

Voraussetzungen für erfolgreiche Wirkungsuntersuchungen in Evaluationen

Ergebnisse der Frühjahrstagung 2024 des AK Methoden in der Evaluation

Franziska Heinze¹ und Alexander Kocks²

1. Anlass

Wirkungsuntersuchungen stehen im Zentrum vieler Evaluationen. Sie sollen beispielsweise den kausalen Nachweis erbringen, ob die zu evaluierenden Maßnahmen (z. B. Programme und Projekte) effektiv und wirksam sind, oder aufklären, welche Wirkmechanismen wie dazu beitragen, dass Wirkungen entfaltet werden. Dem Wirkungsbegriff eingeschrieben ist dabei die Annahme einer Kausalbeziehung zwischen (mindestens) einer Ursache und einer Wirkung. Dieser kausale Zusammenhang lässt sich u. a. entlang des jeweils zugrundeliegenden Kausalitätskonzepts näher bestimmen oder kann unter Verwendung von konkreten Theorien des Evaluationsgegenstandes plausibilisiert werden (Reichardt, 2022).

Wirkungsuntersuchungen sind abhängig vom konkreten Erkenntnisinteresse der beteiligten Stakeholder:innen, den vorgefundenen Bedingungen und Eigenheiten des Evaluandums sowie der für Wirkungsuntersuchungen bereitgestellten Ressourcen. Das vor diesem Hintergrund gewählte Wirkungsevaluationsdesign muss sowohl das Problem kausaler bzw. plausibler Ursache-Wirkungs-Zusammenhänge methodologisch adressieren und methodisch bearbeiten als auch (theoriegeleitet) relevante Wirkdimensionen und Einflussfaktoren bestimmen können. Für die Untersuchung von Wirkungen steht den Evaluierenden grundsätzlich ein breites Set an Methoden der empirischen Sozialforschung und weiterer Bezugsdisziplinen (z. B. der Ökonometrie) zur Verfügung: Rigorose quantitative Kausalmissverfahren wie randomisierte Kontrollgruppendesigns (RCTs) und quasi-experimentelle Methoden können dabei ebenso zum Einsatz kommen wie qualitative Methoden zur Eruiierung von Wirkmechanismen, zur Aufklärung von Wirkzusammenhängen oder zur Erhebung von plausiblen Wirkungen (z. B. Process Tracing, Kontributionsanalyse oder Outcome Harvesting).

1 Deutsches Jugendinstitut (DJI), Außenstelle Halle (Saale); Sprecherin des AK Methoden

2 Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval); Sprecher des AK Methoden

Mit den hier nur beispielhaft angeführten unterschiedlichen Methoden gehen jeweils spezifische Vor- und Nachteile einher: Sie befriedigen verschiedene Erkenntnisinteressen (z. B. Wirkungsfeststellung/-messung, Aufzeigen/Verstehen von Ursache-Wirkungs-Zusammenhängen oder Wirkweisen). Einige Methoden eignen sich im Sinne des verstehenden Erklärens nach Weber (1972), um theoriegeleitet zu erklären, warum etwas wirkt oder nicht wirkt. Manche Methoden erlauben eine höhere Generalisierbarkeit der Ergebnisse, andere sind hinsichtlich ihrer Aussagekraft auf das untersuchte Sample beschränkt. Zudem können verschiedene Methoden im Rahmen von Mixed-Methods-Designs und Multi-Methods-Designs miteinander kombiniert werden, um ihre jeweiligen methodischen Stärken umfassend in Wert zu stellen bzw. um sowohl fallübergreifende als auch fallspezifische (vertiefende) Evidenz im Sinne des Erkenntnisinteresses zu generieren.

Mit Blick auf Wirkungsuntersuchungen lassen sich gegenwärtig zwei Tendenzen beobachten: Auf der einen Seite sehen wir eine deutliche Professionalisierung in den letzten Jahren, was den Einsatz verschiedenster Methoden der Wirkungsuntersuchung im Rahmen von Evaluationen betrifft. Auf der anderen Seite konstatieren wir aber auch eine gewisse Unübersichtlichkeit – mithin Verunsicherung –, welche Evaluationsdesigns und -methoden angesichts der steigenden Ansprüche an und Herausforderungen von Evaluationen und Evaluationsgegenständen, nicht zuletzt im Kontext der Legitimationsfunktion von Evaluation, der Berücksichtigung allgemeiner Evaluationsstandards und vor dem Hintergrund begrenzter Ressourcen, die „richtigen“ bzw. die „gegenstandsangemessenen“ sind. Dies bezieht sich nicht nur, aber z. B. auch auf Evaluationen von komplexen (Mehrebenen-)Programmen oder in fragilen, volatilen Konflikt-Kontexten, welche den möglichen Methodeneinsatz restringieren bzw. nach kontextrobusten oder adaptiven Methodendesigns verlangen.

Verschiedenartige Untersuchungs- bzw. Evaluationsgegenstände sowie -fragestellungen verlangen unterschiedliche, auf sie angepasste Evaluationsdesigns. Dabei ist das jeweilige Erkenntnisinteresse der Stakeholder:innen einer Evaluation in Bezug auf Programmwirkungen zu berücksichtigen (z. B. Wirkungsnachweis vs. Erklären (nicht-)intendierter Wirkungen sowie Lernerfahrungen zu „gescheiterten“ Wirkungsannahmen, Rekonstruktion von Wirkmechanismen, Ex-post-Wirkungsnachweis vs. Ex-ante-Wirkungsabschätzung u. v. m.) (Bischoff et al., 2021). Die Wahl des jeweils „richtigen“ Evaluations- bzw. Methodendesigns, das für die Wirkungsuntersuchung herangezogen wird, muss zudem die Kontextbedingungen und die Eigenheiten des jeweiligen Evaluationsgegenstandes berücksichtigen. Beispielsweise ist zu klären, ob bzw. welche offenen Fragen hinsichtlich der Wirksamkeit eines Programms bestehen, ob ein Programm bereits „reif“ und elaboriert ist oder noch entwickelt und erprobt wird. In diesem Zusammenhang ist auch entscheidend, inwiefern bereits in der Programmplanung eine fundierte Wirklogik zugrunde gelegt wurde und die Bedingungen für kausale Wirkungsmessungen gegeben sind, oder ob – wie in multizentrischen Programmen häufig vorzufinden – viele unterschiedliche Wirklogiken nur allgemein vorgegebene Programmziele konkretisieren. Zu berücksichtigen sind dabei auch gegenstandstheoretische Annahmen, d. h., wie eigentlich der zu untersuchende Gegenstand (z. B. das Programm, die Maßnahme) „beschaffen“ ist und welche Implikationen dies für die Untersuchung von Wirkungen hat. Darüber hinaus sind nicht zuletzt auch ethische Fragen zu bedenken, beispielsweise ob der Nutzen einer Evaluation die Kos-

ten für die Beteiligten bzw. Betroffenen aufwiegt oder ob Wirkungsuntersuchungen in ethisch vertretbarer Weise umgesetzt werden können. Die vor diesem Hintergrund zu treffenden Entscheidungen zur methodischen Umsetzung von Wirkungsuntersuchungen können sich daher weder nach wissenschaftstheoretischen und methodologischen Dogmen noch allein nach den methodischen Kompetenzen der Evaluierenden richten. Entscheidend ist zuvorderst, ob das gewählte Design dem Erkenntnisinteresse und dem Gegenstand angemessen ist.

Vor diesem Hintergrund hat der AK Methoden seine Frühjahrstagung am 6. und 7. Juni 2024 an der Universität des Saarlandes dem Thema „*Wissen was wirkt? Voraussetzungen für erfolgreiche Wirkungsuntersuchungen in Evaluationen*“ gewidmet. Die Tagung wurde in Zusammenarbeit mit dem Weiterbildungsstudiengang Master Evaluation an der Universität des Saarlandes in Kooperation mit der Hochschule für Technik und Wirtschaft des Saarlandes (htw saar) realisiert. Ziel der Tagung war es, Voraussetzungen für erfolgreiche Wirkungsevaluationen auf verschiedenen Ebenen zusammenzutragen, unterschiedliche methodische Herangehensweisen und Erfahrungen von Wirkungsuntersuchungen zu diskutieren sowie *good practices* im Umgang mit den Herausforderungen von Wirkungsevaluationen gemeinsam herauszuarbeiten. Zur Strukturierung der Tagung wurde hierbei idealtypisch zwischen drei Ebenen von Voraussetzungen unterschieden, die mit verschiedenen diskussionsanleitenden Fragen verbunden waren:

Stakeholder:innenbezogene Voraussetzungen: Wie kommen wir zu geteilten Standards und einheitlichen Verständnissen unter den Stakeholder:innen darüber, welche Art der Wirkungsuntersuchung für die jeweilige Evaluation angemessen ist? Wie lassen sich Perspektiven von Stakeholder:innen einfangen, beispielsweise, wenn Beratungen mit Stakeholder:innen im Vorfeld nicht möglich sind?

Gegenstands- und kontextbezogene Voraussetzungen: Wann können und sollten wir mit welchen Evaluationsansätzen, -designs und -methoden Wirkungen erfassen? Wie leiten wir aus Evaluationsgegenstand, -fragen und -kontext systematisch ein Design für die Wirkungsuntersuchung ab? Wie können wir ein solches Design (z. B. mit Blick auf Evaluation in fragilen Kontexten) möglichst adaptiv oder kontextrobust gestalten? Unter welchen Bedingungen können welche Methoden zum Einsatz kommen? Was ist das „richtige“ Maß zwischen Erkenntnisnutzen und methodischem Anspruch auf der einen Seite und Machbarkeit (Ressourcen, Timing, Kontext) auf der anderen Seite? Wie identifizieren wir vorhandene Evidenzen und Evidenzlücken für die richtige Methodenwahl?

Methodische Voraussetzungen: Welche methodischen Voraussetzungen braucht es, um wirkungsorientierte Evaluationsdesigns umzusetzen? Was sind dabei wichtige Anwendungsvoraussetzungen, Mindeststandards und Gütekriterien? Welche Fehler und Biases treten bei der Anwendung bestimmter Methoden auf und wie kann man diesen begegnen? Wie gehen wir in methodenintegrierten Designs damit um, wenn verschiedene Methoden zu unterschiedlichen Ergebnissen führen?

2. Fünf Impulsvorträge

In den fünf Impulsvorträgen der Tagung konnten viele dieser Fragen – mit unterschiedlicher Schwerpunktsetzung und anhand unterschiedlicher empirischer Anwendungsbeispiele – adressiert werden.

Marie Gaarder, Direktorin der International Initiative for Impact Evaluation (3ie) in Washington DC/Delhi/London stellte in ihrer Keynote, „*Unveiling the Dynamics of Evidence Utilization: Lessons from 3ie’s ‘Balloon-Squeezing’-Approach*“, die implizite Theorie des Wandels für die Nutzung von Evidenz in der internationalen Entwicklung in den letzten 15 Jahren aus der Perspektive der Erfahrungen von 3ie vor. Indem sie die Voraussetzungen für erfolgreiche Wirkungsevaluationen nicht nur auf der Durchführungsebene, sondern auch auf Ebene von deren Nutzung seitens der Auftraggebenden und im politischen Raum verortet, ergänzte sie das Tagungsthema um eine wichtige Perspektive und adressierte somit dessen Fluchtpunkt: Der Erfolg einer Wirkungsevaluation bemisst sich nicht nur am erfolgreichen Umgang mit Erfolgshemmnissen auf der Planungs- und Durchführungsebene, sondern auch mit Hemmnissen auf der Nutzungsebene. In ihrem Vortrag legte sie dar, wie 3ie seine ursprüngliche Konzentration auf die Produktion immer neuer Erkenntnisse in Form von Wirkungsevaluationen in verschiedenen Entwicklungssektoren weiterentwickelt hat und sich nunmehr stärker auf Qualität und Relevanz konzentriert sowie die Evaluation an die Art der politischen Fragen anpasst. Damit verbunden zeigte sie, wie es der Organisation gelungen ist, die Fülle der zunehmend verfügbaren Evidenz sinnvoll zu nutzen – in Form eines gut kuratierten und einfachen Zugangs zu relevanter Evidenz über das „3ie Development Evidence Portal“.³ Sodann stellte sie in ihrem Vortrag die jüngste Strategie von 3ie dar, in der sich die Organisation darauf fokussiert, bestehende Hindernisse für die Nutzung von Evidenz in internationalen Entwicklungsinstitutionen zu beseitigen. Im Kern beinhaltet diese Strategie drei zentrale Maßnahmen: die Förderung langfristiger Partnerschaften, den Aufbau lokaler Kapazitäten und die Verbesserung einer institutionellen Evidenzkultur. Zur Stärkung letzterer präsentierte sie den „TRIPS framework“ von 3ie, der fünf Maßnahmen umfasst: (1) Trainings zur besseren interorganisationalen Nutzung von Evidenz während des gesamten Programmzyklus, (2) Bereitstellung angemessener Ressourcen, um die Erhebung und Verwendung geeigneter Daten und Evidenz sicherzustellen, (3) Schaffung von Anreizen in Organisationen, um herkömmliche Annahmen in Frage zu stellen und verfügbare Evidenz zu nutzen, (4) Etablierung von Prozessen, die sicherstellen, dass Analysen und Evidenz kontinuierlich in die Programmentwicklung einfließen und (5) Klare Signale von der Führung an die Belegschaft, die auf die Relevanz des interorganisationalen Lernens und die Nutzung (evaluativer) Evidenz abzielen.⁴ Am Ende ihres Vortrags führte Marie Gaarder eine Reihe von Beispielen an, in denen Organisationen wie USAID und die Weltbank diese fünf Maßnahmenbereiche schon mit konkreten Umsetzungsschritten hinterlegt haben.

3 Das Portal bietet Zugang zu den Befunden aus aktuell 14.447 Wirkungsevaluationen, 1.206 Systematic Reviews und 38 Evidence Gaps Maps. Siehe: <https://development.evidence.3ieimpact.org/>

4 Für eine ausführliche Darstellung des TRIPS frameworks siehe: <https://www.3ieimpact.org/sites/default/files/2024-05/TRIPS-guidance-note.pdf>

Stefan Silvestrini, vom CEval – Centrum für Evaluation GmbH in Saarbrücken, ging in seinem Input auf „*Erfolgsfaktoren für (Wirkungs-)Evaluationen – Worauf es wirklich ankommt*“ ein. Ausgangspunkt des Vortrags bildete eine Operationalisierung zur Frage, was gute Evaluationen ausmache. Diese lieferte das Grundgerüst für die Auswertung von zehn Meta-Evaluationen von insgesamt rund 400 Evaluationen. Im Rahmen dieser Auswertung wurden verschiedene Datenquellen (z. B. Evaluationsberichte, Leistungsbeschreibungen) entlang ausgewählter Bewertungsdimensionen (u. a. Qualität der Leistungsbeschreibung, der Methodik und Bewertung der Nützlichkeit der Evaluation) und Einflussfaktoren (z. B. Typ/Zeitpunkt, Sektor, Budget) untersucht. Im Fokus des Vortrags standen die methodische Qualität sowie die, seitens der Adressat:innen wahrgenommene, Nützlichkeit der daraus hervorgehenden Ergebnisse, die zwei wesentliche Faktoren erfolgreicher Wirkungsevaluationen darstellen. Nach einem Einblick in konkrete Operationalisierungen der Bewertungsdimensionen berichtete Stefan Silvestrini einige Ergebnisse zu stakeholder:innen-, gegenstands- und kontextbezogenen sowie methodischen Faktoren, die den Erfolg von Evaluationen beeinflussen. Die Studie stellte heraus, dass beispielsweise der Zeitpunkt einer Evaluation oder das dafür verfügbare Budget als Kontextfaktoren sowohl die methodische Qualität der Evaluation als auch (in etwas geringerem Maße) die wahrgenommene Nützlichkeit einer Evaluation beeinflussen. Als Einflussfaktor seitens der Stakeholder:innen bestimmt vor allem die Qualität der Leistungsbeschreibung die methodische Qualität von Evaluationen mit. Ein Zusammenhang zwischen der methodischen Qualität einer Evaluation und der wahrgenommenen Nützlichkeit einer Evaluation konnte hingegen nicht gefunden werden. Diese Ergebnisse erläuterte Stefan Silvestrini anschließend anhand zweier Fallbeispiele aus seiner eigenen Arbeitspraxis und verdeutlichte daran, wodurch sich eine erfolgreiche von einer weniger erfolgreichen Evaluation unterscheidet. Hierfür stellte er Erfolge und Erfolgsfaktoren auf der einen Seite (z. B. Interesse an Lernen aus Evaluation, proaktive Unterstützung und Beteiligung des Managements, Vertrauen) den Misserfolgen und hinderlichen Faktoren (z. B. Evaluation als Rechenschaftslegung, geringes Vertrauen, Kommunikationsdefizite) auf der anderen Seite gegenüber. Insbesondere dieser Vergleich und die Einblicke in eine gescheiterte Evaluation verdeutlichten im Zusammenspiel mit den Erkenntnissen aus den Meta-Evaluationen, dass Erfolgsfaktoren gelingender Wirkungsevaluationen weit über Fragen der Methodik und des Designs hinausgehen.

Den dritten Input des ersten Tages trug *Katharina Kaepfel* vom Abdul Latif Jameel Poverty Action Lab (J-PAL) Europe in Paris bei. In ihrem Vortrag „*Rigoreuse Wirkungsevaluierung – Methodische Voraussetzungen und kontextsensitive Gestaltung*“ thematisierte Katharina Kaepfel methodische und weitere Voraussetzungen für rigorose, v. a. randomisierte Wirkungsevaluationen. Zunächst erläuterte sie das kontrafaktische Wirkungsverständnis dieses Ansatzes als den Effekt eines Programms zu einem bestimmten Zeitpunkt, der im Vergleich zu den ohne das Programm entstandenen Effekten zum selben Zeitpunkt zu bestimmen ist, damit eine Ursache-Wirkungs-Beziehung (kausaler Zusammenhang) zwischen Programm und Programmwirkung gemessen werden kann. Da in der Regel nicht beobachtet werden kann, welche Entwicklungen sich ohne das Programm vollziehen, bedarf es demzufolge zur Lösung dieses Prob-

lems der Simulation des Kontrafaktischen durch Kontrollgruppen. Sie stellte kurz dar, wie quasi-experimentelle (in der Diktion von J-PAL: „nicht-experimentelle Designs“) ex-post versuchen, Kontrollgruppen nachzubilden bzw. Vergleiche zu ermöglichen. Anschließend erläuterte sie die Ex-ante-Konstruktion von Kontrollgruppen in Randomized Controlled Trials (RCTs) und stellte exemplarisch dar, welche Fragen zu kausalen Zusammenhängen entsprechende Designs in Abhängigkeit von verschiedenen Erkenntnisinteressen beantworten können. Am Beispiel von Wirkungsuntersuchungen zu Unterstützungsleistungen (Geld- vs. Sachleistungen) stellte sie anschließend dar, wie konkrete Kontrollgruppendesigns realisiert wurden und thematisierte auch den Umgang mit damit verbundenen ethischen Fragen. Im Ergebnis der Studien konnten unterschiedliche Effektstärken der jeweiligen Programme festgestellt werden. Diese gemessenen Effekte wurden zusätzlich in eine Kosten-Effektivitäts-Analyse eingespeist. Hieran illustrierte Katharina Kaepfel den Zusammenhang zwischen Erkenntnisinteresse und Aussagekraft von RCTs im Hinblick auf die untersuchten Kontexte: Werden weitere Aussagen (hier: die Effektstärke pro investierten 100\$ Budget) herangezogen, war nicht mehr das Programm mit den höchsten Wirkungseffektstärken als besonders erfolgreich zu klassifizieren, sondern jenes, das unter Kosten-Nutzen-Gesichtspunkten die bestmöglichen Effektstärken aufwies. Abschließend thematisierte Katharina Kaepfel notwendige Voraussetzungen für randomisierte Wirkungsevaluationen. Jene bauen u. a. auf einer Ausgangsanalyse des zu bearbeitenden Problems sowie einer guten Programmplanung und -durchführung auf, beziehen Kenntnisse zur Programmimplementierung und -umsetzung ein und können weiterführende Untersuchungen wie Kosten-Effektivitäts-Analysen enthalten. Zum Schluss resümierte die Referentin die Bedingungen, unter denen eine rigorose Wirkungsevaluation sinnvoll und nützlich sein kann: Dies sei u. a. dann besonders gegeben, wenn ein Programm bereits elaboriert und damit „bereit“ für eine Überprüfung ist, wenn offene Fragen zur Wirksamkeit des Programms bestehen und wenn es randomisierbare Programmelemente enthält, die mit einem entsprechenden Design auf ihre Effekte hin evaluiert werden können.

Den zweiten Tagungstag eröffnete *Christina Kaps* von Camino – Werkstatt für Fortbildung, Praxisbegleitung und Forschung im sozialen Bereich gGmbH in Berlin mit einem Input über die „*Kontextsensible Rekonstruktion von Wirkmechanismen – Möglichkeiten und Grenzen qualitativ vergleichender Analysen (QCA) in Evaluationen*“. Zunächst stellte die Referentin die Bedeutung des Kontextes für die Umsetzung von Projekten heraus. Jener sei, ebenso wie die spezifischen Umsetzungsstrategien von Projekten, in Wirkungsevaluationen zu berücksichtigen, ohne dabei den Anspruch eines systematischen Vergleichs von Projekten und der Ableitung übertragbarer Aussagen aufzugeben. Hierfür eigne sich die Qualitative Comparative Analysis, mit der sich theoriebasiert vergleichende inhaltliche Analysen von Datensätzen bereits mit kleineren und mittleren Fallzahlen durchführen lassen. Die Methode dient dazu, Erfolgspfade zur Erreichung eines erwünschten Zustands in einer vorab definierten Zielgröße zu identifizieren. Zur Veranschaulichung des Vorgehens stellte Christina Kaps die zuvor kurz erläuterte Vorgehensweise am Beispiel der Untersuchung von Öffentlichkeitswirksamkeit von kommunalen Netzwerken, sogenannten Partnerschaften für De-

mokratie im Bundesprogramm „Demokratie leben!“, vor. Die bei einem qualitativen Sample durchgeführte Untersuchung analysierte, inwieweit die Maßnahmen der kommunalen Partnerschaften für Demokratie zur Erreichung von Öffentlichkeit erfolgreich sind und was sie bewirken. Neben der Frage, ob Öffentlichkeit erreicht wird, interessierte, mit welchen Strategien und Maßnahmen(-kombinationen) dies besonders gut gelingt. Mittels explorativer Interviews und deren inhaltsanalytischer Auswertung wurden zunächst Kriterien für das Ziel „Erreichung von Öffentlichkeit“ erarbeitet. Im zweiten Schritt wurden leitfadengestützte Interviews eingesetzt und inhaltsanalytisch ausgewertet, um auf Einzelfallebene die Zielerreichung (entlang der vorab bestimmten Kriterien) und die Kontextfaktoren (Bedingungen) zu ermitteln. Jene Bedingungen wurden dann theoriegeleitet unter Einbezug von Dokumentenanalysen als hinreichend oder notwendig zur Erreichung der Zielgröße klassifiziert und alle möglichen Kombinationen von Bedingungen erfasst. Im Anschluss an Einzelfallauswertungen (entlang des Musters: Bedingungen „treffen zu“ bzw. „treffen nicht zu“ und führen zum (Nicht-)Erreichen der Zielgröße) wurde die fallübergreifende Relevanz von Bedingungen für das Auftreten der Zielgröße mittels Konsistenzprüfung und Grad der Abdeckung analysiert. Auf dieser Basis ermöglichte das Vorgehen, konsistente und häufig auftretende Bedingungen zu identifizieren, die das Auftreten der Zielgröße beeinflussen. Im Ergebnis wurden die Bedingungskonstellationen mit einem Pfadmodell grafisch aufbereitet. Abschließend diskutierte Christina Kaps Potenziale und Grenzen dieser Vorgehensweise und zeigte auf, dass diese Methode systematische Vergleiche, die Identifikation von Typologien sowie den fortlaufenden Rückbezug auf den Einzelfall ermögliche. Sie hob die Stärken der Verbindung unterschiedlicher Methoden im Rahmen von QCA hervor. Die Betrachtung unterschiedlicher Ebenen (Einzelfall und aggregierte Ebene) sowie die Ableitung kontextsensibler Empfehlungen stiften einen wichtigen Nutzen. Zugleich reflektierte Christina Kaps sowohl die methodischen Ansprüche (umfassende und vergleichbare Informationen zu allen Fällen) als auch den hohen Arbeitsaufwand einer QCA-Evaluation.

Felix Leßke, vom Deutschen Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) in Bonn und *Gabriella Camacho Garland*, wissenschaftliche Mitarbeiterin an der Universität in Aarhus und ehemalige Gastwissenschaftlerin am DEval, hielten einen Vortrag zum Thema „*Qualitative Evaluierung mit Kausalkraft? – Erfahrungen mit Process Tracing in der Evaluierung von Rückkehr und Reintegration*“. Dabei zeigten sie am Beispiel der Evaluation der „Unterstützung des BMZ zu Rückkehr und Reintegration“ (Leitung: Johannes Schmitt, DEval), wie die Methode des Process Tracing dazu beitragen kann, Wirkungsevaluationen in herausfordernden Kontexten umzusetzen. Beim Process Tracing handelt es sich um eine Methode, die es ermöglicht, kausale Mechanismen zu identifizieren, indem nachgewiesen wird, was und warum etwas in einem bestimmten Prozess geschehen ist. Für das Evaluationsteam bot sich diese Methode u. a. deshalb an, weil keine Vergleichsgruppenstudien möglich waren und nicht auf vorhandene Daten, wie beispielsweise die Datenbank der Internationalen Organisation für Migration (IOM), zugegriffen werden konnte. Konkret wurde Process Tracing eingesetzt, um zu untersuchen, inwieweit die BMZ-finanzierten Reintegrationsangebote in den Herkunftsländern zur Stärkung der wirtschaftlichen und sozialen Teilhabe von Rückkehrer:innen beitragen. Dafür wurde zunächst gemeinsam mit den Stakeholder:innen eine Prozesstheorie des Wandels (pToC) entwickelt, in de-

ren Konstruktion und Revision auch Erkenntnisse aus einem Rapid Evidence Review und einer explorativen Fallstudie des Teams eingeflossen sind. Die Datenerhebung zur Überprüfung der pToC (und der darin hypothetisierten Kausalmechanismen) erfolgte auf Basis qualitativer, leitfadengestützter Interviews mit Rückkehrer:innen aus drei Fallstudienländern Ghana, Marokko und Nordirak – flankiert durch Interviews mit weiteren relevanten Stakeholder:innen. Um die Methode bestmöglich durchführen zu können, musste eine Reihe von Voraussetzungen geschaffen werden: Auf Ebene der gegenstands- und kontextbezogenen Voraussetzungen gehörten hierzu spezifische Interviewtrainings für den sensiblen Umgang mit Rückkehrer:innen als vulnerabler Zielgruppe und die Schaffung einer angemessenen Interviewsituation (u. a. neutraler Ort, weibliche Consultants bei Interviews mit Frauen). Um trotz der Heterogenität der Länderkontexte und individuellen Hintergründe der Befragten möglichst verallgemeinerbare Aussagen treffen zu können, wurde zudem eine kriterienbasierte Länderauswahl mit einer geschichteten Zufallsstichprobe zur Auswahl der Befragten kombiniert. Da das Vorgehen einen hohen Detailierungsgrad der erhobenen Daten zur Überprüfung der Kausalmechanismen erfordert, wurde das Process Tracing zur Überprüfung nur bestimmter Teilbereiche der Wirkungslogik genutzt und zu diesen Teilbereichen der pToC die Informationen aus den Tiefeninterviews mit Rückkehrer:innen und Expert:innen mit Informationen aus anderen Datenquellen trianguliert. Angesichts der großen Menge erhobener Daten wurden sogenannte Evidence Grids erstellt, um die Informationen zu fokussieren und einen schnellen Zugriff zu erlangen. Im Ergebnis konnte die Evaluierung so Aufschluss darüber geben, wie wirksam die Reintegrationsangebote in den Herkunftsländern für die Rückkehrer:innen sind.

3. Ergebnisse aus den Arbeitsgruppen

Die Impulse aus den fünf Vorträgen fanden zum Ende der Tagung Eingang in Arbeitsgruppen, in denen Voraussetzungen für erfolgreiche Wirkungsevaluationen vertiefend diskutiert wurden. Jede der drei Arbeitsgruppen befasste sich mit einer der im Tagungsprogramm aufgeworfenen Ebenen und präsentierte die nachfolgend skizzierten Ergebnisse im Plenum.

Die erste Arbeitsgruppe beschäftigte sich mit den *stakeholder:innenbezogenen Voraussetzungen* für erfolgreiche Wirkungsevaluationen. Aufseiten der Stakeholder:innen bildet die Leistungsbeschreibung zur Umsetzung der Wirkungsevaluation das vermittelnde Glied zwischen Auftraggebenden und Evaluierenden. Es wurde festgehalten, dass in der Leistungsbeschreibung dementsprechend alle zentralen Anforderungen, Erkenntnisinteressen und Rahmenbedingungen einer Wirkungsevaluation festgehalten und so weit wie möglich konkretisiert werden sollten. Weitere wichtige Voraussetzungen stellen hierbei auch die Verständigung über Ziele der Wirkungsevaluation und Vereinbarungen zur Nutzung der Evaluationsergebnisse dar. In diesem Zusammenhang wurden ebenso die Verständigung über und Abstimmung der Erwartungshaltungen und Interessen aller Beteiligten an einer Wirkungsevaluation diskutiert sowie Fragen zur Beteiligung und Mitwirkung von Stakeholder:innen in den verschiedenen Phasen einer Wirkungsevaluation. Die Leistungsbeschreibung als ge-

meinsamer Bezugsrahmen von Auftraggebenden und Evaluierenden dient im Verlauf einer Wirkungsevaluation sowohl als Instrument der wechselseitigen Verständigung als auch zur Kommunikation mit weiteren in die Evaluation Involvierten (siehe hierzu vertiefend den DeGEval ...Info-Beitrag „Ausschreibung von Evaluationsaufträgen“ in diesem Heft). Neben der Leistungsbeschreibung stellen allgemein eine Kultur des Lernens und des konstruktiven Umgangs mit Fehlern, ein von wechselseitigem Vertrauen geprägtes Miteinander und eine gewisse Offenheit in Bezug auf die Ergebnisse einer Wirkungsevaluation bedeutsame Voraussetzungen für erfolgreiche Wirkungsevaluationen dar. Sie sind nicht zuletzt auf der Mikroebene der Organisationen mitbestimmend für den Erfolg oder das Scheitern einer Wirkungsevaluation.

Gegenstands- und kontextbezogene Voraussetzungen für erfolgreiche Wirkungsevaluationen diskutierte die zweite Arbeitsgruppe, unter dem Eindruck des letzten Vortrags auch mit Blick auf Herausforderungen, die sich in fragilen Kontexten stellen. Kennzeichnend für solche Kontexte sind unter anderem soziale Spannungen (bis hin zu gewaltsamen Auseinandersetzungen), der erschwerte Zugang zu Zielgruppen, die Vulnerabilität der Zielgruppen und ein häufiger Wandel der Kontextbedingungen. Es wurde diskutiert, dass gerade dynamische oder fragile Kontexte eine holistische Herangehensweise bei der Erfassung von Wirkungen erfordern, die am besten durch Mixed- und Multi-Methods-Designs gewährleistet werden. Dabei sind restringierende Faktoren wie zeitliche und finanzielle Ressourcen zu berücksichtigen. Dies spräche nach Ansicht der Arbeitsgruppe u. a. für den Rückgriff auf (in fragilen Kontexten) bereits bewährte Methoden und sollte durch gute Feld- sowie Vorkenntnisse u. a. zum Forschungsstand flankiert sein. In schwer zugänglichen Kontexten oder (z. B. ländlichen) Regionen können dabei auch remote digitale Daten (z. B. Geodaten) und Methoden genutzt werden, um Veränderungen/Wirkungen zu erfassen. Zudem ist in der Regel eine Zusammenarbeit mit Gatekeepern oder lokalen Consultants vonnöten, die nicht nur den jeweiligen (Konflikt-)Kontext gut kennen, sondern auch Zugänge zu vulnerablen Gruppen schaffen können. Methoden sind kontextabhängig zu wählen, nicht alle Methoden sind in allen Kontexten anwendbar. So kann sich z. B. gerade in fragilen Kontexten ein ethisches Dilemma mit Blick auf die Kontrollgruppen im Rahmen von RCTs stellen, die trotz bestehender Bedarfe nicht am (potenziellen) Programmnutzen teilhaben können. Zudem kann dieser Ausschluss soziale Spannungen fördern oder verstärken. Die Einhaltung des *Do-No-Harm*-Prinzips ist hier also höchstes Gebot, auch um vulnerable Zielgruppen nicht durch die Evaluation zu gefährden. Verschiedene Maßnahmen können dazu beitragen, um dies bei der Datenerhebung im Rahmen von Wirkungsevaluationen sicherzustellen: Dazu zählen z. B. Schulungen in kontext- und zielgruppensensibler Interviewführung (inklusive der Fähigkeit zum Konfliktmanagement), die Schaffung von ‚safe places‘ während der Interviewsituation und die situationsangemessene Auswahl von Interviewenden (z. B. unter Berücksichtigung von Alter und Geschlecht). Ebenso ist die spätere (datenschutzkonforme) Weiterverarbeitung von Daten stets auch unter dem Gesichtspunkt des *Do-No-Harm*-Prinzips vorzunehmen. Im Rahmen der Diskussion wurde zudem festgehalten, dass es wichtig ist, von Beginn an Klarheit darüber herzustellen, welche genauen Wirkungen mit der zu evaluierenden Maßnahme überhaupt erzielt werden soll(t)en. Um hierüber Aufschluss zu erlangen, sind mitunter frühzeitig Expert:innen mit Feldkenntnissen einzubinden (u. a. bei der Konstruktion einer ToC im Rahmen einer the-

oriebasierten Wirkungsevaluation). Es wurde schließlich festgehalten, dass am Ende der jeweiligen Evaluation Empfehlungen zu formulieren sind, die – dem Kontext angemessen – realistisch umsetzbar sein sollen.

Die dritte Arbeitsgruppe reflektierte die *methodischen Voraussetzungen* für erfolgreiche Wirkungsevaluationen und hob dabei vor allem auf die präzise Bestimmung des zugrundeliegenden Erkenntnisinteresses und die Klärung des jeweiligen Kausalitätsverständnisses ab. Der Wirkungsbegriff weist ein inhärentes Qualitätsversprechen auf, dessen alltagsweltliche Deutung nicht immer im Einklang mit epistemologischen und ontologischen Implikationen des Begriffs steht. Demnach enthält der Begriff Wirkung im engen Sinne die Vorstellung einer Ursache, die in einem kausalen Verursachungszusammenhang mit einem Effekt steht. Diese vereinfachte Formulierung unterschlägt, dass Kausalität unterschiedlich gedacht werden kann (z. B. monokausal, konfigural oder generativ). Darüber hinaus finden sich in der Evaluation und darüber hinaus häufig auch Wirkungsverständnisse, die das Verhältnis von Ursache und Wirkung als plausibel begründbare Beziehung konzeptualisieren und von einem „weichen“ Kausalitätsverständnis ausgehen. Vor diesem Hintergrund sind die interessierenden bzw. untersuchten Fragestellungen eine wichtige Einflussgröße für die Wahl eines angemessenen methodischen Designs einer Wirkungsevaluation. Jene Entscheidungen sollten dabei aus einer umfangreichen und fundierten Kenntnis unterschiedlicher Methoden und Evaluationsdesigns heraus erfolgen und die Komplementarität unterschiedlicher Methoden nutzen. Dabei ist auch der zu evaluierende Kontext mit zu berücksichtigen und beispielsweise die konkreten Voraussetzungen für die Umsetzung methodischer Designs (z. B. sind Kontroll- oder Wartegruppendedesigns möglich? Welche Zugänge zum untersuchten Feld sind vorhanden?) und damit verbundene ethische Fragen zu klären. Institutionen und Stakeholder:innen sind hierbei nicht nur als Kontext, sondern auch als zugangsermöglichende Instanzen mitzudenken und einzubeziehen. Für Evaluierende, Auftraggebende und Beteiligte bzw. Betroffene von Evaluationen von besonderem Interesse ist auch die konkrete Darlegung methodischer Vorgehensweisen und ihrer jeweiligen Erkenntnismöglichkeiten und -grenzen. Dies unterstützt sie dabei, die Spezifik und Reichweite von Befunden angemessen einschätzen zu können und den bestmöglichen Nutzen aus einer Wirkungsevaluation zu ziehen. Idealerweise werden konkrete Rahmenbedingungen von Wirkungsevaluationen einschließlich ihrer Implikationen für die Wahl entsprechender Methoden bereits im Kontext von Programmplanung mitgedacht und dokumentiert. Dies ermöglicht potenziell, dass sich – mit Blick auf den von Stefan Silvestrini vorgestellten Befund – zukünftig ein positiver Zusammenhang zwischen der methodischen Qualität einer Wirkungsevaluation und deren Nutzen entwickelt und finden lässt.

4. Fazit, Ausblick und Dank

Die Beschäftigung mit den Voraussetzungen für erfolgreiche Wirkungsevaluationen ist weder neu noch abgeschlossen. Die Vorträge verdeutlichten, wie vielfältig und vielschichtig je nach Kontext, Gegenstand, Interessen der Stakeholder:innen und des Evaluationsdesigns die zu berücksichtigenden Voraussetzungen für erfolgreiche Wir-

kungsevaluationen sein können. Zudem verlangen neue Evaluationsgegenstände und Einsatzfelder von Evaluationen sowie ihre sich verändernden Rahmenbedingungen und methodische Innovationen nach einer kontinuierlichen Auseinandersetzung mit den methodischen sowie gegenstands-, kontext- und stakeholderbezogenen Herausforderungen von Wirkungsevaluationen und geeigneten Maßnahmen zur Lösung dieser Herausforderungen. Hierzu beizutragen war das Ziel der diesjährigen Frühjahrstagung des AK Methoden, indem aktuelle Impulse zum Thema mit dem Ziel des wechselseitigen (politikfeldübergreifenden) Lernens zusammengetragen wurden. Die Bereitstellung und Bündelung von Methodenwissen für die verschiedenen, in der DeGEval versammelten Evaluationskontexte ist eine zentrale Aufgabe des AK Methoden in seiner Querschnittsfunktion. Durch die Tagung konnten Teilnehmende aus verschiedenen Arbeitsbereichen und -kreisen der DeGEval und von außerhalb miteinander vernetzt werden und es gab ein großes Interesse an der weiteren Beschäftigung mit dem Thema. Diese Dokumentation und die Möglichkeit, die Präsentationen der Vorträge und die Aufzeichnung des Einführungsvortrags auf der Website des AK Methoden abzurufen, sollen auch über die zwei Tage und den Kreis der Beteiligten hinaus dazu beitragen.⁵ Der AK Methoden wird diesen Austausch auch in Zukunft fördern und vertiefen. Wir danken allen Referent:innen und Tagungsteilnehmer:innen für die anregenden Impulse und Diskussionen, dem lokalen Organisationsteam – Sandra Schopper und Maike Scheipers – für die sehr gute Zusammenarbeit bei der Tagungsvorbereitung und -durchführung sowie den beiden gastgebenden Institutionen – der Universität Saarbrücken (vertreten durch Wolfgang Meyer) und der htw saar (vertreten durch Dieter Filsinger) für die Gastfreundschaft.

Literatur

- Bischoff, U., Zimmermann, E. & König, F. (2021). Erkennen, was wirkt. Die Erprobung von Ansätzen der Wirkungsuntersuchung in der Evaluation von Bundesprogrammen der Demokratieförderung und Extremismusprävention und die damit gemachten Erfahrungen. In B. Milbradt, F. Greuel, S. Reiter & E. Zimmermann (Hrsg.), *Evaluation von Programmen und Projekten der Demokratieförderung, Vielfaltgestaltung und Extremismusprävention* (S. 244–268). Beltz Juventa.
- Reichardt, C. S. (2022). The Counterfactual Definition of a Program Effect. *American Journal of Evaluation*, 43(2), 158–174. <https://doi.org/10.1177/1098214020975485>
- Weber, M. (1972). *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*. Mohr.

Franziska Heinze | Deutsches Jugendinstitut (DJI), Außenstelle Halle (Saale) | Franckeplatz 1 | D-06110 Halle (Saale) | E-Mail: ak-methoden@degeval.org

Alexander Kocks | Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval) | Fritz-Schäffer-Str. 26 | D-53113 Bonn | E-Mail: ak-methoden@degeval.org

5 Abruf unter: <https://www.degeval.org/arbeitskreise/methoden-in-der-evaluation/aktuelles/>