebene der Vorhaben im Vergleich zu Faktoren auf Landesebene einen relativ stärkeren Einfluss auf den Gesamterfolg von Vorhaben haben (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995). Wichtige Einflussfaktoren für den Gesamterfolg einer Maßnahme sind demnach das Finanzvolumen sowie die Laufzeit und die sektorale Zuordnung einer Maßnahme. Dabei ist anzumerken, dass längere und kostspieligere Vorhaben nicht zwangsläufig besser bewertet werden (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995). Möglicherweise besteht ein Zusammenhang zwischen Dauer und Finanzvolumen eines Vorhabens und dessen Komplexität. So könnte die Gesamtbewertung vor allem durch die Komplexität der Zielsysteme bedingt sein (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995). Eine längere Vorbereitungszeit vor Umsetzung einer Maßnahme (Dollar und Levin, 2005; Kilby, 2013) sowie eine erhöhte Managementkompetenz erhöhen deren Erfolgsaussichten (Chauvet et al., 2010; Denizer et al., 2013; Dollar und Levin, 2005). Dagegen wirkt sich eine Verzögerung der Implementierung der Maßnahme negativ auf den Gesamterfolg aus (Chauvet et al., 2010; Denizer et al., 2013; Dollar und Levin, 2005).

Die Bewertung des Gesamterfolges wird auch durch Merkmale der Evaluierung bestimmt. Denizer et al. (2013) zeigen, dass die vergebene Note mit zunehmendem zeitlichen Abstand zwischen Vorhabenende und Evaluierungsdatum schlechter wird. Bezogen auf die Nachhaltigkeit von Vorhaben ist ein ähnlicher Zusammenhang plausibel: Je später nach Vorhabenende die Wirkung evaluiert wird, desto wahrscheinlicher ist ein Rückgang dieser Wirkung. Darüber hinaus besteht möglicherweise ein Zusammenhang zwischen Benotungspraxis und methodischer Berichtsqualität. Es ist nicht auszuschließen, dass methodisch bessere (schlechtere) Evaluierungen die Nachhaltigkeit von Vorhaben kritischer (unkritischer) prüfen und somit eine schlechtere (bessere) Bewertung abgeben.

Alle bisher genannten Faktoren bilden ausschließlich den Einfluss des Implementierungskontextes und der deskriptiven Charakteristika eines Vorhabens und seiner Evaluierung ab. Die Bewertung von Nachhaltigkeit wird jedoch auch durch inhaltliche Ergebnisse eines Vorhabens und seiner Implementierung bestimmt. In den hier zitierten Studien werden diese Aspekte bisher nur bedingt berücksichtigt. Dies ist vermutlich der mangelnden Verfügbarkeit relevanter Informationen geschuldet. Daten zu inhaltlichen Errungenschaften einzelner EZ-Maßnahmen, die im Zusammenhang mit deren Nachhaltigkeitsbewertung stehen, lassen sich nur unmittelbar aus Projektdokumenten ableiten. Um diese Lücke zu schließen, greift die vorliegende Evaluierung auf die Ergebnisse der begleitenden Meta-Evaluierung zurück (Noltze et al., 2018). In dieser wurde ein konzeptioneller Analyserahmen zur Erfassung der für die Nachhaltigkeitsbewertung herangezogenen Kriterien entwickelt. Danach wird die Nachhaltigkeitsbewertung maßgeblich durch sieben Bereiche bestimmt: den Kontext der Maßnahme, ihre Implementierung, den erzielten Wirkungen/Outcome, die Kapazitäten vor Ort, nicht intendierte Wirkungen (Impact)[7] der Maßnahme, die Absehbarkeit des Erhalts von Wirkungen sowie das Zusammenspiel der Dimensionen.[8]

[7] *Zu dem Bereich „Impact" zählen grundsätzlich sowohl „intendierte" als auch „nicht intendierte" Wirkungen. Da die „intendierten Wirkungen" jedoch integraler Bestandteil der Bewertung des OECD-DAC Kriteriums „Impact" sind, beschäftigt sich die vorliegende Untersuchung nur mit den „nicht-intendierten" Wirkungen, denen konzeptionell eine besondere Rolle in der Nachhaltigkeitsbewertung zukommt. Eine Auseinandersetzung mit den intendierten Wirkungen findet sich in Noltze et al. (2018).*

[8] *Für eine ausführliche Diskussion des Analyserahmens der Nachhaltigkeitsbewertung siehe Noltze et al. (2018).*

2.3 Evaluierungspraxis von GIZ und KfW

Die zentralen inhaltlichen Prüfkriterien der Evaluierungen von GIZ und KfW sind seit 2006 durch die Orientierungshilfe des BMZ verbindlich vorgegeben (BMZ, 2006). Die Wahl des konkreten Berichtsformats und die Umsetzung der Evaluierungen liegen jedoch in der Verantwortung der jeweiligen DO. Bei der Bewertung einzelner Vorhaben verwenden GIZ und KfW unterschiedliche Evaluierungstypen.

Die GIZ hat für die Evaluierung einzelner Vorhaben ab 2006 zentrale und dezentrale Evaluierungen genutzt. Die zentralen Evaluierungen wurden von der Stabsstelle Evaluierung der GIZ gesteuert und bis einschließlich 2014 angewandt. Ihre Umsetzung erfolgte unabhängig von der Durchführung der zu evaluierenden Vorhaben (Unabhängige Evaluierungen, UE). Zu den UE gehörten sogenannte Ex-ante-, Zwischen-, Schluss- und Ex-post-Evaluierungen. Die Ex-ante- und Zwischen-Evaluierungen wurden vor bzw. im Verlauf einer Maßnahme durchgeführt, während die Schluss-Evaluierungen in der Regel sechs Monate vor bzw. nach Vorhabenende und Ex-post-Evaluierungen zwei bis fünf Jahre nach Vorhabenende erstellt wurden. UE wurden im jährlichen Wechsel für Maßnahmen eines bestimmten Sektors durchgeführt. Die dezentralen Evaluierungen beinhalten sogenannte Projektfortschrittskontrollen (PFK), die bis März 2014 zur Anwendung kamen. Diese werden seit April 2014 durch sogenannte Projektevaluierungen (PEV) ersetzt. PEV sind der nunmehr einzig verbleibende Evaluierungstyp zur Bewertung einzelner Maßnahmen. Im Gegensatz zu den zentralen Evaluierungstypen liegt die Durchführungsverantwortung dezentraler Evaluierungen bei den Auftragsverantwortlichen der einzelnen Vorhaben. PFK und PEV kommen sechs bis zwölf Monate vor Ende der Vorhaben zum Einsatz.[9]

Im Unterschied zur GIZ evaluiert die KfW einzelne Maßnahmen seit 2006 durchgängig mit Ex-post-Evaluierungen. Ex-post-Evaluierungen der KfW werden in der Regel drei bis fünf Jahre nach Vorhabenende durchgeführt. In der KfW werden die Evaluierungen von der unabhängigen Evaluierungseinheit der KfW-Entwicklungsbank organisiert. Die Auswahl der Vorhaben für eine Evaluierung basiert seit 2006 auf einer jährlich festzulegenden Stichprobe von abgeschlossenen Vorhaben, in die jeweils die Hälfte der Vorhaben eines Sektors einbezogen wird.[10]

Hinsichtlich ihrer Nachhaltigkeitsbewertung sind die einzelnen Evaluierungsformate nur bedingt miteinander vergleichbar. PFK, PEV und Schluss-Evaluierungen werden unmittelbar zu Ende eines Vorhabens durchgeführt. Die Bewertung der Nachhaltigkeit erreichter Wirkungen eines Vorhabens ist de facto eine Einschätzung zukünftiger Entwicklungen. Im Gegensatz dazu basiert die Nachhaltigkeitsbewertung in Ex-post-Evaluierungen auf Beobachtungen und tatsächlichen Entwicklungen, die mindestens drei Jahre über das Ende eines Vorhabens hinausgehen. Diese Unterschiede müssen bei der Analyse der Einflussfaktoren auf die Nachhaltigkeit von Vorhaben berücksichtigt werden.

2.4 Datengrundlage und Portfolioanalyse

Die Datengrundlage (Beobachtungen) des vorliegenden Berichtes besteht aus den Vorhaben der GIZ und KfW, die zwischen 2006 und 2016 entlang der DAC-Kriterien bewertet wurden.[11] Zum Zeitpunkt der Datenerhebung im Oktober 2016 flossen insgesamt 1.015 evaluierte Maßnahmen in die Grundgesamtheit ein.[12] Davon stammen 462 aus dem Bereich der finanziellen Zusammenarbeit (KfW) und 553 aus dem Bereich der technischen Zusammenarbeit (GIZ). Während es sich bei allen Evaluierungen der KfW um Ex-post-Evaluierungen handelt, unterteilen sich die Evaluierungen der GIZ in 56 Ex-post- und 44 Schluss-Evaluierungen, 110 PEV sowie 343 PFK. Neben bilateralen Vorhaben sind in der Grundgesamtheit auch sogenannte Sektor-, Regional- und Globalvorhaben enthalten.[13]

Abbildung 1 zeigt die vergebene Nachhaltigkeitsnote nach DO für alle in der Grundgesamtheit enthaltenen Evaluierungen.[14]

9 *Für eine ausführliche Beschreibung des Evaluierungssystems der GIZ siehe: https://www.giz.de/de/ueber_die_giz/265.html*

10 *Für eine ausführliche Beschreibung des Evaluierungssystems der KfW Entwicklungsbank siehe: https://www.kfw-entwicklungsbank.de/Internationale-Finanzierung/KfW-Entwicklungsbank/Evaluierungen/*

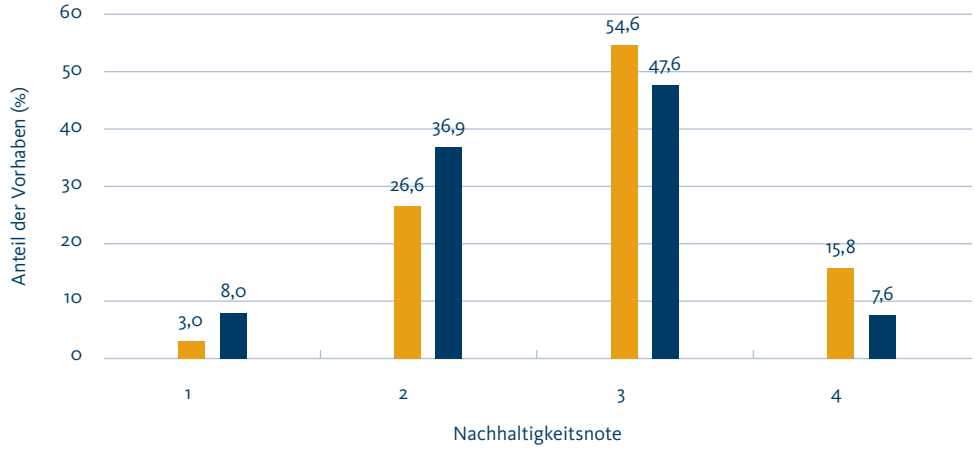
11 *Da Ex-ante- und Zwischen-Evaluierungen relativ früh während eines laufenden Vorhabens durchgeführt werden, erscheinen sie ungeeignet, die Nachhaltigkeit im Sinne der Dauer und Stabilität von Wirkungen zu bewerten. Beide Evaluierungstypen wurden daher aus der Grundgesamtheit ausgeschlossen.*

12 *Es ist 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. kein 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.*

13 *Die Grundgesamtheit enthält 99 Regionalvorhaben (87 der GIZ, 12 der KfW), 52 Sektorvorhaben (35 der GIZ, 17 der KfW) sowie 6 Globalvorhaben (der GIZ).*

14 *Zusätzlich zu den im Folgenden gezeigten Abbildungen werden in Tabelle 8 die Merkmale der Grundgesamtheit nach DO beschrieben.*

Abbildung 1: Vergebene Nachhaltigkeitsnote nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind die relativen Häufigkeiten der vergebenen Nachhaltigkeitsnoten nach DO. Die Abbildung enthält alle evaluierten Vorhaben der Grundgesamtheit. N = 1.015

Die Abbildung ist vor dem Hintergrund der in Abschnitt 2.1 beschriebenen Notenstufen zu interpretieren. Demnach wird einem Vorhaben bei einer Bewertung bis zur Notenstufe 3 bescheinigt, dass dessen positive entwicklungspolitische Wirkungen entweder auf absehbare Zeit überwiegen werden oder nachweislich nach Ende eines Vorhabens bestehen bleiben. Diese Einschätzung wird bei 93 Prozent aller GIZ-Vorhaben und bei 85 Prozent aller KfW-Vorhaben getroffen. Das heißt, dass rund neun von zehn EZ-Vorhaben durch deren Evaluierungen als „nachhaltig" eingestuft werden.[15]

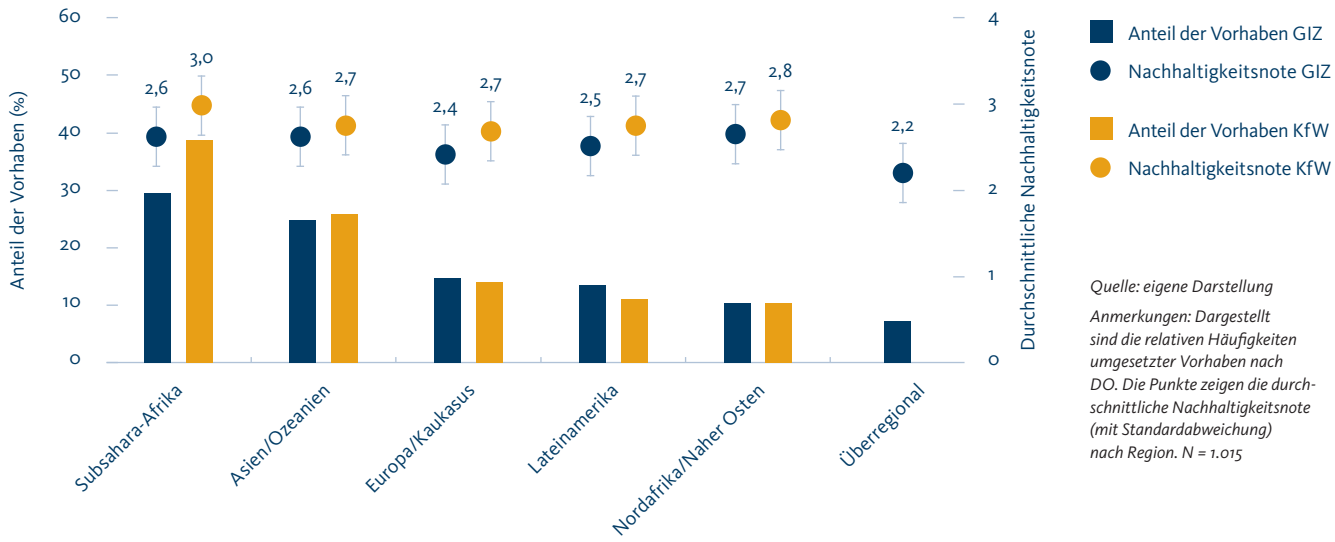
Das hier dargestellte Portfolio von GIZ und KfW enthält Vorhaben aus vier Kontinenten und zehn Sektoren. Abbildung 2 zeigt die Verteilung dieser Vorhaben über verschiedene Regionen sowie die durchschnittlich vergebene Nachhaltigkeitsnote je Region und DO. Die Balken zeigen die relative Häufigkeit der umgesetzten Vorhaben, die Punkte stellen die durchschnittlich vergebene Note dar. Demnach setzten beide DO die Mehrheit ihrer Vorhaben in Subsahara-Afrika um. Der Anteil der KfW-Vorhaben in Afrika ist im Vergleich zur GIZ signifikant größer.[16] Vorhaben in den Regionen Asien/Ozeanien, Europa/Kaukasus, Lateinamerika und Nordafrika machen einen ähnlichen Anteil am Portfolio beider DO aus. Die GIZ bearbeitet einen geringen Teil ihrer Vorhaben auf globaler, d. h. regionsübergreifender Ebene (Sektor- und Globalvorhaben).

Im Hinblick auf die Bewertung der Nachhaltigkeit der Vorhaben zeigt sich, dass die durchschnittlich vergebene Nachhaltigkeitsnote für Maßnahmen der KfW im Vergleich zu Maßnahmen der GIZ über alle Regionen hinweg schlechter ist. Statistisch signifikante Unterschiede zwischen den Benotungen beider DO bestehen allerdings nur in den Regionen Subsahara-Afrika und Europa/Kaukasus.[17] Innerhalb des Portfolios der GIZ erhalten überregionale Vorhaben die beste Nachhaltigkeitsbewertung. Diese werden im Vergleich zu GIZ-Vorhaben aus Subsahara-Afrika, Asien/Ozeanien und Nordafrika/Naher Osten signifikant besser bewertet.[18] Überregionale Vorhaben unterscheiden sich von bilateralen Maßnahmen insofern, als sie keinem konkreten Partnerland zugeordnet werden können. Sie sind daher weniger abhängig von Umsetzungsstrukturen. Innerhalb des Portfolios der KfW werden Vorhaben in Subsahara-Afrika signifikant schlechter bewertet als in Vorhaben in Europa/Kaukasus bzw. in Asien/Ozeanien.[19]

Abbildung 3 zeigt die sektorale Verteilung der Maßnahmen sowie die je Sektor durchschnittlich vergebene

[15] *Kennzeichnet man – wie in Kapitel 2.1 beschrieben – Vorhaben mit der Notenstufe 3 als „nicht nachhaltig", sind nur rund 45 Prozent aller GIZ- und rund 30 Prozent aller KfW-Vorhaben „nachhaltig".*
[16] *Die Aussage basiert auf einem Zwei-Gruppen-Proportionen-Test.*
[17] *Aufgrund der Normalverteilung der Note sowie homogener Varianz innerhalb der Region beruht die Aussage auf einer Varianzanalyse (ANOVA).*
[18] *Die Aussage beruht auf einer Varianzanalyse (ANOVA).*
[19] *Die Aussage beruht auf einer Varianzanalyse (ANOVA).*

Abbildung 2: Regionale Verteilung der Vorhaben und deren Nachhaltigkeitsbewertung nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind die relativen Häufigkeiten umgesetzter Vorhaben nach DO. Die Punkte zeigen die durchschnittliche Nachhaltigkeitsnote (mit Standardabweichung) nach Region. N = 1.015

Nachhaltigkeitsnote. Erneut wird bei der Darstellung zwischen Vorhaben der GIZ und der KfW unterschieden. Die Ergebnisse zeigen, dass für beide DO die Sektoren Wirtschaft und Demokratie von besonderer Bedeutung sind. Außerdem werden relativ häufig Vorhaben in den Sektoren Wasser und Gesundheit umgesetzt. Im Hinblick auf das Sektor-Portfolio bestehen signifikante Unterschiede zwischen GIZ und KfW.[20] So ist der Anteil der GIZ-Vorhaben in den Sektoren Wirtschaft, Demokratie, Umwelt und Bildung im Vergleich zum Portfolio der KfW signifikant größer. Die KfW hingegen setzt einen signifikant höheren Anteil ihrer Vorhaben in den Sektoren Wasser, Gesundheit, Energie, Landwirtschaft und Transport um. Diese Differenzen spiegeln die unterschiedlichen Kernkompetenzen beider DO wider. So betreibt die GIZ Vorhaben der technischen Zusammenarbeit (TZ) und ist in der Regel aktiv an der Umsetzung im Partnerland beteiligt. Die KfW hingegen betreibt vorwiegend finanzielle Zusammenarbeit (FZ) und konzentriert sich größtenteils auf (die Förderung von) Investitionen und den Dialog mit Partnern.

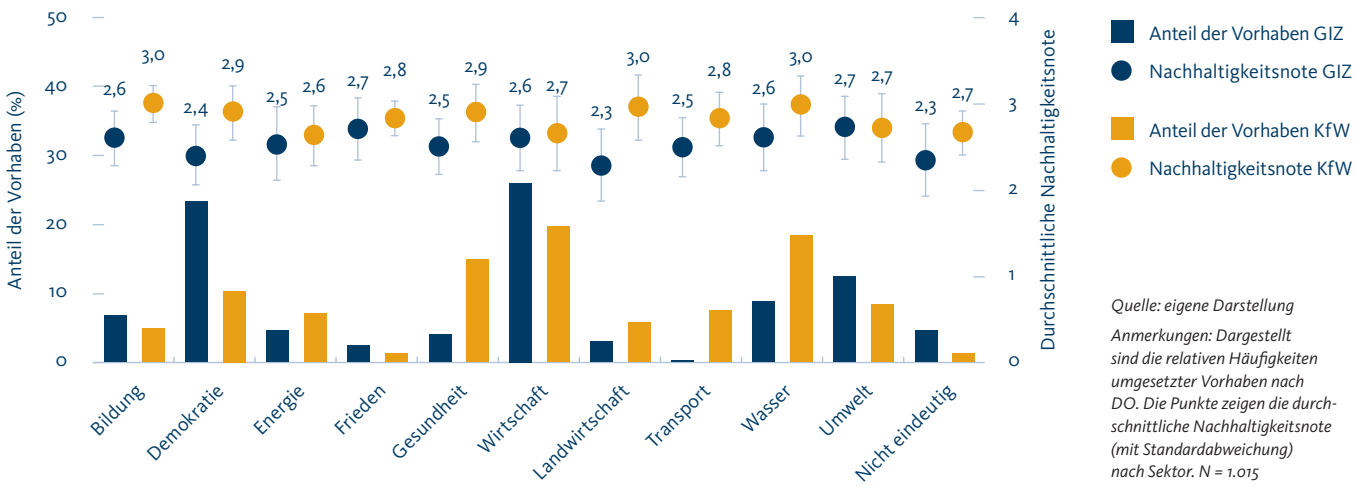
Abbildung 3 zeigt außerdem, dass die Nachhaltigkeitsnote zwischen den Sektoren nur mäßig variiert. Innerhalb des Portfolios der GIZ erhalten Vorhaben in den Sektoren Demokratie und Landwirtschaft die beste Bewertung. Bei der KfW werden Vorhaben im Energiesektor als besonders „nachhaltig" bewertet. Am schlechtesten schneiden bei der GIZ Vorhaben in den Sektoren Frieden und Umwelt ab, bei der KfW Vorhaben in den Sektoren Bildung, Landwirtschaft und Wasser. Innerhalb des Portfolios beider DO finden sich keine signifikanten Unterschiede in der Benotung zwischen den einzelnen Sektoren.

2.5 Stichprobenziehung

In der Analyse der Einflussfaktoren auf die vergebene Nachhaltigkeitsnote wäre es denkbar, alle Beobachtungen der Grundgesamtheit zu berücksichtigen. Eine größere Zahl von Datenpunkten erlaubt es, Zusammenhänge zwischen der vergebenen Nachhaltigkeitsnote und einzelnen Faktoren mit höherem Maß an statistischer Sicherheit zu bestimmen. Allerdings liegen für die Grundgesamtheit lediglich Meta-Daten zu den Merkmalen der Vorhaben und zu den Evaluierungsberichten vor. Der Einfluss der in den Berichten herangezogenen Bewertungskriterien und der methodischen Berichtsqualität auf die Nachhal-

[20] *Die Aussage basiert auf einem Zwei-Gruppen-Proportionen-Test.*

Abbildung 3: Sektorale Verteilung der Vorhaben und deren Nachhaltigkeitsbewertung nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind die relativen Häufigkeiten umgesetzter Vorhaben nach DO. Die Punkte zeigen die durchschnittliche Nachhaltigkeitsnote (mit Standardabweichung) nach Sektor. N = 1.015

tigkeitsnote kann mit den vorhandenen Meta-Daten nicht bestimmt werden. Die eingangs genannten Evaluierungsfragen können damit allein anhand der Meta-Daten nicht vollständig beantwortet werden.

Das Evaluierungsteam greift daher auf Informationen aus der begleitenden Meta-Evaluierung von Noltze et al. (2018) zurück. In dieser werden die Kriterien erfasst, die in den jeweiligen Berichten zur Bewertung von Nachhaltigkeit herangezogen wurden. Dies erfolgt unter Zuhilfenahme eines Analyserasters, welches aus sieben Bereichen besteht. Die einzelnen Bereiche gliedern sich wiederum in insgesamt 18 Kriterien und 48 differenzierte Kriterien. Außerdem wird die methodische Qualität der Berichte bewertet. Auch dies geschieht anhand eines Analyse-Rasters. Die Raster zum Erfassen der Kriterien zur Nachhaltigkeitsbewertung sowie zur Bewertung der methodischen Qualität befinden sich im Anhang (Tabelle 6 und Tabelle 7). Die Meta-Evaluierung wurde für eine Stichprobe der vorliegenden Evaluierungsberichte von GIZ und KfW durchgeführt. Aufgrund der in Abschnitt 2.3 diskutierten Unterschiede in der Nachhaltigkeitsbewertung von Vorhaben erfolgte die Stichprobenziehung getrennt nach Berichtstyp.[21]

Tabelle 1 zeigt die Anzahl der Beobachtungen in der Grundgesamtheit je Evaluierungstyp sowie die durch Noltze et al. (2018) bearbeitete Stichprobe. Bei der Bestimmung der Stichprobengröße wurde die Notenverteilung der Grundgesamtheit eines Evaluierungstyps berücksichtigt. Basierend auf der Verteilung der Nachhaltigkeitsnote und auf dem Anteil „nachhaltiger" (Note 1 bis 3) und „nicht nachhaltiger" (Note 4) Maßnahmen wurden dabei zunächst zwei verschiedene Stichprobengrößen berechnet. Für jeden Evaluierungstyp fand die jeweils größere Stichprobe Eingang in die Meta-Evaluierung (Noltze et al., 2018). Die durchschnittlich vergebene Nachhaltigkeitsnote, der Anteil als „nachhaltig" bewerteter Vorhaben je Evaluierungstyp sowie die einzelnen Stichprobengrößen sind in Tabelle 9 im Anhang aufgeführt.

Die Stichprobe der Meta-Evaluierung bildet auch die Grundlage für die hier umgesetzten empirischen Analysen. Insgesamt

[21] *Bei Meta-Evaluierungen, die von mehr als einer Person durchgeführt werden, können die Ergebnisse durch Unterschiede in der subjektiven Bewertungspraxis beeinflusst sein. Um zu prüfen, ob eine systematische Verzerrung der Ergebnisse vorliegt, wurden in der begleitenden Meta-Evaluierung 10 Prozent der Stichprobe je Evaluierungstyp von mindestens zwei Personen gelesen und bewertet. Mit Hilfe statistischer Verfahren wurde im Anschluss der sogenannte Kappa-Interkodierer-Reliabilitätskoeffizient nach Cohen gebildet. Dieser gibt Auskunft über den Grad der Übereinstimmung bei der Bewertung einzelner Kriterien zwischen verschiedenen Personen. In der begleitenden Meta-Evaluierung ergibt sich ein Kappa-Wert von 0,63, wonach eine substanzielle Übereinstimmung bei der Bewertung der Kriterien zwischen den beteiligten Personen vorliegt. Für eine ausführliche Darstellung zum methodischen Vorgehen der Meta-Evaluierung siehe Noltze et al. (2018).*

enthält die Stichprobe 513 evaluierte Maßnahmen, davon 341 Maßnahmen der GIZ und 172 der KfW. Aufgrund der relativen Häufigkeiten der Evaluierungstypen in der Grundgesamtheit sowie der jeweiligen Notenverteilungen besteht die Stichprobe aus unterschiedlichen Anteilen der verschiedenen Evaluierungstypen. So sind überwiegend PFK und KfW-Ex-post-Evaluierungen enthalten. PEV, GIZ-Ex-post- und Schluss-Evaluierungen finden sich hingegen weniger häufig.

Tabelle 1: Grundgesamtheit evaluierter Maßnahmen und Stichprobengröße nach Evaluierungstyp

Evaluierungstyp	Anzahl evaluierter Maßnahmen	Anzahl evaluierter Maßnahmen in der Stichprobe
GIZ Ex-post	56	47
GIZ Schluss	44	38
GIZ PFK	343	174
GIZ PEV	110	82
Zwischensumme	553	341
KfW Ex-post	462	172
Total	1.015	513

Quelle: eigene Darstellung
Anmerkungen: Die Größe der jeweiligen Stichprobe hängt von der Größe der Grundgesamtheit sowie der Varianz der Note bzw. des Anteils „nachhaltiger" und „nicht nachhaltiger" Vorhaben ab. Detaillierte Angaben hierzu sind in Tabelle 9 im Anhang enthalten.

3.

METHODISCHE VORGEHENSWEISE

Im folgenden Kapitel werden das in den Analysen angewandte Regressionsmodell spezifiziert sowie die darin enthaltenen Variablen operationalisiert. Im Anschluss werden die Limitationen des methodischen Vorgehens diskutiert.

3.1 Empirische Strategie

Wie bereits in Abschnitt 2.1 erwähnt, besteht ein konzeptioneller Zusammenhang zwischen der Bewertung von Nachhaltigkeit und der Bewertung der übrigen DAC-Kriterien. Auch die Ergebnisse der begleitenden Meta-Evaluierung zeigen, dass die Nachhaltigkeit einer Maßnahme anhand von Kriterien bewertet wird, die auch für die Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact einer Maßnahme herangezogen werden können (Noltze et al., 2018). Dieser Zusammenhang wird in Abbildung 4 verdeutlicht. Demnach erfolgt der Einfluss einer Variablen auf die Nachhaltigkeitsnote entweder direkt oder indirekt – über die Beeinflussung der anderen DAC-Kriterien, die dann wiederum die Nachhaltigkeitsbewertung beeinflussen. Bei der Modellierung der Einflussfaktoren ergeben sich verschiedene Möglichkeiten, diese Zusammenhänge zu berücksichtigen. So kann die Durchschnittsnote aller DAC-Kriterien (außer Nachhaltigkeit) als zusätzliche Kontrollvariable in die Modelle aufgenommen werden. Dadurch ist es möglich, zwischen dem Einfluss einer Variablen auf die Nachhaltigkeit eines Vorhabens und deren Einfluss auf die übrigen DAC-Kriterien zu unterscheiden. Dieses Vorgehen ist allerdings insofern problematisch, als für Faktoren, die die Nachhaltigkeit vor allem über die anderen DAC-Kriterien beeinflussen, kein Effekt auf die Nachhaltigkeitsnote festgestellt werden kann. Um sowohl den direkten als auch den indirekten Effekt eines Faktors zu erfassen, muss die Durchschnittsnote der DAC-Kriterien aus den Modellen ausgeschlossen werden. Die im Folgenden vorgestellten Modelle werden deshalb sowohl mit als auch ohne DAC-Durchschnittsnote geschätzt; so lässt sich abschätzen, ob neben direkten auch indirekte Einflüsse bestehen.

Die Analyse der Einflussfaktoren auf die vergebene Nachhaltigkeitsnote erfolgt durch multivariate Regressionsmodelle. In multivariaten Regressionsmodellen wird der Einfluss mehrerer erklärender Variablen auf eine zu erklärende Variable bestimmt. Im vorliegenden Fall ist die vergebene Nachhaltigkeitsnote die zu erklärende Variable. Dabei können jeder Note bestimmte Ausprägungen einer Reihe erklärender Variablen zugeordnet werden. Im Zusammenspiel beider Größen über eine Vielzahl von Berichten hinweg kann so der marginale Effekt einer bestimmten erklärenden Variablen auf die Nachhaltigkeitsnote statistisch ermittelt werden. Dabei wird zwischen der Effektstärke (Wie groß ist der Effekt der Variable unter der Annahme, dass die anderen Variablen konstant gehalten werden?) und dem statistischen Signifikanzniveau des errechneten Effekts (Mit welcher Wahrscheinlichkeit tritt das beobachtete Ergebnis unter der Annahme keines Zusammenhangs auf?) unterschieden. Da die Nachhaltigkeitsnote ordinal skaliert ist – das heißt einer Rangordnung mit 1 als bester und 4 als schlechtester Ausprägung folgt –, wird ein geordnetes logistisches Regressionsmodell geschätzt. Das Modell hat die allgemeine Form:

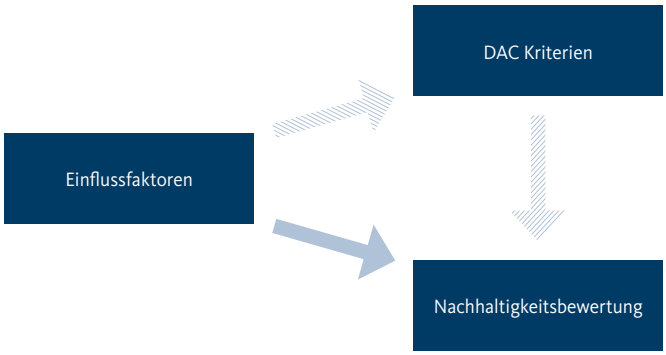
$$N_i^* = \beta_l \boldsymbol{X}_i + \gamma \boldsymbol{DAC}_i + \varepsilon_i$$

Dabei stellt N_i^* die latente, nicht beobachtete Nachhaltigkeitsbewertung in Bericht i dar.[22] In der geschätzten Modellspezifikation ist $\boldsymbol{X}$ eine Matrix mit erklärenden Variablen. Zur Überprüfung der Robustheit der Ergebnisse werden neben der hier beschriebenen Modellspezifikation weitere Modelle geschätzt. Diese unterscheiden sich in Bezug auf die in $\boldsymbol{X}$ enthaltenen Variablen. Die Modifikationen werden in Abschnitt 3.2 beschrieben. Nach den in Kapitel 2.2 diskutierten Einflussfaktoren auf die Nachhaltigkeitsbewertung einzelner Maßnahmen sind in $\boldsymbol{X}$ bestimmte **Merkmale des Vorhabens, Merkmale der Evaluierung, der Implementierungskontext einer Maßnahme** sowie **die Bewertungskriterien der Nachhaltigkeit** enthalten. Der Vektor β beinhaltet demnach die zu schätzenden Koeffizienten. Diese geben den Effekt der jeweiligen erklärenden Variablen auf die Nachhaltigkeitsbewertung an.

[22] *Dabei besteht zwischen der latenten Variablen N_i^* und der vergebenen Nachhaltigkeitsnote N_i folgender Zusammenhang:*

$$N_i = \begin{cases} 1 \text{ wenn } Ni* \leq \mu 1 \\ 2 \text{ wenn } \mu 1 < Ni* \leq \mu 2 \\ 3 \text{ wenn } \mu 2 < Ni* \leq \mu 3 \\ 4 \text{ wenn } \mu 3 < Ni* \end{cases}$$

Abbildung 4: Zusammenhang zwischen Einflussfaktoren, DAC-Kriterien und Nachhaltigkeitsbewertung



Quelle: eigene Darstellung

Anmerkungen: Die schraffierten Pfeile zeigen indirekte Effekte eines Einflussfaktors auf die Nachhaltigkeitsbewertung. Der ausgefüllte Pfeil zeigt direkte Effekte.

Die in X enthaltenen **Merkmale eines Vorhabens** sind dessen Laufzeit (Jahre) und Finanzvolumen (Logarithmus der Kosten in Mio. €) sowie dessen entwicklungspolitische Oberziele nach sozialer, politischer, ökonomischer und ökologischer Dimension (Anzahl Oberziele). Zusätzlich wird im Modell erfasst, welche DO die Maßnahme umsetzt (GIZ oder KFW). Weiter wird abgebildet, ob Verzögerungen bei der Implementierung der Maßnahme vorliegen (Indikator-Variable) und ob ein Vorhaben der wichtigsten Umsetzungs-Region – Subsahara-Afrika – sowie dem wichtigsten Umsetzungs-Sektor – nachhaltige Wirtschaftsentwicklung – angehört (Indikator-Variablen).

Die **Merkmale der Evaluierung** beinhalten den Zeitpunkt der Evaluierung relativ zum Vorhabenende (Jahre vor bzw. nach Vorhaben-Ende) sowie den Evaluierungstyp (PFK, PEV, Schluss-Evaluierung).

Der **Implementierungskontext einer Maßnahme** wird durch das Bruttoinlandsprodukt (BIP) eines Landes pro Kopf abgebildet (aktuelle US-Dollar). Darüber hinaus werden die von einem Land erhaltenen Zahlungen an Official Development Assistance (ODA) für den Zeitraum der Implementierung einer Maßnahme berücksichtigt. Um die Vergleichbarkeit empfangener Transfers zwischen verschiedenen Ländern zu gewährleisten, wird der Anteil von ODA-Transfers am BIP eines Landes ermittelt (ODA/BIP in %). Daten zum wirtschaftlichen Entwicklungsstand und zu ODA-Transfers stammen aus der Datenbank der Weltbank (World Bank, 2017). Der politische Kontext eines Landes wird in den Modellen durch den „Freedom in The World"-Index von Freedom House dargestellt (Skala 1 bis 7).[23] Dieser gibt Auskunft über das Maß an politischen Rechten und zivilen Freiheiten einer Gesellschaft (Freedom House, 2016). Zur Integration der genannten Variablen in die Modelle werden die Mittelwerte der jeweiligen Variablen über die Laufzeit einer Maßnahme gebildet.

Die hier genutzten Makro-Indikatoren sind auf Ebene einzelner Länder aggregiert. EZ-Maßnahmen betreffen hingegen selten das gesamte Gebiet eines Staates, sondern sind geographisch eingegrenzter. Innerhalb eines Landes kann es signifikante Unterschiede in den wirtschaftlichen, politischen, sozialen und ökologischen Rahmenbedingungen geben. Diese regionalen Unterschiede werden in den vorhandenen Makro-Daten nicht abgebildet (Denizer et al., 2013). So kann beispielsweise das mittlere Wirtschaftswachstum eines Landes deutlich über dem Wachstum der wirtschaftlich schwächsten Region liegen. Im Modell wird daher zusätzlich zu den beschriebenen Makro-Indikatoren der Einfluss des **projektspezifischen Kontextes** abgebildet. Dazu wird auf Ergebnisse zu den Bewertungskriterien der Nachhaltigkeit aus der begleitenden Meta-Evaluierung zurückgegriffen (Noltze et al., 2018). In Anlehnung an die in den Leitfragen zur Nachhaltigkeitsbewertung vorgegebene Überprüfung der Stabilität des Kontextes (BMZ, 2006) werden die in den Berichten getroffenen Kontext-Erwähnungen im Modell erfasst.

[23] *1 = beste Bewertung und 7 = schlechteste Bewertung.*

Dabei wird zwischen einem negativen Einfluss des Kontextes auf die Nachhaltigkeit von Vorhaben, keinem Einfluss des Kontextes und einem positiven Einfluss des Kontextes unterschieden.

Neben dem projektspezifischen Kontext werden weitere Bewertungskriterien der Nachhaltigkeit aus der Meta-Evaluierung abgeleitet (Noltze et al., 2018). Wie bereits in Abschnitt 2.2 beschrieben, gliedert sich das darin angelegte Raster zur Erfassung nachhaltigkeitsrelevanter Kriterien in sieben Bereiche: 1.) Kontext, 2.) Implementierung, 3.) Outcome, 4.) Kapazitäten vor Ort, 5.) Nicht intendierte Wirkungen (Impact), 6.) Absehbarkeit des Erhalts von Wirkungen sowie 7.) Zusammenspiel der Dimensionen (siehe Tabelle 6). Durch die systematische Überprüfung jedes Berichts nach diesem Raster ergibt sich ein umfassendes Bild der inhaltlichen Stärken und Schwächen hinsichtlich der Nachhaltigkeit einer Maßnahme. In den Modellen ist der Einfluss der Bereiche auf die Nachhaltigkeit eines Vorhabens enthalten. Dabei kann dieser „negativ", „neutral" oder „positiv" sein. Anhand der im Bericht getroffenen Aussagen wird jedem der 48 differenzierten Kriterien ein numerischer Wert zugeschrieben. Die numerischen Werte für negativen (-1), neutralen (0) oder positiven (+1) Einfluss auf die Nachhaltigkeit des Vorhabens werden anschließend innerhalb der 18 Kriterien bzw. innerhalb der sieben Bereiche zu einem Wert aggregiert. Je positiver (negativer) dieser Wert ausfällt, desto förderlicher (hemmender) ist der Einfluss eines bestimmten Bereichs auf die Nachhaltigkeitsnote.

Der Vektor $\boldsymbol{DAC}$ beinhaltet die Durchschnittsnote aller DAC-Kriterien mit Ausnahme des Kriteriums Nachhaltigkeit. Zur Überprüfung der Ergebnisse wird ein Modell ohne den Vektor $\boldsymbol{DAC}$ geschätzt. ε_{ij} ist der normalverteilte Fehlerterm mit dem Erwartungswert 0 und konstanter Varianz. Eine vollständige Liste aller erklärenden Variablen inklusive Definition und Quelle befindet sich im Anhang (Tabelle 10).

Neben den hier beschriebenen Variablen besteht möglicherweise ein Zusammenhang zwischen der methodischen Berichtsqualität und der vergebenen Nachhaltigkeitsnote. Um diesen zu erfassen, wird auf die im Rahmen der begleitenden Meta-Evaluierung durchgeführte Qualitätsbewertung der Berichte zurückgegriffen (siehe Tabelle 7).[24] Die Qualitätsbewertung der Berichte findet sich nicht als Kontrollvariable in Matrix $\boldsymbol{X}$, sondern dient als analytisches Gewicht einzelner Beobachtungen.[25] In den Regressionen werden Berichte durchschnittlicher Qualität demnach einfach, Berichte überdurchschnittlicher Qualität stärker und Berichte unterdurchschnittlicher Qualität schwächer gewichtet.[26] Die Gewichtung einzelner Beobachtungen soll sicherstellen, dass die glaubwürdigsten Ergebnisse den größten Einfluss in der Synthese erhalten. Während die Gewichtung von Beobachtungen gängige Praxis im Rahmen quantitativer Meta-Analysen ist (Borenstein et al., 2009), wird die methodische Berichtsqualität in keiner der hier zitierten vergleichbaren Studien explizit bei der Modellierung des Projekterfolges berücksichtigt (Assefa et al., 2014; Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005).

Die existierenden Querschnittsauswertungen der deutschen TZ nehmen keine Gewichtung von Beobachtungen vor. Stattdessen wird in einer Reihe von Synthesestudien im Auftrag der GIZ die methodische Qualität von Evaluierungen als Ausschlusskriterium für inhaltliche Querschnittsauswertungen genutzt (Caspari, 2014; Huber et al., 2014). In diesen Studien wird a priori ein Schwellenwert methodischer Güte festgelegt, bei dessen Unterschreitung eine Beobachtung nicht in die Querschnittsauswertung einfließt. Dieses Vorgehen scheint zwar plausibel, doch besticht eine Gewichtung der Beobachtungen nach methodischer Qualität durch drei zentrale Vorteile: Erstens wird hierfür kein willkürlicher Schwellenwert benötigt, zweitens werden auch Berichte in der Analyse berücksichtigt, die nur leicht unterhalb eines Schwellenwertes liegen, und drittens erlaubt die Gewichtung aller Beobachtungen, zwischen Berichten besserer und schlechterer Qualität zu differenzieren.

Neben der methodischen Güte einzelner Berichte muss auch der Tatsache Rechnung getragen werden, dass KfW und GIZ unterschiedliche Evaluierungstypen einsetzen. So beurteilen einige Berichtstypen die Nachhaltigkeit einer Maßnahme aufgrund tatsächlicher Beobachtungen (KfW-Ex-post und GIZ-Ex-post-Evaluierungen), während andere Typen ihre

[24] *Für eine ausführliche Beschreibung der Erfassung und Bewertung der methodischen Berichtsqualität im Rahmen der Meta-Evaluierung siehe Noltze et al. (2018).*
[25] *Die methodische Berichtsqualität wird als standardisierter Qualitätsindex erfasst. Dieser hat den Mittelwert = 1 und eine Standardabweichung = 0,5.*
[26] *Aufgrund der nicht integren Ausprägungen des standardisierten Qualitätsindex erfolgt die Gewichtung der Beobachtungen mittels analytischer Gewichtung. Die Gewichtung ist dabei invers proportional zu der Varianz einer Beobachtung.*

Bewertung auf Einschätzungen zu erwartender Entwicklungen gründen (PFK, PEV, Schluss-Evaluierungen). Das Evaluierungsteam nimmt an, dass sich beide Berichtstypen bezüglich der Nachhaltigkeitsbewertung systematisch unterscheiden. Das Modell wird daher für zwei verschiedene Gruppen der Stichprobe geschätzt. Eine Unterteilung der Beobachtungen erhöht die Vergleichbarkeit der Bewertungen innerhalb einer Gruppe. Die erste Gruppe besteht aus Ex-post-Evaluierungen der KfW und GIZ, die zweite Gruppe enthält PFK, PEV und Schluss-Evaluierungen (alle GIZ).

3.2 Sensitivitätschecks

Um die Robustheit der Ergebnisse zu überprüfen, wurden neben den beschriebenen Modellspezifikationen weitere Modelle geschätzt. Diese unterscheiden sich vor allem in Bezug auf die in $\boldsymbol{X}$ enthaltenen erklärenden Variablen. Einzelne Spezifikationen variieren – aufgrund der Datenverfügbarkeit zu bestimmten Variablen – auch hinsichtlich der Anzahl der im Modell enthaltenen Beobachtungen.

Durch die Schätzung alternativer Modelle kann unter anderem ermittelt werden, ob die Ergebnisse abhängig von der gewählten Operationalisierung bestimmter Variablen sind. Beispielsweise können die politische Stabilität eines Landes sowie die Qualität seiner Institutionen durch den Rule-of-Law-Index der Weltbank oder durch den Freedom-House-Index erfasst werden. Außerdem können bestimmte Variablen ausführlicher betrachtet werden. Während die Hauptmodelle etwa nur die Region enthalten, in der die deutsche EZ ihren Schwerpunkt hat, wurde in den zusätzlichen Modellen der Einfluss aller weiteren Regionen geprüft.[27] Eine vergleichbare Vorgehensweise erfolgte auch im Umgang mit den Sektoren.[28] Da sich regionale und sektorale Effekte zwischen den DO unterscheiden können, wurden zusätzlich Interaktionsterme zwischen DO und Region sowie zwischen DO und Sektor berücksichtigt. Auch der Einfluss der einzelnen Oberzieldimensionen wurde in zusätzlichen Modellspezifikationen untersucht.[29]

Schließlich wurde auch überprüft, ob alle notwendigen Informationen im Modell enthalten sind. Dabei wurden auch Variablen in zusätzliche Modelle aufgenommen, zu denen zwar nicht für alle Beobachtungen ausreichend Informationen vorliegen, die jedoch möglicherweise dennoch einen Einfluss auf die Nachhaltigkeitsbewertung haben könnten. In weiteren Modellen wurden als Merkmale die an der Evaluierung beteiligten Personen (Anzahl) sowie das Datum der Evaluierung (Jahr) berücksichtigt. Weitere Einflussfaktoren, wie die Dauer der Evaluierung (Tage) oder die Dauer der Feldmission (Tage), konnten aufgrund der geringen Datenverfügbarkeit letztendlich in keines der Modelle aufgenommen werden. Als Kenngrößen des Implementierungskontextes wurden in weiteren Modellen zusätzlich das jährliche Wirtschaftswachstum (in %), der Rule-of Law-Index der Weltbank (-4 bis +4), die Lebenserwartung bei Geburt (in Jahren), die Bevölkerungszahl eines Landes (in Millionen) sowie die Einschulungsrate (in % der relevanten Altersgruppe) berücksichtigt.

Neben Vorhaben, die in einem spezifischen Land umgesetzt werden, enthält die Stichprobe auch Vorhaben, die in mehreren Ländern realisiert werden. Diesen sogenannten Regional- und Sektorvorhaben lassen sich keine auf Landes-Ebene aggregierten Indikatoren zuordnen. Um diese Beobachtungen dennoch zu berücksichtigen, wurden zusätzliche Modelle ohne Indikatoren auf Landes-Ebene geschätzt.

Tabelle 11 (im Anhang) enthält alle Variablen, die in die Zusatzmodelle aufgenommen wurden. Bei der Darstellung und Diskussion der Ergebnisse der Hauptmodelle wird auf relevante Ergebnisse der Zusatzmodelle verwiesen. Letztere stehen nicht im Widerspruch zu den Ergebnissen der Hauptmodelle.

3.3 Limitationen des methodischen Vorgehens

Generell ist zu beachten, dass die Ergebnisse der Regressionsmodelle als statistische Zusammenhänge über alle betrachteten Evaluierungsberichte hinweg zu verstehen sind. Besonderheiten einzelner Vorhaben können mit den hier angewendeten

[27] *In der Grundgesamtheit finden sich Vorhaben in den Regionen Subsahara-Afrika, Nordafrika/Naher Osten, Asien/Ozeanien, Europa/Kaukasus und Lateinamerika sowie überregionale Vorhaben.*

[28] *Zu den Sektoren der deutschen EZ gehören Bildung, Demokratie/Zivilgesellschaft und öffentliche Verwaltung, Energie, Friedensentwicklung und Krisenprävention, Gesundheit/Familienplanung/HIV/Aids, Nachhaltige Wirtschaftsentwicklung, Sicherung der Ernährung/Landwirtschaft/Fischerei, Transport und Kommunikation, Trinkwasser/Wassermanagement/Abwasser/Abfallentsorgung, Umweltpolitik/Schutz und nachhaltige Nutzung natürlicher Ressourcen.*

[29] *Eine weitere denkbare Darstellung der Zielsysteme eines Vorhabens ist die Berücksichtigung der Anzahl der DAC-Haupt- und -Nebenziel-Kennungen. Diese werden allerdings nicht systematisch in den Meta-Daten der GIZ-Evaluierungen abgebildet. Sie sind daher nicht Bestandteil der Analyse.*

Analyseverfahren nicht explizit untersucht werden. Hierzu wäre eine weiterführende Studie nötig, in der gezielt eine begrenzte Anzahl an Vorhaben – beispielsweise aus einzelnen Sektoren – betrachtet wird.

Ferner ist bei der Interpretation der Ergebnisse zu berücksichtigen, dass die im Modell enthaltenen erklärenden Variablen möglicherweise endogen sind. So ist denkbar, dass unbeobachtete Faktoren bestimmte in $\boldsymbol{X}$ enthaltene Variablen beeinflussen und dabei gleichzeitig einen Einfluss auf die vergebene Nachhaltigkeitsnote haben. Beispielsweise kann der Grad der Zielerreichung zu einem bestimmten Zeitpunkt im Projektzyklus die Dauer eines Vorhabens beeinflussen. Erfolgreiche Projekte könnten daher eher verlängert werden. Gleichzeitig kann sich der Projekterfolg auch direkt auf die Nachhaltigkeit einer Maßnahme auswirken. Darüber hinaus ist nicht auszuschließen, dass einzelne Kriterien eher in einer bestimmten Ausprägung beobachtet werden. Beispielsweise können negative politische Rahmenbedingungen leichter wahrnehmbar sein als positive. Es ist denkbar, dass der negative politische Kontext eher Eingang in die Bewertung findet, wenn die Nachhaltigkeit einer Maßnahme ohnehin kritisch betrachtet wird. Demnach bestimmt nicht der politische Kontext die Nachhaltigkeit, sondern die Leichtigkeit, mit der er bei gegebener Bewertung zu beobachten ist. In beiden hier genannten Fällen würde der durch das Modell ermittelte Effekt verzerrt, was bei der Interpretation der Ergebnisse zu berücksichtigen ist.[30]

Darüber hinaus kann die Bewertung der Nachhaltigkeit einer Maßnahme deren tatsächliche Nachhaltigkeit nur ungenau widerspiegeln. Der Bewertungsprozess ist stets subjektiv. Außerdem hat die begleitende Meta-Evaluierung gezeigt, dass die Bewertung von Nachhaltigkeit in der deutschen EZ weitestgehend unsystematisch und uneinheitlich erfolgt. Auch ist unklar, wie die im jeweiligen Bericht genannten Kriterien bei der Notenvergabe gewichtet werden. Diesem teilweise intransparenten Bewertungsverfahren steht eine Notenverteilung (von 1 bis 4) gegenüber, die eine Messgenauigkeit vorgibt, die in dieser Form nicht existiert. Durch die Berücksichtigung der methodischen Berichtsqualität in den Regressionsmodellen wird diesem Sachverhalt zumindest teilweise Rechnung getragen. Die Gewichtung einzelner Berichte erlaubt, dass Zusammenhänge zwischen erklärenden Variablen und der Nachhaltigkeitsnote in methodisch überdurchschnittlich guten Berichten stärker berücksichtigt werden.[31]

[30] *Der Endogenität bestimmter Variablen kann durch die Nutzung von Instrumentalvariablen begegnet werden. Die Instrumentalvariablen werden dabei so gewählt, dass sie die exogene Varianz der erklärenden Variablen isolieren. Weiter muss sichergestellt sein, dass Instrumentalvariablen die zu erklärende Variable nur durch die endogene erklärende Variable beeinflussen.*

[31] *Alternativ könnte die methodische Berichtsqualität auch als Kontrollvariable ins Modell aufgenommen werden. Dieses Vorgehen impliziert, dass die methodische Qualität eines Berichtes die Nachhaltigkeitsnote lediglich durch eine Verschiebung des Y-Achsenabschnitts beeinflusst. Es ist jedoch nicht auszuschließen, dass die methodische Berichtsqualität auch einen direkten Einfluss auf den Zusammenhang zwischen erklärenden Variablen und der Nachhaltigkeitsnote ausübt. Durch die Gewichtung der Beobachtungen nach Berichtsqualität können diese Zusammenhänge im Modell abgebildet werden.*

4.

ERGEBNISSE

Im folgenden Kapitel wird die Stichprobe zunächst im Hinblick auf die im Modell enthaltenen erklärenden Variablen beschrieben. Anschließend wird der in Abschnitt 2.1 dargestellte konzeptionelle Zusammenhang zwischen den DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact und dem Kriterium Nachhaltigkeit empirisch überprüft. Darauf aufbauend werden die Regressionsergebnisse, gegliedert nach Evaluierungsfragen, dargestellt und diskutiert. Abschließend werden aus den einzelnen Ergebnisteilen übergeordnete Ergebnisse synthetisiert.

4.1 Verteilung der erklärenden Variablen nach Notenstufe

Eine deskriptive Darstellung aller im Modell enthaltenen erklärenden Variablen erleichtert die Interpretation der Regressionsergebnisse. In Tabelle 2 sind deren Mittelwerte und Standardabweichungen für die Stichprobe dargestellt. Die Mittelwerte sind dabei nach der vergebenen Nachhaltigkeitsnote unterteilt. Bei der Interpretation der Mittelwerte ist zu beachten, dass Unterschiede einzelner Werte zwischen den Notenstufen nicht als kausaler Zusammenhang zwischen der Variablen und der Nachhaltigkeitsnote interpretiert werden können. Es ist nicht auszuschließen, dass die hier dargestellten Variablen mit weiteren Variablen korrelieren, die ihrerseits wiederum die Nachhaltigkeitsnote beeinflussen.[32]

Die Ergebnisse zeigen, dass eine schlechtere durchschnittliche Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact tendenziell mit einer schlechteren Nachhaltigkeitsbewertung einhergeht. Darüber hinaus kann festgehalten werden, dass kürzere Vorhaben tendenziell eine bessere Nachhaltigkeitsbewertung erhalten. Weiterhin wird mit steigendem Finanzvolumen einer Maßnahme deren Nachhaltigkeit tendenziell schlechter bewertet. Maßnahmen in der Region Subsahara-Afrika sowie im Sektor „nachhaltige Wirtschaftsförderung“ werden tendenziell schlechter bewertet. Ihr relativer Anteil innerhalb einer Notenstufe steigt mit schlechter werdender Bewertung. Dies ist insofern bemerkenswert, als Vorhaben in der Region Subsahara-Afrika sowie im Sektor nachhaltige Wirtschaftsförderung im Portfolio der GIZ und KfW dominieren.

Mit Blick auf den Implementierungskontext einer Maßnahme zeigt sich, dass ein Anstieg des BIP pro Kopf eines Landes tendenziell mit einer besseren Nachhaltigkeitsbewertung eines Vorhabens einhergeht. Es findet sich hingegen kein Zusammenhang zwischen dem Anteil erhaltener ODA-Transfers (% des BIP eines Landes) und der Nachhaltigkeitsbewertung eines Vorhabens. Auch zwischen dem Rule-of-Law-Index und der Nachhaltigkeitsbewertung ist kein Zusammenhang festzustellen.

Hinsichtlich der im Rahmen der Meta-Evaluierung erfassten Bewertungskriterien von Nachhaltigkeit wird deutlich, dass sich die Nachhaltigkeitsnote verbessert, je positiver der Gesamteinfluss aller Kriterien auf die Nachhaltigkeit bewertet wird. Dieses Muster zeigt sich ebenso innerhalb der sieben Bereiche (Kontext, Implementierung, Outcome, Kapazitäten vor Ort, nicht intendierte Wirkungen (Impact), Absehbarkeit des Erhalts von Wirkungen, Zusammenspiel der Dimensionen).[33]

[32] *So können beispielsweise besonders kurze (oder lange) Vorhaben besonders häufig mit einem bestimmten Evaluierungstyp bewertet werden. Die bessere Note kürzerer Vorhaben kann dadurch begründet sein, dass bestimmte Evaluierungstypen bessere (oder schlechtere) Noten vergeben und dass diese Evaluierungstypen gleichzeitig besonders häufig bei kurzen (oder langen) Vorhaben zum Einsatz kommen. Der vermeintliche kausale Zusammenhang zwischen Dauer und Note existiert in diesem Fall nicht.*

[33] *Der Einfluss eines Kriteriums ist entweder positiv, neutral oder negativ. Durch die Zusammenfassung einzelner Kriterien zu inhaltlichen Blöcken kann so der Einfluss der verschiedenen Bereiche auf die Nachhaltigkeitsbewertung ermittelt werden.*

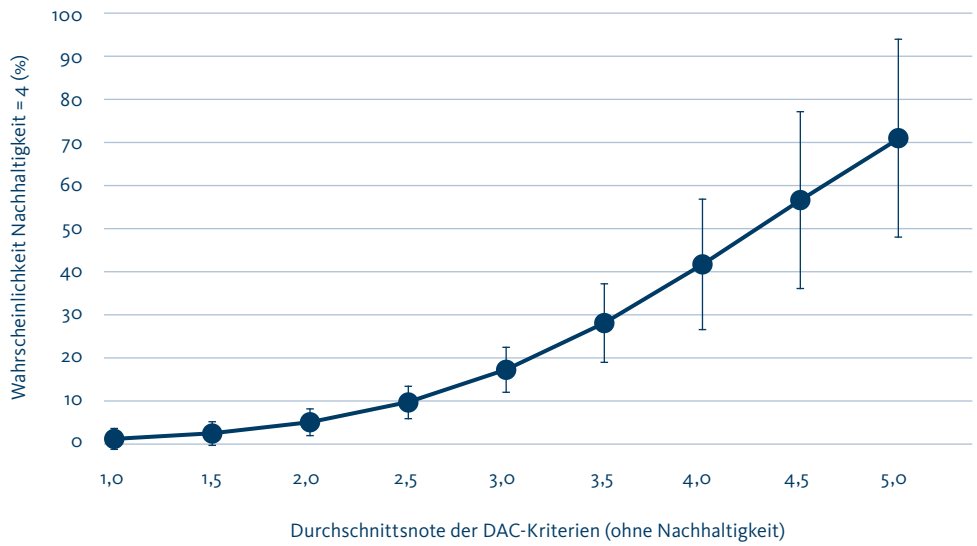
Tabelle 2: Deskriptive Statistiken der erklärenden Variablen nach Nachhaltigkeitsnote

	Nachhaltigkeitsnote			
	1 (n = 30)	**2 (n = 166)**	**3 (n = 256)**	**4 (n = 61)**
Merkmale der Vorhaben				
Note DAC-Kriterien ohne Nachhaltigkeit (Durchschnitt)	1,8 (0,4)	2,0 (0,5)	2,4 (0,5)	3,2 (0,7)
Dauer des Vorhabens (Jahre)	3,6 (2,1)	4,1 (2,6)	4,6 (3,2)	5,8 (3,9)
Wert des Vorhabens (Mio. €) (GIZ n = 297)	7,5 (7,8)	11,3 (15,0)	11,0 (13,9)	11,4 (20,5)
Oberziel-Dimensionen (Anzahl)	1,7 (0,8)	1,7 (0,7)	1,7 (0,7)	1,7 (0,6)
Anteil Vorhaben in Region Subsahara-Afrika (%)	21	30	35	44
Anteil Vorhaben im Sektor nachhaltige Wirtschaftsförderung (%)	17	25	23	36
Verzögerte Implementierung (%)	23	24	34	25
Implementierungskontext der Vorhaben				
BIP pro Kopf (aktuelle USD) (GIZ n = 245, KfW n = 166)	3.203 (2.600)	2.575 (2.671)	2.243 (2.309)	1.868 (1.830)
Netto-ODA (% am BIP) (GIZ n = 241, KfW n = 165)	5,7 (6,0)	6,0 (7,9)	7,1 (9,0)	6,8 (6,1)
Freedom-House-Index (GIZ n = 234, KfW n = 158)	4,1 (1,6)	3,8 (1,7)	4,0 (1,5)	4,1 (1,4)
Merkmale der Evaluierungen				
Zeitpunkt relativ zu Vorhabenende (Jahre)	0,1 (1,8)	1,0 (2,3)	1,3 (2,5)	3,0 (3,2)
Bewertungskriterien der Nachhaltigkeit (Summe positiver und negativer Einflüsse)				
Gesamtkriterien	2,9 (3,9)	2,8 (3,3)	-0,1 (3,8)	-4,9 (4,3)
Kriterien Kontext	-0,3 (0,8)	-0,3 (1,1)	-0,6 (0,9)	-1,1 (1,0)
Kriterien Planung und Implementierung	0,5 (1,3)	0,6 (1,0)	0,2 (1,2)	-0,5 (0,9)
Kriterien Outcome	2,1 (2,3)	2,1 (1,9)	0,8 (2,1)	-1,2 (2,5)
Kriterien Kapazitäten der Partner	0,7 (1,7)	0,5 (1,7)	-0,5 (1,9)	-2,1 (1,9)
Kriterien nicht intendierte Wirkungen (Impact)	0,1 (0,6)	0,2 (0,6)	0,1 (0,5)	-0,1 (0,5)
Kriterien Absehbarkeit des Erhalts von Wirkungen	0,5 (0,5)	0,5 (0,6)	0,1 (0,7)	-0,4 (0,6)
Kriterien Zusammenspiel der Dimensionen	0,2 (0,6)	0,3 (0,6)	0,3 (0,7)	<0,1 (0,6)

Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind Mittelwerte und Standardabweichungen für die Stichprobe (n = 513). Diese enthält 341 Beobachtungen der GIZ und 172 Beobachtungen der KfW. Die Angaben in Klammern zeigen, für wie viele der Beobachtungen Informationen bezüglich der jeweiligen Variablen vorliegen. Informationen bezüglich einzelner Variablen ohne Klammer sind vollständig.

Abbildung 5: Nachhaltigkeitsbewertung in Abhängigkeit der Bewertung der DAC-Kriterien



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte und Konfidenzintervalle (95 %) für die Vergabe der Nachhaltigkeitsnote 4 nach DAC-Durchschnittsnote. Marginale Effekte geben die Wahrscheinlichkeit an, mit der die Nachhaltigkeitsnote 4 für Vorhaben mit bestimmter Durchschnittsnote der DAC-Kriterien (ohne Nachhaltigkeit) vergeben wird. Die Ergebnisse basieren auf der Hauptspezifikation des in Abschnitt 3.1 vorgestellten Regressionsmodells. Das Modell enthält 352 Beobachtungen (KfW-Ex-post, GIZ-Ex-post, Schluss-Evaluierungen, PFK, PEV). Die Beobachtungen sind nach methodischer Qualität gewichtet.

4.2 Empirischer Zusammenhang zwischen Nachhaltigkeit und anderen DAC-Kriterien

Der konzeptionelle Zusammenhang zwischen dem Kriterium Nachhaltigkeit und den anderen DAC-Kriterien wurde bereits in Abschnitt 2.1 diskutiert. Die in Tabelle 2 dargestellten Mittelwerte zeigen, dass möglicherweise auch ein empirischer Zusammenhang zwischen der Nachhaltigkeitsbewertung und der Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact besteht. Allerdings können die Mittelwerte durch andere Variablen beeinflusst werden. Daher kann aus Tabelle 2 nur eine Korrelation zwischen den Benotungen abgeleitet werden.

Anhand des in Abschnitt 3.1 vorgestellten Regressionsmodells kann festgestellt werden, ob die beobachteten Korrelationen auch bei Anwesenheit aller in $\boldsymbol{X}$ enthaltenen Variablen statistisch signifikant sind. Abbildung 5 zeigt den Einfluss der Durchschnittsnote der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact auf die Nachhaltigkeitsbewertung. Die Abbildung basiert auf den Ergebnissen des Regressionsmodells. Die dargestellten Datenpunkte sind marginale Effekte. Diese geben an, mit welcher Wahrscheinlichkeit die Nachhaltigkeitsnote 4 für verschiedene Ausprägungen der Durchschnittsnote aller DAC-Kriterien (außer Nachhaltigkeit) vergeben wird.

Demnach nimmt mit schlechter werdender Bewertung der übrigen DAC-Kriterien die Wahrscheinlichkeit der schlechtesten Nachhaltigkeitsbewertung zu. Dabei sind alle marginalen Effekte ab einer guten DAC- Durchschnittsnote (2) statistisch signifikant. So liegt die Wahrscheinlichkeit einer schlechten Nachhaltigkeitsbewertung bei guter Bewertung der DAC-Kriterien (2) bei rund 6 Prozent. Bei ausreichender (4) bis unzureichender (5) Bewertungen der Vorhaben steigt sie auf rund 43 Prozent bzw. 70 Prozent. Zusammenfassend ist

festzuhalten, dass, wenn Maßnahmen bezüglich der anderen DAC-Kriterien unterdurchschnittlich eingestuft werden, eine sehr hohe Wahrscheinlichkeit besteht, dass auch ihre Nachhaltigkeit als nicht ausreichend bewertet wird. Der konzeptionelle Zusammenhang der DAC-Kriterien besteht somit auch statistisch. Alle im Modell enthaltenen erklärenden Variablen können sowohl einen direkten als auch einen indirekten Einfluss auf die Nachhaltigkeitsnote haben. Wie die in Abbildung 5 dargestellten Ergebnisse zeigen, kann ein indirekter Einfluss immer dann bestehen, wenn ein Faktor eines der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact beeinflusst, die wiederum die Nachhaltigkeitsbewertung bestimmen.

4.3 Regressionsergebnisse

4.3.1 Darstellung der Ergebnisse

Die hier dargestellten Ergebnisse basieren auf dem in Abschnitt 3.1 diskutierten Regressionsmodell. Dabei wird der Einfluss der Nachhaltigkeits-Bewertungskriterien sowohl aggregiert über alle sieben Bereiche (reduziertes Modell) als auch getrennt nach einzelnen Bereichen (volles Modell) betrachtet. Außerdem werden die Modelle sowohl mit der DAC-Durchschnittsnote (ohne Nachhaltigkeit) als Kontrollvariable als auch ohne die DAC-Durchschnittsnote geschätzt. Die Ergebnisse sind nach Ex-post-Evaluierungen (Tabelle 3) und PFK, PEV und Schluss-Evaluierungen (Tabelle 4) unterteilt. Jede Tabelle enthält demnach die Ergebnisse von vier unterschiedlichen Modellspezifikationen. Dargestellt sind die Regressionskoeffizienten einzelner erklärender Variablen. Nach der Darstellung der Ergebnisse in der Übersicht werden anschließend die Ergebnisse bezüglich einzelner Variablen genauer beleuchtet. Die Diskussion der Ergebnisse erfolgt entlang der Evaluierungsfragen.

4.3.2 Einfluss vorhabenspezifischer Merkmale

Inwieweit beeinflussen programm- und projektspezifische Faktoren die Nachhaltigkeitsbewertung von Vorhaben? Abbildung 6 und Abbildung 7 zeigen die marginalen Effekte aller im Modell enthaltenen programm- und projektspezifischen Variablen. Marginale Effekte leiten sich direkt aus den Regressionskoeffizienten ab (siehe Tabelle 3 und Tabelle 4). Sie zeigen den Einfluss einer erklärenden Variablen auf die Wahrscheinlichkeit, dass eine bestimmte Notenstufe vergeben wird. Dabei können marginale Effekte für jede der vier Nachhaltigkeits-Notenstufen ermittelt werden. Allerdings zeigt sich, dass die Mehrheit der Vorhaben mit den Noten 2 und 3 bewertet wird (siehe Abbildung 6). Daher werden die marginalen Effekte
hier überwiegend anhand der Notenstufe 2 diskutiert. Für alle Ergebnisse werden jedoch auch die Effekte bezüglich der anderen Notenstufen überprüft.

Die Ergebnisse zeigen, dass Ex-post-Evaluierungen die Nachhaltigkeit von Vorhaben mit längerer Dauer tendenziell besser bewerten. Dieser Zusammenhang besteht vor allem bei der Vergabe der Notenstufe 2. Dabei ist die Wahrscheinlichkeit einer guten Nachhaltigkeitsbewertung für Vorhaben mit rund 13 Jahren Projektdauer am höchsten. Mit zunehmender Dauer (> 13 Jahre) geht die Wahrscheinlichkeit zurück, bleibt aber insgesamt positiv. Bei der Interpretation dieses Effektes ist anzumerken, dass die Dauer eines Vorhabens möglicherweise mit unbeobachteten Faktoren korreliert, die ihrerseits die Nachhaltigkeitsbewertung beeinflussen. So hängt die Dauer eines Vorhabens auch von den durch das Vorhaben erzielten Wirkungen in der Vergangenheit ab. Je positiver die erzielten Wirkungen in der Vergangenheit, desto wahrscheinlicher die Bewilligung einer Folgephase. Gleichzeitig erhöht sich jedoch auch die Wahrscheinlichkeit einer besseren Nachhaltigkeitsbewertung. Im Modell der PFK, PEV und Schluss-Evaluierungen kann kein Effekt der Vorhabendauer auf die Bewertung der Nachhaltigkeit festgestellt werden.

Tabelle 3: Ergebnisse der Regressionsmodelle (Ex-post-Evaluierungen)

	Reduziertes Modell		Volles Modell	
	mit DAC	ohne DAC	mit DAC	ohne DAC
Merkmale der Vorhaben				
DAC-Bewertung (Durchschnitts-Note)	2,01*** (0,43)		2,41*** (0,49)	
Dauer (Jahre)	-0,01 (0,07)	-0,03 (0,06)	-0,64* (0,33)	-0,49* (0,28)
Dauer (Jahre quadriert)			0,03* (0,02)	0,02* (0,01)
Wert (Logarithmus der Kosten in Mio. €)	-0,18 (0,19)	-0,24 (0,19)	-0,15 (0,18)	-0,22 (0,42)
Oberziel-Dimensionen (Anzahl)	0,06 (0,43)	0,37 (0,39)	-0,18 (0,43)	0,25 (0,42)
Subsahara-Afrika (Dummy)	-0,07 (0,55)	0,55 (0,52)	-0,14 (0,62)	0,61 (0,56)
Nachhaltige Wirtschaftsförderung (Dummy)	0,18 (0,62)	0,57 (0,52)	0,24 (0,63)	0,68 (0,53)
Verzögerte Implementierung (Dummy)	-0,63 (0,46)	-0,47 (0,46)	-0,62 (0,46)	-0,38 (0,46)
GIZ (Dummy)	-1,69*** (0,68)	-2,35*** (0,66)	-2,66*** (0,87)	-2,88*** (0,83)
Implementierungskontext der Vorhaben				
BIP pro Kopf (aktuelle USD)	2E-04*** (8E-05)	2E-04** (8E-05)	3E-04*** (9E-05)	2E-04** (9E-05)
Netto-ODA (% am BIP)	5E-03 (0,03)	0,02 (0,03)	-0,02 (0,03)	-0,03 (0,03)
Freedom-House-Index (1–7)	-0,25* (0,13)	-0,27** (0,12)	-0,19 (0,16)	-0,20 (0,13)
Merkmale der Evaluierungen				
Zeitpunkt relativ zu Vorhabenende (Jahre)	0,20*** (0,08)	0,22*** (0,08)	0,22** (0,09)	0,23*** (0,09)
Bewertungskriterien der Nachhaltigkeit (Summe positiver und negativer Einflüsse)				
Gesamteinfluss	-0,37*** (0,06)	-0,45*** (0,06)		
Kriterien Kontext			-0,09 (0,21)	-0,18 (0,16)
Kriterien Implementierung			-0,74** (0,32)	-0,61** (0,30)
Kriterien Outcome			-0,14 (0,12)	-0,30*** (0,11)
Kriterien Kapazitäten vor Ort			-0,60*** (0,14)	-0,61*** (0,12)
Kriterien nicht intendierte Wirkungen (Impact)			-0,04 (0,39)	-0,20 (0,35)
Kriterien Absehbarkeit des Erhalts von Wirkungen			-0,80** (0,39)	-0,59* (0,34)
Kriterien Zusammenspiel der Dimensionen			-0,43 (0,39)	-0,45 (0,32)
Cut 1	-5,95 (3,35)	-11,23 (3,53)	-11,00 (3,95)	-16,47 (3,97)
Cut 2	-1,11 (3,21)	-6,35 (3,27)	-5,91 (3,85)	-11,31 (3,84)
Cut 3	4,25 (3,25)	-1,77 (3,29)	0,66 (3,78)	-6,13 (3,80)
Anzahl Beobachtungen	184			
Pseudo R2	0,46	0,39	0,52	0,43

	Reduziertes Modell		Volles Modell	
	mit DAC	ohne DAC	mit DAC	ohne DAC
AIC	246,92	273,18	238,44	270,57
BIC	298,36	321,40	312,38	341,30
Log. Likelihood	-107,46	-121,59	-96,22	-113,28
Chi-square	96,52	81,36	93,48	110,81

Quelle: eigene Darstellung

*Anmerkungen: Dargestellt sind Koeffizienten mit zugehörigen Standardfehlern. ***, **, * zeigen, dass die Koeffizienten auf einem Signifikanzniveau von 1, 5 bzw. 10 Prozent ungleich null sind. Signifikanzniveaus basieren auf gruppierten Standardfehlern auf Ebene eines Evaluierungsberichtes. Cut 1 bis 3 sind Schwellenwerte, welche die Übergänge zwischen den einzelnen vorhergesagten Notenstufen anzeigen. Pseudo R2 ist ein Pseudo-Bestimmtheitsmaß des Modells dessen Werte zwischen 0 (keine Vorhersage der Nachhaltigkeitsnote) und 1 (perfekte Vorhersage der Nachhaltigkeitsnote) liegen. Das Akaike-Informationskriterium (Akaike information criterion, AIC) und das Bayessche Informationskriterium (Bayesian information criterion, BIC) sind Qualitätsmaße der Modelle. Je niedriger ihr Wert, desto geringer die Wahrscheinlichkeit eines Informationsverlustes. Log. Likelihood basiert auf der Addition der Wahrscheinlichkeiten der vorhergesagten und tatsächlichen Ergebnisse und ist ein Maß der Modellgüte. Die chi-square Statistik ist ein Maß der Modellgüte.*

Tabelle 4: Ergebnisse der Regressionsmodelle (PFK, PEV, Schluss-Evaluierungen)

	Reduziertes Modell		Volles Modell	
	mit DAC	ohne DAC	mit DAC	ohne DAC
Merkmale der Vorhaben				
DAC-Bewertung (Durchschnitts-Note)	1,35*** (0,46)		1,19*** (0,44)	
Dauer (Jahre)	0,16 (0,14)	0,14 (0,14)	0,10 (0,15)	0,09 (0,16)
Wert (Logarithmus der Kosten in Mio. €)	-0,42* (0,23)	-0,48** (0,23)	-0,52** (0,23)	-0,59*** (0,23)
Anzahl Oberziel-Dimensionen	0,27 (0,25)	0,25 (0,24)	0,14 (0,26)	0,10 (0,25)
Subsahara Afrika (Dummy)	0,02 (0,44)	-0,09 (0,43)	0,05 (0,44)	-0,05 (0,44)
Nachhaltige Wirtschaftsförderung (Dummy)	0,93* (0,50)	0,76* (0,45)	0,91* (0,47)	0,79* (0,43)
Verzögerte Implementierung (Dummy)	-0,18 (0,44)	0,34 (0,36)	0,02 (0,50)	0,43 (0,43)
PEV (Dummy)	0,21 (0,83)	-0,46 (0,76)	-0,20 (0,81)	-0,78 (0,78)
PFK (Dummy)	0,86 (0,66)	0,25 (0,56)	0,44 (0,68)	-0,12 (0,59)
Implementierungskontext der Vorhaben				
BIP pro Kopf (aktuelle USD)	7E-05 (1E-04)	1E-04 (1E-04)	6E-05 (1E-04)	8E-05 (1E-04)
Netto-ODA (% am BIP)	4E-03 (0,03)	0,02 (0,03)	2E-03 (0,03)	0,02 (0,03)
Freedom-House-Index (1-7)	0,11 (0,13)	0,06 (0,13)	0,02 (0,14)	-0,20 (0,13)
Merkmale der Evaluierungen				
Zeitpunkt relativ zu Vorhabenende (Jahre)	0,60 (0,47)	0,50 (0,40)	0,46 (0,47)	0,36 (0,40)

	Reduziertes Modell		Volles Modell	
	mit DAC	ohne DAC	mit DAC	ohne DAC
Bewertungskriterien der Nachhaltigkeit (Summe positiver und negativer Einflüsse)				
Gesamteinfluss	-0,18*** (0,05)	-0,24*** (0,04)		
Kriterien Kontext			-0,60*** (0,21)	-0,61*** (0,19)
Kriterien Implementierung			-0,26* (0,16)	-0,31** (0,15)
Kriterien Outcome			-0,17* (0,10)	-0,20** (0,10)
Kriterien Kapazitäten vor Ort			-0,05 (0,08)	-0,11 (0,09)
Kriterien nicht intendierte Wirkungen (Impact)			0,54* (0,30)	0,50* (0,30)
Kriterien Absehbarkeit des Erhalts von Wirkungen			-0,76** (0,33)	-0,98*** (0,35)
Kriterien Zusammenspiel der Dimensionen			-0,02 (0,26)	-0,03 (0,25)
Cut 1	-4,98 (3,74)	-9,27 (3,56)	-7,96 (3,90)	-12,17 (3,65)
Cut 2	-2,11 (3,74)	-6,54 (3,55)	-5,02 (3,87)	-9,35 (3,60)
Cut 3	1,73 (3,82)	-2,95 (3,56)	-0,86 (3,90)	-5,38 (3,56)
Anzahl Beobachtungen	168			
Pseudo R2	0,20	0,16	0,24	0,21
AIC	330,66	343,05	330,25	338,32
BIC	383,77	393,03	402,25	407,05
Log. Likelihood	-148,33	-155,52	-142,12	-147,16
Chi-squared	52,50	46,88	65,34	59,28

Quelle: eigene Darstellung

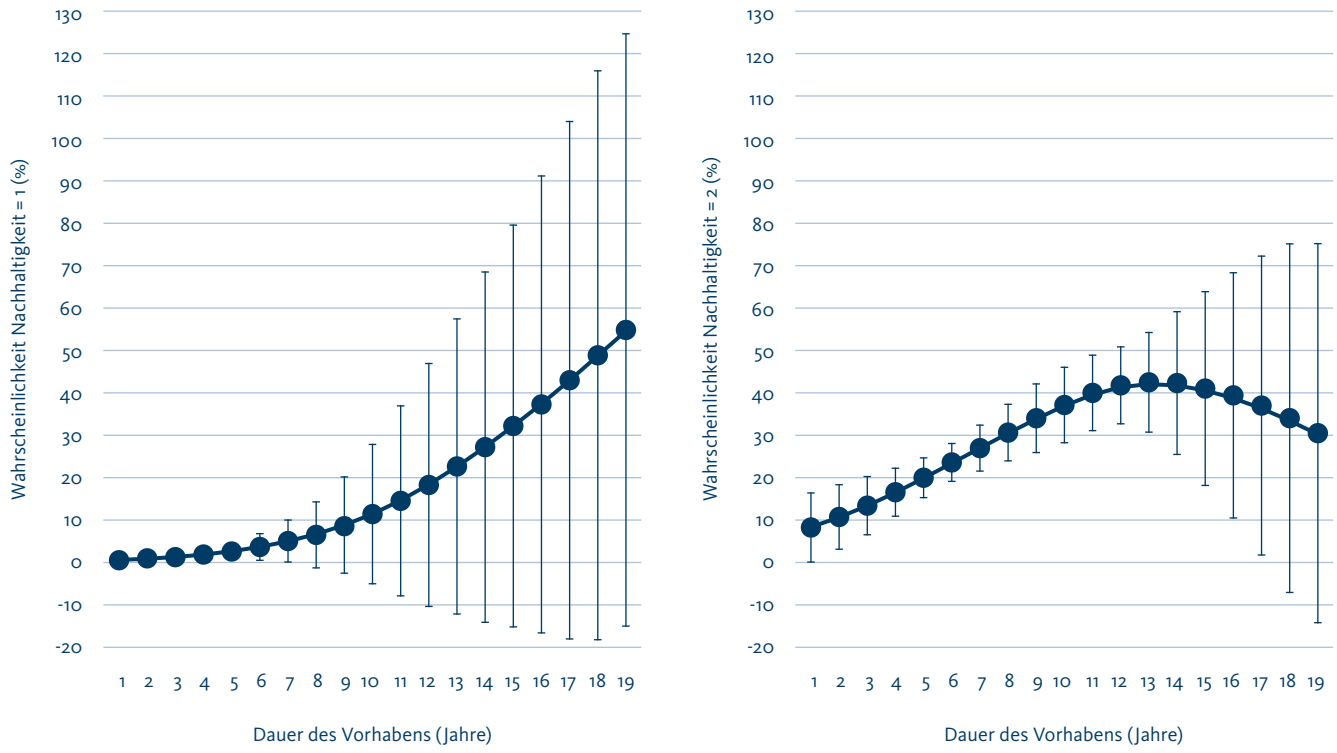
*Anmerkungen: Dargestellt sind Koeffizienten mit zugehörigen Standardfehlern. ***, **, * zeigen, dass die Koeffizienten auf einem Signifikanzniveau von 1, 5 bzw. 10 Prozent ungleich null sind. Signifikanzniveaus basieren auf gruppierten Standardfehlern auf Ebene eines Evaluierungsberichtes. Cut 1 bis 3 sind Schwellenwerte, welche die Übergänge zwischen den einzelnen vorhergesagten Notenstufen anzeigen. Pseudo R2 ist ein Pseudo-Bestimmtheitsmaß des Modells dessen Werte zwischen 0 (keine Vorhersage der Nachhaltigkeitsnote) und 1 (perfekte Vorhersage der Nachhaltigkeitsnote) liegen. Das Akaike-Informationskriterium (Akaike information criterion, AIC) und das Bayessche Informationskriterium (Bayesian information criterion, BIC) sind Qualitätsmaße der Modelle. Je niedriger ihr Wert, desto geringer die Wahrscheinlichkeit eines Informationsverlustes. Log. Likelihood basiert auf der Addition der Wahrscheinlichkeiten der vorhergesagten und tatsächlichen Ergebnisse und ist ein Maß der Modellgüte. Die chi-square Statistik ist ein Maß der Modellgüte.*

Abbildung 7 zeigt die marginalen Effekte weiterer Vorhaben-Merkmale. Hier sind die Effekte sowohl für Ex-post-Evaluierungen als auch für PFK, PEV und Schluss-Evaluierungen enthalten.

Die zunehmende finanzielle Ausstattung einer Maßnahme geht in PFK, PEV und Schluss-Evaluierungen mit einer signifikant besseren Note einher, nicht aber in Ex-post-Evaluierungen. In alternierenden Modell-Spezifikationen ist dieser Effekt nicht robust.[34] Demnach kann kein positiver Zusammenhang zwischen dem Wert einer Maßnahme und deren Nachhaltigkeit festgestellt werden. Diese Befunde decken sich mit Erkenntnissen empirischen Analyse von Evaluierungsberichten der Weltbank (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995). Diese zeigen, dass längere und kostspieligere Vorhaben nicht zwangsläufig zu besseren Bewertungen des Gesamterfolgs einer Maßnahme führen. Die Ergebnisse zeigen außerdem, dass die zunehmende Anzahl an Oberziel-Dimensionen keinen Einfluss auf die Nachhaltigkeitsnote hat. Auch eine Verzögerung bei der Implementierung wirkt sich nicht signifikant auf die Nachhaltigkeitsbewertung

34 *Werden Makro-Indikatoren ausgeschlossen, erhöht sich im PFK-, PEV- und Schluss-Modell die Anzahl der Beobachtungen von 168 auf 247. Zusätzliche Beobachtungen betreffen vor allem Regional- und Sektorvorhaben. In diesem Modell kann kein signifikanter Einfluss des Wertes einer Maßnahme auf die Nachhaltigkeitsnote festgestellt werden.*

Abbildung 6: Einfluss der Dauer eines Vorhabens auf die Nachhaltigkeitsbewertung in Ex-post-Evaluierungen



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte mit zugehörigen Konfidenzintervallen (95%). Marginale Effekte geben die Wahrscheinlichkeit an, mit der die Nachhaltigkeitsnote 1 (linkes Schaubild) bzw. die Nachhaltigkeitsnote 2 (rechtes Schaubild) für Vorhaben mit bestimmter Dauer vergeben wird. Die Ergebnisse beruhen auf den Modellen für Ex-post-Evaluierungen (siehe Tabelle 3).

aus. Nimmt die Zeitspanne zwischen Durchführung der Evaluierung und Vorhaben-Ende hingegen zu, verringert sich in Ex-post-Evaluierungen die Wahrscheinlichkeit einer guten Nachhaltigkeitsnote.[35] Auch dieser Befund deckt sich mit Erkenntnissen empirischer Studien (Bulman et al., 2015; Denizer et al., 2013; Dollar und Levin, 2005; Isham et al., 1995).

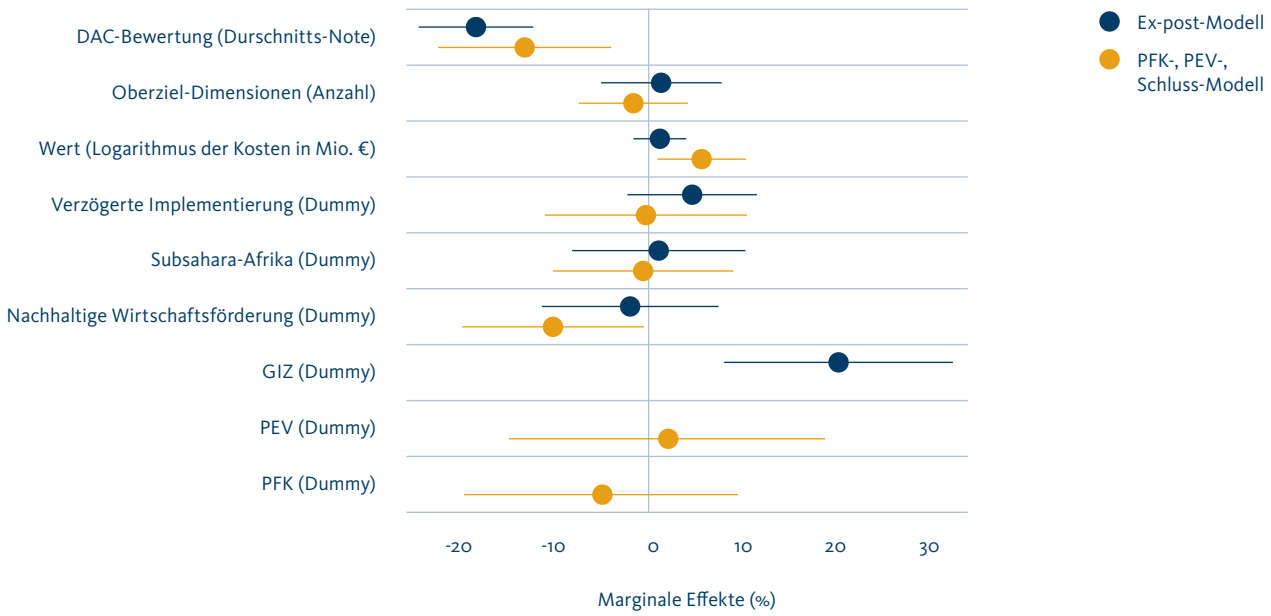
Maßnahmen, die in Subsahara-Afrika umgesetzt werden, erhalten keine bessere oder schlechtere Nachhaltigkeitsbewertung. Hinsichtlich des Umsetzungssektors zeigt sich, dass Maßnahmen im Bereich der nachhaltigen Wirtschaftsförderung in PFK, PEV und Schluss-Evaluierungen signifikant schlechter bewertet werden. Dies ist insofern bemerkenswert, als in diesen Sektoren die meisten Vorhaben der GIZ umgesetzt werden und man durchaus komparative Vorteile gegenüber Maßnahmen in anderen Sektoren vermuten würde.[36]

Weiter zeigen die Ergebnisse, dass im Modell der Ex-post-Evaluierungen für Vorhaben der GIZ eine signifikant höhere Wahrscheinlichkeit besteht, eine bessere Nachhaltigkeitsnote

[35] *Marginale Effekte sind nicht in der Abbildung enthalten. Die Aussage ergibt sich aus den Ergebnissen, die in Tabelle 3 dargestellt sind.*

[36] *Zusätzlich zu den hier dargestellten Vorhaben-Merkmalen wurde der Einfluss getestet, den jede einzelne Umsetzungsregion sowie jeder einzelne Umsetzungssektor einer Maßnahme auf die Nachhaltigkeitsnote hat. Auch Interaktionsterme zwischen DO und Region sowie zwischen DO und Sektor wurden in alternative Modelle integriert. Außer den hier dargestellten Zusammenhängen wurden dabei keine weiteren signifikanten Effekte ermittelt.*

Abbildung 7: Einfluss der Merkmale eines Vorhabens auf die Nachhaltigkeitsnote



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte mit zugehörigen Konfidenzintervallen (95%). Marginale Effekte zeigen an, wie sich die Erhöhung einer erklärenden Variablen um eine Einheit auf die Wahrscheinlichkeit auswirkt, dass die Nachhaltigkeitsnote 2 vergeben wird. Die Ergebnisse sind getrennt dargestellt für das Modell mit Ex-post-Evaluierungen sowie das Modell mit PFK, PEV und Schluss-Evaluierungen. Die Ergebnisse beruhen auf den Gesamt-Modellen (siehe Tabelle 3 und Tabelle 4). Die Referenzkategorie des GIZ-Dummys sind KfW-Vorhaben, die Referenzkategorie für PEV und PFK sind Schluss-Evaluierungen.

zu erhalten. Dieser Aspekt wird bei den Ergebnissen des Einflusses der Bewertungskriterien auf die Nachhaltigkeitsnote (siehe Abschnitt 4.3.4) ausführlicher diskutiert. Im Modell der PFK, PEV und Schluss-Evaluierungen bestehen zwischen den Evaluierungstypen keine signifikanten Unterschiede bezüglich der Notenvergabe.

4.3.3 Einfluss des Implementierungskontextes

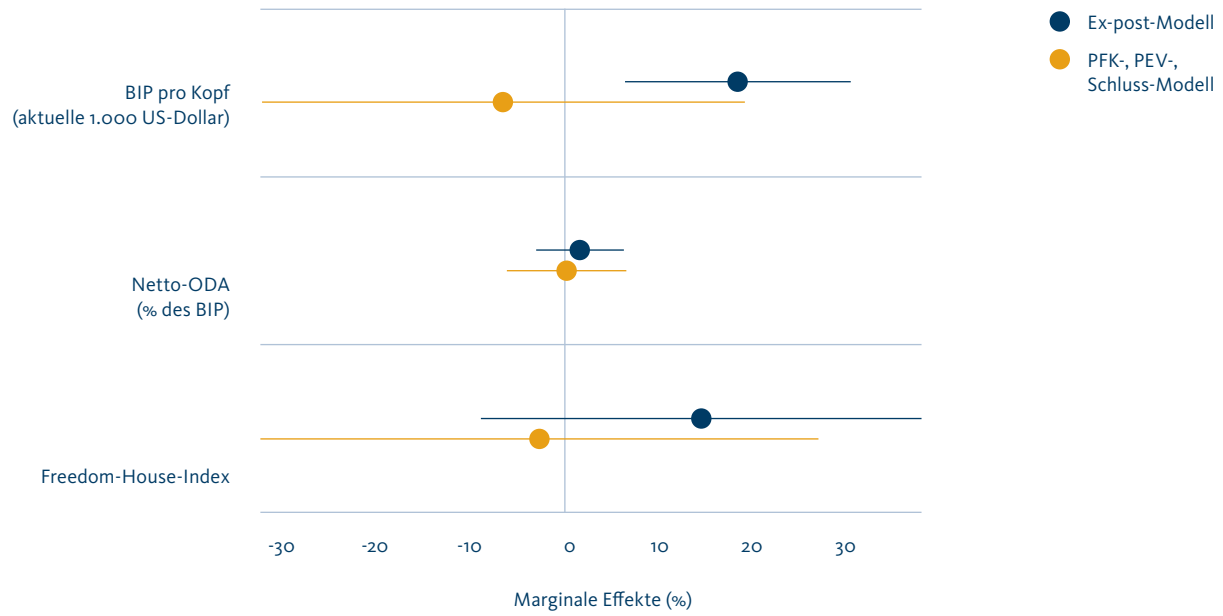
Inwieweit und in welchem Maße beeinflussen kontextspezifische Faktoren die Nachhaltigkeitsbewertung von Vorhaben? Der Einfluss des nationalen Implementierungskontextes einer Maßnahme auf dessen Nachhaltigkeitsnote wird anhand einzelner Makro-Indikatoren ermittelt. Abbildung 8 zeigt die marginalen Effekte aller im Modell enthaltenen Kontext-Variablen.

Die Ergebnisse verdeutlichen, dass im Ex-post-Modell ein positiver Zusammenhang zwischen dem wirtschaftlichen Entwicklungsstand eines Landes (dargestellt als BIP pro Kopf) und der Nachhaltigkeitsbewertung von Vorhaben besteht. Demnach führt die Erhöhung des BIP pro Kopf um 1.000 US-Dollar zu einer rund 2 Prozent höheren Wahrscheinlichkeit, dass die Note 2 vergeben wird. Auch Denizer et al. (2013) zeigen, dass der Erfolg einer Maßnahme positiv vom wirtschaftlichen Entwicklungsstand und der wirtschaftlichen Stabilität eines Landes beeinflusst wird.

Es findet sich allerdings kein Zusammenhang zwischen dem nationalen politischen Kontext (dargestellt als Freedom-House-Index) und der Nachhaltigkeitsnote.[37] Das steht im Widerspruch zu Erkenntnissen aus der Literatur. Diese zeigen, dass ein höheres Maß an Rechtsstaatlichkeit und Demokratie auf nationaler Ebene förderlich für den Gesamterfolg von Vorhaben ist (Chauvet et al., 2010; Denizer et al., 2013; Dollar und Levin, 2005).

[37] *Auch wenn der politische Kontext durch den Rule-of-Law-Index in den Modellen abgebildet wird, lässt sich kein signifikanter Zusammenhang ermitteln.*

Abbildung 8: Einfluss des Implementierungskontextes auf die Nachhaltigkeitsnote



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte mit zugehörigen Konfidenzintervallen (95%). Diese zeigen an, wie sich die Erhöhung einer erklärenden Variablen auf die Wahrscheinlichkeit auswirkt, dass die Nachhaltigkeitsnote 2 vergeben wird. Die Ergebnisse sind getrennt dargestellt für das Modell mit Ex-post-Evaluierungen sowie für das Modell mit PFK, PEV und Schluss-Evaluierungen. Die Ergebnisse beruhen auf den Gesamt-Modellen (siehe Tabelle 3 und Tabelle 4).

Demgegenüber gibt es Erkenntnisse, die darauf hinweisen, dass der Anteil von Hilfszahlungen am BIP eines Landes zu einer Verschlechterung der Ergebnisse von Vorhaben führen kann (Dollar und Levin, 2005). So können mit zunehmendem Erhalt von ODA-Transfers die Kapazitäten von Partnerländern überfordert werden (KfW Entwicklungsbank, 2003). Dies kann hier nicht bestätigt werden. Der Anteil von ODA-Mitteln am BIP eines Landes hat keinen signifikanten Einfluss auf die vergebene Nachhaltigkeitsnote. Hierbei ist jedoch zu berücksichtigen, dass die Mittelallokation möglicherweise von unbeobachteten Faktoren bestimmt wird, die ihrerseits die Nachhaltigkeitsbewertung beeinflussen. Es ist nicht auszuschließen, dass ODA-Mittel vor allem in jene Länder fließen, in denen der Bedarf besonders groß und die Rahmenbedingungen für die Umsetzung von Vorhaben besonders schwierig sind (Dollar und Levin, 2005).

Dass der nationale Kontext so wenig Einfluss zu haben scheint, mag überraschend sein, wird aber von Ergebnissen von Denizer et al. (2013) bestätigt. Diese machen deutlich, dass der Erfolg von Vorhaben innerhalb eines Landes stärker variiert als zwischen Ländern. Demnach sind projektspezifische Faktoren bedeutender für die Erklärung des Gesamterfolges einer Maßnahme.[38] Möglicherweise bilden die im Modell enthaltenen Indikatoren auf Landes-Ebene aufgrund ihres hohen Aggregationsniveaus den unmittelbaren Implementierungskontext einer Maßnahme nur unzureichend ab.

4.3.4 Einfluss der Bewertungskriterien

Inwieweit beeinflussen die in der Meta-Evaluierung erfassten Bewertungskriterien die Nachhaltigkeitsnote von Vorhaben? Die Bewertungskriterien spiegeln die durch das Vorhaben erzielten Leistungen und Wirkungen wider. Dabei wird jedem

[38] *Zusätzlich zu den hier dargestellten Kontext-Merkmalen wurde der Effekt getestet, den das jährliche Wirtschaftswachstums eines Landes (%), der Rule-of-Law-Index der Weltbank, die Lebenserwartung bei Geburt (Jahre), die Bevölkerungszahl eines Landes und die Einschulungsrate auf die vergebene Note haben. Dabei konnte für keinen dieser Faktoren ein signifikanter Zusammenhang festgestellt werden.*

berichteten Kriterium ein positiver, neutraler oder negativer Einfluss auf die Nachhaltigkeit eines Vorhabens zugeschrieben. Wie in Abschnitt 3.1 erläutert, sind die Kriterien in insgesamt sieben inhaltliche Bereiche unterteilt. Innerhalb dieser Bereiche wird der Einfluss der Einzelkriterien aggregiert. Positive Werte bedeuten, dass ein Bereich durch die Evaluierung als überwiegend förderlich für die Nachhaltigkeit eines Vorhabens eingeschätzt wird. Negative Werte zeigen an, dass ein Bereich als überwiegend hemmend für die Nachhaltigkeit von Vorhaben wahrgenommen wird.

Abbildung 9 zeigt die marginalen Effekte der im Modell enthaltenen sieben Bereiche der Bewertungskriterien.

Die Ergebnisse machen deutlich, dass bestimmte Bereiche in allen Evaluierungstypen ähnliche Einflüsse auf die Nachhaltigkeitsnote haben. So führt eine zunehmend positive Einschätzung des Bereichs Implementierung in beiden Modellen zu einer signifikant höheren Wahrscheinlichkeit, dass ein Vorhaben mit der Nachhaltigkeitsnote 2 bewertet wird (+6 Prozent im Ex-post-Modell und +3 Prozent im PFK-, PEV-, Schluss-Modell, wenn sich die Einschätzung des Bereichs um einen Wert verbessert). Im Bereich Implementierung wird der Einfluss der Kriterien „Alignment", „Partizipation" und „Steuerung" auf die Nachhaltigkeit einer Maßnahme bewertet. Diese Kriterien sind demnach besonders relevant für die Nachhaltigkeitsbewertung eines Vorhabens. Der ermittelte Zusammenhang kann jedoch auch dadurch bedingt sein, dass die Beobachtung dieser Kriterien besonders häufig in positiver Ausprägung erfolgt. In diesem Fall würden nicht die Kriterien per se, sondern die Einfachheit, mit der ihre positive Ausprägung festgestellt werden kann, die Nachhaltigkeitsnote beeinflussen. Die Ergebnisse der Meta-Evaluierung (Noltze et al., 2018) legen allerdings nahe, dass keine einseitige Berichterstattung bezüglich der Einflüsse der Kriterien im Bereich Implementierung erfolgt. Auch die Zunahme positiver Eindrücke im Bereich „Absehbarkeit des Erhalts von Wirkungen" erhöht in beiden Modellen die Wahrscheinlichkeit, dass die Nachhaltigkeitsnote 2 vergeben wird (+5 Prozent in Ex-post-Evaluierungen und +8 Prozent in PFK-, PEV- und Schluss-Evaluierungen). Die Absehbarkeit des Erhalts von Wirkungen ist ein zentraler Bestandteil der Bewertung von Nachhaltigkeit (BMZ, 2006). Daher ist es wenig überraschend, dass diesem Bereich in allen Evaluierungstypen ein wichtiger Einfluss auf die Nachhaltigkeitsnote zugesprochen wird. Gleichzeitig zeigt sich jedoch, dass Nachhaltigkeit auch von Faktoren, die über die reine Dauerhaftigkeit von Wirkungen hinausgehen, beeinflusst wird.

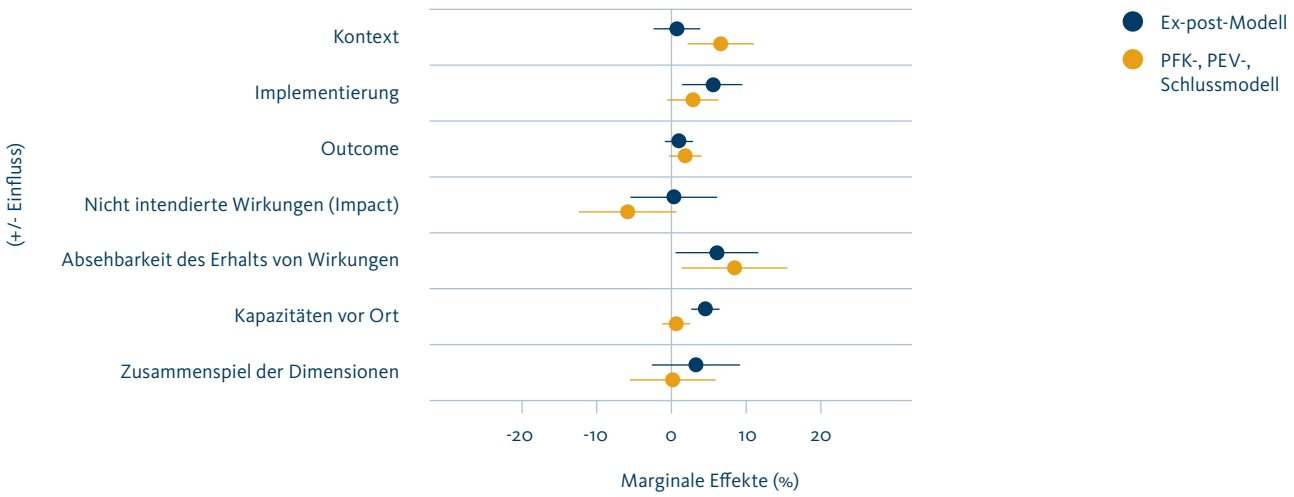
Die Ergebnisse verdeutlichen außerdem, dass sich je nach Evaluierungstyp einige Bereiche bezüglich ihres Einflusses auf die Nachhaltigkeitsbewertung unterscheiden. So wird die Vergabe der Notenstufe 2 mit zunehmend positiver Einschätzung des Implementierungskontextes einer Maßnahme nur im PFK-, PEV- und Schluss-Modell wahrscheinlicher. Während die Ergebnisse der begleitenden Meta-Evaluierung zeigen, dass der Kontext besonders häufig zur Bewertung der Nachhaltigkeit einer Maßnahme herangezogen wird (Noltze et al., 2018), machen die Regressionsergebnisse deutlich, dass sich diese insgesamt häufigere Berücksichtigung nur in PFK, PEV und Schluss-Evaluierungen signifikant in der Note niederschlägt. Aufgrund des Zeitpunktes ihrer Durchführung bewerten PFK, PEV und Schluss-Evaluierungen Nachhaltigkeit vor allem über eine Einschätzung zukünftiger Entwicklungen. Der unmittelbare Kontext eines Vorhabens ist demnach ein wichtiger Aspekt, auf dessen Grundlage eine Einschätzung über die Nachhaltigkeit der Wirkungen erfolgt.

Auch führt die bessere Einschätzung des Bereichs „Outcome" nur im PFK-, PEV-, Schluss-Modell zu einer signifikant höheren Wahrscheinlichkeit, eine Nachhaltigkeitsbenotung der Stufe 2 zu erhalten (+2 Prozent wenn sich die Einschätzung des Bereichs um einen Wert verbessert). Auch hier gilt, dass die Kriterien im Bereich Outcome als Grundlage für die Einschätzung herangezogen werden. In der retrospektiven Bewertung von Nachhaltigkeit, wie sie in Ex-post-Evaluierungen erfolgt, spielt der Bereich Outcome jedoch eine untergeordnete Rolle.

Die Einschätzung zu nicht intendierten Wirkungen hat in PFK, PEV und Schluss-Evaluierungen einen leicht signifikanten Einfluss auf die vergebene Nachhaltigkeitsnote. Die begleitende Meta-Evaluierung zeigt, dass dieser Bereich eher selten Berücksichtigung findet.

In Ex-post-Evaluierungen hingegen werden vor allem lokale Kapazitäten beleuchtet. Hier findet sich ein signifikanter Einfluss auf die Nachhaltigkeitsbewertung. So führt eine positive Einschätzung der entsprechenden Kriterien zu einer um

Abbildung 9: Einfluss der Bewertungskriterien auf die Nachhaltigkeitsnote



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte mit zugehörigen Konfidenzintervallen (95%). Diese zeigen an, wie sich die Erhöhung einer erklärenden Variablen auf die Wahrscheinlichkeit auswirkt, dass die Nachhaltigkeitsnote 2 vergeben wird. Der +/- Einfluss eines Bereichs zeigt den aggregierten Einfluss aller Kriterien eines Bereichs auf die Nachhaltigkeitsbewertung einer Evaluierung. Die Ergebnisse sind getrennt dargestellt für das Modell mit Ex-post-Evaluierungen sowie für das Modell mit PFK, PEV und Schluss-Evaluierungen. Die Ergebnisse beruhen auf den Gesamt-Modellen (siehe Tabelle 3 und Tabelle 4).

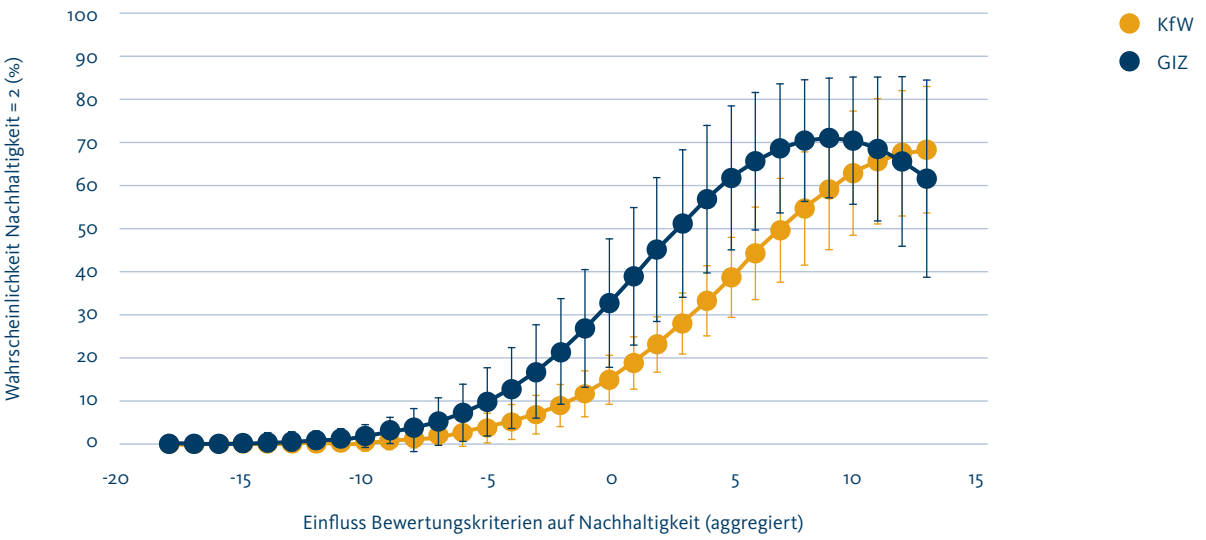
5 Prozent erhöhten Wahrscheinlichkeit, dass die Nachhaltigkeitsnote 2 vergeben wird. Lokale Kapazitäten bilden finanzielle, technische und institutionelle Fähigkeiten der Partner vor Ort ab. Es scheint plausibel, dass sich diese Faktoren besonders in der Benotung von Ex-post-Evaluierungen widerspiegeln. Da Ex-post-Evaluierungen einige Jahre nach Vorhabenende durchgeführt werden, sind die Partner alleine für die Umsetzung und Fortführung einer Maßnahme verantwortlich und stehen daher wahrscheinlich eher im Fokus der Evaluierung.

Das Zusammenspiel der Dimensionen hat in keinem der Modelle einen signifikanten Einfluss auf die vergebene Note. Dieser Bereich wird bei der Bewertung von Nachhaltigkeit generell selten betrachtet (Noltze et al., 2018).

Die Ergebnisse in Abbildung 7 haben gezeigt, dass im Ex-post-Modell für Vorhaben der GIZ eine signifikant höhere Wahrscheinlichkeit besteht, die Nachhaltigkeitsnote 2 zu erhalten (+22 Prozent) als für Vorhaben der KfW. Wie in Abbildung 10 veranschaulicht, hängt dies unmittelbar mit dem Einfluss der Bewertungskriterien auf die Nachhaltigkeit von Vorhaben zusammen. Die X-Achse zeigt den in einem Bericht festgestellten Gesamteinfluss auf die Nachhaltigkeit eines Vorhabens an (der aggregierte Einfluss aller 7 Bereiche). Der negative Wertebereich der X-Achse (-18 bis -1) repräsentiert Vorhaben mit einem Überhang negativ bewerteter Kriterien. Der positive Wertebereich (+1 bis +17) präsentiert Vorhaben mit einem Überhang positiv bewerteter Kriterien. Die Y-Achse gibt die geschätzte Wahrscheinlichkeit an, mit der ein Vorhaben mit der Nachhaltigkeitsnote 2 bewertet wird.

Die Ergebnisse machen deutlich, dass Evaluierungen der GIZ im Vergleich zu Evaluierungen der KfW bei identischen Werten mit höherer Wahrscheinlichkeit die Note 2 vergeben. Die Unterschiede zwischen den DO sind im Wertebereich von -3 bis +8 statistisch signifikant. So liegt beispielsweise die Wahrscheinlichkeit, dass eine GIZ-Maßnahme bei einem Wert von +5 mit der Note 2 bewertet wird, bei rund 61 Prozent. Dagegen liegt die Wahrscheinlichkeit, dass ein KfW-Vorhaben

Abbildung 10: Einfluss der Bewertungskriterien auf die Nachhaltigkeitsnote nach Durchführungsorganisation



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind durchschnittliche marginale Effekte und Konfidenzintervalle (95 %). Die Ergebnisse basieren auf der Hauptspezifikation des in Abschnitt 3.1 vorgestellten Regressionsmodells. Das Modell enthält alle Ex-post-Evaluierungen (n = 184). Die Beobachtungen sind nach methodischer Qualität gewichtet.

bei dem gleichen Wert mit 2 bewertet wird, bei nur rund 41 Prozent. Diese Ergebnisse legen nahe, dass sich in GIZ-Evaluierungen positive Ausprägungen der Kriterien insgesamt deutlich stärker in positiven Notenstufen ausdrücken. Im negativen Wertebereich finden sich hingegen keine signifikanten Unterschiede zwischen den DO.[39]

4.3.5 Einfluss der methodischen Qualität

Inwieweit beeinflusst die methodische Qualität der Evaluierungen die Nachhaltigkeitsbewertung? Allen Ergebnissen liegt zwar eine Gewichtung der Beobachtungen nach methodischer Qualität zugrunde, der direkte Zusammenhang zwischen Berichtsqualität und Notenvergabe wurde jedoch nicht explizit untersucht. Abbildung 11 zeigt den Zusammenhang zwischen dem Qualitätsindex und der Nachhaltigkeitsnote (siehe Noltze et al., 2018).

Wie aus Abbildung 11 ersichtlich, lässt sich kein Zusammenhang zwischen der methodischen Berichtsqualität und der vergebenen Nachhaltigkeitsnote feststellen. Das heißt, Evaluierungen mit überdurchschnittlicher methodischer Güte vergeben keine besseren oder schlechteren Nachhaltigkeitsnoten.

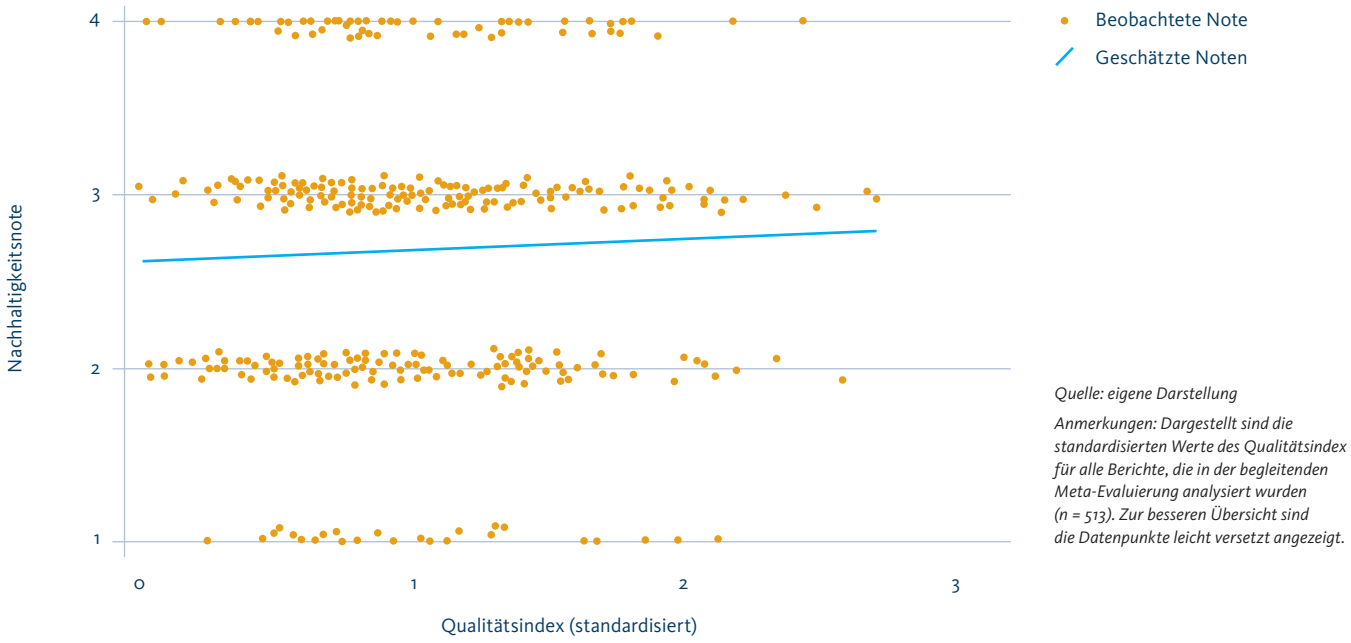
4.3.6 Übergeordnete Erkenntnisse

Tabelle 5 fasst den Erklärungsgehalt einzelner Variablen bezüglich der vergebenen Nachhaltigkeitsnote zusammen. Die erklärenden Variablen sind getrennt aufgeführt, und zwar nach DAC-Durchschnittsnote, Merkmalen der Vorhaben, Merkmalen des Implementierungskontextes, Merkmalen der Evaluierung sowie Bewertungskriterien von Nachhaltigkeit. Diese Variablen werden dabei sukzessive einem Basismodell (ohne erklärende Variablen) hinzugefügt.

Die Ergebnisse zeigen, dass allein durch das Basismodell bereits 52 Prozent aller vergebenen Noten richtig vorhergesagt werden.[40] Dies ist damit zu begründen, dass in beiden Modellen (Ex-post- und PFK-, PEV-, Schluss-Modell) rund 52 Prozent aller Beobachtungen mit der Notenstufe 3 bewertet wurden.

39 *Auch zwischen PFK, PEV und Schluss-Evaluierungen lassen sich keine Unterschiede bezüglich der Positivität der Bewertungskriterien und der Notenvergabe feststellen.*
40 *Als Basismodell wird das Modell ohne erklärende Variablen bezeichnet.*

Abbildung 11: Einfluss der methodischen Qualität auf die Nachhaltigkeitsnote



Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind die standardisierten Werte des Qualitätsindex für alle Berichte, die in der begleitenden Meta-Evaluierung analysiert wurden (n = 513). Zur besseren Übersicht sind die Datenpunkte leicht versetzt angezeigt.

Wird diesem Basis-Modell die Durchschnittsnote der übrigen DAC-Kriterien zugefügt, erhöht sich der Anteil der richtigen Vorhersagen bezüglich der Nachhaltigkeitsnote auf 64 Prozent (Ex-post-Modell) bzw. 60 Prozent (PFK-, PEV-, Schluss-Modell). Weiter zeigt sich, dass, wenn die Merkmale eines Vorhabens, des Implementierungskontextes sowie der Evaluierungen zusätzlich berücksichtigt werden, der Erklärungsgehalt beider Modelle nur geringfügig ansteigt. Dies scheint plausibel, da Merkmale der Vorhaben streng genommen lediglich den Weg zur nachhaltigen Wirkung ebnen. Wirkungen selbst werden durch sie nicht erfasst. Werden die in der Meta-Evaluierung von Noltze et al. (2018) gewonnenen Bewertungskriterien hinzugezogen, verbessert sich die Vorhersagekraft des Ex-post-Modells auf 75 Prozent. Im PFK-, PEV-, Schluss-Modell verbessert sich die Vorhersage der Notenstufe hingegen kaum. Möglicherweise ist dies auf das Analysedesign der dezentralen Evaluierungen zurückzuführen: Hier basiert der Wirkungs- und somit auch der Nachhaltigkeitsnachweis ausschließlich auf Zukunftsabschätzungen. Eine echte Messung ist schon durch den Zeitpunkt der Evaluierungen, der deutlich vor Ende der Vorhaben liegt, nicht möglich.

Tabelle 5: Anteil richtiger Vorhersagen und Akaike-Informationskriterium (AIC) nach Modell-Spezifikation

	Ex-post-Modelle		PFK-, PEV-, Schluss-Modelle	
	% richtiger Vorhersagen	AIC	% richtiger Vorhersagen	AIC
Basismodell	52	405,19	52	376,00
+ DAC-Durchschnittsnote	64	303,64	60	339,91
+ Merkmale der Vorhaben	61	305,37	59	340,10
+ Merkmale des Implementierungskontextes	62	306,83	61	345,00
+ Merkmale der Evaluierung	65	305,64	61	346,69
+ Bewertungskriterien der Nachhaltigkeit	74	238,44	65	330,25

Quelle: eigene Darstellung

Anmerkungen: Die hier dargestellten Modell-Spezifikationen bestehen aus Basismodellen (ohne erklärende Variablen), welche sukzessive um die in Abschnitt 3.1 eingeführten erklärenden Variablen erweitert werden. Spalten zwei und vier zeigen die durch das jeweilige Modell richtig vorhergesagten Notenstufen. Spalten drei und fünf zeigen das Akaike-Informationskriterium (Akaike information criterion, AIC). Dabei handelt es sich um ein Qualitätsmaß des Modells. Je niedriger der Wert, desto geringer die Wahrscheinlichkeit eines Informationsverlustes.

5.

SCHLUSSFOLGERUNGEN UND EMPFEHLUNGEN

Bevor die Ergebnisse der vorliegenden Evaluierungssynthese diskutiert werden, soll noch einmal auf die Besonderheiten der Evaluierung von Nachhaltigkeit hingewiesen werden. Diese sind für das Verständnis der Schlussfolgerungen und Empfehlungen von Bedeutung. In Evaluierungen der deutschen EZ erfolgt die Bewertung von Nachhaltigkeit zusammen mit der Bearbeitung der DAC-Kriterien Relevanz, Effektivität, Effizienz und übergeordnete entwicklungspolitische Wirkungen (Impact). Die Nachhaltigkeit von Vorhaben soll demnach anhand der Dauerhaftigkeit positiver Wirkungen, der Stabilität des Umfelds sowie der Risiken und Potenziale bewertet werden (BMZ, 2006). Die Vorgaben zur Nachhaltigkeitsbewertung haben einen engen konzeptionellen Bezug zu allen anderen DAC-Kriterien. Die Ergebnisse der begleitenden Meta-Evaluierung belegen, dass sich dieser konzeptionelle Zusammenhang auch in der Evaluierungspraxis wiederfindet. In dieser erfolgt die Nachhaltigkeitsbewertung anhand einer Vielzahl unterschiedlicher Bewertungskriterien (Noltze et al., 2018). Die hier vorgestellten Ergebnisse belegen, dass auch eine bessere Bewertung der DAC-Kriterien Relevanz, Effektivität, Effizienz und Impact mit einer besseren Nachhaltigkeitsbewertung einhergeht. Die Nachhaltigkeitsbewertung kann somit nicht isoliert von den anderen Erfolgskriterien betrachtet werden. In diesem Sinne betreffen die im Folgenden ausgesprochenen Empfehlungen auch Bereiche, die anderen DAC-Kriterien zugeordnet werden können.

Zunächst werden Empfehlungen zur Stärkung der Nachhaltigkeit von Vorhaben ausgesprochen (Abschnitt 5.1). Im Anschluss folgen einige übergeordnete Empfehlungen zur Vergleichbarkeit der Nachhaltigkeitsbewertungen. Die Empfehlungen sollen ein systematisches Lernen hinsichtlich der Erkenntnisse zu Vorhaben der deutschen staatlichen TZ und FZ fördern (Abschnitt 5.2).

5.1 Einflussfaktoren der Nachhaltigkeitsbewertung

In der Evaluierungspraxis der deutschen EZ variiert die Nachhaltigkeitsbewertung von Vorhaben nur geringfügig: Über 84 Prozent der untersuchten Evaluierungen bewerten die Nachhaltigkeit mit der Notenstufe 2 oder 3. Die Durchschnittsnote aller DAC-Kriterien (ohne Nachhaltigkeit) ist hierbei – nach statistischem Signifikanzniveau und nach Effektstärke – in allen Regressionsmodellen der wichtigste Einflussfaktor. Durch die geringe Varianz der Nachhaltigkeitsnote und die Existenz einer erklärenden Variablen mit hoher statistischer Signifikanz ist es schwer, weitere relevante Einflussfaktoren zu identifizieren. Dennoch zeigen die Regressionsmodelle, dass bestimmte Faktoren bei der Notenvergabe besonders gewichtet werden. Eine Erklärung hierfür liefern die in der begleitenden thematischen Meta-Evaluierung gewonnenen Informationen. Zwar ist die Erhebung solch zusätzlicher Informationen mittels quantitativer Inhaltsanalyse sehr aufwändig, doch steigt mit ihnen die Aussagekraft der Evaluierungssynthese deutlich.

Die Ergebnisse zeigen, dass nur wenige Faktoren sowohl im Modell der Ex-post-Evaluierungen als auch im Modell der PFK, PEV und Schluss-Evaluierungen einen statistisch signifikanten Einfluss auf die Notenvergabe haben. Demnach hängt es stark vom Zeitpunkt der Evaluierung ab, in welchem Ausmaß bestimmte Variablen bei der Notenvergabe berücksichtigt werden. Im Folgenden werden die wichtigsten Ergebnisse diskutiert und anschließend Empfehlungen abgeleitet.

5.1.1 Einfluss von Leistungen und Wirkungen der Vorhaben

In der Gesamtschau der Regressionsmodelle zeigt sich, dass neben der Durchschnittsnote der DAC-Kriterien (ohne Nachhaltigkeit) vor allem die in der Meta-Evaluierung von Noltze et al. (2018) ermittelten Kriterien der Nachhaltigkeitsbewertung einen signifikanten Einfluss auf die vergebene Nachhaltigkeitsnote ausüben. Generell kann festgehalten werden, dass in Ex-post-Evaluierungen die Rolle und Beiträge der entwicklungspolitischen Partner und Zielgruppen von besonderer Bedeutung für die Nachhaltigkeitsbewertung von Vorhaben sind. Demgegenüber werden in PFK, PEV und Schluss-Evaluierungen vor allem die direkten Leistungen und die Umsetzung einer Maßnahme sowie der unmittelbare Implementierungskontext berücksichtigt. Die unterschiedliche Gewichtung der einzelnen Bereiche ist wahrscheinlich durch den Zeitpunkt, zu dem die jeweiligen Evaluierungstypen eingesetzt werden, bedingt. Während Ex-post-Evaluierungen ihre Bewertung auf Beobachtungen gründen, geschieht die Nachhaltigkeitsbewertung im Rahmen von PFK, PEV und Schluss-Evaluierungen in Form einer Prognose. Drei bis fünf Jahre nach Vorhaben-Ende sind vor

allem die Kapazitäten der Partner und weniger die Umsetzungsstrukturen eines Vorhabens beobachtbar. Wird eine Prognose noch während der Umsetzung einer Maßnahme abgegeben, dienen dementsprechend eher die Aktivitäten einer Maßnahme und der unmittelbare Kontext als Bewertungsgrundlage.

Es finden sich jedoch auch Gemeinsamkeiten bezüglich der ermittelten Einflussfaktoren. So wird der Absehbarkeit des Erhalts von Wirkungen sowohl in Ex-post-Evaluierungen als auch in PFK, PEV und Schluss- Evaluierungen ein deutlich positiver Einfluss auf die Nachhaltigkeit von Vorhaben zugesprochen. Dies zeigt, dass die Dauer von Wirkungen – ein Kernelement der Nachhaltigkeitsbewertung – in allen Evaluierungstypen einen signifikanten Einfluss auf die Notenvergabe hat.

Mit Blick auf die Leistungen und Wirkungen der Vorhaben zeigt sich, dass sich die Nachhaltigkeit vor allem durch Stellschrauben, die im direkten Einflussbereich der Vorhaben liegen, deutlich erhöhen lässt.

Im Folgenden werden nun die Empfehlungen genannt, die sich aus den Ergebnissen und Schlussfolgerungen der Evaluierungssynthese ergeben. Diese werden in den jeweiligen Unterpunkten durch Anregungen und Gedanken, die sich vornehmlich auf die Umsetzung beziehen, ergänzt.

1. Dem BMZ und den DO wird empfohlen, die Kapazitäten der Partner und Träger vor Ort bei der Planung und Durchführung von Vorhaben stärker zu berücksichtigen und systematisch zu fördern.
   - In diesem Sinne könnte eine explizite Abschätzung der Kapazitäten aller relevanten Partner und Träger bereits bei der Planung von Vorhaben in das Votum zur Förderwürdigkeit eines Moduls einfließen. Dabei sollte sichergestellt werden, dass auf Seiten der Partner und Träger die technischen, finanziellen und institutionellen Voraussetzungen für die Fortführung der vormals durch die Maßnahme erbrachten Leistungen gegeben sind.
   - Darüber hinaus könnte die Prüfung der Partner- und Trägerkapazitäten in regelmäßigen Abständen während eines laufenden Vorhabens wiederholt werden. Die Übertragung der Leistungen auf die Partner zum Ende eines Vorhabens könnte zudem durch die Entwicklung langfristiger Exit-Strategien sichergestellt werden.
   - Durch die Stärkung des Partnersystems könnte die Eigenverantwortung der Partnerländer hinsichtlich der Umsetzung der Agenda 2030 sichergestellt werden.
2. GIZ und KfW wird empfohlen, steuerungsrelevante Faktoren eines Vorhabens zukünftig nicht nur im Hinblick auf die Wirksamkeit, sondern auch im direkten Bezug zur Nachhaltigkeit zu verstehen und zu berücksichtigen.
   - Hierzu zählen insbesondere die Nutzung institutioneller Strukturen vor Ort, die Aufbereitung von Lernerfahrungen sowie das Entwickeln von Upscaling- und Exit-Strategien.

5.1.2 Einfluss von Merkmalen der Vorhaben

Die Evaluierungssynthese zeigt, dass einzelne Merkmale der Vorhaben die Nachhaltigkeitsnote signifikant beeinflussen. Im Vergleich zu den Leistungen und Wirkungen einer Maßnahme ist der Einfluss dieser Merkmale jedoch geringer – und damit auch ihr Aussagewert im Rahmen der Modellierung. Dies ist durchaus plausibel, denn Merkmale eines Vorhabens haben keinen direkten Einfluss auf dessen Nachhaltigkeit. Vielmehr bilden sie den Rahmen für die Umsetzung eines Vorhabens und dessen Wirkungserreichung. Beispielsweise steigt zwar mit zunehmendem Finanzvolumen der Spielraum eines Vorhabens; die Wirkung auf die Nachhaltigkeit hängt aber weniger mit der Höhe der Mittel zusammen als vielmehr damit, was mit den (begrenzten) Mitteln erreicht wird. Dennoch können aus den Ergebnissen einige Erkenntnisse abgeleitet werden.

Zu den Kern-Merkmalen eines Vorhabens gehören dessen Dauer und Finanzvolumen. Im Hinblick auf die Effektivität von EZ-Maßnahmen haben Denizer et al. (2013) gezeigt, dass längere und kostspieligere Vorhaben nicht zwangsläufig besser bewertet werden. Die Ergebnisse der vorliegenden Evaluierungssynthese sind diesbezüglich ambivalent. In Ex-post-Evaluierungen zeigt sich ein positiver Zusammenhang zwischen der Dauer einer Maßnahme und deren Nachhaltigkeit. Dieser zeigt sich nicht in PFK, PEV und Schluss-Evaluierungen. Die finanzielle Ausstattung eines Vorhabens hingegen hat in PFK, PEV und Schluss-Evaluierungen einen positiven Einfluss auf dessen Nachhaltigkeit. In Ex-post-Evaluierungen findet sich dieser Zusammenhang nicht. Der mögliche Einfluss von Dauer und Finanzvolumen scheint vielmehr kontextspezifisch zu sein.

Bemerkenswert ist darüber hinaus, dass Regional- oder Sektor-Expertise die Nachhaltigkeit von Vorhaben nicht positiv beeinflussen: GIZ und KfW sind in Regionen beziehungsweise in Sektoren, in denen sie viel Arbeitserfahrung haben, ebenso „nachhaltig" wie in Regionen und Sektoren, in denen sie weniger aktiv sind.

5.1.3 Einfluss des Implementierungskontextes

Eine wesentliche Beeinflussung der Nachhaltigkeit durch externe Kontextfaktoren kann nach den Ergebnissen der Evaluierungssynthese weitestgehend ausgeschlossen werden. Neben makro-ökonomischen und politischen Kennzahlen auf nationaler Ebene wurden auch spezifische Informationen zum lokalen Kontext einer EZ-Maßnahme in die Modelle aufgenommen. Weder der nationale noch der lokale Implementierungskontext einer Maßnahme haben den Ergebnissen der Regressionsmodelle zufolge einen hohen Erklärungswert für die Nachhaltigkeit eines Vorhabens. Einen nachweisbar positiven Einfluss zeigt hier lediglich der wirtschaftliche Entwicklungsstand eines Landes, und zwar in Ex-post-Modellen. Die geringe Aussagekraft von Makro-Indikatoren auf nationaler Ebene zeigt sich auch in vergleichbaren Studien zur Effektivität von Entwicklungsmaßnahmen und überrascht insofern auch nicht, wenn es um die Nachhaltigkeit von Vorhaben geht (Bulman et al., 2015; Denizer et al., 2013). Aus Sicht der Vorhaben ist dies zunächst eine gute Nachricht, da sie Kontextfaktoren nicht unmittelbar beeinflussen können und deren Einfluss mehr oder weniger als gegeben hingenommen werden muss. Die Nachhaltigkeit liegt vielmehr in der Hand der DO, die gemeinsam mit den Partnern, Trägern und Zielgruppen vor Ort für die Ausgestaltung nachhaltiger Strukturen und Prozesse verantwortlich sind.

5.2 Systematisches, strategisches und institutionenübergreifendes Lernen aus Evaluierungen

Die Vielfalt an Bewertungskriterien, uneinheitliche Evaluierungstypen sowie unterschiedliche Formate und Inhalte bei der Aufbereitung von Meta-Daten von Vorhaben und Evaluierungen erschweren die Vergleichbarkeit der Ergebnisse und somit ein systematisches Lernen. Dies hat verschiedene Gründe:

Erstens dienen die Leitfragen zur Bewertung von Nachhaltigkeit (BMZ, 2006) zwar als Orientierung bei der Notenvergabe, geben aber keine ausreichenden Vorgaben zur Operationalisierung. Die konkreten Bewertungskriterien, die sich hinter jeder einzelnen Note verbergen, sind vielfältig und nicht eindeutig zu benennen. Eine grundsätzliche Flexibilität bei der Bewertung ist zwar aufgrund des diversen Portfolios an umgesetzten Maßnahmen notwendig; dennoch muss die Nachhaltigkeitsbewertung auch für Außenstehende nachvollziehbar und vergleichbar sein. Zu Beginn der Jahrtausendwende stand der Harmonisierungsgedanke im Zentrum des Konzeptes „Evaluierung aus einem Guss". Mit der Agenda 2030 drückt sich dieser Gedanke im Prinzip der gemeinsamen Rechenschaftslegung aus.

Zweitens weisen die hier betrachteten DO systematische Unterschiede in Bewertungspraxis und Evaluierungsmanagement auf. Die Ergebnisse belegen, dass Evaluierungen der GIZ – bei gleicher Anzahl positiv bewerteter Kriterien – im Vergleich zu denen der KfW signifikant bessere Nachhaltigkeitsnoten vergeben. Darüber hinaus führt die Anwendung verschiedener Evaluierungstypen zu strukturellen Unterschieden in der Bewertung von Nachhaltigkeit. Je nach angewandtem Evaluierungstyp erfolgt diese entweder in Form einer Zukunftseinschätzung (PFK, PEV, Schluss-Evaluierung) oder in Form einer Retrospektive (GIZ- und KfW-Ex-post-Evaluierung). Ferner existieren Unterschiede im Management und in der Überprüfung der Evaluierungsergebnisse. Bei der KfW werden alle Ex-post-Evaluierungen von der Evaluierungsabteilung inhaltlich geprüft und abgenommen. Dabei wird die Bewertung einzelner Maßnahmen in den Kontext der Bewertung vergleichbarer Maßnahmen gesetzt. Eventuell auftretende Diskrepanzen können so vermieden werden. Demgegenüber wurden bzw. werden PFK und PEV der GIZ dezentral beauftragt und abgenommen. Dabei liegt die Durchführung im Verantwortungsbereich des oder der Auftragsverantwortlichen der jeweiligen Maßnahme. Während bei der KfW ein Kernteam an Mitarbeitenden alle Berichte kontrolliert und somit ein Mindestmaß an Vergleichbarkeit schafft, ist ein organisationsweiter Abgleich einzelner Berichte bei der GIZ durch das System dezentraler Evaluierungen nicht möglich. Es ist daher zu vermuten, dass Bewertungen von GIZ-Vorhaben insgesamt heterogener sind und stärker als bei der KfW von Eigenschaften der Autoren abhängen.

Drittens sind die von den DO aufbereiteten Meta-Daten von Evaluierungen und Vorhaben nur teilweise deckungsgleich. Für die vorliegende Analyse relevante Informationen wurden teilweise unvollständig oder aber nur von einer DO systematisch erfasst.[41]

Vor dem Hintergrund der unzureichenden Systematik in der bisherigen Evaluierungs- und Bewertungspraxis zu Nachhaltigkeit, aber auch zur entwicklungspolitischen Wirksamkeit insgesamt, unterstützen die Ergebnisse der vorliegenden Evaluierungssynthese die an das BMZ gerichtete Empfehlung der begleitenden Meta-Evaluierung, die Evaluierungspraxis von GIZ und KfW zu harmonisieren (siehe Empfehlung 8 in Noltze et al., 2018). Darüber hinaus ergibt sich eine Reihe von weiteren Empfehlungen für die Weiterentwicklung der Evaluierungspraxis. Auch die folgenden beiden Empfehlungen werden jeweils durch Anregungen und Gedanken, die sich vornehmlich auf die Umsetzung beziehen, ergänzt.

3. Um die systematische Bewertung von Nachhaltigkeit zu gewährleisten, wird dem BMZ und den DO empfohlen, einheitliche und verbindliche Kriterien zu entwickeln. Diese sollten als Grundlage der Notenvergabe dienen und hierfür transparent gewichtet werden.
   - Um dabei dem heterogenen Portfolio deutscher FZ und TZ gerecht zu werden, sollte auf angemessene sektor- oder regionalspezifische Flexibilität der Kriterien geachtet werden. Ein verbindlicher Umgang mit den Kriterien könnte gegebenenfalls auch sektoral oder für TZ-/FZ-Module getrennt festgelegt werden.
4. Dem BMZ und den DO wird empfohlen, Meta-Daten zu Vorhaben und deren Evaluierungen – soweit möglich – zwischen den DO harmonisiert und zentral zu erfassen.
   - Eine systematische und zentrale Erfassung der Meta-Daten von Vorhaben und Evaluierungen würde institutionsübergreifende, aggregierte Analysen erheblich erleichtern und somit beschleunigen.
   - Vor diesem Hintergrund könnten das BMZ und die DO prüfen, wie den Anforderungen der gemeinsamen Rechenschaftspflicht im Sinne der Agenda 2030 durch die Erfassung und Aufbereitung von Meta-Daten Rechnung getragen werden kann.

[41] *Beispielsweise sind die OECD-DAC-Kennungen der Haupt- und Nebenziele eines Vorhabens in den Meta-Daten der GIZ unvollständig. Beide DO machen keine Angaben zur Dauer der Evaluierung (Arbeitstage) sowie zur Dauer eines Vor-Ort-Aufenthaltes der Evaluierungsmission.*

6.

LITERATUR

Assefa, Y. et al. (2014), „Macro and micro determinants of project performance", *African Evaluation Journal*, Vol. 2, Nr. 1.

Benoit, S. et al. (2017), *„Evaluation: a missed opportunity in the SDG's first set of Voluntary National Reviews"*, IIED Briefing Paper, IIED, EvalSDG, EvalPartners, London.

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.

Borenstein, M. et al. (2009), „Introduction to Meta-Analysis", Wiley, West Sussex, United Kingdom.

Bulman, D. et al. (2015), *„Good Countries or Good Projects?"*, Nr. 7245, Policy Research Working Paper, World Bank Group, Washington, DC.

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

Caspari, A. (2014), *„Sektorbezogene Querschnittsauswertung: Meta-Evaluierung Ländliche Entwicklung"*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Eschborn.

Chauvet, L. et al. (2010), *„What Explains Aid Project Success in Post-Conflict Situations?"*, Nr. 5418, Policy Research Working Paper, World Bank Group, Washington, DC.

Denizer, C. et al. (2013), „Good countries or good projects? Macro and micro correlates of World Bank project performance", *Journal of Development Economics*, Vol. 105, S. 288–302.

Dollar, D. und V. Levin (2005), *„Sowing and Reaping: Institutional Quality and Project Outcomes in Developing Countries"*, Nr. 3524, Policy Research Working Papers, World Bank Group, Washington, DC.

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

Hemmer, H.-R. und A. Lorenz (2003), „What determines the success or failure of German bilateral financial aid?", *Review of World Economics*, Vol. 139, Nr. 3, S. 507–549.

Huber, S. et al. (2014), *„Querschnittsauswertung Bildung: Meta-Evaluierung und Synthese"*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn.

Isham, J. et al. (1995), „Does participation improve performance? Establishing causality with subjective data", *The World Bank Economic Review*, Vol. 9, Nr. 2, S. 175–200.

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.

Kilby, C. (2013), „The political economy of project preparation: An empirical analysis of World Bank projects", *Journal of Development Economics*, Vol. 105, S. 211–225.

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.

Lucks, D. et al. (2016), *„Counting critically: SDG „follow-up and review" needs interlinked indicators, monitoring and evaluation"*, IIED Briefing Paper, IIED, EvalSDG, EvalPartners, London.

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

OECD (1991), „The DAC Principles for Evaluation of Development Assistance", OECD Publishing, Paris.

Ofir, Z. et al. (2016), „*Five considerations for national evaluation agendas informed by the SDGs*“, IIED Briefing Paper, International Institute for Environment and Development, London.

Schwandt, T. et al. (2016), „*Evaluation: a crucial ingredient of SDG success*“, IIED Briefing Paper, IIED, EvalSDG, EvalPartners, London.

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.

World Bank (2017), „World Development Indicators“, *http://data.worldbank.org/*.

7.

ANHANG

7.1 Tabellen

Tabelle 6: 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 Nebenwirkungen 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 ausreichend bzw. eventuelle Nebenwirkungen (im Sinne von Zielkonflikten zwischen Dimensionen) „hinnehmbar" sind.

Quelle: eigene Darstellung

Anmerkungen: Für eine ausführliche Diskussion des Analyserasters siehe Noltze et al. (2018).

Tabelle 7: Analyseraster der Qualitätsbewertung

Bereiche	Nr.[42]	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 Wirkungszusammenhänge	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. Wirkungshypothesen 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.

[42] *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.*

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. Datenerhebungsmethoden		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.
5. Evaluierungsdesign	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.[43]
7. Auswertung und Schlussfolgerungen	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

Anmerkungen: Für eine ausführliche Diskussion des Analyserasters siehe Noltze et al. (2018).

Tabelle 8: Merkmale der Vorhaben, Evaluierungsmissionen und Evaluierungen nach DO

	GIZ (n = 553)	KfW (n = 462)	% Unterschied
Regionale Verteilung (% der Vorhaben)			
Subsahara-Afrika	29,48	38,74	-31 ***
Asien/Ozeanien	24,77	25,76	-4ns
Europa/Kaukasus	14,65	14,07	4ns
Lateinamerika	13,56	11,04	19ns
Nordafrika/Naher Osten	10,31	10,39	<1ns
Überregional	7,23	Keine Vorhaben	
Sektorale Verteilung (% der Vorhaben)			
Wirtschaft	26,04	19,70	24**
Demokratie	23,33	10,39	55***
Wasser	8,86	18,40	-108***
Gesundheit	7,05	14,94	-112***
Umwelt	12,48	8,44	32**
Andere	22,24	28,13	-26**

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

Merkmale der Vorhaben			
Start (Jahr)	2008 (3,53)	2002 (4,70)	<1***
Dauer (Jahre)	3,38 (1,28)	7,25 (3,24)	-114***
Wert (Mio. €) (GIZ n = 473, KfW n = 458)	7,38 (7,31)	42,70 (211,0)	-479***
Kennungen (Anzahl) (GIZ n = 383, KfW n = 434)	2,28 (1,89)	2,65 (1,42)	-16***
Merkmale der Evaluierungen			
Zeitpunkt relativ zu Vorhabenende (Jahre)	0,04 (1,72)	3,41 (2,37)	-8.425***
Feldmission (%) (GIZ n = 512, KfW n = 417)	97	79	18***
Evaluator (Anzahl) (GIZ n = 537, KfW n = 417)	3,28 (1,37)	3,24 (0,81)	1ns
Berichtete Nachhaltigkeits-Kriterien (Anzahl)	6,19 (0,27)	4,12 (0,28)	33***
Positivität der Nachhaltigkeits-Kriterien	0,33 (0,14)	0,03 (0,13)	91*
Nachhaltigkeitsnote	2,55 (0,75)	2,83 (0,72)	-11***
GIZ Ex-post	2,75 (0,86)		-3ns
GIZ Schluss	2,80 (0,63)		-1ns
PFK	2,56 (0,65)		-11***
PEV	2,30 (0,92)		-23***

Quelle: eigene Darstellung

*Anmerkungen: Dargestellt sind Mittelwerte und Standardabweichungen für die Grundgesamtheit nach DO (n = 1.015). Die in Spalte vier dargestellten Werte sind prozentuale Unterschiede zwischen den DO bezüglich einzelner Variablen. **, *** zeigen, dass sich die Werte auf einem Signifikanzniveau von 5 Prozent bzw. 10 Prozent unterscheiden. „ns" bedeutet, dass es keine signifikanten Unterschiede gibt. Die Angaben in hochgestellten Klammern zeigen, für wie viele der Beobachtungen Informationen zu den jeweiligen Variablen vorliegen. Informationen zu einzelnen Variablen ohne Klammer sind vollständig.*

Tabelle 9: Nachhaltigkeitsnote und Stichprobenumfang nach Evaluierungstyp

Evaluierungstyp	Anzahl	Nachhaltigkeitsnote (Standardabweichung)	Stichprobe Note	„Nachhaltige" Vorhaben (%) (Note 1–3)	Stichprobe Anteil	Anzahl Beobachtungen Stichprobe
GIZ Ex-post	56	2,75 (0,86)	47	80,4	46	47
GIZ Schluss	44	2,80 (0,63)	34	88,6	38	38
GIZ PFK	343	2,56 (0,65)	110	95,9	174	174
GIZ PEV	110	2,30 (0,93)	82	89,0	80	82
Zwischensumme	553		273	92,4	338	341
KfW Ex-post	462	2,83 (0,72)	140	84,2	172	172
Total	1.015		413		509	513

Quelle: eigene Darstellung

Anmerkungen: Die Größe der Stichprobe wird durch die durchschnittlich vergebene Nachhaltigkeitsnote nach Evaluierungstyp bestimmt (Stichprobe Note) bzw. durch den Anteil der als „nachhaltig" bewerteten Vorhaben je Evaluierungstyp (Stichprobe Anteil). Die verwendete Formel lautet: $sd^2/((\epsilon^2)/(z^2))+sd^2/N$). Mit sd = Standardabweichung (Stichprobe Note) bzw. Anteil „nachhaltiger" Vorhaben (Stichprobe Anteil), N = Grundgesamtheit, z = t-Verteilungswert von $1-0{,}05/2$ und ϵ= maximaler Fehler. Annahmen: $\epsilon = 0{,}1$ und $z = 1{,}96$ ($\alpha = 0.05$).

Tabelle 10: Kontrollvariablen im Hauptmodell

Variable	Definition	Einheit	Quelle	Mittelwert (Standardabweichung)	
				Ex-post (n = 184)	PFK, PEV, Schluss (n = 164)
DAC-Bewertung	Durchschnittliche Bewertung	Note	Meta-Daten GIZ und KfW	2,64 (0,68)	2,19 (0,55)
Dauer	Länge der Vorhaben von Start bis Ende	Jahre	Meta-Daten GIZ und KfW	6,94 (3,43)	2,77 (1,26)
Finanzvolumen	Gesamtwert der Vorhaben	Logarithmus Mio.€	Meta-Daten GIZ und KfW	16,03 (1,21)	15,44 (0,95)
Anzahl Oberziel- Dimensionen	Durch das Vorhaben verfolgte Oberziele	Anzahl	Evaluierungsberichte GIZ und KfW	1,46 (0,54)	1,72 (0,79)
Subsahara-Afrika	Vorhaben wird in Subsahara-Afrika umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW	0,43	0,35
Nachhaltige Wirtschaftsförderung	Vorhaben wird im Sektor nachhaltige Wirtschaftsförderung umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW	0,22	0,29
Verzögerte Implementierung	Vorhaben wird mit Verzögerung implementiert	Dummy	Evaluierungsberichte GIZ und KfW	0,46	0,19
GIZ	Vorhaben wird von der GIZ umgesetzt	Dummy	Meta-Daten GIZ und KfW	0	1
BIP pro Kopf	Bruttoinlandsprodukt pro Kopf	aktuelle 1.000 USD	Weltbank	2,32 (2,58)	2,25 (2,29)
Netto-ODA	Anteil Official Development Assistance am Bruttoinlands-produkt	Prozent	Weltbank	5,85 (7,31)	5,99 (7,94)
Freedom-House-Index	Index zur Erfassung des Freiheitstatus' eines Landes	Index-Wert	Freedom in the World	3,98 (1,51)	3,90 (1,50)
Zeitpunkt der Evaluierung relativ zum Vorhabenende	Zeitraum zwischen Vorhabenende und Datum der Evaluierung	Jahre	Meta-Daten und Evaluierungsberichte GIZ und KfW	3,82 (2,30)	-0,40 (0,64)
Gesamteinfluss der Kriterien	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien	Aggregierter Einfluss aller Kriterien (-48 bis +48)	Evaluierungsberichte GIZ und KfW	-0,22 (5,17)	0,54 (4,19)
Kriterien Kontext	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Kontext	Aggregierter Einfluss der Kriterien (-4 bis +4)	Evaluierungsberichte GIZ und KfW	-0,62 (1,13)	-0,5 (1,00)
Kriterien Implementierung	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Implementierung	Aggregierter Einfluss der Kriterien (-8 bis +8)	Evaluierungsberichte GIZ und KfW	0,11 (0,97)	0,30 (1,29)
Kriterien Outcome	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Outcome	Aggregierter Einfluss der Kriterien (-13 bis +13)	Evaluierungsberichte GIZ und KfW	0,35 (2,54)	1,25 (2,20)
Kriterien Kapazitäten vor Ort	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Kapazitäten vor Ort	Aggregierter Einfluss der Kriterien (-9 bis +9)	Evaluierungsberichte GIZ und KfW	-0,06 (2,40)	-0,51 (1,79)

Variable	Definition	Einheit	Quelle	Mittelwert (Standardabweichung)	
Kriterien Nicht intendierte Wirkungen	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches nicht intendierte Wirkungen	Aggregierter Einfluss der Kriterien (-4 bis +4)	Evaluierungsberichte GIZ und KfW	0,14 (0,58)	0,12 (0,63)
Kriterien Absehbarkeit des Erhalts von Wirkungen	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Absehbarkeit des Erhalts von Wirkungen	Aggregierter Einfluss der Kriterien (-4 bis +4)	Evaluierungsberichte GIZ und KfW	0,14 (0,68)	0,21 (0,66)
Kriterien Zusammenspiel der Dimensionen	Einfluss der in der begleitenden Meta-Evaluierung erfassten Bewertungskriterien des Bereiches Zusammenspiel der Dimensionen	Aggregierter Einfluss der Kriterien (-6 bis +6)	Evaluierungsberichte GIZ und KfW	0,30 (0,69)	0,34 (0,76)

Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind alle im Hauptmodell enthaltenTabelle 11: en Kontrollvariablen (siehe Abschnitt 3.1).

Tabelle 11: Kontrollvariablen zusätzlicher Modelle

Variable	Definition	Einheit	Quelle
Regionalvorhaben	Das Vorhaben ist ein Regionalvorhaben	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Sektorvorhaben	Das Vorhaben ist ein Sektorvorhaben	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Asien/Ozeanien	Vorhaben wird in Asien/Ozeanien umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Europa/Kaukasus	Vorhaben wird in Europa/Kaukasus umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Lateinamerika	Vorhaben wird in Lateinamerika umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Nordafrika/Naher Osten	Vorhaben wird in Nordafrika/Naher Osten umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Bildung	Vorhaben wird im Sektor Bildung umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Demokratie	Vorhaben wird im Sektor Demokratie umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Energie	Vorhaben wird im Sektor Energie umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Frieden	Vorhaben wird im Sektor Frieden umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Gesundheit	Vorhaben wird im Sektor Gesundheit umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Landwirtschaft	Vorhaben wird im Sektor Landwirtschaft umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Transport	Vorhaben wird im Sektor Transport umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Wasser	Vorhaben wird im Sektor Wasser umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Umwelt	Vorhaben wird im Sektor Umwelt umgesetzt	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
GIZ und Region	Vorhaben der GIZ werden Vorhaben der KfW in verschiedenen Regionen gegenübergestellt	Interaktionsterm	Meta-Daten und Evaluierungsberichte GIZ und KfW
GIZ und Sektor	Vorhaben der GIZ werden Vorhaben der KfW in den verschiedenen Sektoren gegenübergestellt	Interaktionsterm	Meta-Daten und Evaluierungsberichte GIZ und KfW
Ökonomisches Oberziel	Vorhaben hat ein ökonomisches Oberziel	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Soziales Oberziel	Vorhaben hat ein soziales Oberziel	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Politisches Oberziel	Vorhaben hat ein politisches Oberziel	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Ökologisches Oberziel	Vorhaben hat ein ökologisches Oberziel	Dummy	Meta-Daten und Evaluierungsberichte GIZ und KfW
Rule of Law	Index zur Erfassung des Niveaus der Rechtsstaatlichkeit	Index	Weltbank

BIP-Wachstum	Jährliche Änderungsrate des Bruttoinlandsprodukts	Prozent	Weltbank
Lebenserwartung	Lebenserwartung bei Geburt	Jahre	Weltbank
Bevölkerung	Bevölkerung eines Landes	Logarithmus der Einwohnerzahl	Weltbank
Einschulungsrate	Einschulungsrate Grundschule	Prozent Schüler an Altersklasse	Weltbank
Anzahl Evaluatoren	Anzahl an Erstellung der Evaluierung beteiligter Personen	Anzahl	Meta-Daten und Evaluierungsberichte GIZ und KfW
Datum der Evaluierung	Datum der Fertigstellung der Evaluierung	Jahr	Meta-Daten und Evaluierungsberichte GIZ und KfW
Dauer der Evaluierung	Dauer zwischen Start und Ende der Evaluierung	Tage	Meta-Daten und Evaluierungsberichte GIZ und KfW
Dauer Feldmission	Dauer der Vor-Ort-Mission	Tage	Meta-Daten und Evaluierungsberichte GIZ und KfW

Quelle: eigene Darstellung

Anmerkungen: Dargestellt sind alle in alternativen Modellspezifikationen verwendeten Variablen (siehe Abschnitt 3.2).

7.2
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

7.3 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
	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