

Working Paper

Handreichung für Evaluationen in der Umweltpolitik

Öko-Institut Working Paper 7/2019

Franziska Wolff, Nele Kampffmeyer, Katja Schumacher (Öko-Institut)



Öko-Institut e.V. / Oeko-Institut e.V.
Geschäftsstelle Freiburg / Freiburg Head Office

Postfach / P.O. Box 17 71
79017 Freiburg. Deutschland / Germany
Tel.: +49 761 45295-0
Fax: +49 761 45295-288

Büro Darmstadt / Darmstadt Office

Rheinstraße 95
64295 Darmstadt. Deutschland / Germany
Tel.: +49 6151 8191-0
Fax: +49 6151 8191-133

Büro Berlin / Berlin Office

Schicklerstraße 5-7
10179 Berlin. Deutschland / Germany
Tel.: +49 30 405085-0
Fax: +49 30 405085-388

info@oeko.de
www.oeko.de

Working Paper

Guide for evaluations in environmental policy

Franziska Wolff, Nele Kampffmeyer, Katja Schumacher

Working Paper 7/2019 Öko-Institut e.V. / Oeko-Institut e.V.

November 2019

Download: www.oeko.de/fileadmin/oekodoc/WP-Evaluation-Umweltpolitik.pdf



Dieses Werk bzw. Inhalt steht unter einer Creative Commons Namensnennung, Weitergabe unter gleichen Bedingungen 4.0 Lizenz. Öko-Institut e.V. 2019

This work is licensed under Creative Commons Attribution-Share Alike 4.0. Oeko-Institut e.V. 2019

Die Working Paper Series des Öko-Instituts ist eine Sammlung wissenschaftlicher Beiträge aus der Forschungsarbeit des Öko-Instituts e.V. Sie präsentieren und diskutieren innovative Ansätze und Positionen der aktuellen Nachhaltigkeitsforschung. Die Serie ist offen für Arbeiten von Wissenschaftlerinnen und Wissenschaftlern aus anderen Forschungseinrichtungen. Die einzelnen Working Paper entstehen in einem sorgfältigen wissenschaftlichen Prozess ohne externes Peer Review.

Oeko-Institut's Working Paper Series is a collection of research articles written within the scope of the institute's research activities. The articles present and discuss innovative approaches and positions of current sustainability research. The series is open to work from researchers of other institutions. The Working Papers are produced in a scrupulous scientific process without external peer reviews.

Zusammenfassung

Wirksame Umweltpolitik mindert Umweltprobleme. Evaluation hilft zu beurteilen, ob ein politisches Programm oder eine Maßnahme tatsächlich wirkt oder nicht, und welche Faktoren ihre Wirkungen hemmen oder fördern. Indem Evaluation gute Praxisbeispiele und Erfolgsbedingungen von Politik ermittelt, bereitet sie positive Lernerfahrungen auf und macht sie breiter verfügbar.

Dieses Arbeitspapier dient als Handreichung zur Durchführung von Evaluationen in der Umweltpolitik. Es soll einen schnellen Einstieg in die Theorie und Praxis von Evaluation bieten. Kapitel 2 gibt zunächst einen Überblick über relevante Aspekte der Evaluationsforschung: eine Definition von „Evaluation“, unterschiedliche Funktionen von Evaluation, Evaluationsstandards, die Kategorisierung von Evaluationstypen, -gegenständen und -kriterien, eine Diskussion wesentlicher Evaluationsansätze und -methoden und einen knappen Überblick über Untersuchungsdesigns und Ansätze der Datenerhebung. Kapitel 3 fokussiert auf die Evaluierungspraxis. Es wertet Analyseleitfäden und -raster aus, die in der Politikevaluation zum Einsatz kommen, und stellt ein eigenes Analyseraster vor, das im Rahmen eines Forschungsvorhabens für die Evaluation des „Nationalen Programms für Nachhaltigen Konsum“ (2016) entwickelt wurde. Kapitel 4 fasst abschließend forschungsimmanente und praktische Herausforderungen von Evaluationen zusammen und zeigt Möglichkeiten auf, wie diese überwunden werden können.

Das Arbeitspapier basiert auf Vorarbeiten aus einem Vorhaben, das im Auftrag des Umweltbundesamtes im Rahmen des Umweltforschungsplanes – Forschungskennzahl 3717 16 311 0 – erstellt und mit Bundesmitteln finanziert wurde.

Abstract

Effective environmental policy reduces environmental problems. Evaluation helps to assess whether a policy programme or measure actually works, and which factors inhibit or promote its effects. By identifying good practices and conditions for policy success, evaluation prepares positive learning experiences and makes them more widely available.

This working paper serves as a guide for carrying out environmental policy evaluations. It is intended to provide a quick introduction to the theory and practice of evaluation. Chapter 2 first gives an overview of relevant aspects of evaluation research: a definition of “evaluation”; different functions and types of evaluation; evaluation standards; a categorisation of evaluation objects and criteria; a discussion of essential evaluation approaches and methods; and a brief overview of research designs and data collection approaches. Chapter 3 focuses on evaluation practices. It screens analytical guidelines and approaches that are used in policy evaluation and presents an own analytical approach that was developed as part of a research project for the evaluation of the “National Programme for Sustainable Consumption” (2016). Chapter 4 concludes by summarising the research-immanent and practical challenges of evaluations and outlines ways in which these can be overcome.

This Working Paper is based on work within a project commissioned by the German Environment Agency within the Environmental Research Plan – project code (FKZ) 3717 16 311 0 – and financed by federal funds.

Inhaltsverzeichnis

Zusammenfassung	5
Abstract	5
Abbildungsverzeichnis	8
1. Einleitung	9
2. Evaluationsforschung: Ziele, Kriterien, Methoden, Designs	10
2.1. Was ist Evaluation?	10
2.2. Zwecke bzw. Funktionen von Evaluation	11
2.3. Evaluationsstandards	12
2.4. Typen und Zeitpunkte von Evaluation	13
2.5. Gegenstände von Evaluation	15
2.6. Evaluationskriterien	15
2.6.1. Relevanz	15
2.6.2. Kohärenz	15
2.6.3. Wirksamkeit (Effektivität) bzw. Erfolgsaussichten	16
2.6.4. Wirkungen (Impact) und Nebenwirkungen	16
2.6.5. Dauerhaftigkeit von Programm- bzw. Maßnahmewirkungen	16
2.6.6. Wirtschaftlichkeit bzw. (Kosten-)Effizienz	16
2.6.7. Verteilungswirkungen	17
2.6.8. Rechtmäßigkeit	17
2.6.9. Politische Durchsetzbarkeit	18
2.6.10. Soziale Akzeptanz	18
2.7. Evaluationsansätze und -methoden	18
2.7.1. Experimentelle und quasi-experimentelle Evaluation	18
2.7.2. Empirische Schätzungen und Modellierungen	20
2.7.3. Theoriebasierte Evaluation	22
2.8. Untersuchungsdesigns	28
2.9. Datenerhebung	29
4. Evaluationspraxis	30
4.1. In der Politikevaluation genutzte Analyseleitfäden und -raster	30
4.2. Exemplarischer Analyserahmen für die Evaluation eines Programms	31
4.2.1. Rekonstruktion und Analyse des Wirkungsmodells	32
4.2.2. Leitfragen für die Evaluierung	35
4.2.3. Analyseraster	38

5. Herausforderungen von Evaluation in Theorie und Praxis	46
Literaturverzeichnis	47

Abbildungsverzeichnis

Abbildung 1: Programm-/ Maßnahmenumsetzung („Aktionsmodell“)	23
Abbildung 2: Programm-/ Maßnahmenwirkungen („Veränderungsmodell“)	23
Abbildung 3: Wirkungsmodell mit Aktions- und Veränderungsmodell	24
Abbildung 4: Programm-/ Maßnahmenwirkungen (II)	33
Abbildung 5: Programm-/ Maßnahmenwirkungen (III)	33
Abbildung 6: Programm-/ Maßnahmewirkungen (IV)	34
Abbildung 7: Erweitertes Wirkungsmodell	35

Tabellenverzeichnis

Tabelle 1: Elemente von Evaluationen	10
Tabelle 2: Typen von Evaluation	14
Tabelle 3: Evaluationskriterien und -fragen bei der Evaluation des NPNK	36

1. Einleitung

Wirksame Umweltpolitik mindert Umweltprobleme. Evaluation hilft zu beurteilen, ob ein politisches Programm oder eine Maßnahme tatsächlich wirkt oder nicht, und welche Faktoren ihre Wirkungen hemmen oder fördern. Indem Evaluation gute Praxisbeispiele und Erfolgsbedingungen von Politik ermittelt, bereitet sie positive Lernerfahrungen auf und macht sie breiter verfügbar. Neben wirkungsbezogenen Fragen kann Evaluation auch Aufschluss über die Relevanz, Kohärenz, Wirtschaftlichkeit, sozialen Effekte oder Akzeptanz einer Maßnahme geben, sowohl vorab (ex ante) als auch begleitend und nach Umsetzung (ex post). Evaluation – gut durchgeführt und von den Adressaten ernstgenommen – ist daher ein zentrales Instrument, um Umweltpolitik relevanter, wirksamer, kostengünstiger, fairer oder kohärenter zu machen. Sie kann so einen wichtigen Beitrag dazu leisten, die Überschreitung der planetaren Grenzen aufzuhalten.

Dieses Arbeitspapier dient als Handreichung zur Durchführung von Evaluationen in der Umweltpolitik. Es soll einen schnellen Einstieg in die Theorie und Praxis von Evaluation bieten. Seine Zielgruppe sind Politikforscherinnen und -forscher sowie -praktikerinnen und -praktiker, die sich auf Grundlage bestehender Kenntnisse der empirischen Sozialwissenschaften Orientierung im Anwendungsfeld der Evaluation verschaffen oder Anregungen für eine eigene Evaluation holen wollen. Für eine vertiefte Auseinandersetzung mit den Konzepten und Methoden der empirischen Sozialforschung bietet sich die parallele Lektüre von entsprechenden Standardwerken an (Häder 2015; Kromrey et al. 2016; Schnell et al. 2011).

Kapitel 2 gibt zunächst einen Überblick über relevante Aspekte der Evaluationsforschung. Dies umfasst eine Definition von „Evaluation“, die Aufschlüsselung unterschiedlicher Funktionen von Evaluation, eine Zusammenschau wichtiger Evaluationsstandards und die Kategorisierung von Evaluationstypen, -gegenständen und -kriterien. Vor diesem Hintergrund werden drei wesentliche Evaluationsansätze und -methoden dargestellt und ihre Stärken und Schwächen bewertet: die experimentelle und quasi-experimentelle Evaluation, empirische Schätzungen und Modellierungen sowie die theoriebasierte Evaluation. Abschließend werden unterschiedliche Untersuchungsdesigns und Ansätze der Datenerhebung aufgeführt.

Kapitel 3 fokussiert auf die Evaluierungspraxis. Zum einen werden in der Evaluierungspraxis genutzte Analyseleitfäden und -raster ausgewertet. Zum anderen wird ein Analyseraster vorgestellt, das im Rahmen eines Forschungsvorhabens für die Evaluation des „Nationalen Programms für Nachhaltigen Konsum“ (BMUB; BMJV; BMEL 2017) entwickelt wurde.

Kapitel 4 fasst abschließend forschungsimmanente und praktische Herausforderungen von Evaluationen zusammen und verweist auf Möglichkeiten, diese zu überwinden.

Das Papier basiert auf Vorarbeiten, die im Rahmen des Forschungs- und Entwicklungsvorhaben „Nachhaltigen Konsum weiterdenken: Evaluation und Weiterentwicklung von Maßnahmen und Instrumenten“ durchgeführt wurden. Das Vorhaben wurde im Auftrag des Umweltbundesamtes im Rahmen des Umweltforschungsplanes – Forschungskennzahl 3717 16 311 0 – erstellt und mit Bundesmitteln finanziert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen.

2. Evaluationsforschung: Ziele, Kriterien, Methoden, Designs

2.1. Was ist Evaluation?

Dass es sich beim Begriff „Evaluation“ um ein „vieldeutiges Wort“ handelt, wurde schon früh thematisiert (Weiss 1974, S. 19). Evaluation kann Unterschiedliches bedeuten. Dabei hängt die jeweils verwendete Definition stark vom Kontext ab. Grundsätzlich wird sowohl der *Prozess*, also die Durchführung, als auch das *Ergebnis* eines Evaluationsprozesses als „Evaluation“ bezeichnet (Stockmann 2007a, S. 25).

Über den alltagspraktischen Prozess des „Bewertens“ hinaus beinhaltet das Konzept der „Evaluation“ die „systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Beurteilung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer Interventionsprogramme“ (Rossi et al. 1988, S. 3). Evaluationen werden in der Regel vorgenommen, um Erfahrungswissen als Grundlage für die Entscheidungsfindung zu generieren. Hierzu werden Informationen gesammelt und entlang begründeter, auf den Sachverhalt bezogener Kriterien bewertet. Dieser Prozess der Evaluation erfolgt in einem objektivierten Verfahren (Kromrey 2001, S. 3; Stockmann 2006b, S. 15).

Dieses Verständnis zusammenfassend definiert die Gesellschaft für Evaluation (DeGEval) Evaluation als die systematische Untersuchung des Nutzens bzw. der Güte eines Evaluationsgegenstandes (z.B. Instruments, Programms) auf Basis von empirisch gewonnenen Daten; sie impliziert eine Bewertung anhand offengelegter Kriterien für einen bestimmten Zweck (DeGEval 2016, S. 33). Abgrenzend dazu wird *Monitoring* als die routinemäßige, regelmäßige und kriteriengeleitete Sammlung von Vergleichsdaten verstanden, die den Zweck verfolgt, rechtzeitig Steuerungsbedarfe zu erkennen. Sie ist im Unterschied zur Evaluation nicht bewertend und immer längsschnittlich angelegt (ebd.).

Die OECD versteht Evaluation als die systematische und objektive Bewertung laufender oder abgeschlossener Projekte, Programme oder Politiken im Hinblick auf ihr Design, ihre Umsetzung und Ergebnisse. Ziel sei die Bestimmung der Relevanz und Erreichung von Zielen, Effizienz, Effektivität, Wirkungen („Impacts“) sowie der Dauerhaftigkeit solcher Wirkungen. Dabei solle eine Evaluation glaubwürdige und nützliche Informationen bereitstellen, um Organisationen in die Lage zu versetzen, Lehren zu ziehen und diese in Entscheidungsprozesse aufzunehmen (OECD 2002).

Mit Kromrey (2001) lassen sich die Elemente einer Evaluation folgendermaßen darstellen:

Tabelle 1: Elemente von Evaluationen

Alltagssprache	Wissenschaftlicher Sprachgebrauch	Präzisierung	Klärungsbedarf
Irgendetwas wird...	Programme, Maßnahmen, Organisation etc. werden	Existierend; in Planung/Entwicklung; bereits implementiert; Feldversuch/Pilotprojekt; Programmumfeld etc.	Was ist das Programm und seine Ziele? Was ist der „Gegenstand“ der Evaluierung? Was sind die Evaluationsziele?
von irgendetwem ...	durch Personen, die zur Bewertung besonders befähigt sind	unabhängige Wissenschaftler, Auftragsforscher, im Programm Mitwirkende, externe Berater, engagierte Betroffene etc.	Wer hat welche Funktionen/Kompetenzen? Informanten/Informationsquellen Informationsbeschaffung und -aufbereitung Evaluierende

Alltagssprache	Wissenschaftlicher Sprachgebrauch	Präzisierung	Klärungsbedarf
in irgendeiner Weise ...	in einem objektivierten Verfahren	Hearing, qualitative/quantitative Forschungslogik, experimentell/nicht-experimentell, formativ/summativ etc.	Methoden und Verfahren der Informationsbeschaffung Methoden und Verfahren des Bewertens Legitimation zum Bewerten
nach irgendetwelchen Kriterien bewertet.	nach explizit auf den Sachverhalt bezogenen und begründeten Kriterien (und gegebenenfalls Standards) bewertet.	Zielerreichung/Effekte/Nebenwirkungen, Effizienz/Effektivität, Sozialverträglichkeit, Zielgruppenbezug etc.	Ziele (wessen Ziele?) Kriterien Standards

Quelle: (Kromrey 2001, S. 108).

Die Ursprünge der wissenschaftlichen Evaluation liegen in Bewertungen öffentlicher Programme und Einrichtungen (Schulen, Gefängnisse etc.), wie sie die Regierung der USA im 18. Jahrhundert erstmals beauftragte. Einen deutlichen Aufschwung erhielt die Evaluationspraxis in den 1930er und 1940 Jahren, ebenfalls in den USA. Ihre Professionalisierung und Institutionalisierung wird in den frühen 1960ern verortet (Meyer 2002). Bis in die 1980er herrschte in Evaluationswissenschaften bzw. -praxis ein positivistisches Begriffsverständnis vor, bei dem Evaluationsergebnisse als „objektive“ Fakten erachtet wurden, die es durch angemessene (tendenziell quantifizierende) Methoden zu erzeugen galt. Im „Idealfall“ versuchte man mittels experimenteller Designs direkte kausale Zusammenhänge zwischen einer Maßnahme und ihren Wirkungen zu identifizieren (siehe hierzu u.a. Scriven 1972; 1991; Thomas D. und G. E. 1990; Campbell 1969). Nicht zuletzt aufgrund praktischer Schwierigkeiten in der Umsetzung wurde dieser „methodische Rigorismus“ zunehmend in Frage gestellt. Eine Gruppe von Autorinnen und Autoren stellte insbesondere den anvisierten Nutzen (und damit die instrumentelle Dimension) von Evaluation in den Vordergrund: Evaluation soll demnach vor allen Dingen die Informationsbedarfe der jeweils relevanten Akteure befriedigen (Stockmann 2006b, S. 22–25). Eine Erweiterung der Perspektive brachten ab Ende der 1980er Jahre konstruktivistische Ansätze. Aus ihrer Sicht handelt es sich bei den Ergebnissen einer Evaluation nicht um Fakten, sondern um soziale Konstrukte, die aus einem interaktiven Prozess zwischen Evaluierern und Evaluierten entstehen (Guba und Lincoln 2003; Wilhelm 2015).

2.2. Zwecke bzw. Funktionen von Evaluation

Evaluationen können unterschiedliche Ziele bzw. Funktionen haben, dabei auch mehrere dieser kombinieren. Vergleichsweise umfassend und dabei pragmatisch ist die Unterteilung in fünf Funktionen von Evaluationen bei Bortz & Döring (2016, S. 987–988)¹:

- **Erkenntnisfunktion:** Evaluationen tragen dazu bei, Erkenntnisse zu den Eigenschaften und Wirkungen von Evaluationsgegenständen zu sammeln.
- **Lern- und Dialogfunktion:** Im Rahmen einer Evaluation können unterschiedliche Stakeholder miteinander in Dialog treten und Lernprozesse bei allen Beteiligten stattfinden. Damit dies

¹ In der Literatur finden sich vielfältige alternative Kategorisierungen, siehe z.B. Shadish et al. (1991); Stockmann (2004), S. 3–5; Fitzpatrick et al. (2011).

geschieht, müssen die Stakeholder allerdings in geeigneter Weise in den Evaluationsprozess einbezogen werden.

- **Optimierungsfunktion:** Evaluationen sollen oft Erkenntnisse liefern, die zur zielgerichteten Verbesserung des Evaluationsgegenstandes beitragen können. Damit dieses Ziel erreicht wird, müssen im Laufe der Evaluation entsprechende Verbesserungsvorschläge erarbeitet werden.
- **Entscheidungsfunktion:** Die Ergebnisse von Evaluationen können die Grundlage für Entscheidungen bilden, etwa ob Maßnahme A oder B umgesetzt oder ob eine Maßnahme eingestellt bzw. weitergeführt wird. Zu beachten ist hier, dass diejenigen, die die Evaluation durchführen, nicht die Entscheidungsträger sind.
- **Legitimationsfunktion:** Evaluationen können dazu dienen, die Entwicklung und Durchführung von Interventionen bzw. die Existenz von Programmen nach außen zu legitimieren. Um diese Funktion zu erfüllen, muss die Evaluation sowohl transparent als auch glaubwürdig unparteilich bzw. überparteilich sein.²

2.3. Evaluationsstandards

Evaluationsstandards legen Merkmale fest, die gute Evaluierungen kennzeichnen und sollen Orientierung geben, wie Evaluationen professionell zu gestalten sind. Eine hohe Verbreitung bei Evaluationen im Bereich der Entwicklungsevaluation haben die Qualitätsstandards für Evaluierung in der Entwicklungszusammenarbeit des Entwicklungsausschusses (DAC) der OECD. Sie umfassen übergreifende Aspekte wie Unabhängigkeit und Evaluationsethik sowie Standards zur Zielsetzung, Planung und dem Design von Evaluationen, zur Umsetzung und Berichterstattung, zum Follow-up, der Nutzung und dem Lernen aus Evaluationen (OECD DAC 2010).

In Deutschland hat die Gesellschaft für Evaluation (DeGEval) 2002 Evaluationsstandards definiert, die 2016 überarbeitet wurden. Ihnen zufolge sollen Evaluationen vier grundlegende Eigenschaften aufweisen: Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit. Diese Standardgruppen umfassen wiederum jeweils mehrere Einzelstandards (s. Kasten).

Evaluationsstandards der DeGEval

- **Nützlichkeit:** Identifizierung der Beteiligten und Betroffenen (N1), Klärung der Evaluationszwecke (N2), Kompetenz und Glaubwürdigkeit des Evaluators/der Evaluatorin (N3), Auswahl und Umfang der Informationen (N4), Transparenz von Werthaltungen (N5), Vollständigkeit und Klarheit der Berichterstattung (N6), Rechtzeitigkeit der Evaluation (N7), Nutzung und Nutzen der Evaluation (N8);
 - **Durchführbarkeit:** Angemessene Verfahren (D1), diplomatisches Vorgehen (D2), Effizienz von Evaluation (D3);
 - **Fairness:** Formale Vereinbarungen (F1), Schutz individueller Rechte (F2), umfassende und faire Prüfung (F3), unparteiische Durchführung und Berichterstattung (F4), Offenlegung von Ergebnissen und Berichten (F5);
 - **Genauigkeit:** Beschreibung des Evaluationsgegenstandes (G1), Kontextanalyse (G2), Beschreibung von Zwecken und Vorgehen (G3), Angabe von Informationsquellen (G4), valide und reliable Informationen (G5), systematische Fehlerprüfung (G6), angemessene Analyse qualitativer und quantitativer Informationen (G7), begründete Bewertungen und Schlussfolgerungen (G8), Meta-Evaluation (G9).
-

Quelle: DeGEval (2016).

² Gerade in Bezug auf diese Funktion besteht die Gefahr der Instrumentalisierung von Evaluationen bspw. zur Verantwortungsdelegation, zum Herauszögern von wirksamen Maßnahmen oder zur Symbolpolitik (vgl. Bortz und Döring 2016, S. 987–988; Stockmann 2004, S. 4).

Auf den OECD- und DeGEval-Standards aufbauend, beinhalten die Evaluationsstandards des Deutschen Evaluierungsinstituts der Entwicklungszusammenarbeit (DEval) die Kriterien der Nützlichkeit, Evaluierbarkeit, Fairness, Unabhängigkeit und Integrität, Genauigkeit, Wissenschaftlichkeit und Nachvollziehbarkeit sowie der Vergleichbarkeit (DEval 2018).

2.4. Typen und Zeitpunkte von Evaluation

Die Frage nach **Typen** von Evaluation steht in engem Zusammenhang mit dem **Zeitpunkt** der Evaluation. Unterschieden wird in:

- **ex ante** Evaluationen (vor der Programmumsetzung);
- **begleitende** Evaluationen (während der Programmumsetzung); oder
- **ex post** Evaluationen (nach abgeschlossener Programmumsetzung).

Typen von Evaluationen lassen sich auch nach weiteren Aspekten bilden, wie Evaluationszweck, -gegenstand oder -kriterien (vgl. unten).

In Bezug auf unterschiedliche **Evaluationszwecke** lassen sich formative von summativen Evaluationen unterscheiden (Scriven 1991; Stockmann 2004):

- **Formative Evaluation** ist aktiv-gestaltend, prozessorientiert, konstruktiv und kommunikationsfördernd angelegt und erlaubt Veränderungen am Untersuchungsgegenstand. Sie findet entweder in der Phase der Programmentwicklung oder begleitend zum jeweiligen Umsetzungsprozess statt;
- **Summative Evaluation** ist zusammenfassend, bilanzierend und ergebnisorientiert. Sie erfolgt nach Abschluss einer Maßnahme oder eines Programms und vergleicht den angestrebten mit dem tatsächlichen Zustand.

Typisiert man nach dem **Gegenstand**, lassen sich u.a. unterscheiden:

- Programmevaluation: Auswertung eines Programms, i.d.R. zu formativen Zwecken
- Prozessevaluation: Auswertung eines (Umsetzungs-)Prozesses
- Metaevaluation: Auswertung bereits existierender Evaluationen

Die Unterscheidung nach **Evaluationskriterium** unterliegt der gängigen Differenzierung zwischen

- Ergebnis-/Wirkungsevaluation
- Effizienzanalyse

Je nach Typus lassen sich unterschiedliche Forschungsfragen verfolgen und unterschiedliche Formen der Datenerhebung einsetzen. Die folgende Tabelle führt einige Evaluationstypen auf.

Tabelle 2: Typen von Evaluation

Typus	Zeitpunkt	Forschungsfragen	Evaluationsziele	Datenerhebung
Programm-evaluation, formative Evaluation	ex ante	Welche Schwachpunkte gibt es schon prospektiv?	Konzeptualisierung Analyse zur Programmentwicklung, Entwicklung des Programms (Defizite ergründen und beseitigen)	Qualitative Erhebung (Literaturrecherche, Gruppendiskussionen, Interviews)
	ex ante (während der Erprobung)	Welche Schwachpunkte gibt es?	Kontrolle und Beratung (weitere Entwicklung, Verbesserung)	
Prozessevaluation	begleitend	Erreicht ein Programm die richtige Zielgruppe? Gibt es Abweichungen von den Planungen?	Kontrolle und Beratung (Aufdecken von Programmfehlern und Abweichungen des Konzepts)	Qualitative, quantitative Erhebung (Erläuterung zum Programmkonzept, Protokoll, Beobachtung, Befragung)
Wirkungsevaluation (Ergebnisevaluation, summative Evaluation, Outcome Evaluation, Impact Evaluation)	begleitend	Wurden die Ziele erreicht?	Prüfung von Wirkung und Nutzen	Qualitative und quantitative Erhebung
	ex post	Inwieweit sind die Veränderungen Folgen der Maßnahmen?	(Aufdeckung was wie unter welchen Bedingungen wirkt)	
Effizienzanalyse (Kosten-Nutzen-Analyse, Kosten-Wirkungs-Analyse)	ex ante	In welchem Verhältnis stehen Kosten und Nutzen? (materiell, immateriell)	Prüfung / Quantifizierung von Kosten und Nutzen	Quantitative Erhebung
	ex post			
Meta-Evaluation	ex ante	Lassen sich die Ergebnisse der einzelnen Studien übertragen?	Zusammenführung der Ergebnisse von mehreren thematisch gleichen Programmen	Quantitative Auswertung
	nach der Wirkungsanalyse (Programme)			

Quelle: Brandt (2004, S. 13), mit leichten Anpassungen.

Bei in der Realität durchgeführten Evaluationen handelt es sich häufig um Mischformen. So kann beispielsweise auch eine formative Evaluation begleitend durchgeführt werden, und im Rahmen einer Wirkungsanalyse können mögliche (für Wirkungsdefizite verantwortliche) Schwachstellen identifiziert oder Effizienz Aspekte analysiert werden.

2.5. Gegenstände von Evaluation

Evaluationen werden in nahezu allen Disziplinen und an einer unüberschaubaren Zahl von Gegenständen durchgeführt. Bortz & Döring (2016, S. 980) formulieren es so: „Museen werden ebenso evaluiert wie Serviceroboter“. Eine grundsätzliche Einschränkung und einheitliche Systematisierung von Evaluierungsgegenständen ist daher nahezu unmöglich (Stockmann 2007a, S. 26).

Für Evaluationen im politischen Kontext ist die Unterscheidung u.a. in Programme, Projekte, Instrumente und Maßnahmen(-bündel/ -cluster), aber auch in Prozesse als Evaluationsgegenstände möglich. Dabei bilden Instrumente die kleinste Handlungseinheit. Im Verständnis von Stockmann setzen sich Projekte aus mehreren Instrumenten zusammen, Programme wiederum bestehen aus aufeinander bezogenen Projekten (Stockmann 2007b, S. 13).

2.6. Evaluationskriterien

Eine Evaluation misst den Nutzen und die Güte eines Evaluationsgegenstandes an einem oder mehreren vorab zu definierenden Evaluationskriterien. Diese orientieren sich in der Regel an den jeweiligen Zielen, die relevante Anspruchsgruppen mit der Evaluation verbinden. Abhängig vom Erkenntnisinteresse, dem Zeitpunkt der Evaluation und dem Evaluationstyp kann also die Wahl zwischen einer Vielzahl möglicher Kriterien getroffen werden.

Im Folgenden werden einige, im Zusammenhang mit der Evaluation von politischen Programmen und Maßnahmen häufig genutzte Kriterien kurz erläutert. Eine besondere Rolle nehmen dabei die fünf Evaluierungskriterien des Entwicklungsausschusses der OECD – Relevanz, Effektivität, Effizienz, Wirkung/Impact und Nachhaltigkeit – ein (OECD 2010b, S. 13–14). Erstmals 1991 verfasst, haben sie eine weltweite und über die Entwicklungszusammenarbeit hinausgehende Anwendung erfahren, auch wenn immer wieder über ihre Weiterentwicklung diskutiert wird (z.B. Faust und Verspohl 2019).

Die folgenden Definitionen gängiger Evaluationskriterien sind nicht allgemeingültig, und oft ist es sinnvoll, sie für den Kontext des konkreten Programms oder Projekts anzupassen.

2.6.1. Relevanz

In der Entwicklungspolitik wird „Relevanz“ oft definiert als das „Ausmaß, in dem die Ziele einer Entwicklungsmaßnahme mit den Bedürfnissen der Begünstigten, den Anforderungen eines Landes, den globalen Prioritäten und den Politiken der Partner und Geber im Einklang stehen“ (OECD 2009a).

Themenfeldübergreifend kann Relevanz als das Ausmaß verstanden werden, in dem die Ziele, aber auch der thematische Gegenstandsbereich eines Programms mit Erkenntnissen und Erwartungen von Stakeholdern (einschließlich der Wissenschaft) in Einklang stehen. Bei umweltpolitischen Programmen kann unter „Relevanz“ beispielsweise geprüft werden, ob das Programm Handlungsoptionen umfasst, die gemäß dem Stand der Literatur besonders hohe Umweltentlastungseffekte versprechen (so genannte „big points“; vgl. Bilharz 2008). Ein Abgleich kann auch mit in der (öffentlichen-, Fach-) Debatte geäußerten Erwartungen bezüglich von Programmzielen und Gegenstandsbereich erfolgen.

2.6.2. Kohärenz

Das Kriterium der „Kohärenz“ lässt sich in interne und externe Kohärenz untergliedern. Interne Kohärenz betrifft das Maß, in dem ein Programm bzw. eine Maßnahme in sich widerspruchsfrei ist –

sich seine Ziele, Grundsätze, Mechanismen und Maßnahmen also einander stützen oder einander zumindest nicht zuwiderlaufen.

Externe Kohärenz bezieht sich auf das Maß, in dem ein Programm oder eine Maßnahme widerspruchsfrei zu anderen bzw. ausgewählten einzelnen politischen Interventionen ist (EEA 2016). Dabei ist zu berücksichtigen, dass Ziele zunächst miteinander kohärent erscheinen können (Bsp. Klima- und Naturschutz), Maßnahmen zur Erreichung des einen Ziels aber in der Umsetzung der Erreichung des anderen Ziels zuwiderlaufen können (Bsp. biodiversitätsarme, monokulturelle Energieholzplantagen). Für die Evaluation, inwieweit beispielsweise ein politisches Instrument mit bestimmten anderen (Umwelt-)Zielen kohärent ist, ist es daher notwendig, die Wirkungsketten vom Instrument über seine Umsetzung durch Zielgruppen bis hin zu möglichen Effekten auf die Erzielung anderer Umweltprobleme zu analysieren. Hierfür liegen Methodiken vor (Wolff et al. 2016).

2.6.3. Wirksamkeit (Effektivität) bzw. Erfolgsaussichten

Das Kriterium der Wirksamkeit bezieht sich auf die Frage der Zielerreichung. Zielerreichung zu messen setzt voraus, dass die Ziele eines Programms, einer Maßnahme etc. klar benannt, gegebenenfalls sogar quantifiziert sind. Die OECD definiert Wirksamkeit als Ausmaß, in dem die Ziele eines Programms unter Berücksichtigung ihrer relativen Bedeutung erreicht worden sind oder voraussichtlich erreicht werden. Bei der ex ante Bewertung von (potenzieller) Wirksamkeit kann auch von „Erfolgsaussichten“ gesprochen werden.

Zielerreichung lässt sich beispielsweise anhand einer fünfstufigen Skala (niedrig; niedrig bis moderat; moderat; moderat bis hoch; hoch) bewerten. Die Bewertung sollte auf Grundlage vorab spezifizierter Kriterien erfolgen, die sich aus der inhaltlichen Ausrichtung des zu bewertenden Ziels bzw. eines erwarteten Ergebnisses herleiten (DEval 2017).

2.6.4. Wirkungen (Impact) und Nebenwirkungen

Während „Wirksamkeit“ das tatsächlich Erreichte in Bezug zu den (Programm-)Zielen setzt, geht es beim Kriterium der „Wirkungen“ (Impact) offener um alle – positiven und negativen, erwarteten oder unerwarteten, beabsichtigten und unbeabsichtigten – Wirkungen eines Programms. Nebenwirkungen sind explizit erfasst. Eine „zielfreie“ Bewertung im Hinblick auf Wirkungen ist insbesondere dann vorteilhaft, wenn Ziele wenig konkret oder nur implizit benannt sind („goal-free evaluation“, cf. Scriven 1972). Hier können beispielsweise „relative Verbesserungen“ gemessen werden, d.h. Fortschritte gegenüber einer Baseline, ohne diese Fortschritte auf ein gegebenes Ziel zu beziehen, wie dies bei der Messung von „Zielerreichung“ getan wird (vgl. Kriterium der Wirksamkeit, oben).

2.6.5. Dauerhaftigkeit von Programm- bzw. Maßnahmewirkungen

Das Kriterium der „Dauerhaftigkeit“ bezieht sich auf das (voraussichtliche) Fortbestehen positiver (Langzeit-) Wirkungen eines Programms oder einer Maßnahmen über dessen bzw. deren unmittelbare Laufzeit hinaus. Es ist damit verwandt mit den Kriterien „Wirkungen“ und „Wirksamkeit“, erfasst aber explizit auch die langfristige Widerstandsfähigkeit positiver Wirkungen gegenüber Risiken (OECD 2009a).

2.6.6. Wirtschaftlichkeit bzw. (Kosten-)Effizienz

In den OECD-DAC-Evaluationsstandards wird „Effizienz“ definiert als Maß dafür, wie sparsam Ressourcen/Inputs (Finanzmittel, Fachwissen, Zeit usw.) in qualitative und quantitative Ergebnisse (Outputs, Outcomes, Impacts) umgewandelt werden (OECD 2009a; 2010b). Die OECD schlägt folgende

Leitfragen für die Bewertung der Effizienz eines Programms oder einer Maßnahme vor: Waren die Aktivitäten kosteneffizient? Wurden die Ziele innerhalb des Zeitrahmens erreicht? Wurde die Maßnahme auf die kostengünstigste Art umgesetzt – verglichen zu einem alternativen Vorgehen? Hier können insbesondere Indikatoren wie die Haushalts- oder Fördermitteleffizienz (eingesetzte Haushaltsmittel³ bzw. Fördermittel im Verhältnis zur Wirkung), der Hebeleffekt (gesamte Investitionen im Verhältnis zu den eingesetzten Fördermitteln), Kosten-Nutzen-Bewertungen, Nutzwertanalysen oder Vermeidungskosten (Netto-Investitionen und Netto-Einsparungen im Verhältnis zur Wirkung) oder auch Beschäftigungswirkungen erfasst werden.

2.6.7. Verteilungswirkungen

Jenseits der Effizienz eines Programms bzw. einer Maßnahme können seine bzw. ihre Verteilungswirkungen (d.h. distributiven Effekte) evaluiert werden. Dabei konkurrieren verschiedene Definitionen und Operationalisierungen von Verteilungswirkungen sowie unterschiedliche Wirkungskategorien (Jacob et al. 2016). Betrachtet werden können beispielsweise die ökonomischen Auswirkungen von Maßnahmen sowie deren Wirkungen auf materielles Wohlergehen (z.B. Einkommen, Vermögenswerte, Beschäftigung, Arbeitsbelastung, Lebensstandard, ökonomische Abhängigkeit), aber auch auf Gesundheit und Wohlbefinden, Lebensumgebung, Familie und Gemeinschaft etc.

Zur Erfassung von Verteilungseffekten werden dann die jeweiligen Auswirkungen auf unterschiedliche Bevölkerungsgruppen analysiert. Dies kann beispielsweise nach Einkommensgruppen differenziert werden (z.B. Quartilen der Einkommensverteilung), Haushaltstypen (z.B. Haushalte mit oder ohne Kinder), nach der Stellung im Wirtschaftssystem (Produzenten / Konsumenten, Arbeitnehmer / Arbeitgeber etc.), nach Eigentumsverhältnissen (z.B. Mieter / Vermieter) oder räumlichen Faktoren (z.B. Regionen, Stadt / Land). Im Hinblick auf ökonomische Verteilungswirkungen kann u.a. von Interesse sein, ob sich eine Maßnahme progressiv oder regressiv auswirkt, also ob einkommensschwache oder einkommensstarke Haushalte relativ stärker oder schwächer entlastet werden (vgl. Schumacher et al. 2016).

2.6.8. Rechtmäßigkeit

Wenn politische Systeme vom Legalitätsprinzip geprägt sind, beruht Politik in den allermeisten Fällen auf einer rechtlichen Grundlage. In der Praxis ist die Überprüfung der Rechtmäßigkeit eines Maßnahmenvorschlags wichtiger Ausgangspunkt für seine (ex ante) Evaluation.

Die Betrachtung der Rechtmäßigkeit einer Maßnahme – im Sinne ihrer rechtlichen und materiellen Kohärenz – erfolgt in der Regel im Vorfeld. In einem föderalen System müssen bei dieser Überprüfung die Kompetenzen und Zuständigkeiten der einzelnen Ebenen berücksichtigt werden (Knoepfel et al. 1997, S. 94–95). Beim Vorschlag für eine bundesrechtliche Maßnahme beinhaltet das die Frage nach seiner Konformität mit unterschiedlichen anderen Rechtsmaterien und seiner Vereinbarkeit mit dem materiellen Verfassungsrecht; beispielsweise kann eine Umweltregulierung in Grundrechte wie etwa die allgemeine Handlungsfreiheit eingreifen und ist dann rechtfertigungsbedürftig. Daneben ist die Prüfung erforderlich, ob der Maßnahmenvorschlag europarechtskonform ist und ob ihm ggf. völkerrechtliche Normen entgegenstehen.

³ Haushaltsmittel umfassen Mittel für den Vollzug einer Maßnahme. Sofern diese Maßnahme eine Fördermaßnahme ist, umfassen Haushaltsmittel auch Fördermittel.

2.6.9. Politische Durchsetzbarkeit

Mit der politischen Durchsetzbarkeit eines Programmes werden seine Aussichten auf praktische Umsetzung bezeichnet (Scharpf 1973). Sie hängt eng damit zusammen, wie gut sich um das Programm herum Mehrheiten im demokratischen Willensbildungsprozess – und damit auch in der öffentlichen Debatte – organisieren lassen.

Die politische Durchsetzbarkeit eines Evaluierungsgegenstandes zu bewerten ist äußerst anspruchsvoll, da diese in hohem Maße multikausal und kontingent ist.⁴ Scharpf zufolge kann sie daher nie abstrakt, sondern nur konkret diskutiert werden: „wenn bekannt ist, welcher Akteur welche Ziele mit welchen Mitteln und bei welchen Rahmenbedingungen mit wessen Unterstützung und gegen wessen Widerstand zu verfolgen sucht“ (Scharpf 1973, S. 3).

2.6.10. Soziale Akzeptanz

Die soziale Akzeptanz eines (bereits „politisch durchgesetzten“) Programmes (vgl. vorheriger Abschnitt) bezeichnet das Ausmaß, mit dem konkrete Zielgruppen oder die weitere Gesellschaft eine tolerierende bzw. im besten Fall befürwortende Einstellung gegenüber dem Programm einnehmen. Sie beeinflusst u.a. die Inanspruchnahme von Programmleistungen durch Zielgruppen, ihre Reaktion auf Anreize, die das Programm setzt, und die Durchsetzbarkeit regulativer Programmelemente bei den Zielgruppen. Soziale Akzeptanz ist das Ergebnis eines vielschichtigen und voraussetzungsreichen Prozesses (Lucke 2003). Als empirisch zu bestimmender Sachverhalt ist „Akzeptanz“ (eines Programms) von seiner normativen „Legitimität“ abzugrenzen.

2.7. Evaluationsansätze und -methoden

Im Folgenden stellen wir drei unterschiedliche Evaluationsansätze dar, die auch mit spezifischen Methoden einhergehen: die experimentelle und quasi-experimentelle Evaluation; empirische Schätzungen und Modellierungen; und die theoriebasierte Evaluation. Diese Aufteilung in drei grobe Ansätze ist nur ein Zugang unter vielen; tatsächlich finden sich verwirrend viele Systematisierungsversuche in der Literatur. Einen Überblick über diese bieten Meyer & Stockmann (2014).

2.7.1. Experimentelle und quasi-experimentelle Evaluation

Einordnung: Bei experimentellen Ansätzen wird versucht eindeutige kausale Zusammenhänge zwischen einer Intervention und beobachteten Veränderungen herzustellen. Wie schon in Kapitel 2.1 beschrieben, wurde dieser Anspruch besonders in der früheren Evaluationsforschung geltend gemacht und ab den 1980er Jahren von stärker qualitativen und konstruktivistischen Ansätzen abgelöst bzw. durch diese ergänzt. Gerade bei der Evaluation von politischen Instrumenten und Programmen stoßen experimentelle Evaluationen aus forschungspraktischen Gründen häufig an ihre Grenzen, da Versuchs- und Kontrollgruppen schwer oder gar nicht voneinander getrennt werden können (Stockmann 2006b, S. 22–23; Bussmann et al. 1997, S. 195). Allerdings werden in der

⁴ Einfluss auf die politische Durchsetzbarkeit können u.a. die Interessenlagen und Organisationsfähigkeit von politischen Parteien, Gebietskörperschaften, Verbandsinteressen und sonstigen potenziellen „Vetoplayern“ haben; das institutionelle Gefüge, innerhalb dessen die politische Entscheidung zu treffen ist (u.a. der Grad von Politikverflechtung in Mehrebenensystemen); das Maß der Vertrautheit von Politik, Verwaltung und Öffentlichkeit mit dem Instrument(entypus); die Anschlussfähigkeit der zugrundeliegenden Regulierungsidee an gesellschaftlich akzeptierte Normen und Diskurse; Be- oder Entlastungswirkungen auf öffentliche Haushalte; erwartete Auswirkungen auf Wirtschaft und Arbeitsmarkt; die Kompatibilität mit den gegebenen rechtlichen Rahmenbedingungen; das Auftreten bzw. aktive Nutzen von Gelegenheitsfenstern; politische Entrepreneurship etc. (vgl. Böcher und Töller 2007, 2012; Jänicke et al. 2003, 1976; Jordan et al. 2003; Sabatier 1987; Scharpf 1973).

Entwicklungszusammenarbeit teilweise Evaluationen mit experimentellem oder quasi-experimentellem Charakter durchgeführt (World Bank 2016).

Kerninhalte & Methoden: Wie oben beschrieben gelten experimentelle Anordnungen als geeignete Methode, um kausale Wirkungszusammenhänge zwischen einer oder mehreren unabhängigen Variablen (Instrument, Projekt, Programm etc.) und einer oder mehreren abhängigen Variablen (das oder die jeweiligen Ziele) zu identifizieren.

Beim experimentellen Design werden potenzielle Adressaten einer Maßnahme – Personen, Haushalte, Firmen, Gemeinden, etc. – vor deren Implementation nach dem Zufallsprinzip („randomisiert“) in zwei Gruppen geteilt: Eine Gruppe (die so genannte Zielgruppe) nimmt an der geplanten Maßnahme teil, die andere (die so genannte Kontrollgruppe) nicht. Da aufgrund der zufälligen Auswahl der beiden Gruppen – abgesehen von der zu evaluierenden Intervention – keine wesentlichen Unterschiede zwischen den beiden Gruppen bestehen sollten (interne Validität), kann davon ausgegangen werden, dass Veränderungen in der oder den abhängigen Variablen (Effekte) von der Intervention ausgelöst wurden (Caspari und Barbu 2009, S. 9). So wird das Problem der Zuordnung einer beobachteten Veränderung zu einer Maßnahme (kausale Attribution) gelöst.

Eine solche Evaluation kann sowohl unter „künstlichen“ Bedingungen als Laborexperiment, als auch als Feldexperiment stattfinden, wobei Feldexperimente schwieriger zu kontrollieren sind (Busmann et al. 1997, S. 194–195; IEG 2013, xiv).

Da in der Regel keine „sozialen Experimente“ evaluiert werden, ist es in der Praxis nur äußerst selten möglich, die für eine echte experimentelle Evaluation notwendigen Bedingungen zu schaffen. Deshalb wird auf quasi-experimentelle Evaluationsdesigns zurückgegriffen, die versuchen, den Anforderungen experimenteller Evaluation so nahe wie möglich zu kommen. Sie unterscheiden sich meist insofern, als keine Randomisierung erfolgt, d.h. die Einteilung in Versuchs- und Vergleichsgruppe nicht zufällig erfolgt. Hier kann die Evaluation auch ex post oder begleitend erfolgen, da es möglich ist, auch im Nachhinein eine Vergleichsgruppe zu bilden, die der Versuchsgruppe möglichst ähnlich ist. Um die Nachteile der nicht erfolgten Randomisierung zumindest teilweise auszugleichen, werden u.a. statistische Verfahren angewendet (Busmann et al. 1997, S. 194–195; Kromrey 2001, S. 122–123).

Experimentelle Evaluationen gelten insofern methodisch als „Goldstandard“, als sie den Anforderungen einer quantitativen Forschungslogik an wissenschaftliches Vorgehen (Validität, Reliabilität, Replizierbarkeit etc.) am nächsten kommen (vgl. King et al. 1994). Auch verhindert der zufallsgesteuerte Auswahlprozess von Teilnehmer/innen am Laborexperiment systematische Auswahlverzerrungen. Im Feldexperiment können systematische Auswahlverzerrungen allerdings weiterhin auftreten, da es nicht möglich ist, alle Eigenschaften der Teilnehmer so zu kontrollieren, dass keine Unterschiede zwischen Ziel- und Kontrollgruppe existieren (Caspari & Barbu 2009, S. 9).

Die Umsetzung experimenteller Designs, und in etwas geringerem Maße quasi-experimenteller Designs in der Praxis, ist allerdings schwierig. Ein Grund ist, dass sie schon bei der Konzipierung, spätestens aber bei der Umsetzung politischer Maßnahmen berücksichtigt werden müssen. Dies ist in der politischen Realität häufig nicht umsetzbar. Außerdem ist der Ausschluss bestimmter Gruppen, der Voraussetzung für das Vorhandensein einer Kontrollgruppe ist, bei politischen Instrumenten in der Regel nicht möglich oder politisch/ ethisch problematisch. Auch ist die Kontrolle exogener Einflüsse in den meisten sozialen Kontexten erschwert bis unmöglich. Weiterhin können einige Faktoren in experimentellen Anordnungen zu systematischen Schwierigkeiten führen. Dazu gehören Effekte, die sich aus der Evaluation selbst ergeben, wie die Reaktion auf die Gruppenzuweisung, Rivalität zwischen den Gruppen, unterschiedliche Abbruchraten unter Probanden je nach wahrgenommener Attraktivität der Versuchsbedingungen („selektiver Drop-out“), oder die Sensitivierung

durch Anfangsmessungen („Prämessungen“). Nicht zuletzt wird aus konstruktivistischer Perspektive argumentiert, dass Evaluierende die Evaluation beeinflussen, insbesondere das Untersuchungsdesign und die Interpretation der Daten (Kromrey 2001, S. 119–125; Pawson und Tilley 2005).

Stärken & Schwächen der experimentellen und quasi-experimentelle Evaluation:

Stärken	Schwächen
Methode löst das Problem der kausalen Attribution („Welche Veränderung hat die Maßnahme bewirkt? In welchem Umfang sind beobachtbare Veränderungen auf die Maßnahme zurück zu führen?“)	Hohe konzeptionelle und forschungspraktische Anforderungen (teils: Kosten)
„Gold Standard“ der Evaluation: kommt den Anforderungen einer quantitativen Forschungslogik an wissenschaftliches Vorgehen (Validität, Reliabilität, Replizierbarkeit etc.) am nächsten	Der Ausschluss der Kontrollgruppe aus dem Programm, der Maßnahme etc. kann in realen Kontexten auf politische oder ethische Vorbehalte stoßen
Der zufallsgesteuerte Auswahlprozess von Teilnehmer/innen zumindest am Laborexperiment (nicht aber im Feldexperiment) verhindert systematische Auswahlverzerrungen	Exogene Faktoren sind bei politischen Instrumenten, Programmen etc. schwer oder nicht zu kontrollieren
	Die Evaluation selbst kann (u.a. wegen der künstlichen Situation) zu Effekten bei den „Beforschten“ führen
	Die Konstruktion des Experiments ist nicht unabhängig von der oder dem Forschenden
	Experimentelle Designs sind nur ex ante möglich
	Experimentelle Designs erfordern inhaltlich eng abgesteckte ‚Treatments‘ und eignen sich weniger für breite Programme mit einer Vielfalt von Maßnahmen (in denen die Zahl der abhängigen und intervenierenden Variablen zu hoch ist, um die kausale Rolle der unabhängigen Variablen zu identifizieren)

Quelle: Eigene Zusammenstellung.

2.7.2. Empirische Schätzungen und Modellierungen

Einordnung: Empirische Schätzungen und Modellierung können ein geeigneter Ansatz sein, um Informationen (empirisch erhoben oder aus Statistiken) hinsichtlich der Evaluationsfragestellung auszuwerten bzw. zu bewerten (vgl. Stockmann 2004). Die Bewertung richtet sich in der Regel nach vorher definierten Kriterien. Die Informationen können qualitativer Natur sein, sind aber insbesondere in Modellierungsansätzen zumeist quantitativ. Unter dem Begriff der „empirischen Sozialforschung“ verbergen sich eine Vielzahl an Datenerhebungs- und -auswertungsansätzen, die für eine Evaluation hilfreich und sinnvoll sind. Modelle werden ebenso für ex ante, begleitende und ex post Bewertungen eingesetzt. Modelle können rein qualitativ sein und Beziehungen zwischen Elementen eines Systems darstellen. Sie können aber auch quantitativ sein, indem sie Zustände und Veränderungen in Zahlen darlegen. Die Zahlenwerte können auf empirischen Daten beruhen bzw. empirische Daten mit theoriebasierten Annahmen verknüpfen. Modelle können aber auch rein analytisch (theoriebasiert) sein.

Kerninhalte & Methoden: Empirische Schätzungen oder Modellierungen beruhen auf einer Konkretisierung von Abhängigkeiten oder Zusammenhängen in einem Modell. Dies kann entweder funktional (also durch Funktionen) oder zunächst qualitativ (durch Verbindungspfeile) ausgedrückt

werden. Zusammen mit einer guten Datenerhebung oder Datengrundlage bildet die Spezifizierung der Zusammenhänge in einem Modell den Kern von empirischen Schätzungen und Modellierungen.

Statistische oder ökonometrische Methoden können dann eingesetzt werden, um Zusammenhänge auf Basis der Daten zu ermitteln. Während statistische oder ökonometrische Methoden in der Regel auf gegenwärtige oder vergangene Daten aufsetzen, werden Modelle auch für Projektionen in die Zukunft (ex ante) verwendet.

Für den Einsatz von empirischen Schätzmethoden bzw. Modellierung im Rahmen einer Evaluation ist die Formulierung einer Hypothese und ihrer Einflussfaktoren essentiell. Diese bestimmen abhängige und unabhängige Variablen, deren Parametrisierung mithilfe von statistischen Ansätzen (beispielsweise Regressionsanalysen, Simulationsanalysen) erfolgen kann. Konkret kann auf diese Weise untersucht werden, welche Rolle die Entwicklung verschiedener Faktoren auf einen bestimmten Zustand (oder eine Veränderung eines Zustands) hatte oder haben wird. Damit ist es möglich, den Einfluss einzelner Faktoren voneinander zu entflechten. In der Regel werden Informationen aus Kontrollgruppen, die nicht von der zu evaluierenden Intervention betroffen sind, als Vergleichsinformation hinzugenommen, um den Einfluss der zu evaluierenden Intervention zu separieren. Alternativ wird die Entwicklung mit der eines „Counterfactual“-Szenarios verglichen. Das Counterfactual beschreibt entweder rückblickend oder in die Zukunft schauend eine Entwicklung *ohne* die zu evaluierende Intervention. (Bei der Evaluierung einer Energiesteuer beispielsweise würde das Counterfactual die Entwicklung ohne die Energiesteuer simulieren). In diesem Sinne lässt sich nicht nur das Ausmaß des Einflusses, sondern auch die Ursächlichkeit bewerten.

Stärken & Schwächen von empirischen Schätzungen und Modellierungen:

Stärken	Schwächen
Nutzbar in allen Formen der Evaluation (ex ante, begleitend, ex post), v.a. mit Fokus auf Kriterium der Wirkung / Wirksamkeit und der Ursächlichkeit	Benötigt ein Counterfactual-Szenario oder eine Kontrollgruppe
Erlaubt eine quantitative Abschätzung der Effekte, insbesondere Wirkungen, aber auch Effizienz	Sehr datenintensiv, Befragungen oder Erhebungen können ressourcenintensiv sein. Unsicherheiten in den Daten beeinflussen die Ergebnisse.
Erlaubt den Vergleich der Effekte verschiedener Interventionen.	Bei sehr unterschiedlichen Wirkmechanismen der Interventionen (weiche Maßnahmen versus Investitionsförderung) ist ein quantitativer Vergleich wenig aussagekräftig.
Unterfüttert qualitative Einschätzungen	
Modellspezifikation erfordert gutes Verständnis der Wirkmechanismen	Modellspezifikation kann unvollständig oder unpassend sein.
Statistische Methoden weisen die Güte der Ergebnisse aus.	Ergebnisse können falsch interpretiert werden.

Quelle: Eigene Zusammenstellung.

2.7.4. Theoriebasierte Evaluation

Einordnung: Theoriebasierte Evaluation entstand Ende der 1970er (Rossi und Freeman 1979). Sie entfaltete sich ab den 1990ern (Donaldson 2007; u.a. Chen 1990; 2015; Rogers und Weiss 2007; Weiss 1997; Stame 2004) und hat sich inzwischen zu einem weit verbreiteten Ansatz entwickelt, seit Mitte der 2000er auch in der deutschsprachigen Literatur (Giel 2013; Haubrich 2006; 2009; Hense und Kriz 2005; Stockmann 2006a; Silvestrini und Reade 2008). Der Ansatz wurde und wird auch häufig in Evaluationen staatlicher Behörden bzw. internationaler Organisationen eingesetzt (vgl. World Bank 2004; OECD 2010a; Caspari und Barbu 2009; Leeuw und Vaessen 2009; Treasury Board of Canada 2012; European Commission 2013; 2015; 2017b, S. 334–337); dies reicht von der Gesundheitsforschung über die Kinder- und Jugendarbeit bis hin zur Entwicklungszusammenarbeit. Der Ansatz ist auch bekannt unter den Bezeichnungen „Interventionslogik-Analyse“ (Leeuw 2003) und „Programmtheorie-Evaluation“ (Donaldson 2007). Die „Realistic Evaluation“ (Pawson und Tilley 1997; 2005) ist eine Unterart des Ansatzes,⁵ der auch mit dem „Theory of Change“-Modell (Mayne 2015) und dem „Logical Framework“ („Log Frame“)-Ansatz verwandt ist. Der Namensteil „theoriebasiert“ verweist auf das Überprüfen von inhaltlichen Annahmen und reflektiert die Abgrenzung von der *methoden*fokussierten Evaluation der 1960er und 1970er. Diese konzentrierte sich auf die Messung von Effektivität, ohne hinreichend nach den inhaltlichen Ursachen, Mechanismen und Kontextbedingungen von (mangelnder) Effektivität zu suchen („black box assessments“, Chen 1990; Schmitt 2018).

Kerninhalte & Methoden: Bei der theoriebasierten Evaluation handelt es sich um einen weitgehend qualitativen Ansatz. Er beruht auf der narrativen (gegebenenfalls auch grafisch aufbereiteten) Rekonstruktion und Plausibilitätsprüfung der Wirkannahmen („Wirkungsmodell“, auch: „Programmtheorie“, „Interventionslogik“, „Theorie des Wandels“), die einem Programm, einem Politikinstrument, einer Maßnahme etc. zugrundeliegen. Im Fall der begleitenden bzw. ex post Evaluierung werden diese Wirkannahmen mit qualitativen und (wo vorhanden) quantitativen empirischen Daten abgeglichen.

Theoriebasierte Evaluation lässt sich in drei Schritte unterteilen.

1) Der *erste Schritt* ist die Rekonstruktion des Wirkungsmodells, also der Ursache-Wirkungs-Annahmen bzw. „Wirkungsketten“ zwischen Programm / Maßnahme, seinen / ihren konkretisierten Dienstleistungen und Produkten („Outputs“) und den Programm- bzw. Maßnahmenwirkungen. Die Wirkungen werden in „Outcomes“ und „Impacts“ unterschieden. Mithilfe von Ressourcen („Inputs“) setzen die zuständigen Akteure das Programm, Instrument oder die Maßnahme durch eine Reihe von Aktivitäten um. Im Folgenden erläutern wir die verwendeten Fachbegriffe, angelehnt an die Definitionen der Deutschen Gesellschaft für Evaluation (DeGEval 2016)

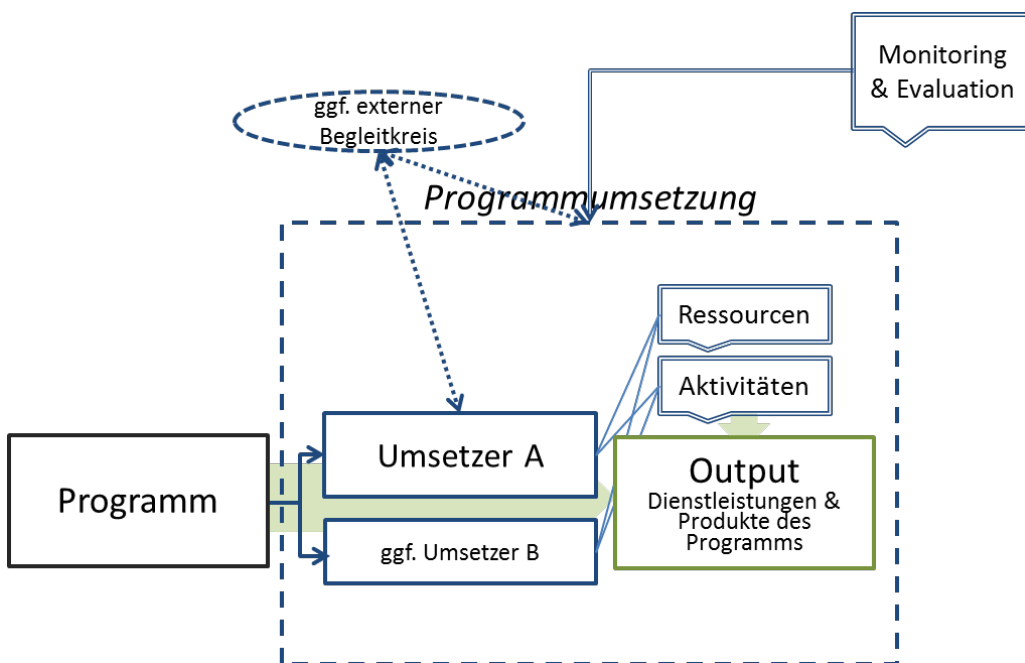
- „**Ressourcen**“ (oder „Inputs“) sind die finanziellen, personellen, materiellen, administrativen, organisationalen und anderen Mittel, die in ein Programm oder eine Maßnahme investiert werden, um dessen bzw. deren Ziele zu erreichen;
- „**Aktivitäten**“ sind die im Zuge der Umsetzung eines Programms oder einer Maßnahme durchgeführten Arbeitsschritte, Tätigkeiten, Leistungen etc.;
- „**Outputs**“ sind die zählbaren Dienstleistungen und Produkte eines zu evaluierenden Programms bzw. einer Maßnahme, über die Wirkungen erreicht werden sollen;

⁵ Sie ist in besonderem Maße bemüht, die Eingebettetheit, das Eigenleben und die Nicht-Linearität von Programmen bei ihrer Evaluation zu berücksichtigen (vgl. Pawson und Tilley 1997; 2005). In der Analyse wird auf Wirkmechanismen, Ergebnismuster (Outcome Patterns), Kontext sowie die Verknüpfungen zwischen diesen geachtet.

- **„Outcomes“** sind die Auswirkungen des Programms bzw. der Maßnahme auf Ebene seiner Zielgruppen. Dies kann eigene Umsetzungsmaßnahmen bei den Zielgruppen umfassen, von den Zielgruppen in Anspruch genommene Leistungen, Änderung von Wissensbeständen, Bewusstsein/Einstellungen wie auch Verhaltensänderungen bei der Zielgruppe. Wir unterscheiden daher zwischen Outcomes verschiedener „Ordnungen“ (Outcome I, Outcome II, Outcome III etc.);
- **„Impacts“** sind die Einwirkungen des Programms jenseits seiner unmittelbar adressierten Zielgruppe(n), in unserem Fall vor allem auf die natürliche Umwelt, aber auch auf die Gesellschaft im weiteren Sinne.

Ein Wirkungsmodell kann unterschieden werden in ein so genanntes „Aktionsmodell“ und ein „Veränderungsmodell“. ⁶ Im **Aktionsmodell** wird die Umsetzung eines Programms oder einer Maßnahme dargestellt, bzw. „welche Organisationen und Akteure im Hinblick auf welche Zielgruppen was genau unter welchen Bedingungen tun müssen, damit die Maßnahme realisiert wird“ (Döring et al. 2016, S. 1011) (vgl. Abbildung 1, dort die blauen Elemente).

Abbildung 1: Programm-/ Maßnahmenumsetzung („Aktionsmodell“)



Quelle: eigene Darstellung.

Das **Veränderungsmodell** bildet den unterstellten Wirkungspfad des Programms oder der Maßnahme auf die unmittelbaren Zielgruppen, die weitere Gesellschaft (über die Zielgruppen hinaus) und die Umwelt ab. Es modelliert den Zusammenhang zwischen Programm, seinen Outputs, Outcomes und Impacts. Eine erste, besonders vereinfachte Schematisierung bietet Abbildung 2:

Abbildung 2: Programm-/ Maßnahmenwirkungen („Veränderungsmodell“)



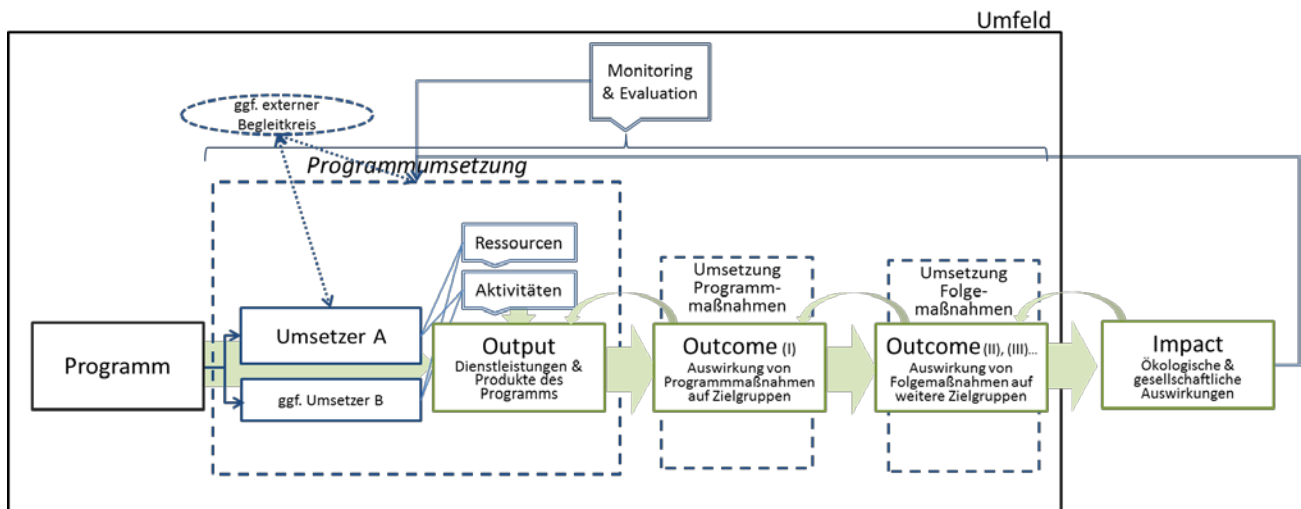
Quelle: Eigene Darstellung.

⁶ Die Begriffe sind Chen (2015) entlehnt („action model“, „change model“)

Vertreterinnen und Vertreter der theoriebasierten Evaluierung weisen allerdings darauf hin, dass Wirkungsketten in der Wirklichkeit deutlich komplexer, weniger linear und mit Rückkopplungsschleifen versehen sind (Pawson und Tilley 1997; Weiss 1997). Auch führen sie keineswegs immer zu den erhofften Programmwirkungen. Ein entsprechend ausdifferenzierteres Schema entwickeln wir in Kapitel 3.2.1.

Aktionsmodell (blau) und Veränderungsmodell (grün) ergeben das vollständige Wirkungsmodell (Abbildung 3).

Abbildung 3: Wirkungsmodell mit Aktions- und Veränderungsmodell



Quelle: eigene Darstellung.

2) Ist das Wirkungsmodell etabliert – also die grundlegende Logik identifiziert, wie das Programm oder die Maßnahme seine/ ihre Ziele erreichen soll –, werden in einem *zweiten Schritt* die dem Programm bzw. der Maßnahme zugrundeliegenden Annahmen über die Wirkungsbedingungen und Wirkungsketten in größerem Detail betrachtet und ausgewertet (Bussmann et al. 1997, S. 81–82):

- Sind die (expliziten und impliziten) Annahmen *logisch kohärent* bzw. in sich schlüssig? Passen die einzelnen Bestandteile des Wirkungsmodells – auch in den Größenordnungen der Wirkungsbeziehungen – zueinander? Ist das zugrundeliegende Konzept stimmig, steht das Programm gegebenenfalls in Widerspruch zu bestehenden bzw. geplanten anderen Politiken?
- Sind die (expliziten und impliziten) Annahmen *empirisch konsistent*? D.h., sind die wichtigsten Bestandteile des Wirkungsmodells empirisch ausreichend abgestützt oder stehen zumindest nicht in Widerspruch zu empirischen Beobachtungen in der Literatur?

Für die Überprüfung und Hinterfragung des Wirkungsmodells stellen sich folgende Fragen:

- Wie kann die Maßnahme die erwarteten Wirkungen erzeugen – welche Outputs tragen (über welche Mechanismen) und warum, unter welchen Bedingungen für welche Zielgruppen zu den intendierten und gegebenenfalls auch nicht-intendierten Outcomes und Impacts bei?
- Welche Faktoren fördern, welche hemmen die Entstehung von Wirkungen? (Erfolgsfaktoren, Hemmnisse)

Zur Beantwortung der Fragen sollte zunächst die aktuelle wissenschaftliche Literatur zum Themenfeld herangezogen werden (beispielsweise: Konsumforschung, Forschung zu Politikeffektivität). Dabei können aus der Literatur Hypothesen zu Erfolgsfaktoren oder Hemmnissen abgeleitet werden, mit deren Hilfe die Analyse des Wirkungsmodells strukturiert werden kann (Wolff & Schönherr 2011).

Darüber hinaus sollte die Richtigkeit des Wirkmodells und der Kausalitäten mit Stakeholdern diskutiert werden (in Interviews, Workshops, Fokusgruppen etc.) (Caspari und Barbu 2009, S. 19).

Wird das Wirkungsmodell eines Programms oder einer Maßnahme explizit benannt und überprüft, so lässt sich zum einen das Problemlösungspotenzial des Programms (qualitativ) abschätzen. Zum anderen kann das Programm verbessert werden, weil Wirkprozesse und mögliche Hemmnisse besser verstanden werden.

LogFrames als ex ante Evaluierungen

Vor allem in der Entwicklungszusammenarbeit ist es üblich, *bereits bei der Entwicklung von Projekten* Wirkungsannahmen zu diskutieren. Im Sinne einer theoriebasierten ex ante Evaluierung wird ein so genanntes Logical Framework (LogFrame) entwickelt. In ihm werden die angenommenen Wirkungszusammenhänge zwischen Output, Outcome und Impact explizit gemacht. Zusätzlich werden die dahinterstehenden Annahmen offengelegt, einschließlich der Risiken, die das Zustandekommen der unterschiedlichen Wirkungen gefährden können. Ergänzend dazu können jeweils Ziele zu den Outputs, Outcomes und Impacts gelistet, Indikatoren gebildet, Baselines und gegenwertige Werte der Indikatoren bestimmt sowie Datenquellen benannt werden. Beim Monitoring des Projektes sowie bei seiner ex post Evaluierung wird das LogFrame wieder aufgegriffen und überprüft, inwieweit die erwarteten Wirkungen eingetroffen und gegebenenfalls die Annahmen anzupassen sind.

3) Der *dritte Schritt* ist nur für begleitende bzw. ex post Evaluierungen relevant, nicht jedoch für ex ante Auswertungen. Hier findet der Abgleich der unterstellten Wirkungskette mit den empirischen Daten statt. Dies kann sich auf a) Programm- bzw. Maßnahmenwirkungen oder auf b) Erfolgsfaktoren und Hemmnisse (Ursache-Wirkungs-Zusammenhänge) beziehen.

a) Zu den Programm- bzw. Maßnahmenwirkungen werden Daten zu folgenden Fragen aufbereitet:

- Hat die Maßnahme, das Instrument oder Programm die erwarteten konkreten Dienstleistungen und Produkte (Outputs) erzeugt?
- Hat sie zu Änderungen im Verhalten von Zielgruppen geführt (Outcomes)?
- Hat sie unbeabsichtigte Nebenwirkungen gezeitigt?
- Inwiefern haben Änderungen im Zielgruppenverhalten zu Änderungen in ökologischen und sozialen Indikatoren geführt (Impacts)?

Um die Fragen zu beantworten, empfiehlt es sich, auf Ebene der unterschiedlichen Wirkungen vorab *Indikatoren* zu bilden: Output-, Outcome- und Impact-Indikatoren (siehe Kasten).

Beispiele für Wirkungsindikatoren

Beispiel-Maßnahme „Einführung einer verpflichtenden Energieverbrauchskennzeichnung“

Mögliche Outputs:

- Rechtliche Verpflichtung

Mögliche Output-Indikatoren:

- Verabschiedung Gesetz, Verordnung
-

Mögliche Outcomes:

- Kauf von Produkten, die mit der besten Kategorie („A+++“) gekennzeichnet sind

Mögliche Outcome-Indikatoren:

- Marktanteil der mit „A+++“ gekennzeichneten Produkte x Jahre nach Einführung der Kennzeichnungspflicht
 - Anzahl der mit „A+++“ gekennzeichneten Produkte p.a.
 - Marktdurchdringung der mit „A“ gekennzeichneten Produkte auf Ebene von Haushalten
-

Beispiele für Wirkungsindikatoren

<ul style="list-style-type: none"> • Rückgang des Kaufs weniger effizienter Geräte • Minderung des Stromverbrauchs durch Geräte • Minderung des absoluten Stromverbrauchs 	<ul style="list-style-type: none"> • Änderungen in der Zusammensetzung der Sortimente des Handels • Änderungen in der durchschnittlichen und spezifischen Energieeffizienz von Geräten (Kilowatt pro Jahr) • Änderungen im gerätebezogenen Stromverbrauch von Haushalten • Änderungen im absoluten Stromverbrauch
<p>Mögliche Impacts:</p> <ul style="list-style-type: none"> • Umwelt: Minderung von Treibhausgas-Emissionen • Sozial/Wirtschaft: Beschäftigungseffekte 	<p>Mögliche Impact-Indikatoren:</p> <ul style="list-style-type: none"> • THG-Minderungen (kumulativ jährlich) im Evaluierungszeitraum (Mio t CO₂ Äquivalente) • Direkte Bruttobeschäftigungseffekte bei Geräteherstellern • Indirekte Beschäftigungseffekte in vor- und nachgelagerten Sektoren

Quelle: vgl. Wolff & Schönherr (2011).

Die Indikatoren werden mit dem vorhandenen Datenmaterial abgeglichen. Lassen sich für bestimmte geeignete Indikatoren keine Daten finden oder generieren, sind Proxy-Variablen⁷ zu entwickeln.

Wichtige Prüffragen bei der Analyse von (Programm-, Maßnahmen-) Wirkungen sind:

- Lassen sich die beobachteten Änderungen tatsächlich auf das zu evaluierende Programm bzw. die zu evaluierende Maßnahmen zurückführen? (= Kausale Attribution)
- Beziehungsweise: In welchem Umfang lassen sich die Änderungen auf das Programm / die Maßnahme zurückführen? (= Kausale Kontribution; Identifizierung der „Nettowirkung“ einer Maßnahme)⁸.

Methodisch können kausale Attribution und kausale Kontribution mithilfe der Prozessanalyse („process tracing“) erreicht werden. Dabei werden (z.B. politische Umsetzungs-) Prozesse detailliert empirisch nachvollzogen, um so die Frage zu klären, ob bzw. wie stark sich bestimmte Änderungen (z.B. im Verhalten von Zielgruppen) auf den Evaluationsgegenstand (hier: Politikmaßnahme, Programm) zurückführen lassen, oder ob andere Faktoren für die Änderung verantwortlich sind.⁹

Alternativ helfen Kontrollgruppen (siehe Kapitel 2.7.1) oder kontrafaktische Szenarien („Counterfactuals“, siehe Kapitel 2.7.2) zu erkennen, ob beobachtete Änderungen die Wirkung der evaluierten Maßnahme oder auf alternative Faktoren zurückzuführen sind (kausale Attribution).

b) Ergänzend zum Abgleich der unterstellten Wirkungskette mit empirischen Daten kann die Frage nach *Erfolgsfaktoren* und *Hemmnissen* (und damit nach Ursache-Wirkungszusammenhängen) angegangen werden:

⁷ D.h. eine Variable, die eine Eigenschaft misst, die nicht (oder: nicht objektiv, reliabel, valide oder mit vertretbarem Aufwand) *direkt* gemessen werden kann.

⁸ Zur Unterscheidung zwischen kausaler Attribution und Kontribution (siehe Caspari & Barbu 2009, S. 20-21).

⁹ Beim Process Tracing handelt es sich damit um eine „Untersuchungsmethode zur kausalen Erklärung, bei der vielfältige empirische Beobachtungen innerhalb eines oder mehrerer Fälle als potenzielle Implikationen theoretischer Kausalmechanismen verstanden werden“ (Starke 2015).

- *Warum* hat das Programm oder die Maßnahme die identifizierten Wirkungen erzeugt?
- Was waren fördernde Faktoren?
- Warum konnten andere, erwartete Effekte nicht beobachtet werden? Was waren hemmende Faktoren?

Zur strukturierten Beantwortung dieser Fragen können die im zweiten Schritt entwickelten Hypothesen genutzt werden.

Die empirischen *Daten*, die zur Beantwortung der Fragen nach a) Wirkungen und b) Ursachen herangezogen werden, können sowohl qualitativer als auch quantitativer Natur sein. Ihre Quellen können Interviews (mit Umsetzungsakteuren, Zielgruppen, Intermediären, Expertinnen und Experten), Umfragen, Marktstatistiken, Monitoringdaten, Stoffstromanalysen, ökonometrische Modellierungen etc. sein. Sie können primär erhoben oder aus der bestehenden Literatur entnommen werden. Interviews eignen sich in besonderem Maße, um zu verstehen, warum bestimmte angenommene Wirkungspfade funktioniert haben oder warum nicht, und ob sich Änderungen tatsächlich ursächlich auf die Maßnahme zurückführen lassen.

Was die konkreten **Methoden** der theoriebasierten Evaluation betrifft, so stellt die narrative Analyse von Wirkungsmodellen (Programmtheorien, Interventionslogiken etc.) das Rückgrat des Ansatzes dar. Insbesondere in der begleitenden und ex post-Evaluation gilt es, dieses Rückgrat mit Daten zum Stand und den Wirkungen der Umsetzung anzureichern. Wie unter (3) erwähnt, kann hierfür auf die gesamte Breite qualitativer und quantitativer Daten zurückgegriffen werden. Es wird empfohlen, Akteure, die für die Programmentwicklung und Umsetzung zuständig sind, aber auch gesellschaftliche Stakeholder partizipativ in die Evaluation einzubeziehen. So kann auch erfasst werden, wenn unterschiedliche Wahrnehmungen einer Interventionslogik existieren. Dies erhöht die Qualität und Akzeptanz der Analyse (Brulin und Svensson 2012).

Stärken & Schwächen: Die Offenheit theoriebasierter Evaluation für unterschiedliche Fragestellungen und (qualitative, quantitative) Methoden ermöglicht die Betrachtung unterschiedlicher Evaluationskriterien, von Effizienz über Effektivität und Wirkungen bis hin zu Akzeptanz. Die Rekonstruktion von Wirkungsmodellen gibt sowohl auf Programm- als auch auf Maßnahmenebene Hinweise auf mögliche Risiken und Herausforderungen bei der Erzielung gewünschter Wirkungen und beim Entstehen möglicher unerwarteter Nebenwirkungen. Wirkungsmodelle als im Wesentlichen narrative (gegebenenfalls graphisch unterstützte) Rekonstruktionen von Interventionslogiken können auch die Komplexität multizentrischer Programme erfassen. Sie sind sowohl für die ex ante als auch ex post Analyse einsetzbar. Theoriebasierte Evaluation greift auch für die Wirkungsevaluation von Evaluationsgegenständen, die nicht mit klaren Zielen versehen sind, da das zugrundeliegende Wirkungsverständnis über die Zielerreichung hinausgeht und möglichst alle Wirkungen – einschließlich Nebenwirkungen – umfasst.

Theoriebasierte Evaluation wird in der Evaluationsforschung gerade für solche Fälle empfohlen, in denen eine quantitative Abschätzung von Effekten schwierig ist; dies gilt u.a. für „weiche“ Instrumente und solche, die noch nicht lange implementiert sind und für die daher noch keine Daten vorliegen (Astbury und Leeuw 2010; Gysen und Bachus 2006; Leeuw und Vaessen 2009; Persson und Nilsson 2007). Mittels qualitativer Methoden wird versucht, sowohl „weiche“ Wirkungen (auf Wissensbestände, Umweltbewusstsein) als auch „härtere“ Wirkungen (auf Umwelthandeln, Umweltbelastungen) und Langfristwirkungen zumindest theoriebasiert zu erfassen. Die Rekonstruktion von Wirkungsmodellen, die in der Regel den Erfahrungsschatz der Literatur zu den betreffenden Wirkungszusammenhängen berücksichtigt, kann – wenn empirische Primärdaten schlecht verfügbar sind – zumindest strukturierte Plausibilitätsaussagen treffen.

Theoriebasierte Evaluation fragt zudem explizit nach *Ursachen* des gemessenen Erfolgs oder Misserfolgs eines Programms oder einer Maßnahme, und kann so formativ für deren lernende Weiterentwicklung nutzbar gemacht werden. Die Wirkungskettenanalyse kann (durch die Aufbereitung der einschlägigen Literatur, Einsatz von Interviews und ergänzenden Datenerhebungsverfahren) durchaus zeit- und ressourcenintensiv sein. Die Achillesferse der theoriebasierten Evaluation (wie auch der von empirischen Schätzungen und Modellierungen) ist die kausale Zuordnung von Wirkungen zu einer Maßnahme oder einem Programm. Hier können kausale Evidenzen aber über Prozesstracing oder Entwicklung kontrafaktischer Szenarien zumindest verdichtet werden.

Weitere Stärken, aber auch Schwächen der theoriebasierten Evaluation listet die folgende Tabelle (u.a. basierend auf Treasury Board of Canada 2012; Giel 2013; Pawson et al. 2004).

Stärken & Schwächen der theoriebasierten Evaluation:

Stärken	Schwächen
Nutzbar in allen Formen der Evaluation (ex ante, begleitend, ex post), v.a. mit Fokus auf Kriterium der Wirkung / Wirksamkeit, aber auch andere Kriterien (multikriteriell)	Ermöglicht nicht unbedingt eine quantitative Angabe zur Wirkung eines Programms
Nutzbar in vielen Settings, auch, wo andere Methoden (z.B. Experimente, Counterfactuals) nicht genutzt werden können	Voraussetzungsvoll, weil neben Evaluationsexpertise vertiefte Kenntnis der inhaltlichen Materie und Synthese unterschiedlicher Datenquellen und Stakeholder-Perspektiven erforderlich sind
Nutzbar für Interventionen, in denen eine quantitative Abschätzung von Effekten schwierig ist (z.B. weil noch nicht lange implementiert; weil „weiches“ Instrument; weil Langfristwirkung)	Aufwändig, wenn Durchführung in vollem Umfang (Kontext spezifizieren, Hypothesen-Testen); es sind aber pragmatische Zugänge möglich (mit weniger detaillierten Wirkungsmodellen und weniger Hypothesentesten)
Erschließt die „black box“ einer Intervention und gibt Antworten auf die Fragen „ <i>Warum</i> hat eine Maßnahme Wirkungen entfaltet (oder nicht)? <i>Wie</i> hat die Maßnahme gewirkt; unter welchen Bedingungen?“ (Ursache-Wirkungs-Beziehungen); ermöglicht so Lernen und Programmverbesserung (formativ)	Daten, die auf kausale Zusammenhänge und Mechanismen hinweisen, sind schwierig zu finden
Ermöglicht Evaluatoren und Umsetzern, ein Wirkungs-Narrativ zu erzählen, das für involvierte Gruppen / Stakeholder nachvollziehbar ist; bindet i.d.R. Umsetzer ein	Setzt Offenheit und Kooperationsbereitschaft bei Programmteiligen voraus
Ermöglicht die sinnvolle Strukturierung und Interpretation existierender Datenbestände	

Quelle: Eigene Zusammenstellung.

2.8. Untersuchungsdesigns

Evaluationen können unterschiedliche Untersuchungsdesigns nutzen, die auch kombinierbar sind. Da es sich im Wesentlichen um die üblichen Untersuchungsdesigns der sozialwissenschaftlichen Forschung handelt, zu der umfassend Literatur existiert (u.a. Bussmann et al. 1997; Stockmann 2000, Böttcher et al. 2014; Diller 2016), führen wir die Designs hier nur stichwortartig an:

- Experimentelle versus nicht-experimentelle Untersuchung (vgl. Kapitel 2.7.1)
- Einzelfalluntersuchung versus (vergleichende) Multifallstudie
- Komparatives Design (z.B. Vorher-Nachher Vergleich, Mit-Ohne Vergleich) versus nicht-komparatives Design
- Querschnitts-, Längsschnitt- und Trendanalyse
- Vollerhebung versus Auswahl
- Partizipatives Design, z.B. im Rahmen von Planungszellen, multikriteriellen Ansätzen, Delphi-Methoden, Fokusgruppen.

2.9. Datenerhebung

Für die unterschiedlichen Evaluationstypen und Untersuchungsdesigns eignen sich teils qualitative (nicht-standardisierte), teils quantitative (numerische) Verfahren der Erhebung empirischer Daten. In Evaluationen werden zunehmend qualitative und quantitative Methoden kombiniert, d.h. es wird ein „Multimethodenansatz“ verfolgt (Stockmann 2006c, S. 22). Welche Methoden der Datenerhebung gewählt werden, hängt damit zusammen, welche Forschungsfragen gestellt werden, welche Daten zu ihrer Beantwortung nötig sind und welche Indikatoren sich eignen (v.a. Outcome- und Impact-Indikatoren, siehe grüner Kasten in Kapitel 2.7.3). Nicht zuletzt spielt eine Rolle, welche Ressourcen für die Datenerhebung zur Verfügung stehen.

Mögliche **Methoden** sind unter anderem:

- nicht-/standardisierte Beobachtung
- standardisierte oder leitfadengestützte Interviews
- Befragungen
- Fokusgruppen und Gruppendiskussion
- Dokumenten- und Inhaltsanalysen
- Auswertung von Monitoringdaten oder Sekundärstatistiken (beispielsweise Marktdaten)
- Stoffstromanalysen
- Ökonomische und ökologische Modellierung

4. Evaluationspraxis

4.1. In der Politikevaluation genutzte Analyseleitfäden und -raster

In der Evaluationspraxis sind Analyseleitfäden oder -raster das „zentrale(...) Steuerungsinstrument einer Evaluation“ (Hupfer 2007, S. 24). Sie dienen zur Strukturierung des Informationsbedarfs und sind Grundlage der Datenerhebung, -analyse und -interpretation. Ihre Konstruktion leitet sich aus dem der Evaluation zugrunde liegenden Theoriegerüst sowie den spezifischen Nutzenerwartungen der Beteiligten ab.

Wir haben eine Reihe von in der Politikevaluation genutzte Analyseleitfäden und -raster gescreent und ausgewertet. Die Dokumente wurden oder werden in unterschiedlichen Politikfeldern und auf verschiedenen politischen Ebenen für die ex ante, begleitende und ex post Evaluierung politischer Instrumente genutzt. Sie umfassen sowohl generische Leitfäden (z.B. der Bundesregierung und der EU Kommission zur ex ante Gesetzesfolgenabschätzung oder der OECD und Weltbank zur ex post Evaluierung von Projekten der Entwicklungszusammenarbeit) als auch (in geringerem Umfang) Leitfäden, die in konkreten Evaluationen zum Einsatz kamen. Es wurden zum einen solche Leitfäden ausgewählt, die von zentralen politischen Akteuren der nationalen, EU und internationalen Ebene eingesetzt werden. Zum anderen wurden Leitfäden gewählt, die für Evaluationen von Umweltpolitik inhaltlich und / oder methodisch von Interesse sein können. Im Ergebnis wurden folgende Leitfäden erfasst:

Deutschland:

- Vorgaben zur Gesetzesfolgenabschätzung in der Gemeinsamen Geschäftsordnung der Bundesministerien (GGO); Arbeitshilfe zur Gesetzesfolgenabschätzung (BMI 2009)
- Gesellschaft für Internationale Zusammenarbeit: „Das Evaluierungssystem der GIZ. Zentrale Projektevaluierung im BMZ-Geschäft“ (GIZ 2018b), „Das Evaluierungssystem der GIZ. Theorie des Wandels für Evaluierungen der GIZ“ (GIZ 2018a), „Die Evaluierungspolicy der GIZ. Prinzipien, Leitlinien und Anforderungen an unsere Evaluierungspraxis“ (GIZ 2018c)
- Exemplarische Forschungsprojekte: Evaluation der Nationalen Klimaschutzinitiative (NKI) (Öko-Institut et al. 2017); Wirkungsanalyse bestehender Klimaschutzmaßnahmen und -programme (Wuppertal Institut für Klima, Umwelt und Energie und Ecofys 2014); Erfolgsfaktoren für die Förderung zur Anpassung an den Klimawandel (Ecolo und Bioconsult 2017); Evaluation der Umweltberatungsprojekte des Bundesumweltministeriums und des Umweltbundesamtes – Nachhaltige Wirkungen der Förderung von Bundesverbänden (CeVal 2002)

Europäische Ebene:

- EU Kommission: Better Regulation „Guidelines on impact assessment“ (European Commission 2017a), Better Regulation Toolbox 47: „Evaluation Criteria and Questions“ (European Commission 2017b)
- EU Kommission DG DEVCO & EuropeAid: Guidelines for project and programme evaluation (EU Joint Evaluation Unit 2006) & LogFrame Matrix (EuropeAid 2015)
- Europäisches Parlament: Impact Assessment Handbook & European Implementation Assessment des Wissenschaftlichen Diensts des EP (European Parliament 2012; 2017)
- Europäische Umweltagentur: „Framework for environment and climate policy evaluation“ (EEA 2016)
- Exemplarische Forschungsprojekte: „Policies to Promote Sustainable Consumption Patterns“ (EUPOPP 2011), „Sustainable Consumption Policies Effectiveness Evaluation“ (SCOPE2 2008)

Internationale Organisationen und Beispiele aus dem internationalen Ausland:

- UNEP: „UNEP Evaluation Policy“ (UNEP 2016), „UNEP Evaluation Manual“ (UNEP EOU 2008); Anwendungsbeispiel: „Terminal Evaluation der Green Economy Initiative“ (UNEP EOU 2017)
- OECD: „Regulatory Impact Analysis: A Tool for Policy Coherence“ (OECD 2009b)
- OECD: „DAC Network on Development Evaluation“ (OECD 2010a)
- World Bank: „Evaluating Behavior Change in International Development Operations: A New Framework“ (Flanagan und Tanner 2016) sowie „World Bank Group Impact Evaluation“ (World Bank 2012; 2016)
- UK: „The Magenta Book - Guidance for evaluation“ (HM Treasury 2011)
- Kanada: „Supporting Effective Evaluations: A Guide to Developing Performance Measurement Strategies“ (Government of Canada 2010)

Eine **Analyse** der Dokumente ergibt folgende Einblicke:

- Grundsätzlich es ist üblicher, dass staatliche Akteure Leitlinien und Leitfäden für die *ex ante* Evaluation (Gesetzesfolgenschätzung) ihrer eigenen Regelungen spezifizieren als für deren *ex post* Evaluation. Ausnahmen finden sich vor allem im Kontext internationaler Entwicklungszusammenarbeit (z.B. Weltbank, OECD DAC, GIZ) bzw. bei UN Organisationen (UNEP) sowie im Europäischen Kontext beim Europäischen Parlament (Leitlinien für „European Implementation Assessments“ des Wissenschaftlichen Dienstes des Europäischen Parlamentes).
- Viele der Leitlinien nutzen die theoriebasierte Evaluierung oder mit ihr verwandte Ansätze wie Logical Framework Analysis bzw. „theories of change“. Demgegenüber gibt es kaum „echte“ experimentelle Designs, wenn auch in einigen Fällen quasi-experimentelle Ansätze angestrebt werden (beispielsweise bei der Weltbank).
- In der Regel wird ein Mix aus qualitativen und quantitativen Methoden angestrebt bzw. angewendet, wobei der Anteil qualitativer Methoden überwiegt.
- Viele der erfassten Leitfäden definieren zwar Evaluationskriterien, führen darüber hinaus das Analyseraster aber nur wenig aus. Eine Konkretion erfolgt dann erst auf Ebene spezifischer Evaluationen (UNEP 2010; z.B. European Commission 2016).

4.2. Exemplarischer Analyserahmen für die Evaluation eines Programms

Im Folgenden wird ein Analyseraster vorgestellt, das für die *ex ante* Evaluation des „Nationalen Programms für Nachhaltigen Konsum“ (BMUB; BMJV; BMEL 2017) entwickelt wurde. Das Analyseraster wurde im Rahmen des Ufoplan-Vorhabens „Nachhaltigen Konsum weiterdenken“ erstellt, innerhalb dessen dieses Arbeitspapier entstand. Es wurde dann zunächst im Rahmen einer *ex ante* Programmevaluation – basierend auf der Druckfassung des verabschiedeten Programms – angewendet (Muster et al. 2018). Anschließend wurde es einer (begleitenden bzw. *ex post*) Evaluation ausgewählter Maßnahmen des Programms zugrunde gelegt, die sich zum Zeitpunkt der Evaluation in der Umsetzung befanden oder bereits umgesetzt waren (Muster et al. 2019).

Das Nationale Programm für Nachhaltigen Konsum (NPNK) wurde 2016 als ressortübergreifendes Programm entwickelt, um nachhaltigen Konsum zu stärken. Als Evaluationsgegenstand ist das Programm in mehrfacher Hinsicht interessant: Es wird von mehreren Ministerien getragen und gesteuert („multizentrisches Programm“); es besteht aus einem übergreifenden Programmteil und mehreren inhaltlichen Teilen (je zu unterschiedlichen Bedürfnisfeldern des Konsums); und es umfasst überwiegend „weiche“ Maßnahmen (deren Wirkungen schwierig nachzuweisen zu sind).

Als methodische Grundlage für die beiden im Projekt durchgeführten Evaluationen des NPNK bzw. seiner Maßnahmen wurde die theoriebasierte Evaluation gewählt (vgl. Kapitel 2.7.3). Hintergrund dieser Entscheidung war, dass – wie oben beschrieben – theoriebasierte Evaluation eine (ex ante, begleitende) Bewertung auch von weichen und noch nicht (weit) umgesetzten Maßnahmen ermöglicht. Weil sie die „black box“ einer Intervention zu erschließen vermag und nach Ursache-Wirkungs-Beziehungen fragt, ermöglicht sie Schlussfolgerungen („Lernen“) insbesondere im Hinblick auf das Kriterium der Wirkungen bzw. Wirksamkeit, die formativ in die weitere Ausgestaltung von Maßnahmen eingehen können. Sie nutzt qualitative und teils auch quantitative empirische Daten, ist aber sowohl methodisch als auch in Bezug auf die Datenintensität weniger aufwändig als (quasi-) experimentelle Evaluationen, empirische Schätzungen oder Modellierungen. Nicht zuletzt eignet sie sich sowohl für die Evaluation von Programmen als auch von Einzelmaßnahmen.

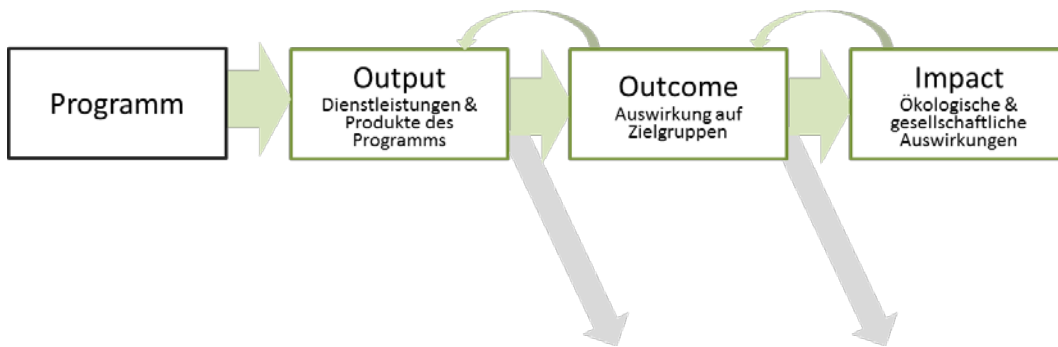
Im Folgenden wird der grundlegende Analyserahmen für die Evaluation des NPNK und ausgewählter Programmmaßnahmen aufgespannt. Zudem werden die Leitfragen und das Analyseraster dargestellt.

4.2.1. Rekonstruktion und Analyse des Wirkungsmodells

Wie unter Kapitel 2.7.3 geschildert, bestehen die ersten beiden Schritte einer theoriebasierten Evaluation in der Rekonstruktion und Analyse bzw. Bewertung des Wirkungsmodells. Für die ex ante Analyse des gesamten NPNK galt es also, die grundlegende Programmtheorie zu rekonstruieren und auf ihre Plausibilität hin zu überprüfen. Für die ex post Evaluierung einzelner Programmmaßnahmen war die spezifische Wirkungslogik dieser Einzelmaßnahmen zu analysieren.

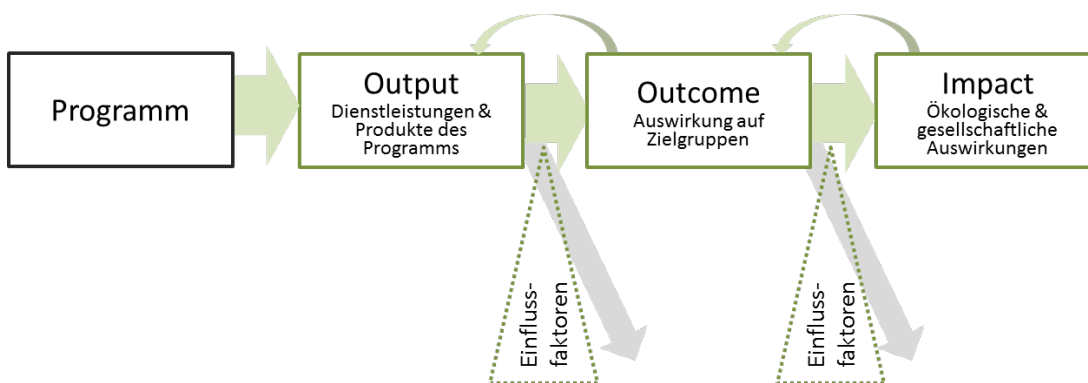
Dabei wurde das in Kapitel 2.7.3 nur angerissene Schema für die Analyse des Wirkungsmodells verfeinert, insbesondere was die Wirkungskettenanalyse zu den Programmwirkungen betrifft („Veränderungsmodell“). Hier erinnern wir an das noch sehr reduktionistische Schema zu Programmwirkungen von oben (Abbildung 2 oben). Es unterstellt, dass ein Programm oder eine Maßnahme zu Outputs („Produkten und Dienstleistungen“) führt, welche Veränderungen im Verhalten von Zielgruppen auslösen (Outcome), und diese wiederum ökologische und soziale Wirkung über die Zielgruppe hinaus (Impact) entfalten.

Während hier aus Abstraktionsgründen ein lineares und „optimistisches“ Bild gezeichnet wird, gilt es zu berücksichtigen, dass nicht alle von einem Programm oder einer Maßnahme ausgelösten Aktivitäten und Handlungen tatsächlich die gewünschten Effekte auslösen (graue, nach unten abzweigende Pfeile in Abbildung 4): Wirkmechanismen können bei Zielgruppen ins Leere laufen, Verhaltensänderungen nicht die erwarteten (Umwelt-)Wirkungen zeitigen, oder nicht-intendierte Wirkungen können entstehen. Zudem erfasst die Abbildung Rückkopplungsschleifen zwischen Impact, Outcome und Output (grüne gebogene Pfeile in der Abbildung): Es können Anpassungen des Programms oder Programm-Outputs erfolgen, wenn z.B. die erwünschten Wirkungen nicht stark genug sind, unerwünschte Wirkungen auftreten oder breitere soziale Wirkungen (Impact) Verhaltensänderungen (Outcome) bei Zielgruppen verstärken.

Abbildung 4: Programm-/ Maßnahmenwirkungen (II)


Quelle: Eigene Darstellung.

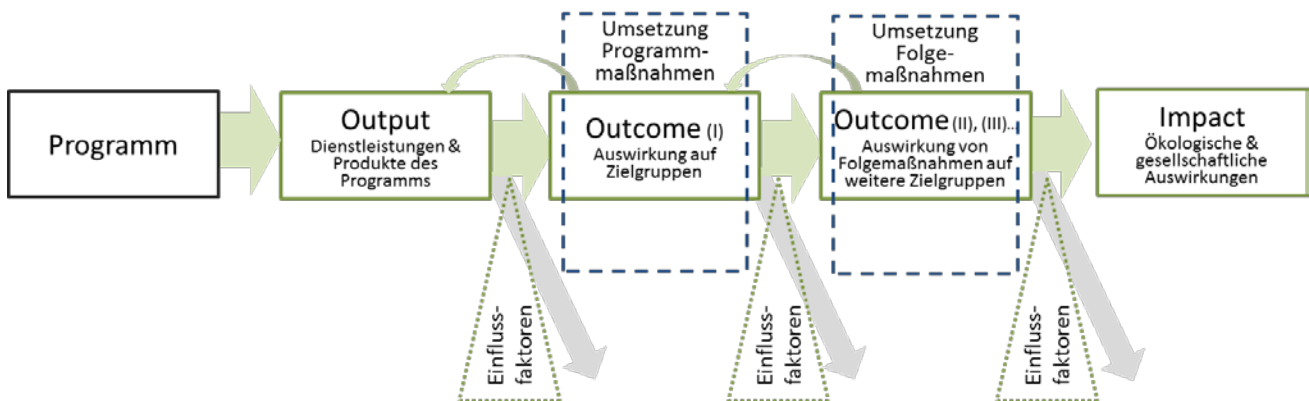
Abbildung 5 berücksichtigt zusätzlich, dass es (programminterne und -externe) Einflüsse auf die Wirkungskette gibt, die sich sowohl fördernd als auch hemmend auf den Programmserfolg auswirken können: Rahmenbedingungen, Fehlannahmen, Budgetkürzungen, Zielgruppenwiderstände etc. In ihrer positiven Ausprägung – d.h. als Erfolgsfaktoren – wurden solche Einflussfaktoren in Kapitel **Fehler! Verweisquelle konnte nicht gefunden werden.** aus der Literatur extrahiert.

Abbildung 5: Programm-/ Maßnahmenwirkungen (III)


Quelle: Eigene Darstellung.

Abbildung 6 schließlich reflektiert, dass aus einem Programm in der Regel nicht nur ein (erwünschtes) Outcome resultiert, sondern oft eine ganze Kaskade von Outcomes (hier: Outcome I-III). So kann ein Programm die Bildung eines Gremiums (Outcome I) durch eine Behörde (faktisch die erste Zielgruppe des Programms) vorschlagen, welches in der Folge einen Leitfaden (Outcome II) für Anwender (als die zweite Zielgruppe) entwickelt. Der Leitfaden soll bei den Anwendern Verhaltensänderungen (Outcome III) hervorrufen. Was hier als Outcome I-III bezeichnet ist, wird an anderer Stelle auf Englisch als „immediate“, „intermediate“ und „ultimate“ Outcomes, oder als kurz-, mittel- und langfristige Outcomes bezeichnet. Es ist offensichtlich, dass eine Outcome-Kaskade recht lang und komplex werden kann, zumal an jeder Schnittstelle neue Einflüsse, Annahmen und Risiken zu bedenken sind.

Abbildung 6: Programm-/ Maßnahmewirkungen (IV)



Quelle: Eigene Darstellung.

Neben dem Kreis der originären Programmumsetzer (großer, blau gestrichelter Kasten) ergeben sich weitere Umsetzungskontexte, wo Zielgruppen – gegebenenfalls gemeinsam mit Stakeholdern – entlang des Wirkungspfad Programmmaßnahmen umsetzen bzw. zu deren Umsetzung Folgemaßnahmen entwickeln und ihrerseits implementieren (vgl. Kasten).

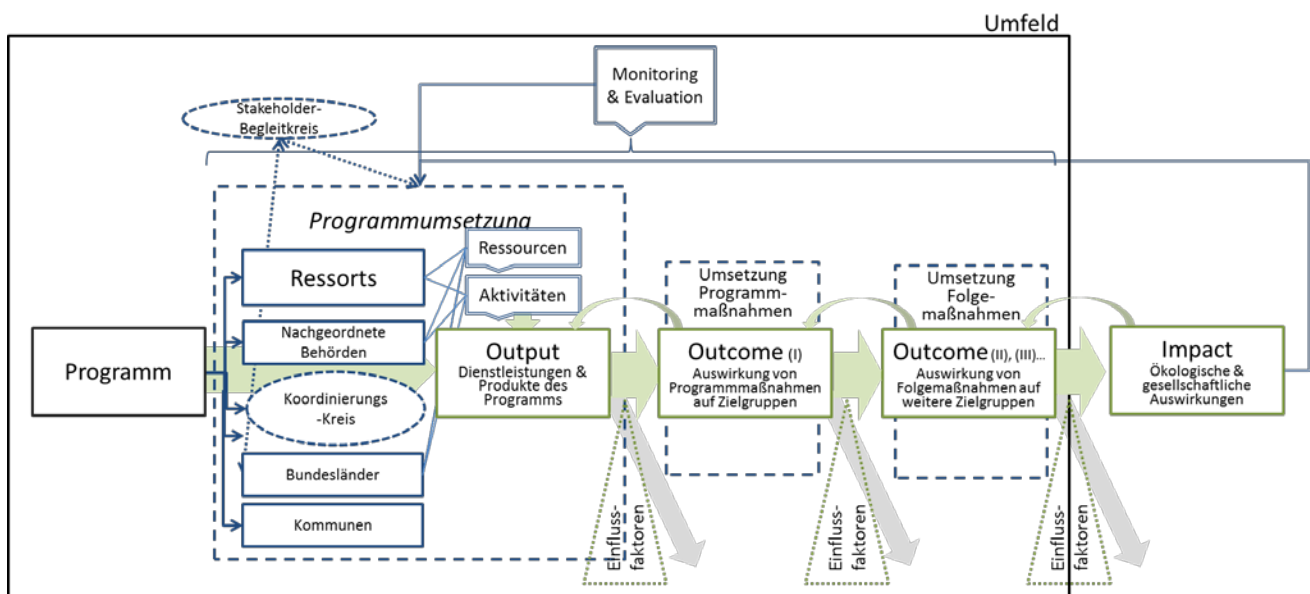
Umsetzungskontexte eines Programms

Nehmen wir das fiktive Beispiel einer im Programm angeregten Bildungskampagne für nachhaltigen Konsum: Sie könnte durch eine **Bund-Länder-Arbeitsgruppe** unter Mithilfe von *Expertinnen und Experten* entwickelt werden; durch **Landeskulturministerien** unter Einbezug von *Umweltverbänden* und der *Landes-Elternvertretung* konkretisiert werden; und schließlich in Form von Einzelmaßnahmen an **Schulen** mithilfe von *Trägerkreisen interessierter Eltern* umgesetzt werden. (Zielgruppen sind jeweils fett gedruckt, Stakeholder kursiv gesetzt.)

Jeder dieser Umsetzungskontexte stellt faktisch eine Art eigenes Aktionsmodell dar. In Abbildung 6 ist dieses durch die Kästen in blau gestrichelter Linie angedeutet.

Kombiniert man das Veränderungs- mit einem Aktionsmodell, das den Umsetzungsprozess abbildet (vgl. Kapitel 2.7.3), ergibt sich das folgende, gegenüber Abbildung 3 oben erweiterte Wirkungsmodell:

Abbildung 7: Erweitertes Wirkungsmodell



Quelle: Eigene Darstellung.

4.2.2. Leitfragen für die Evaluierung

Sind Rekonstruktion und Analyse des grundlegenden Wirkungsmodells eines Programms, einer Maßnahme oder eines Projekts abgeschlossen, geht es um die Bewertung – ex ante oder ex post – der analysierten Zusammenhänge bzw. vorgefundenen empirischen Daten dazu. Zentral hierfür sind die vorab festgelegten Evaluierungskriterien.

Für die Evaluation des Nationalen Programms für Nachhaltigen Konsum und ausgewählter Einzelmaßnahmen des Programms wurden folgende **Evaluationskriterien** vereinbart:

- **Relevanz**: Unter Relevanz wurde das Ausmaß verstanden, in dem die Ziele und der thematische Gegenstandsbereich des Programms bzw. der Maßnahmen mit zentralen Erkenntnissen und Erwartungen von wissenschaftlichen und zivilgesellschaftlichen Stakeholdern in Einklang stehen. Hierfür erfolgt ein Abgleich mit der wissenschaftlichen Literatur sowie der öffentlichen bzw. Fachdebatte.
- **Kohärenz** untergliedert sich in interne und externe Kohärenz. Unter *interner* Kohärenz verstehen wir das Ausmaß, in dem das Programm bzw. die Maßnahme in sich widerspruchsfrei ist – d.h. seine/ ihre Ziele, Grundsätze, Mechanismen und Maßnahmenvorschläge sich gegenseitig stützen oder einander zumindest nicht zuwiderlaufen. *Externe* Kohärenz bezieht sich auf das Ausmaß, in dem das Programm bzw. die Maßnahme widerspruchsfrei zu anderen relevanten Strategien (hier v.a. zur Deutschen Nachhaltigkeitsstrategie) ist.
- **Effizienz**: Hier geht es darum, wie sparsam Ressourcen in qualitative und quantitative Ergebnisse (Outputs, Outcomes, Impacts) umgewandelt werden. Als Leitfrage für die Bewertung der Effizienz einer Maßnahme wurde gefragt, ob die Ziele bzw. die erwarteten Wirkungen mit anderen Maßnahmen kostengünstiger erreicht werden könnten. Effizienz im Kontext des NPNK wurde nur ex post auf Ebene konkreter Maßnahmen eingeschätzt.
- **Erfolgsaussichten bzw. Wirksamkeit**: Unter Erfolgsaussichten wurde das Ausmaß verstanden, in dem die Ziele des Programms bzw. der jeweiligen Maßnahme unter Berücksichtigung ihrer relativen Bedeutung voraussichtlich kurz- und mittelfristig erreicht werden. Dafür wurden die

institutionelle Ausgestaltung bewertet, die Ressourcenausstattung, die Ziele und das Maß, in dem die Programmmaßnahmen dazu beitragen können, die genannten Ziele zu erreichen. Erfolgsaussichten werden aus heutiger Sicht bzw. ex ante analysiert. Als Wirksamkeit wurde das Ausmaß definiert, in dem eine (bereits umgesetzte oder in Umsetzung befindliche) Maßnahme die gewünschten Wirkungen (Outcomes, Impacts) erzielt; Wirksamkeit wird ex post analysiert.

- **Dauerhaftigkeit von (Programm-, Maßnahme-) Wirkungen:** Dieses Kriterium bezieht sich auf das voraussichtliche Fortbestehen längerfristiger positiver Wirkungen des Programms bzw. der Einzelmaßnahme über dessen/ deren unmittelbare Laufzeit hinaus. Hinweise auf längerfristige Wirkungen gibt das Vorhandensein von Mechanismen für Monitoring, Evaluation, Programm-/Maßnahmenfortschreibung sowie Ansätze für Verstetigung.

Diese Evaluationskriterien werden mithilfe von **Evaluationsfragen** operationalisiert. Diese sind angepasst an den Aufbau und die Inhalte des Nationalen Programms für nachhaltigen Konsum, das unter anderem zwischen übergreifenden und bedürfnisfeldspezifischen Maßnahmen unterscheidet. Entsprechend wird das Kriterium der Relevanz auf die übergreifenden Programmziele, aber auch auf bedürfnisfeldspezifische Ziele, auf übergreifende und bedürfnisfeldspezifische Handlungsansätze angewendet. Auch wird im Hinblick auf die Frage nach der internen Kohärenz beispielsweise berücksichtigt, aus welchen Teilen das Programm besteht und welche davon miteinander kohärent sein sollten, um eine möglichst große Steuerungswirkung zu erzeugen.

Die folgende Tabelle zeigt die Evaluationskriterien und -fragen im Überblick.

Tabelle 3: Evaluationskriterien und -fragen bei der Evaluation des NPNK

Evaluationskriterium	Evaluationsfragen
Relevanz	<ul style="list-style-type: none"> • Berücksichtigt das Programm die relevanten Herausforderungen und Hemmnisse von nachhaltigem Konsum und nachhaltigen Konsumpolitiken? (auf übergreifender Programmebene und in Bezug auf spezifische Bedürfnisfelder) • Berücksichtigt das Programm die relevanten Ziele, Leitideen, Bedürfnisfelder, Handlungsansätze und Maßnahmen(bündel) zur Förderung eines nachhaltigen Konsums? (auf übergreifender Programmebene und in Bezug auf spezifische Bedürfnisfelder) Fehlen wichtige Ziele, Leitideen, Bedürfnisfelder, Handlungsansätze und Maßnahmen(bündel)? • <i>[Bei ex post-Evaluierung von Einzelmaßnahmen]:</i> Berücksichtigt die Maßnahme die Bedürfnisse und Erwartungen unterschiedlicher Stakeholder?
Kohärenz	<ul style="list-style-type: none"> • Inwieweit ist das Programm / die Maßnahme kohärent mit anderen nachhaltigkeitsorientierten Politiken bzw. Strategien der Bundesregierung (externe Kohärenz) – hier: mit der Deutschen Nachhaltigkeitsstrategie? • Inwieweit ist das Programm / die Maßnahme in sich kohärent (interne Kohärenz)?
Effizienz	<ul style="list-style-type: none"> • <i>[Bei ex post-Evaluierung von Einzelmaßnahmen]:</i> Könnten die Ziele bzw. die erwarteten Wirkungen mit anderen Maßnahmen kostengünstiger erreicht werden?
Wirksamkeit¹⁰ bzw. Erfolgsaussichten	<ul style="list-style-type: none"> • Institutionelle Ausgestaltung: Sind Zuständigkeiten für die Umsetzung klar benannt? Existieren Koordinierungs- und Kontrollmechanismen? Werden relevante Stakeholder wirksam und sinnvoll eingebunden? • Ressourcen: Welche finanziellen und personellen Kapazitäten sind für die Umsetzung des (Gesamt-) Programmes / der Maßnahme vorgesehen?

¹⁰ Die Frage nach der Wirksamkeit wird ex post nur auf Ebene von Einzelmaßnahmen beantwortet. Es wird also nicht die Frage gestellt, ob / in welchem Maße die Ziele des gesamten Programms erreicht wurden.

Evaluationskriterium	Evaluationsfragen
	<ul style="list-style-type: none"> • Ziele: Inwieweit bauen Ziele auf unterschiedlichen Abstraktionsebenen (Leit-, Mittler-, Handlungsziele) aufeinander auf? Sind die Handlungsziele spezifisch, messbar, angemessen und terminiert (vgl. „SMART“-Konzept)¹¹? • Maßnahmenvorschläge (ex ante): Wie konkret sind Maßnahmenvorschläge ausformuliert? Ist die Umsetzung von Maßnahmen verbindlich? Sind die Maßnahmen(-bündel) geeignet, die Ziele des Programms bzw. seiner Teilbereiche (übergreifende Handlungsansätze und Bedürfnisfelder) zu erreichen? • Maßnahmen (ex post): Wie genau soll die Maßnahme Outcomes und Impacts erzeugen? Ist diese Wirkungslogik plausibel? Wird diese Sicht der Wirkungslogik von unterschiedlichen Stakeholdern geteilt? In welchem Maße wurden die Ziele der Maßnahme erreicht bzw. die erwarteten Outcomes und Impacts erzielt? Warum hat die Maßnahme die beobachtbaren Wirkungen erzielt – was waren fördernde (endogene, exogene) Faktoren? Warum hat die Maßnahme nicht noch weitere Wirkungen erzielt – was waren hemmende (endogene, exogene) Faktoren? In welchem Maße wurden nicht-intendierte Wirkungen (positive oder negative) erzielt?
Dauerhaftigkeit der Programmwirkungen	<ul style="list-style-type: none"> • Monitoring & Evaluation: Wird der Programmfortschritt regelmäßig erfasst (Monitoring)? Ist eine systematische und regelmäßige Untersuchung des Nutzens bzw. der Güte des Programms vorgesehen (Evaluation)? Existieren (Outcome, Impact-) Indikatoren für die Erfassung des Programmfortschritts und der Programmgüte? Existieren „SMARTE“ Ziele und Baselines, um die Programmwirkung messen zu können? • Weiterentwicklung: Ist ein systematischer Prozess zur Weiterentwicklung des Programms vorgesehen („Review“)? • Institutionelle Einbettung: Existiert eine angemessene institutionelle Verankerung des Programms? Existiert ein angemessener Umsetzungsmechanismus für das Programm bzw. Teile des Programms? • Verstetigungsperspektive: In welchem Maße können Strukturen und Erfolge, die im Rahmen des Programms und seiner Maßnahmen potenziell aufgebaut werden, verstetigt werden? Ist die langfristige Finanzierung des Programms und seiner Maßnahmen gesichert?

Quelle: eigene Zusammenstellung.

Auf Basis der Evaluationskriterien und -fragen einerseits und der Struktur des Programms andererseits wurde ein **Analyseraster** entwickelt, das dem konkreten Evaluationsprozess zugrunde gelegt wurde. Über die obige Tabelle hinaus spezifiziert das Analyseraster Hilfsfragen, definiert Datenquellen und bietet Raum für Definitionen oder andere Anmerkungen.

Für die Entwicklung des Rasters wurden unterschiedliche der in der politischen Evaluierungspraxis genutzten Raster ausgewertet (vgl. Kapitel 3.1) und Elemente und Erkenntnisse daraus aufgegriffen, insbesondere aus OECD (2010b) und EEA (2016). Das Analyseraster wird im Folgenden abgebildet.¹²

¹¹ In einem Programm sind in der Regel Ziele unterschiedlicher Abstraktion verankert: *Leitziele* und *Mittlerziele* sind relativ abstrakt und können daher nicht „SMART“ ausformuliert sein; konkrete *Handlungs-* oder *Qualitätsziele* hingegen sollten SMART formuliert sein (vgl. Bertelsmann Stiftung 2014).

¹² An der Entwicklung des Analyserasters waren weitere Mitglieder des Projektteams maßgeblich beteiligt, denen wir an dieser Stelle danken möchten: Viola Muster (ConPolicy/ TU Berlin), Christian Thorun (ConPolicy), Ulf Schrader (TU Wien), Lucia Reisch (Copenhagen Business School und Zeppelin Universität Friedrichshafen) sowie Rainer Griebshammer und Corinna Fischer (Öko-Institut).

4.2.3. Analyseraster

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
I. Wirkungsmodell				
	<ul style="list-style-type: none"> Wie genau soll das Programm Wirkungen (bei wem) erzielen? Über welche Mechanismen bzw. Wirkungsketten? 	<ul style="list-style-type: none"> Was sind die (subjektiven) Annahmen der Programmentwickler/ Stakeholder über die beabsichtigte Funktionsweise des Programms? (Aktions- und Veränderungsmodell) 	Ausarbeitung eines groben Wirkungsmodells für das Gesamtprogramm: Ziele/intendiertes Ergebnis, Aktivitäten und Ressourcen im Programm, programmendogene und -exogene Hemmnisse oder Erfolgsfaktoren, Erfolgsindikatoren, benötigte Informationen zur Bewertung der Durchführung, Datenquellen. Gegencheck mit Programmbeteiligten.	
II. Relevanz				
Identifikation und Bewertung der relevanten Herausforderungen und Hemmnisse	<ul style="list-style-type: none"> Berücksichtigt das Programm die relevanten Herausforderungen und Hemmnisse von nachhaltigem Konsum und nachhaltigen Konsumpolitiken? (Herausforderungen können auch die Form von „Megatrends“ haben) 		Abgleich der im übergreifenden Programmteil gelisteten Hemmnisse und Megatrends mit den in der wissenschaftlichen Literatur bzw. von Stakeholdern identifizierten (wichtigen) Herausforderungen, Hemmnissen und Megatrends	Abgleich der in den BF-Kapiteln gelisteten Herausforderungen und der Hemmnisse mit den in der wissenschaftlichen Literatur bzw. von Stakeholdern identifizierten (wichtigen) Herausforderungen und Hemmnissen

¹³ Bedürfnisfeld wird im Folgenden abgekürzt als „BF“.

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
Identifikation und Bewertung der Ziele für einen nachhaltigen Konsum	<ul style="list-style-type: none"> Berücksichtigt das Programm die relevanten Ziele und Leitideen zur Förderung eines nachhaltigen Konsums? Fehlen wichtige Ziele und Leitideen? 	<ul style="list-style-type: none"> Was sind die Ziele des Programms bzw. der Bedürfnisfelder (BFs)? Fehlen Ziele – jenseits von politisch gesetzten Zielen (diese werden unter → Kohärenz erfasst) solche, die man als „gesellschaftlich anerkannt“ beschreiben könnte? <i>Anmerkung: Ziele werden an unterschiedlichen Stellen im Programm genannt</i> 	<p>Abgleich der Ziele mit dem wissenschaftlichen Diskurs/der gesellschaftlichen und politischen Diskussion</p> <p>Abgleich der in der wissenschaftlichen Literatur identifizierten (aus Umwelt-/ Nachhaltigkeitsicht) prioritären Bedürfnisfelder → Fehlen bestimmte BFs?</p>	<p>Abgleich der BF-spezifischen Ziele (sofern spezifiziert) mit den in der wissenschaftlichen Literatur identifizierten ökologischen und sozialen „Big Points“ dieses BFs --> Fehlen bestimmte Themen / Ziele innerhalb der BFs?</p>
	<ul style="list-style-type: none"> Berücksichtigt das Programm die relevanten Bedürfnisfelder zur Förderung eines nachhaltigen Konsums? Fehlen wichtige Bedürfnisfelder? 			
Identifikation der relevanten Handlungsansätze und Maßnahmen(bündel)?	<ul style="list-style-type: none"> Berücksichtigt das Programm die wesentlichen Handlungsansätze und Maßnahmen(bündel) zur Förderung eines nachhaltigen Konsums? Sind die relevanten Ansätze und Maßnahmen(bündel) genannt? 		<p>Abgleich der in der Literatur / Debatte identifizierten (potentiell wirksamen) übergreifenden „Handlungsansätzen“ mit den im Programm genannten Ansätzen und Maßnahmen(bündeln).</p>	<p>Abgleich der in der Literatur / Debatte identifizierten (potentiell wirksamen) Maßnahmen(bündel) mit den im Programm genannten Maßnahmen(bündeln) je Handlungsansatz.</p>
Identifikation der relevanten Megatrends	<ul style="list-style-type: none"> Werden im Programm die relevanten Megatrends identifiziert? Fehlen wichtige Aspekte? 	<ul style="list-style-type: none"> Werden die Implikationen der Megatrends für nachhaltigen Konsum ausreichend erfasst? 	<p>Abgleich der Megatrends mit den in der Literatur identifizierten/ diskutierten wichtigen Trends und deren Implikationen</p>	<p><i>[wird auf dieser Ebene nicht betrachtet]</i></p>

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
III. Kohärenz				
Externe Kohärenz	<ul style="list-style-type: none"> Inwieweit ist das Programm kohärent mit anderen nachhaltigkeitsorientierten Politiken bzw. Strategien der Bundesregierung (Fokus: deutsche Nachhaltigkeitsstrategie)? 	<ul style="list-style-type: none"> Bestehen Widersprüche (Konflikte) oder Übereinstimmungen (Überlappungen und Synergien) mit den Zielen und Maßnahmen ausgewählter anderer Politiken/Strategien? 	<ul style="list-style-type: none"> Abgleich der Programmziele mit der Deutschen Nachhaltigkeitsstrategie (und damit mittelbar mit den UN Nachhaltigkeitszielen) 	<ul style="list-style-type: none"> Abgleich der BF-Ziele und Themen mit den für das BF relevanten Aspekten der Deutschen Nachhaltigkeitsstrategie (und damit mittelbar mit den UN Nachhaltigkeitszielen)
Interne Kohärenz	<ul style="list-style-type: none"> Inwieweit ist das Programm in sich kohärent? 	<p>Übergreifende Programmebene:</p> <ul style="list-style-type: none"> Inwiefern sind die Querschnittskapitel des Programms in sich stimmig? <p>Bedürfnisfeldebene:</p> <ul style="list-style-type: none"> Inwiefern sind die bedürfnisfeldspezifischen Programmelemente stimmig mit der übergreifenden Programmebene? Inwiefern sind die Programmelemente der einzelnen Bedürfnisfeld-Kapitel untereinander stimmig? Inwiefern sind die bedürfnisfeldspezifischen Programmelemente in sich stimmig? 	<p>Abgleich:</p> <p>Interne Kohärenz innerhalb der Querschnittskapitel:</p> <ul style="list-style-type: none"> Sind die im Programm identifizierten „Querschnitts“-Hemmnisse die relevanten zu überwindenden Hemmnisse, um die übergreifenden Ziele/Leitideen zu erreichen? Nehmen die übergreifenden Ziele/ Leitideen die im Programm identifizierten Megatrends auf? Adressieren die übergreifenden Ziele/ Leitideen bzw. Handlungsansätze die beschriebene internationale Dimension von nachhaltigem Konsum? 	<p>Abgleich:</p> <ul style="list-style-type: none"> Kohärenz zwischen: BF-Kapitel und Hemmnissen; BF-Kapitel und Megatrends; BF-Kapitel und Leitideen; BF-Kapitel und übergreifenden Handlungsansätzen? Interne Kohärenz innerhalb der jeweiligen BF-Kapitel, z.B. zwischen den Abschnitten „Relevanz“ und „Politik“?

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
<p>IV. Effizienz [Nur auf Ebene von Einzelmaßnahmen in der ex post-Analyse geprüft]</p>				
Kosteneffizienz	<ul style="list-style-type: none"> • Wie gut ist das Verhältnis zwischen den Kosten der Maßnahme und den erzielten Wirkungen? • Könnten die Ziele bzw. die erwarteten Wirkungen mit anderen Maßnahmen kostengünstiger erreicht werden? 	<ul style="list-style-type: none"> • Welche Kosten fallen durch die Maßnahme an? • In welchem Umfang werden die Wirkungen, die mit der Maßnahme erzielt werden sollen (Outputs, Outcomes, Impacts), tatsächlich erreicht? • Wurden alternative Ansätze vorgeschlagen/ in Erwägung gezogen, die verglichen werden könnten? 	<ul style="list-style-type: none"> • Sind übergreifende Ziele/ Leitideen und übergreifende Handlungsansätze miteinander stimmig? 	<p>Betrachtung umgesetzter bzw. in Umsetzung befindlicher Maßnahmen: Gegenüberstellung von Informationen zu den Kosten (Interviews, Dokumente) und eigener Einschätzung der erzielten Wirkungen</p>
<p>V. Wirksamkeit bzw. Erfolgsaussichten (i.S.v. „potenzielle Wirksamkeit“)</p>				
Institutionelle Ausgestaltung	<ul style="list-style-type: none"> • Sind Zuständigkeiten für die Umsetzung klar benannt? • Existieren Koordinierungs- und Kontrollmechanismen? • Werden relevante Stakeholder wirksam und sinnvoll eingebunden? 	<ul style="list-style-type: none"> • Inwiefern wird der horizontalen Natur des Programms in seiner institutionellen Verankerung Rechnung getragen? (Funktioniert die horizontale Zusammenarbeit über die Ministerien hinweg?) • Sind die definierten Umsetzungsmechanismen geeignet, die beteiligten Umsetzer zu koordinieren und die Umsetzung zu kontrollieren, gegebenenfalls mangelnde Umsetzung zu 	<p>Experteninterviews</p>	

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
		sanktionieren? Existiert eine Instanz, die den Stand der Umsetzung kontrolliert und im Fall von Umsetzungsproblemen aktiv werden kann (sanktionierend, unterstützend etc.)?		
Finanzielle und personelle Kapazitäten	<ul style="list-style-type: none"> • Welche finanziellen und personellen Kapazitäten sind für die Umsetzung des (Gesamt-) Programmes bzw. der Maßnahme vorgesehen? 		Experteninterviews	
Ziele	<ul style="list-style-type: none"> • Inwieweit bauen Ziele auf unterschiedlichen Abstraktionsebenen (Leit-, Mittler-, Handlungsziele) aufeinander auf? • Sind die Handlungsziele spezifisch, messbar, angemessen & terminiert (vgl. „SMART“-Konzept)? • Sind die Ziele mit Fristen unterlegt? 	<ul style="list-style-type: none"> • Handelt es sich um Qualitäts- oder Handlungsziele, qualitative oder quantifizierte („harte“, „weiche“) Ziele, gegebenenfalls absolute oder relative (Reduktions-)Ziele? Sind die Ziele ambitioniert oder eher schwach (angesichts des Problemdrucks und der gesellschaftlichen Akzeptanz politischer Eingriffe in den fraglichen Bereich)? 	Bewertung der übergreifenden Ziele	
Ex ante: Maßnahmenvorschläge	<ul style="list-style-type: none"> • Wie konkret sind Maßnahmenvorschläge im Allgemeinen ausformuliert? (Können die Erfolgsaussichten der im Programm vorgeschlagenen Maßnahmen bewertet werden?) • Ist die Umsetzung von Maßnahmen im Allgemeinen verbindlich? 	<ul style="list-style-type: none"> • <i>Bewertungsfähigkeit der Maßnahme:</i> Wie konkret sind die im Programm gelisteten Maßnahmen ausformuliert? Inwieweit sind die Ziele der Maßnahmen erkennbar? Werden Instrumente/Umsetzungsmaßnahmen definiert? Werden konkrete 		

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
	<ul style="list-style-type: none"> Sind die Maßnahmen(-bündel) geeignet, die Ziele des Programms bzw. seiner Teilbereiche (übergreifende Handlungsansätze und Bedürfnisfelder) zu erreichen 	<p>Umsetzungsakteure genannt? Werden konkrete Zielgruppen genannt?</p> <ul style="list-style-type: none"> <i>Zielgruppen & Umsetzungsakteure:</i> Adressieren die Maßnahmenbündel die für die Zielerreichung wichtigsten Zielgruppen & strategisch relevanten (Umsetzungs-) Akteure? <i>Instrumententypen:</i> Werden unterschiedliche Instrumententypen zur Umsetzung der Maßnahmen vorgeschlagen? Z.B. kommunikative Instrumente, die Umweltwissen/ Umweltbewusstsein fördern? Auch „härtere“ ordnungsrechtliche und ökonomische Instrumente? 		
Ex post: Wirksamkeit der Maßnahmen(bündel)	<ul style="list-style-type: none"> Wie genau soll die Maßnahme Outcomes und Impacts erzeugen? Ist diese Wirkungslogik plausibel? Wird diese Sicht der Wirkungslogik von unterschiedlichen Stakeholdern geteilt? In welchem Maße wurden die Ziele der Maßnahme erreicht bzw. die erwarteten Outcomes und Impacts erzielt? 			

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
	<ul style="list-style-type: none"> • Warum hat die Maßnahme die beobachtbaren Wirkungen erzielt – was waren fördernde (endogene, exogene) Faktoren? • Warum hat die Maßnahme nicht noch weitere Wirkungen erzielt – was waren hemmende (endogene, exogene) Faktoren? • In welchem Maße wurden nicht-intendierte Wirkungen (positive oder negative) erzielt? 			

VI. Dauerhaftigkeit von Programmwirkungen [nur auf Programmebene; in den BF werden einzelne Aspekte unter „Effektivität“ angesprochen]

Monitoring & Evaluation	<ul style="list-style-type: none"> • Wird der Programmfortschritt regelmäßig erfasst (Monitoring)? • Ist eine systematische und regelmäßige Untersuchung des Nutzens/ der Güte des Programms vorgesehen (Evaluation)? • Existieren (Outcome-, Impact-) Indikatoren für die Erfassung des Programmfortschritts und der Programmgüte? • Existieren „SMARTER“ Ziele und Baselines, um die Programmwirkung messen zu können? • Ex post: Wurden Evaluationsergebnisse in der Weiterentwicklung des Programms systematisch berücksichtigt? 	<ul style="list-style-type: none"> • Indikatoren: qualitativ, quantitativ, valide, reliabel, verfügbar, relevant, finanzierbar? • Daten: Sind Datenquellen, gewünschte Häufigkeit der Datenerhebung und Verantwortlichkeiten für die Datenerhebung benannt? • Evaluation: Transparenz und Unabhängigkeit des Evaluierungsprozesses? 	Experteninterviews
------------------------------------	---	--	--------------------

Kriterium	Leitfragen	Hilfsfragen & Anmerkungen	Vorgehen	
			Übergreifende Programmebene	Bedürfnisfeldebene ¹³
Weiterentwicklungsmechanismen	<ul style="list-style-type: none"> • Ex ante: Ist ein systematischer Prozess zur Weiterentwicklung des Programms vorgesehen? („Review“) • Ex post: Wurde ein systematischer Prozess zur Weiterentwicklung des Programms durchgeführt? („Review“) 	<ul style="list-style-type: none"> • Ist ein systematischer Prozess zur Identifikation neuer „Big Points“ vorgesehen? • Ist das Programm als „lernendes Programm“ angelegt, in dem Ziele, Mechanismen, Maßnahmen etc. im Laufe der Umsetzung angepasst werden können? 	Experteninterviews	
Verstetigungsperspektive	<ul style="list-style-type: none"> • In welchem Maße können Strukturen und Erfolge, die durch das Programm/ Maßnahmen aufgebaut werden, verstetigt werden? • Ist Finanzierung langfrist. gesichert? 			

5. Herausforderungen von Evaluation in Theorie und Praxis

Die praktische Durchführung von Evaluationen ist mit einer Reihe von forschungsimmanenten und praktischen Herausforderungen verknüpft. Einige dieser potentiellen Schwierigkeiten sind allgemeiner Natur und finden sich auch bei der Anwendung sozialwissenschaftlicher Methoden in anderen Kontexten. Dazu gehören beispielsweise eine unzureichende Datengrundlage oder Schwierigkeiten bei der Kausalitätszuschreibung (siehe hierzu z.B. Häder 2015; Kromrey et al. 2016; Schnell et al. 2011).

Andere Herausforderungen resultieren spezifisch aus dem Kontext oder Gegenstand von Evaluationen (Wilhelm 2015, S. 10–11; Stockmann und Meyer 2010, S. 75). Nicht zuletzt ist dies der Fall, weil sich Evaluationen im „Spannungsverhältnis zwischen Wissenschaftlichkeit und Nützlichkeit“ (Stockmann 2007a, S. 29) bewegen. Zu entsprechenden praktischen Herausforderungen gehören:

- Mangelnde Präzision bei der Zielformulierung: Häufig sind die Zielformulierungen der Programme, Projekte oder Maßnahmen unpräzise oder vieldeutig. Dies erschwert es, die Zielerreichung zu bewerten. In einem entsprechenden Fall kann eine „zielfreie“ Bewertung erfolgen, die „relative Verbesserungen“ (d.h. Fortschritte gegenüber einer Baseline) misst, ohne diese auf die gegebenen unpräzisen Ziele zu beziehen (vgl. Kap. 2.6.4);
- Bedeutung strategischer Kalküle und unterschiedlicher Interessen: Bei der Bewertung von Instrumenten, Maßnahmen und Programmen kann die „taktische“ Evaluationsfunktion eine relevante Rolle spielen. Es besteht die Gefahr, dass Akteure die Ergebnisse in ihrem Sinne nutzen, interpretieren und instrumentalisieren. Die verschiedenen beteiligten Akteursgruppen können divergierende Nutzeninteressen hinsichtlich der Evaluationsergebnisse haben. Solchen sozialen, akteursbezogenen Herausforderungen kann durch Einhaltung von Evaluationsstandards (vgl. Kapitel 2.3) begegnet werden, insbesondere durch Gewährleistung von Transparenz und der Einbindung unterschiedlicher Akteure in den Evaluationsprozess. Sie sind Erfolgsvoraussetzungen für das Gelingen der Evaluation und die Legitimität ihrer Ergebnisse.

Wieder andere Herausforderungen verknüpfen sich mit spezifischen Evaluationsgegenständen. So verbinden sich beispielweise mit der Evaluation multizentrischer Programme folgende Schwierigkeiten (Haubrich 2001, S. 1–2):

- Heterogenität der beteiligten und durchführenden Akteure: Wenn die zu einem Programm gehörenden Projekte und Instrumente an verschiedenen Standorten und bei diversen Organisationen angesiedelt sind und dementsprechend unterschiedlich umgesetzt werden, wird die Ableitung allgemeingültiger Aussagen auf Programmebene erschwert. Mit diesem Umstand gilt es zumindest offen umzugehen.
- Konzept versus Implementierung: Teilweise bestehen Programme im Kern nur aus einem politischen Konzept, das erst durch das Zusammenwirken der Programmbeteiligten im Rahmen von Diskussions- und Umsetzungsprozessen konkretisiert und definiert wird. In solchen Fällen sollten Evaluationen als begleitende, formative Evaluationen umgesetzt werden.

Literaturverzeichnis

- Bertelsmann Stiftung (2014): Reformkompass, Zielformulierung: Zielpyramide und SMARTe Ziele. Osnabrück, 2014. Online verfügbar unter https://www.reformkompass.de/uploads/tx_itaio_download/Reformkompass_Werkzeug_Zielformulierung_Beschreibung.docx.
- Bilharz, M. (2008): "Key Points" nachhaltigen Konsums, Ein strukturpolitisch fundierter Strategieansatz für die Nachhaltigkeitskommunikation im Kontext aktivierender Verbraucherpolitik. Marburg: Metropolis.
- BMI (2009): Arbeitshilfe zur Gesetzesfolgenabschätzung. Bundesministerium des Inneren. Berlin, 2009.
- BMUB; BMJV; BMEL (Hg.) (2017): Nationales Programm für Nachhaltigen Konsum, Gesellschaftlicher Wandel durch einen nachhaltigen Lebensstil (2. aktualisierte Auflage). Berlin, 2017.
- Böcher, M.; Töller, A. E. (2007): Instrumentenwahl und Instrumentenwandel in der Umweltpolitik: Ein theoretischer Erklärungsrahmen. In: Jacob, K.; Biermann, F.; Busch, P.-O. und Feindt, P. H. (Hg.): Politik und Umwelt. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 299–322.
- Böcher, M.; Töller, A. E. (2012): Umweltpolitik in Deutschland: eine politikfeldanalytische Einführung (50). Wiesbaden: Springer-Verlag.
- Bortz, J.; Döring, N. (2016): Evaluationsforschung. In: Döring, N.; Bortz, J. und Pöschl, S. (Hg.): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5., vollst. überarb., aktualisierte u. erw. Aufl. Berlin: Springer, S. 975–1036.
- Brandt, D. (2004): Wirkungen situativer Kriminalprävention eine Evaluationsstudie zur Videoüberwachung in der Bundesrepublik Deutschland, Diplomarbeit vorgelegt im Wintersemester 2003/2004 an der Universität Bielefeld Fakultät für Soziologie. Bielefeld, 2004.
- Brulin, G.; Svensson, L. (2012): Managing sustainable development programmes: A learning approach to change: Routledge.
- Bussmann, W.; Klöti, U.; Knoepfel, P. (Hg.) (1997): Einführung in die Politikevaluation: Helbing und Lichtenhahn.
- Campbell, D. (1969): Reforms as Experiments. In: *American Psychologist* 24 (4), S. 409–429.
- Caspari, A.; Barbu, R. (2009): Wirkungsevaluierungen: Zum Stand der internationalen Diskussion und dessen Relevanz für Evaluierungen der deutschen Entwicklungszusammenarbeit. In: *BMZ Evaluation Working Papers*, S. 1–44.
- CeVal (2002): Evaluation der Umweltberatungsprojekte des Bundesumweltministeriums und des Umweltbundesamtes – Nachhaltige Wirkungen der Förderung von Bundesverbänden, Im Auftrag des Umweltbundesamtes. Unter Mitarbeit von Wolfgang Meyer; Klaus-Peter Jacoby und Reinhardt Stockmann (TEXTE, 36/02). Dessau, 2002.
- Chen, H. T. (1990): Theory-driven evaluations. Thousand Oaks, CA: Sage.
- Chen, H. T. (2015): Practical program evaluation, Theory-driven Evaluation and the Integrated Evaluation Perspective. Thousand Oaks, CA: Sage.
- DeGEval (2016): Standards für Evaluation, Erste Revision auf Basis der Fassung 2002, verabschiedet durch die Mitgliederversammlung der DeGEval e.V. am 21. September 2016 (Langfassung). Mainz, 2016.
- DEval (2017): Evaluierung des Aktionsplans des BMZ zur Inklusion von Menschen mit Behinderungen. Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit. Bonn, 2017.
- DEval (2018): Methoden und Standards 2018: Standards für Evaluierungen des DEval. Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit. Bonn, 2018.

- Donaldson, S. I. (2007): Program theory-driven evaluation science: Strategies and applications. New York: Taylor & Francis.
- Döring, N.; Bortz, J.; Pöschl, S. (Hg.) (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften 5., vollst. überarb., aktualisierte u. erw. Aufl. Berlin: Springer.
- Ecolo; Bioconsult (2017): Erfolgsfaktoren für die Förderung zur Anpassung an den Klimawandel. Endbericht, Im Auftrag des Umweltbundesamtes. Unter Mitarbeit von Manfred Born, Lars Galwoschus, Regan Mundhenke, Carolin Scheil und Bastian Schuchardt, S. W. (Climate Change, 11/2017). Dessau, 2017.
- EEA (2016): Environment and climate policy evaluation. European Environment Agency. Copenhagen, 2016.
- EU Joint Evaluation Unit (2006): Evaluation methods for the European Union's External Assistance, Guidelines for Project and Programme Evaluation Volume 3. Directorate General External Relations; Directorate General Development; EuropeAid Co-operation Office. Brussels, 2006.
- EUPOPP (2011): Effects and success factors of sustainable consumption policy instruments: a comparative assessment across Europe. Öko-Institut e.V., NCRC, ECOI, BEF & UCL. Berlin, 2011.
- EuropeAid (2015): Logframe matrix of the project, Annex to Procedures and Practical Guidance (PRAG). Brussels, 2015. Online verfügbar unter <http://ec.europa.eu/europeaid/prag/annexes.do?annexName=E3d&lang=en>.
- European Commission (2013): Evalsed Sourcebook: Methods and technique. DG Regio. Brussels, 2013.
- European Commission (2015): Guidance Document on Monitoring and Evaluation, European Cohesion Fund, European Regional Development Fund. DG Regio. Brussels, 2015.
- European Commission (2016): Commission Staff Working Document: Evaluation of the EU Framework for Metering and Billing of Energy Consumption Accompanying the document Proposal for a Directive of the European Parliament and of the Council amending Directive 2012/27/EU on Energy Efficiency, SWD/2016/0399 final - 2016/0376 (COD). Brussels, 2016.
- European Commission (2017a): Better Regulation Guidelines, Commission Staff Working Paper SWD (2017) 350. Brussels, 2017.
- European Commission (2017b): Better Regulation Toolbox. Brussels, 2017.
- European Parliament (2012): Impact Assessment Handbook, Guidelines for Committees. European Parliament, Conference of Committee Chairs. Brussels, 2012.
- European Parliament (2017): Ex-Post Evaluation in the European Parliament: Method and Process, Working Document. Ex-Post Evaluation Unit, European Parliamentary Research Service, European Parliament. Brussels, 2017.
- Faust, J.; Verspohl, I. (2019): Die DAC-Evaluierungskriterien: zwischen Optimierung und Transformation (DEval-Policy Brief, 9/2018). Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit. Bonn, 2019.
- Fitzpatrick, J. L.; Sanders, J. R.; Worthen, B. R. (2011): Program evaluation, Alternative approaches and practical guidelines 4th ed. Upper Saddle River, N.J., Montreal: Pearson Education.
- Flanagan, A. E.; Tanner, J. C. (2016): Evaluating Behavior Change in International Development Operations, A New Framework (IEG Working Paper, 2016/No. 2). World Bank (Hg.).
- Giel, S. (2013): Theoriebasierte Evaluation, Konzepte und methodische Umsetzungen (Internationale Hochschulschriften, 584). Münster, New York, München, Berlin: Waxmann.
- GIZ (2018a): Das Evaluierungssystem der GIZ, Theorie des Wandels für Evaluierungen der GIZ.

- GIZ (2018b): Das Evaluierungssystem der GIZ, Zentrale Projektevaluierung im BMZ-Geschäft. BMZ (Hg.).
- GIZ (2018c): Die Evaluierungspolicy der GIZ, Prinzipien, Leitlinien und Anforderungen an unsere Evaluierungspraxis.
- Government of Canada (2010): Supporting Effective Evaluations, A Guide to Developing Performance Measurement Strategies. Online verfügbar unter <https://www.canada.ca/en/treasury-board-secretariat/services/audit-evaluation/centre-excellence-evaluation/guide-developing-performance-measurement-strategies.html>, zuletzt geprüft am 17.04.2018.
- Guba, E. G.; Lincoln, Y. S. (2003): Fourth generation evaluation [13. print]. Newbury Park, Calif. [u.a.]: Sage.
- Häder, M. (2015): Empirische Sozialforschung. Wiesbaden: Springer.
- Haubrich, K. (2001): Cluster-Evaluation – ein Modell für einen „dornigen“ Evaluationsgegenstand. DeGEval-Jahrestagung, Block 16: „Modelle der Evaluation“, 2001.
- Haubrich, K. (2006): Die Konstruktion des Untersuchungsgegenstandes in der Evaluation innovativer multizentrischer Programme. In: Rehberg, K.-S. (Hg.): Soziale Ungleichheit, kulturelle Unterschiede. Verhandlungen des 32. Kongresses der Deutschen Gesellschaft für Soziologie in München 2004. Frankfurt, S. 3872–3881.
- Haubrich, K. (2009): Sozialpolitische Innovation ermöglichen, Die Entwicklung der rekonstruktiven Programmtheorie-Evaluation am Beispiel der Modellförderung in der Kinder- und Jugendhilfe: Waxmann Verlag.
- Hense, J.; Kriz, W. C. (2005): Theoriebasierte Evaluation und Bildungscontrolling. In: *Gust, M.*
- HM Treasury (2011): The Magenta Book, Guidance for evaluation.
- Hupfer, B. (2007): Wirkungsorientierte Programmevaluation, Eine Synopse von Ansätzen und Verfahren einschlägiger Institutionen in Deutschland (Schriftenreihe des Bundesinstituts für Berufsbildung Bonn). Bundesinstitut für Berufsbildung. Bonn, 2007.
- IEG (2013): World Bank Group Impact Evaluations, Relevance and Effectiveness. Washington D.C: World Bank.
- Jacob, K.; Guske, A. L.; Weiland, S.; Range, C.; Pestel, N.; Sommer, E. (2016): Verteilungswirkungen umweltpolitischer Maßnahmen und Instrumente, Im Auftrag des Umweltbundesamtes (UBA-Texte, 73/2016). Dessau, 2016.
- Jänicke, M.; Kunig, P.; Stitzel, M. (2003): Lern- und Arbeitsbuch Umweltpolitik, Politik, Recht und Management des Umweltschutzes in Staat und Unternehmen 2., aktualisierte Auflage. Bonn: Verlag J.H.W. Dietz.
- Jordan, A.; Wurzel, Rüdiger K W; Zito, A. R.; Brückner, L. (2003): Policy Innovation or 'Muddling Through'? 'New' Environmental Policy Instruments in the United Kingdom. In: *Environmental Politics* 12 (1), S. 179–200. DOI: 10.1080/09644010412331308344a.
- King, G.; Keohane, R. O.; Verba, S. (1994): Designing Social Inquiry: Scientific inference in qualitative research. Princeton, NJ: Princeton University Press.
- Knoepfel, P.; Varone, F.; Bussmann, W.; Mader, L. (1997): Evaluationsgegenstände und Evaluationskriterien. In: Bussmann, W.; Klöti, U. und Knoepfel, P. (Hg.): Einführung in die Politikevaluation: Helbing und Lichtenhahn, S. 78–118.
- Kromrey, H. (2001): Evaluation - ein vielschichtiges Konzept: Begriff und Methodik von Evaluierung und Evaluationsforschung; Empfehlungen für die Praxis. In: *Sozialwissenschaften und Berufspraxis* 24 (2), S. 105–131.
- Kromrey, H.; Strübing, J.; Roose, J. (2016): Empirische Sozialforschung, Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung mit Annotationen aus qualitativ-

interpretativer Perspektive 13., völlig überarbeitete Auflage (UTB). Konstanz: UVK Verlagsgesellschaft; München; UVK/Lucius.

- Leeuw, F. L. (2003): Reconstructing Program Theories: Methods Available and Problems to be Solved. In: *American Journal of Evaluation* 24 (1), S. 5–20. DOI: 10.1177/109821400302400102.
- Leeuw, F. L.; Vaessen, J. (2009): Impact evaluations and development: NONIE guidance on impact evaluation: Network of networks on impact evaluation.
- Lucke, D. (2003): Akzeptanz. In: Schäfers, B. (Hg.): *Grundbegriffe der Soziologie*. Opladen: Leske & Budrich, S. 5–9.
- Mayne, J. (2015): Useful Theory of Change Models. In: *Canadian Journal of Program Evaluation* 30 (2).
- Meyer, W. (2002): Was ist Evaluation? (CEval-Arbeitspapiere, 5). Centrum für Evaluation. Saarbrücken, 2002.
- Meyer, W.; Stockmann, R. (2014): Evaluation Approaches and Their Fundamental Theoretical Principles. In: Stockmann, R. und Meyer, W. (Hg.): *Functions, methods and concepts in evaluation research*. Houndmills: Palgrave Macmillan, S. 108–174.
- Muster, V.; Griebshammer, R.; Wolff, F.; Thorun, C.; Schrader, U.; Reisch, L. (2019): Nachhaltigen Konsum weiterdenken: Evaluation und Weiterentwicklung von Maßnahmen und Instrumenten, Ex-post Evaluation ausgewählter Maßnahmen. Umweltbundesamt (Hg.). Berlin, 2019.
- Muster, V.; Thorun, C.; Diels, J.; Schrader, U.; Wolff, F.; Kampffmeyer, N.; Griebshammer, R.; Reisch, L. (2018): Ex-ante Evaluation des Nationalen Programms für Nachhaltigen Konsum. Umweltbundesamt. unveröffentlichter Bericht, 2018.
- OECD (2002): Development Assistance Committee Working Party on Aid Evaluation, Glossary of Key terms in Evaluation and Results Based Management. Organisation for Economic Cooperation and Development. Paris, 2002. Online verfügbar unter <http://www.oecd.org/dataoecd/29/21/2754804.pdf>.
- OECD (2009a): Glossar entwicklungspolitischer Schlüsselbegriffe aus den Bereichen Evaluierung und ergebnisorientiertes Management. Arbeitsgruppe Evaluierung der Entwicklungszusammenarbeit des OECD-Entwicklungsausschusses. Paris, 2009.
- OECD (2009b): Regulatory Impact Analysis: A Tool for Policy Coherence, Organisation for Economic Cooperation and Development. Paris.
- OECD (2010a): DAC Guidelines and Reference Series, Quality Standards for Development Evaluation. OECD Development Assistance Committee (Hg.). Paris, 2010.
- OECD (2010b): Evaluating Development Co-Operation, Summary of Key Norms and Standards Second Edition. OECD DAC Network on Development Evaluation (Hg.).
- OECD DAC (2010): Quality Standards for Development, DAC Guidelines and Reference Series. Organisation for Economic Cooperation and Development (Hg.). Paris, 2010.
- Öko-Institut; ifeu; FFU; Hochschule Karlsruhe; Prognos; Ziesing, H.-J.; Klinski, S. (2017): Evaluierung der Nationalen Klimaschutzinitiative, Evaluierungszeitraum 2012-2014. Berlin, 2017. Online verfügbar unter <https://www.klimaschutz.de/sites/default/files/Gesamtbericht%20NKI-Evaluation%202012-2014.pdf>.
- Pawson, R.; Greenhalgh, T.; Harvey, G.; Walshe, K. (2004): *Realist synthesis: an introduction*. Manchester, 2004.
- Pawson, R.; Tilley, N. (1997): *Realistic evaluation*: Sage.
- Pawson, R.; Tilley, N. (2005): *Realistic Evaluation*. In: Mathison, S. (Hg.): *Encyclopedia of evaluation*: Sage, S. 362–367.

- Rogers, P. J.; Weiss, C. H. (2007): Theory-based evaluation: Reflections ten years on: Theory-based evaluation: Past, present, and future. In: *New Directions for Evaluation*; 2007 (114), S. 63–81.
- Rossi, P. H.; Freeman, H. E. (1979): *Evaluation: A systematic approach*: Sage.
- Rossi, P. H.; Freeman, H. E.; Hofmann, G. (1988): *Programm Evaluation, Einführung in die Methoden angewandter Sozialforschung*. Stuttgart: Enke.
- Sabatier, P. A. (1987): Knowledge, Policy-Oriented Learning, and Policy Change: An Advocacy Coalition Framework. In: *Science Communication* 8 (4), S. 649–692.
- Scharpf, F. W. (1973): *Politische Durchsetzbarkeit innerer Reformen im pluralistischdemokratischen Gemeinwesen der Bundesrepublik*. International Institute of Management (Hg.). Berlin, 1973.
- Scharpf, F. W.; Reissert, B.; Schnabel, F. (Hg.) (1976): *Politikverflechtung, Theorie und Empirie des kooperativen Föderalismus in der Bundesrepublik*: Cornelsen Verlag.
- Schmitt, J. (2018): *Schluss mit Schwarzen Boxen. Zur Arbeit mit Kausalmechanismen in Evaluierungen (DEval-Policy Brief, 10(2018))*. Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit. Bonn, 2018.
- Schnell, R.; Hill, P. B.; Esser, E. (2011): *Methoden der empirischen Sozialforschung*. München: Oldenbourg Verlag.
- Schumacher, K.; Cludius, J.; Förster, H. (2016): Energy efficiency vs. renewable energy policies within the German Energiewende – What are the distributional implications for households? (Conference proceedings International Energy Policy and Programme Evaluation Conference, Amsterdam 2016 (www.ieppeec.org)), 2016. Online verfügbar unter <http://www.ieppeec.org/wp-content/uploads/2016/05/Paper-Schumacher-1.pdf>, zuletzt geprüft am 04.05.2017.
- SCOPE2 (2008): *Sustainable Consumption Policies Effectiveness Evaluation - Inventory and assessment of policy instruments*, Deliverable number: D1b-WP1. Köln, Wien, Helsinki, 2008.
- Scriven, M. (1972): Pros and cons about goal-free evaluation. In: *Evaluation comment* 3 (4), S. 1–4.
- Scriven, M. (1991): *Evaluation thesaurus 4th ed.* Newbury Park, Calif: Sage. Online verfügbar unter <http://www.loc.gov/catdir/enhancements/fy0655/91009264-d.html>.
- Shadish, W. R.; Cook, T. D.; Leviton, L. C. (1991): *Foundations of program evaluation, Theories of practice 1st printing*. Newbury Park (California), London: Sage.
- Silvestrini, S.; Reade, N. (2008): *CEval-Ansatz zur Wirkungsevaluation / Stockmann'scher Ansatz*. Centrum für Evaluation. Saarbrücken, 2008.
- Stame, N. (2004): Theory-based evaluation and types of complexity. In: *Evaluation* 10 (1), S. 58–76.
- Starke, P. (2015): Prozessanalyse. In: Wenzelburger, G. und Zohlnhöfer, R. (Hg.): *Handbuch Policy-Forschung*. Wiesbaden: Springer VS (Springer VS Handbuch), S. 453–482.
- Stockmann, R. (2004): *Was ist eine gute Evaluation?, Einführung zu Funktionen und Methoden von Evaluationsverfahren*.
- Stockmann, R. (2006a): *Evaluation und Qualitätsentwicklung: eine Grundlage für wirkungsorientiertes Qualitätsmanagement*: Waxmann Verlag.
- Stockmann, R. (2006b): *Evaluation in Deutschland*. In: Stockmann, R. (Hg.): *Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder*. 3. Aufl. Münster, New York, München, Wien: Waxmann (Sozialwissenschaftliche Evaluationsforschung, Bd. 1), S. 15–46.
- Stockmann, R. (2007a): *Einführung in die Evaluation*. In: Stockmann, R. (Hg.): *Handbuch zur Evaluation: eine praktische Handlungsanleitung*. Münster: Waxmann, S. 24–70.

- Stockmann, R. (Hg.) (2006c): Evaluationsforschung, Grundlagen und ausgewählte Forschungsfelder 3. Aufl. (Sozialwissenschaftliche Evaluationsforschung, Bd. 1). Münster, New York, München, Wien: Waxmann.
- Stockmann, R. (Hg.) (2007b): Handbuch zur Evaluation: eine praktische Handlungsanleitung. Münster: Waxmann.
- Stockmann, R.; Meyer, W. (2010): Evaluation, Eine Einführung (UTB, 8337). Opladen: B. Budrich.
- Thomas D., C.; G. E., M. (1990): Theorien der Programmevaluation: Ein kurzer Abriss. In: Koch, U. und Wittmann, W. (Hg.): Evaluation-Bewertungsgrundlage von Sozial- und Gesundheitsprogrammen.
- Treasury Board of Canada (2012): Theory-based approaches to evaluation, Concepts and practices, 2012. Online verfügbar unter <https://www.tbs-sct.gc.ca/hgw-cgf/oversight-surveillance/ae-ve/cee/tbae-aeat/tbae-aeat-eng.pdf>.
- UNEP (2010): Terminal evaluation of the UNEP project: Supporting the 10 Year Framework Program for Africa on Sustainable Consumption and Production, 2007-2008. UNEP Evaluation Unit. Nairobi, 2010.
- UNEP (2016): United Nations Environment Programme Evaluation Policy. Nairobi, 2016.
- UNEP EOU (2008): Evaluation Manual. United Nations Environment Programme Evaluation and Oversight Unit. Nairobi, 2008.
- UNEP EOU (2017): Terminal Evaluation of the UN Environment Project: "Policy, macro-economic assessments and instruments to empower governments and business to advance resource efficiency and move towards a Green Economy" (61-P3). UNEP Evaluation Office. Nairobi, 2017.
- Weiss, C. H. (1974): Evaluierungsforschung, Methoden zur Einschätzung von sozialen Reformprogrammen. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Weiss, C. H. (1997): Theory-based evaluation: Past, present, and future. In: *New Directions for Evaluation*; (76), S. 41–55.
- Wilhelm, J. L. (2015): Evaluation komplexer Systeme: Ein Feld mit Fragezeichen. In: Wilhelm, J. L. (Hg.): Evaluation komplexer Systeme: Systemische Evaluationsansätze in der deutschen Entwicklungszusammenarbeit: Universitätsverlag Potsdam, S. 7–32.
- Wolff, F.; Jacob, K.; Guske, A. L.; Heyen, D. A.; Hüsing, T. (2016): Kohärenzprüfung umweltpolitischer Ziele und Instrumente. Endbericht (UBA-Texte, 76/2016). Umweltbundesamt, 2016.
- World Bank (2004): Monitoring & Evaluation, Some Tools, Methods & Approaches. Washington D.C., 2004.
- World Bank (2012): Impact Evaluation Toolkit, Measuring the Impact of Results-Based Financing on Maternal and Child Health. Unter Mitarbeit von Vermeersch, C. M. J.; Rothenbühler, E. und Renee Sturdy, J.
- World Bank (2016): Impact Evaluation in Practice. Unter Mitarbeit von Gertler, P. J.; Martinez, S.; Premand, P.; Rawlings, L. B. und Vermeersch, C. M. J. Second Edition.
- Wuppertal Institut für Klima, Umwelt und Energie; Ecofys (2014): Wirkungsanalyse bestehender Klimaschutzmaßnahmen und -programme sowie Identifizierung möglicher weiterer Maßnahmen eines Energie- und Klimaschutzprogramms der Bundesregierung. UBA (Hg.).