Guggisberg, Dorothee: Soziale Dienste in der Schweiz – ein Überblick. In: Riedi, Anna Maria; Zwilling, Michael; Meier Kressig, Marcel; Benz Bartoletta, Petra; Aebi Zindel, Doris (Hrsg.): Handbuch Sozialwesen Schweiz. Bern 2013, S. 228-237
Haller, Dieter: Wirkungsforschung zur Entwicklung der Professionalität, Identität und Legitimation Sozialer Arbeit. In: Eppler, Natalie; Miethe, Ingrid; Schneider, Armin (Hrsg.): Qualitative und quantitative Wirkungsforschung. Ansätze, Beispiele, Perspektiven. Opladen 2011, S. 235-254
Haunberger, Sigrid: Realistic Evaluation als Evaluationsrahmen in der Klinischen Sozialarbeit. In: Hahn, Gernot; Hüttemann, Matthias (Hrsg.): Evaluation psychosozialer Interventionen. Klinische Sozialarbeit Band 7. Köln 2015, S. 47-58
Heiner, Maja (Hrsg.): Selbstevaluation in der sozialen Arbeit. Fallbeispiele zur Dokumentation und Reflexion beruflichen Handelns. Freiburg im Breisgau 1988
Merchel, Joachim: Evaluation in der Sozialen Arbeit. München und Basel 2010
Merchel, Joachim: Evaluation: Motivlagen zwischen Legitimation und Weiterentwicklung einer professionellen Praxis in der Sozialen Arbeit. In: Hahn, Gernot; Hüttemann, Matthias (Hrsg.): Evaluation psychosozialer Interventionen. Klinische Sozialarbeit Band 7. Köln 2015, S. 16-30
Minnig, Christoph; Uebelhart, Beat; Zängl, Peter: Wirkungsmessung in der Sozialen Arbeit – Eine kritische Analyse entlang des Social-Impact-Modells. In: Gmür, Markus; Schauer, Reinbert; Theuvsen, Ludwig (Hrsg.): Performance Management in Nonprofit-Organisationen – Theoretische Grundlagen, empirische Ergebnisse und Anwendungsbeispiele. Bern 2013, S. 206-214
Nuspliger, Kurt: Der Stellenwert der Evaluation am Beispiel des Kantons Bern. In: Leges 1/2005, S. 51-63
Pawson, Ray; Tilley, Nick: Realistic evaluation. London et al. 1997
Schneider, Armin: Evaluation als professionsbildende Forschung in der Sozialen Arbeit. In: Heimgartner, Arno; Loch, Ulrike; Sting, Stephan (Hrsg.): Empirische Forschung in der Sozialen Arbeit. Methoden und methodologische Herausforderungen. Wien 2012, S. 261-272
Schröder, Jan; Kettiger, Daniel: Wirkungsorientierte Steuerung in der sozialen Arbeit: Ergebnisse einer internationalen Recherche in den USA, den Niederlanden und der Schweiz. Schriftenreihe des Bundesministeriums für Familie, Senioren, Frauen und Jugend, Band 229. Stuttgart 2001
Spiegel, Hiltrud von: Aus Erfahrung lernen. Qualifizierung durch Selbstevaluation. Münster 1993

EVALUATIONEN ZUR WIRKUNGSMESSUNG VON NON-PROFIT-ORGANISATIONEN | Erfahrungen aus dem Bereich der Entwicklungszusammenarbeit

Alexandra Caspari

Zusammenfassung | Evaluationen haben in der Entwicklungszusammenarbeit eine lange Tradition. Daher wurde die Diskussion um den Wirkungsnachweis sozialer Maßnahmen stets im Kontext von Evaluationen geführt. Wirkungsevaluationen haben gegenüber reinen Wirkungsstudien einen enormen Mehrwert. Die Erkenntnisse aus der Evaluationsforschung allgemein sowie die praktischen Erfahrungen im Bereich der Wirkungsevaluation von Entwicklungszusammenarbeit scheinen für Non-Profit-Organisationen durchaus hilfreich.

Abstract | Evaluations have a long tradition within development cooperation. Thus, the question on how to proof effects of social programs has always been discussed in the context of evaluation. Impact evaluations have a great surplus value compared to sole impact studies. The findings of evaluation research, as well as the practical experiences relating to impact-evaluations of development cooperation, seem to be appropriate for non-governmental organisations.

Schlüsselwörter ▶ Wirkung ▶ Entwicklungshilfe ▶ Non-Profit-Organisation ▶ Evaluation ▶ Wirksamkeitsforschung

1 Einleitung | Die Frage nach der Wirksamkeit von Maßnahmen in der Entwicklungszusammenarbeit (EZ) ist so alt wie die Entwicklungszusammenarbeit selbst: Hilft die Hilfe? Diese Frage wurde immer wieder teilweise heftig und kontrovers debattiert, denn Geberorganisationen konnten die Wirkungen ihrer Maßnahmen kaum ausreichend nachweisen. Im Mai 2006 veröffentlichte das Center for Global Development (CGD) einen Bericht mit dem bezeichnenden Titel „When Will We Ever Learn?", in dem einleitend festgestellt wird: „Yet after decades in which deve-

lopment agencies have disbursed billions of dollars for social programs, and developing country governments and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impact of most of these social programs" (*CGD* 2006, S. 1). Dieser Bericht löste in der internationalen Entwicklungszusammenarbeit eine intensive Diskussion über Möglichkeiten der Wirkungsmessung von EZ-Maßnahmen aus.

Die Entwicklungszusammenarbeit, ob staatlich oder nicht staatlich, steht bei der Frage der Wirkungsmessung ihrer Projekte und Programme vor dem gleichen Problem wie Non-Profit-Organisationen allgemein: Da die intendierte Wirkung von EZ-Maßnahmen nicht die Gewinnmaximierung oder die Steigerung des Unternehmenswertes (Shareholder-Value) ist, sondern die Behebung oder Milderung sozialer Missstände und die Verbesserung der Lebenssituation unterstützungsbedürftiger Menschen beziehungsweise der Klientinnen und Klienten (Stakeholder-Value), ist die Messung und der Nachweis dieser Wirkungen ungleich schwieriger. In der Entwicklungszusammenarbeit wurde die Diskussion um Wirkungsmessung stets im Kontext von *Evaluationen* geführt. So lautet auch der Untertitel der CGD-Publikation „Improving Lives Through Impact Evaluation" (*CGD* 2006). Dies scheint ein entscheidender Punkt, denn Evaluationen beziehungsweise Wirkungsevaluationen sind mehr als Wirkungsstudien oder reine Wirkungsmessungen und vermögen daher Wirkungen einer sozialen Maßnahme angemessener aufzuzeigen können. Die Erfahrungen der Entwicklungszusammenarbeit im Wirkungsnachweis scheinen daher für NPOs durchaus wertvoll.

2 Wirkungen und Wirkungsmessung | Der englische Begriff Impact Evaluation scheint irreführend, da der Begriff Impact in den Wirkungsmodellen unterschiedlicher Disziplinen teilweise nicht einheitlich definiert und verortet wird beziehungsweise wurde. Allgemein wird heute nicht nur in der Entwicklungszusammenarbeit meist von einer Wirkungslogik ausgegangen, in der durch eingesetzte Ressourcen (Inputs) Leistungen (Outputs) erbracht werden, die zu direkten, intendierten Wirkungen bei den Begünstigten führen (Outcomes), die wiederum zu längerfristigen, übergeordneten, entwicklungspolitischen (oder allgemein gesellschaftlichen) Wirkungen (Impacts) führen (für den Bereich der NPO insbesondere *Phineo* 2014, S. 6, *Phineo* 2013, S. 5, *Stiftung Zewo* 2013, S. 8 f., *DG ECHO* 2007, S. 26, 53). Ungeachtet des Verständnisses von Impact liegt das Hauptaugenmerk bei Wirkungsmessungen aber in erster Linie auf Outcomes, also den kurz- und insbesondere mittelfristigen direkten Wirkungen bei den Begünstigten.

Unter dem Begriff der Wirkung werden hierbei sämtliche Veränderungen verstanden, die eine Maßnahme direkt ausgelöst, das heißt bewirkt hat oder umgekehrt: sämtliche Veränderungen, die eindeutig einer Intervention attribuiert werden können. Diese Veränderungen können intendiert oder nicht intendiert, positiv oder negativ sein. Zentrale Schwierigkeit bei der Attribution, das heißt der kausalen Zuordnung von Wirkungen zu einer Maßnahme, ist die Frage, ob festgestellte Veränderungen bei der Zielgruppe einer Maßnahme tatsächlich Folge dieser Maßnahme sind oder ob sich diese Veränderung auch ohne diese Maßnahme eingestellt hätte. Daher wird zwischen Brutto- und Nettowirkungen unterschieden (*Rossi* u.a. 1988): Bruttowirkungen werden dabei verstanden als die nach Beendigung einer Maßnahme bei der Zielgruppe insgesamt beobachtbaren Veränderungen, die sowohl auf die Maßnahme als auch eine beliebige Anzahl anderer Einflüsse zurückzuführen sind. Nettowirkungen dagegen umfassen nur den isolierten Anteil an den insgesamt auftretenden Veränderungen bei der Zielgruppe, der allein auf die durchgeführte Maßnahme zurückzuführen ist, das heißt der nicht eingetreten wäre, wenn die Maßnahme nicht durchgeführt worden wäre.

Diese Nettowirkungen sind jedoch nicht direkt beobachtbar, da Personen entweder an einer Maßnahme teilnehmen oder nicht. Das heißt, dass die Veränderungen für diejenigen, die nicht an einer Maßnahme teilnehmen, nicht direkt gemessen werden können. Daher werden Wirkungen dadurch ermittelt, dass die bei Teilnehmerinnen und Teilnehmern tatsächlich festgestellten Veränderungen mit Veränderungen bei einer Gruppe von Nichtteilnehmern verglichen werden – einer Vergleichsgruppe. Die Differenz der Veränderungen der beiden Gruppen kann dann als kausaler Nettoeffekt interpretiert werden, der allein der Maßnahme zuzuschreiben ist. Wirkungsmessungen fokussieren folglich auf die Bestimmung der Nettowirkungen einer Maßnahme.

3 Untersuchungsdesigns und Erhebungsmethoden | Um Wirkungen kausal einer Maßnahme zuordnen zu können, müssen also Kontroll- beziehungsweise Vergleichsgruppen gebildet werden. Zentrales Moment bei Wirkungsmessungen ist daher die Wahl des geeigneten Untersuchungsdesigns, denn das Design ist entscheidend für den Grad an Gewissheit, mit dem ein Ursache-Wirkungszusammenhang aufgezeigt werden kann (*Schnell* u.a. 1999, S. 203).

Häufig werden in der Praxis vorexperimentelle Designs angelegt, das heißt es wird ein Vorher-Nachher-Vergleich bei der Zielgruppe beziehungsweise den Teilnehmerinnen und Teilnehmern einer Maßnahme umgesetzt, indem Baselinedaten (erhoben vor der Maßnahme) mit Daten aus einer Erhebung nach der Maßnahme verglichen werden. Die aufgefundenen Veränderungen werden dann der Maßnahme zugeschrieben und daraus die vermeintlichen Wirkungen abgeleitet. Ein solcher Vorher-Nachher-Vergleich zeigt jedoch lediglich die Entwicklung der Zielgruppe über die Zeit hinweg auf und kann demnach keine zuverlässige Antwort auf die Frage nach den Nettowirkungen einer Maßnahme geben – gemessen werden lediglich die Bruttowirkungen. Die hierbei beobachteten Veränderungen sind selten allein, sondern nur teilweise auf die Maßnahme zurückzuführen. Andere, externe Faktoren können die Wirkung der Maßnahme beeinflussen, schwächen oder auch verstärken. Gleichwohl kann sich die Situation der Zielgruppe auch ohne Maßnahme verändert haben. Solche Faktoren, die teilweise oder ganz für die beobachtete Veränderung verantwortlich sein können, bleiben bei einem reinen Vorher-Nachher-Vergleich der Teilnehmenden einer Maßnahme unberücksichtigt.

Dieses Problem kann durch die Bildung einer Kontrollgruppe (KG) beziehungsweise Vergleichsgruppe (VG) gelöst werden, das heißt Personen, die nicht an der Maßnahme teilgenommen haben, ansonsten aber in allen anderen Aspekten identisch beziehungsweise möglichst identisch mit der Zielgruppe sind. Hierdurch wird ein Mit-Ohne-Vergleich möglich (Single-Difference-Methode), das heißt die Wirkungen werden über einen einfachen Vergleich zwischen den Teilnehmern und Nichtteilnehmern einer Maßnahme erhoben. Zentrale Grundannahme hierbei ist, dass die Ausgangssituation der ZG und der VG vor der Maßnahme identisch ist. Diese Annahme ist in der Realität der EZ oder allgemein sozialer Programme nicht immer gegeben, da Maßnahmen häufig gezielt für Personen aufgelegt werden, die entweder besondere Defizite aufweisen oder aber besondere Voraussetzungen erfüllen. Dadurch werden bei diesem Mit-Ohne-Vergleich die berechneten Wirkungen der Maßnahme je nach Situation über- oder unterbewertet. Diesem Problem kann begegnet werden, indem eine Vorher-Messung sowohl bei der ZG als auch der VG durchgeführt wird, das heißt der Mit-Ohne-Vergleich mit dem Vorher-Nachher-Vergleich kombiniert wird (Double-Difference-Methode). Die Wirkung einer Maßnahme ergibt sich aus dem Unterschied zwischen ZG und VG nach der Maßnahme minus dem Unterschied zwischen ZG und VG vor der Maßnahme. Für Wirkungsmessung notwendige Designs sind daher experimentelle oder quasi-experimentelle Designs.

3-1 Experimentelle Designs / Randomized Controlled Trials (RCT) | Beim experimentellen Design werden Personen vor der Implementation einer Maßnahme nach dem Zufallsprinzip (Randomisierung) zwei Gruppen zugeordnet: zum einen einer Gruppe, die an der geplanten Maßnahme teilnimmt, also die Zielgruppe (ZG) sein wird (Teilnehmer), und zum anderen einer Kontrollgruppe (KG), die nicht an der Maßnahme teilnehmen wird (Nichtteilnehmer).

Experimentelle Designs gelten als die rigorosesten Ansätze der Wirkungsmessung, denn durch den zufallsgesteuerten Auswahlprozess werden systematische Gruppenunterschiede eliminiert, das heißt systematische Auswahlverzerrungen liegen nicht vor (*Schnell* u.a. 1999, S. 213 f., *ADB* 2006, S. 5 f., *Bloom* 2006, S. 1, *Baker* 2000, S. 2). Da folglich die Ausgangssituation der ZG und KG vor der Maßnahme übereinstimmen, können alle nach der Maßnahme gemessenen Unterschiede in den Wirkungen zwischen ZG und der KG auch der Maßnahme zugeschrieben werden. Allerdings trifft dies nur in sogenannten Laborexperimenten zu – in sogenannten Feldexperimenten, das heißt der Anwendung einer experimentellen Versuchsanordnung in für die Beteiligten authentischer Alltagsumgebung, ist es annähernd unmöglich, alle Eigenschaften der Personen vorab zu berücksichtigen, so dass Gruppenunterschiede zwischen ZG und KG nach wie vor gegeben sind. Während der einfache Mit-Ohne-Vergleich solche systematischen Gruppenunterschiede nicht berücksichtigt, können diese anhand der Kombination eines Mit-Ohne-Vergleichs und eines Vorher-Nachher-Vergleichs erfasst werden.

3-2 Quasi-experimentelle Designs[1] | Insbesondere im Rahmen einer Wirkungsmessung sozialer Programme sind experimentelle Designs häufig nicht möglich, da eine randomisierte Zuteilung in Zielgruppe und Kontrollgruppe nicht realisierbar ist. Daher ist das quasi-experimentelle Design in der Wirkungsmessung das sicherlich am häufigsten angewandte: Das Quasi-Experiment orientiert sich an der Experimentallogik, allerdings wird hierbei keine Kontrollgruppe nach dem Zufallsprinzip gebildet, sondern es wird eine Vergleichsgruppe (re-)konstruiert. Im Unterschied zum experimentellen Design erfolgt die Bestimmung der Vergleichsgruppen daher beim quasi-experimentellen Design im Nachhinein, das heißt entweder im Laufe der Maßnahme oder gar erst nach Abschluss der Maßnahme im Rahmen einer Wirkungsmessung.

Liegen Daten sowohl für die Zielgruppe als auch einer Vergleichsgruppe für den Zeitpunkt vor Implementierung einer Maßnahme vor, kann die Double-Difference-Methode angewandt werden, was – bei guter Datenlage und optimaler Umsetzung – zu ähnlich robusten Ergebnissen über die Wirkungen einer Maßnahme führt wie ein experimentelles Design.

4 Zum Mehrwert von Wirkungsevaluationen | Die dargestellten experimentellen beziehungsweise quasi-experimentellen Designs, die eine Vergleichsgruppe sowie eine Vorher-Nachher-Messung umfassen und somit eine Double-Difference-Berechnung zulassen, sind zuverlässige Methoden, um Wirkungen eindeutig einer Maßnahme zuzuschreiben und auch deren Umfang darzustellen. Allerdings wird hierdurch lediglich die Frage beantwortet „Hat die Maßnahme Wirkungen entfaltet (oder nicht)?" Nicht aber die Frage „Warum hat eine Maßnahme Wirkungen entfaltet (oder nicht)?" beziehungsweise: „Wie hat die Maßnahme gewirkt, unter welchen Bedingungen?" Untersuchungen, die Ergebnisse über die Wirkungen aufzeigen, aber keine Angaben machen, warum eine Maßnahme die erwarteten Wirkungen gezeigt hat oder nicht, werden in der Evaluations-Community aufgrund fehlender Informationen über die vermittelnden Prozesse und Wirkmechanismen als black box bezeichnet (*Leeuw; Vaessen* 2009, S. 20, *Bloom* 2006, S. 18 f., *White* 2006, S. 9, *Ravallion* 2005, S. 1).

1 Quasi-experimentelle Designs werden insbesondere im Englischen teilweise non-experimental designs genannt. Dies ist jedoch verwirrend, da nicht gleichzusetzen mit vorexperimentellen Designs.

Seit einigen Jahren finden sich im Kontext der Entwicklungszusammenarbeit sehr viele Wirkungsstudien, vornehmlich durchgeführt von Forschungseinrichtungen in den USA, die als Wirkungsevaluation oder *Impact Evaluation* bezeichnet werden, obwohl sie lediglich – wenn auch ausgesprochen elaboriert – die Wirkungen einer Maßnahme mit experimentellen oder quasi-experimentellen Designs nachweisen. Derartige Studien sind ohne Frage außerordentlich nützlich, zeigen sie doch für das schwierige Feld der Wirkungsmessung im sozialen Bereich Umsetzungsmöglichkeiten in der Praxis auf. Doch so förderlich sie für die Diskussion um Möglichkeiten von Wirkungsmessung auch sein mögen, können sie nicht als Wirkungsevaluationen bezeichnet werden – sie sind und bleiben Wirkungsstudien inklusive der *black box*. Denn eine Wirkungsevaluation beinhaltet weitaus mehr als eine Wirkungsmessung – sie umfasst auch die Frage nach dem Warum und Wie.

4-1 Zentrale Elemente von Evaluationen | Für den Begriff der Evaluation findet sich keine einheitliche Definition. Doch die diversen Definitionen in der internationalen Evaluations-Community sind in ihrem Kern letztendlich identisch und beziehen sich meist auf die Ausführungen von *Mertens* (1998, S. 219) „Evaluation is the systematic investigation of the merit or worth of an object (program) for the purpose of reducing uncertainty in decision making" sowie von *Rossi, Freeman* und *Hofmann* (1988, S. 3), nach denen Evaluation die „systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Beurteilung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer Interventionsprogramme" ist und „den gezielten Einsatz sozialwissenschaftlicher Forschungsmethoden zur Verbesserung der Planung und laufenden Überwachung sowie zur Bestimmung der Effektivität und Effizienz von Gesundheits-, Bildungs-, Sozialhilfe- und anderen sozialen Interventionsmaßnahmen" bezeichnet.

Insgesamt lassen sich aus den diversen Definitionen zwei relevante Elemente einer professionellen Evaluation ableiten:

▲ Das Erkenntnisinteresse von Evaluationen ist die Güte und der Nutzen einer Maßnahme beziehungsweise deren Konzeption, Ausgestaltung, Umsetzung und Nutzen mit dem Zweck, die Planung und Umsetzung von Maßnahmen zu verbessern oder allgemein Entscheidungsprozesse zu unterstützen. Evaluationen

beinhalten somit über die Beschreibung beziehungsweise Untersuchung der Maßnahme hinaus auch stets deren nachvollziehbare Bewertung im Sinne von Güte und Nutzen sowie hieraus abgeleitete Handlungsempfehlungen.

▲ Evaluationen sind datengestützt; für die Erhebung notwendiger Informationen und Daten werden empirisch-wissenschaftliche Methoden angewandt, insbesondere quantitative und qualitative Methoden der empirischen Sozialforschung (*DeGEval* 2008, S. 15 f.).

4-2 Die OECD/DAC-Evaluierungskriterien – zentrale Prüfkriterien mit Leitfragen | Um den ersten Aspekt, das Erkenntnisinteresse, angemessen zu berücksichtigen, sind verbindliche Grundlage einer jeden Evaluation in der Entwicklungszusammenarbeit fünf vom Development Assistance Committee (DAC) der OECD 1991 erstmals aufgeführte und 2000 präzisierte Evaluationskriterien: Relevanz, Effektivität, Effizienz, entwicklungspolitische Wirkungen sowie Nachhaltigkeit (*ECD; DAC* 1991 und 2000). Diese DAC-Evaluationskriterien umfassen für jedes der fünf zentralen Prüfkriterien ein Set an konkreten Leitfragen als Orientierungslinie, die situationsspezifisch anzupassen sind, wobei eine Änderung der Substanz nicht erfolgen sollte (*BMZ* 2006).

Das Kriterium *Relevanz* fragt: Tun wir das Richtige? Es umfasst Leitfragen bezüglich des Ausmaßes, „in dem die Ziele der Entwicklungsmaßnahme mit dem Bedarf der Zielgruppen, den Politiken des Kooperationslandes und der Partnerinstitution, den globalen Entwicklungszielen sowie der entwicklungspolitischen Grundausrichtung der Bundesregierung übereinstimmen" (*BMZ* 2006, S. 3). *Effektivität* bedeutet: Erreichen wir die Ziele der Entwicklungsmaßnahme? „Dieses Kriterium bezieht sich auf das Ausmaß, in dem die Maßnahme dazu beiträgt, dass die (direkten) Ziele der Entwicklungsmaßname erreicht werden (Soll-Ist-Vergleich)" (*ebd*., S. 4). Das Kriterium *Effizienz* fragt: Werden die Ziele durch die Entwicklungsmaßnahme wirtschaftlich erreicht? Es „misst die Angemessenheit der für eine Entwicklungsmaßnahme eingesetzten Ressourcen im Hinblick auf die damit erzielten Resultate" (*ebd*., S. 5), auch im Vergleich zu Alternativlösungen. *Impact* fragt: Tragen wir zur Erreichung übergeordneter entwicklungspolitischer Wirkungen bei? Neben der Frage, „ob und inwieweit die Entwicklungsmaßnahme dazu beiträgt, die angestrebten übergeordneten entwicklungspolitischen Ziele zu erreichen" (*ebd*., S. 6), wird hier auch überprüft, ob und welche anderen positiven und negativen Wirkungen eingetreten sind. Das Kriterium *Nachhaltigkeit* fragt: Sind die positiven Wirkungen von Dauer? Hier wird hinterfragt, „inwieweit die positiven Wirkungen der Entwicklungsmaßnahme über das Ende der Unterstützung hinaus fortbestehen" (*ebd*., S. 7).

Diese Evaluationskriterien wurden vom DAC 1999 für die Evaluation humanitärer Hilfe adaptiert, wobei einzelne Kriterien teilweise mit Subkriterien ergänzt wurden, wie zum Beispiel Relevanz mit Angemessenheit, Effektivität mit Rechtzeitigkeit, Effizienz mit Preis-Leistungsverhältnis und Nachhaltigkeit mit Vernetzung (connectedness).[2] Des Weiteren wurden die Kriterien Abdeckung/Reichweite (coverage), Kohärenz und Koordination aufgenommen (*OECD; DAC* 1999, S. 22 f., *ALNAP* 2006, S. 17 ff., *DG ECHO* 2007, S. 50 f.).

Es wird ersichtlich, dass Evaluationen nicht mit Wirkungsstudien gleichzusetzen sind, denn neben den Wirkungen, die vornehmlich über das Kriterium der Effizienz, aber des Impact betrachtet werden, sind bei Evaluationen auch weitere Aspekte zu analysieren, insbesondere die Frage der Relevanz. Hier sind zum Beispiel auch Fragen zu beantworten wie „Ist die Wirkungshypothese plausibel?", „Inwieweit nutzt die Entwicklungsmaßnahme lokale Verfahren und Strukturen?", „Inwieweit ist die Entwicklungsmaßnahme subsidiär zu den Eigenanstrengungen des Partnerlandes?", „Inwieweit greifen verschiedene Instrumente der Entwicklungszusammenarbeit [...] sinnvoll ineinander?" und „Inwieweit waren die Aktivitäten/Maßnahmen in technischer, organisatorischer und finanzieller Hinsicht angemessen für die Zielerreichung, auch im Hinblick auf Nachhaltigkeit?" (*BMZ* 2006, S. 3).

Aber auch unter dem Kriterium der Effektivität ist nicht nur zu beantworten, „inwieweit [...] die (direkten) Ziele der Entwicklungsmaßnahme gemäß des (ggf. angepassten) Zielsystems erreicht" wurden und „welchen (konkreten) Beitrag [...] die Maßnahmen der deutschen EZ zu der Erreichung der Ziele der Entwicklungsmaßnahme" leisteten, sondern auch, „welche Faktoren [...] ausschlaggebend für die bisherige Erreichung beziehungsweise Verfehlung der Projekt-

2 Hiermit soll berücksichtigt werden, dass (kurzfristige) humanitäre Hilfe, die häufig selbst nicht auf Nachhaltigkeit angelegt ist, dennoch langfristige, das heißt nachhaltige (negative) Wirkungen im Umfeld verursachen kann, zum Beispiel auf die Umwelt oder die lokale Beschäftigungssituation (*OECD; DAC* 1999, S. 22).

ziele" waren und „welche sonstigen Effekte – auch negative – […] auf der Ebene der Leistungen und (Projekt-/Programm-)Ziele feststellbar" sind (*BMZ* 2006, S. 4). Evaluationen hinterfragen im Gegensatz zu reinen Wirkungsstudien demnach immer das Warum und Wie. Werden im Rahmen von Evaluationen auch Wirkungsmessungen mit experimentellen oder quasi-experimentellen Designs umgesetzt, so dass die kausal der Maßnahme zuzuschreibenden Nettowirkungen aufgezeigt werden können, werden diese als Wirkungsevaluationen bezeichnet.

5 Erhebungsmethoden in der Evaluation | Ein bedeutender Aspekt ist auch das oben aufgeführte zweite relevante Element einer professionellen Evaluation, nämlich die im Rahmen von Evaluationen angewandten Datenerhebungsverfahren. Im Rahmen der aktuellen Wirkungsdiskussion kommen immer wieder Debatten über Datenerhebungsmethoden auf, die teilweise an den Paradigmenstreit zwischen quantitativer und qualitativer Sozialforschung erinnern (*Caspari* 2008, S. 139 f., *Caspari; Barbu* 2008, S. 20 f., *Caspari* 2004, S. 92 ff., *Caspari* 2006). Hierbei wird allerdings häufig nur auf Wirkungsmessungen fokussiert. Selbstredend werden bei der Vorher- und Nachher-Messung die diversen formulierten Wirkungsindikatoren zumeist über quantitative oder besser standardisierte Methoden erhoben.

Im Kontext der Methodendiskussionen werden jedoch zwei zentrale Aspekte von Wirkungsevaluationen häufig außer Acht gelassen:

▲ Wirkungsevaluationen haben gemäß Definition auch nicht intendierte Wirkungen zu erfassen. Bei reinen Wirkungsstudien scheint die Gefahr groß, dass lediglich die intendierten (positiven) Wirkungen, die zu Maßnahmebeginn über Indikatoren operationalisiert wurden, im Rahmen der Vorher- aber auch Nachher-Messung erhoben werden. Evaluationen dagegen verfolgen quasi einen Trichteransatz: Zunächst werden sämtliche Veränderungen im Umfeld der evaluierten Maßnahme identifiziert, ohne hierbei von den Maßnahmezielen auszugehen. Erst danach wird analysiert, welche der festgestellten Veränderungen der Maßnahme kausal – durch einen Single-Difference-Vergleich der Veränderungen bei Ziel- und Vergleichs- beziehungsweise Kontrollgruppe – oder plausibel – durch einen Vergleich mit dem der Maßnahme zugrunde liegenden Wirkungsmodell – zugeordnet werden können. Wirkungsevaluationen erfassen somit weitaus mehr als lediglich die Zielerreichung: Durch die induktive Vorgehensweise werden auch nicht intendierte positive sowie eventuell negative Wirkungen des Vorhabens aufgezeigt. Um diese Veränderungen im Umfeld der Maßnahme in einem ersten Schritt zu identifizieren, kann auf Dokumente und Sekundärdaten zurückgegriffen werden – insbesondere aber sind qualitative Intensiv- und/oder Fokusgruppeninterviews mit den Beteiligten und Vertreterinnen und Vertretern der Zielgruppe notwendig.

▲ Im Rahmen von Wirkungsevaluationen ist auch die der Maßnahme zugrunde liegende Programmtheorie zu überprüfen sowie die Frage zu beantworten, warum eine Maßnahme ihre Ziele erreicht beziehungsweise nicht erreicht hat und welche Faktoren hierfür ausschlaggebend waren. Hier scheint die Nutzung quantitativer Datenerhebungsmethoden nicht wirklich zielführend. Relevant sind eher nicht reaktive Verfahren wie die Analyse von Programmdokumenten und -berichten, Sekundärdaten sowie wissenschaftlicher Literatur. Sowohl für die Überprüfung des der Maßnahme zugrunde liegenden Ursache-Wirkungs-Modells als auch für die Interpretation der aufgezeigten Wirkungen sind aber auch hier vornehmlich qualitative Datenerhebungsmethoden wie Intensiv-, Experten- oder Fokusgruppeninterviews mit verschiedenen Gruppen von Stakeholdern notwendig.

Es zeigt sich also, dass für angemessene Wirkungsevaluationen die Nutzung des Methodenrepertoires beider Methodologien im Sinne einer Triangulation beziehungsweise eines Methodenmix' notwendig ist, denn „both research traditions and the research methods they are most closely linked to (quantitative vs qualitative), are suited to answering very different types of research question" (*Prowse* 2007, S. 3). Das Erfordernis der Integration von qualitativen und quantitativen Ansätzen bei Wirkungsevaluationen ist in der Entwicklungszusammenarbeit mittlerweile unumstritten (*Bamberger* 2000, *Chung* 2000, *Guijt* 2000, *Aplleton; Booth* 2001, *Ezemenari* et al. 1999, *Kassam* 1998): „A good evaluation necessitates the combining of both qualitative and quantitative approaches to data collection and analysis. Combining both approaches provides quantified results of program impacts as well as explanations of the processes and intervening factors that yielded these outcomes. Combined approaches enrich interpretation or explanation (causality) of outcomes measured by the evaluation" (*Ezemenari* et al. 1999, S. 19).

6 Chancen und Herausforderungen für Non-Profit-Organisationen | Die in der Entwicklungszusammenarbeit seit vielen Jahrzehnten gängige Praxis, die Effektivität und Effizienz sowie gemäß DAC-Kriterien auch die langfristigen übergeordneten Wirkungen, die Nachhaltigkeit und insbesondere die Relevanz ihrer Maßnahmen mithilfe von Evaluationen zu untersuchen, sowie die intensive Diskussion um Möglichkeiten der kausalen Attribution von Wirkungen mittels Wirkungsevaluationen der letzten zehn Jahre hat auch bei NPO im Umfeld der Entwicklungszusammenarbeit zu einer intensiven Auseinandersetzung mit Wirkungsevaluationen geführt. So werden zum Beispiel in der Katastrophen- und in der humanitären Hilfe, aber auch in der entwicklungspolitischen Bildungsarbeit Umsetzungsmöglichkeiten von Wirkungsevaluationen einschließlich der Wirkungsmessung mithilfe (quasi-)experimenteller Designs intensiv diskutiert. Hierzu wurden Handreichungen veröffentlicht, die für eine Wirkungsevaluation notwendige Voraussetzungen erläutern, wie zum Beispiel die Formulierung konkreter Ziele für Maßnahmen auf Projektebene einschließlich der Operationalisierung SMARTer Indikatoren, die im Rahmen eines von Beginn an implementierten Monitorings zu messen sind, und mit teilweise detaillierten Beispielen untermauern (*ALNAP* 2006, *Canteli* 2013, *DWHH* 2008, *VENRO* 2005, *Nelson* 2008, *American Red Cross* et al. 2006, *DG ECHO* 2007, *Massing* u.a. 2010, *Caspari* 2012).

Aber auch außerhalb des Politikfeldes finden sich vermehrt Publikationen, die Wirkungsevaluationen von Maßnahmen spezifisch im NPO-Bereich erläutern, wobei teilweise die verschiedenen Designs zur kausalen Zuordnung der Wirkungen genauestens dargestellt werden (zum Beispiel *Phineo* 2013, *Stiftung Zewo* 2013, S. 13 ff., *Stiftung Zewo* o.J.). Für NPO scheint es sich also zu lohnen, den Blick auf die Erkenntnisse der Evaluationsforschung im Allgemeinen und auf die Erfahrungen im Bereich der Evaluation von Entwicklungszusammenarbeit im Besonderen zu richten. Hier finden sich auch Hinweise, wie quasi-experimentelle oder gar experimentelle Designs in der Evaluationspraxis sozialer Maßnahmen umgesetzt werden können (*Caspari* 2009, *Bamberger* 2006, *ADB* 2006, *White* 2006, *Ravaillon* 2005, *Baker* 2000).

Ist die Anwendung der für die kausale Attribution notwendigen Designs nicht möglich, helfen die Ausführungen zur Nutzung von Methoden-Triangulation, so dass über Cross-Checks der Erkenntnisse aus den diversen Erhebungsverfahren die kausale Kontribution an Plausibilität gewinnt. Kontributionsanalysen untersuchen die zuvor aufgestellte explizite Wirkungstheorie anhand iterativer Prozesse aus Aufstellen, Testen und Verfeinern. Die aktuelle Diskussion um systemische Wirkungsmodelle, die die Komplexität sozialer Programme durch systemisch angelegte Modelle abbilden, wobei externe Faktoren stärker berücksichtigt werden, scheint für die Präzisierung der Wirkungstheorie hilfreich – unabhängig davon, ob diese anschließend anhand von Attributions- oder Kontributionsanalysen überprüft werden.

Auch Non-Profit-Organisationen müssen heute die Erfolge ihre Maßnahmen nachweisen – ebenso wie gewinnorientierte oder öffentliche Organisationen. Öffentliche und private Geldgeber geben sich kaum noch mit dem Nachweis der ordnungsgemäßen Mittelverwendung zufrieden, es werden Belege über die Effektivität und Effizienz der Ressourcenverwendung gefordert. Der Fokus hat sich von einer reinen Input- oder auch Output-Orientierung hin zu einer Outcome-Orientierung verlagert, die Frage der Wirksamkeit ist daher heute auch für NPO von zentraler Bedeutung. Im Politikfeld der Entwicklungszusammenarbeit werden Wirkungen mithilfe von Wirkungsevaluationen untersucht. Wirkungsevaluationen umfassen allerdings mehr als bloßes Messen von Wirkungen, sie beinhalten stets auch nachvollziehbare Bewertungen und Handlungsempfehlungen. Wirkungsevaluationen steigern somit die Transparenz und mithin die Glaubwürdigkeit und Legitimität einer Maßnahme. Darüber hinaus befördern sie den internen Lernprozess auf allen Ebenen, sei es für die Steuerung und Anpassung innerhalb des evaluierten Vorhabens oder auch die Planung und Umsetzung neuer Maßnahmen. Ziel ist immer auch die Qualitätsverbesserung der eigenen Arbeit. Wirkungsevaluationen tragen folglich dazu bei, Maßnahmen effektiver und effizienter zu gestalten und die Wirksamkeit zu erhöhen.

Professor Dr. Alexandra Caspari *lehrt Evaluation, Methoden der empirischen Sozialforschung und Entwicklungspolitik an der Frankfurt University of Applied Sciences. Sie war von 2003 bis 2007 Vorstandsmitglied der DeGEval – Gesellschaft für Evaluation. Seit 2015 ist sie Sprecherin des Arbeitskreises Methoden in der DeGEval.*
E-Mail: caspari@fb4.fra-uas.de

Literatur

ADB – Asian Development Bank: Impact Evaluation. Methodological and Operational Issues. Manila 2006

ALNAP – Active Learning Network for Accountability and Performance in Humanitarian Action: Evaluating humanitarian action using the OECD-DAD criteria. An ALNAP guide for humanitarian agencies. London 2006

American Red Cross et al.: NGO Impact Initiative: An assessment by the international humanitarian NGO community. In: http://www.alnap.org/pool/files/900.pdf (veröffentlicht 2006, abgerufen am 26.4.2016)

Appleton, Simon; Booth, David: Combining Participatory and Survey-based Approaches to Poverty Monitoring and Analysis (Background Paper for the Workshop to be held in Entebbe, Uganda, 30 May-1 June 2001). New York und Oakland 2001 (https://www.odi.org/sites/odi.org.uk/files/odiassets/publications-opinion-files/2185.pdf; abgerufen am 26.4.2016)

Baker, Judy L.: Evaluating the Impact of Development Projects on Poverty: a Handbook for Practitioners. Washington D.C. 2000

Bamberger, Michael (ed.): Integrating Quantitative and Qualitative Research in Development Projects. Washington D.C. 2000

Bamberger, Michael: Conducting Quality Impact Evaluation Under Budget, Time and Data Constraints. Washington D.C. 2006

Bloom, Howard S.: The Core Analytics of Randomized Experiments for Social Research (MDRC Working Papers on Research Methodology). Ohne Ortsangabe 2006

BMZ – Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung: Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen. Bonn 2006

Canteli, Covadonga: A proposal of mixed methods approach to impact evaluations (report on DARA's mixed methods approach). In: http://daraint.org/wp-content/uploads/2013/07/DARA-Impact-Methods.pdf (veröffentlicht 2013, abgerufen am 26.4.2016)

Caspari, Alexandra: Evaluationen der Nachhaltigkeit von Entwicklungszusammenarbeit. Zur Notwendigkeit angemessener Konzepte und Methoden. Wiesbaden 2004

Caspari, Alexandra: Partizipative Evaluationsmethoden – Zur Entmystifizierung eines Begriffs in der Entwicklungszusammenarbeit. In: Flick, Uwe (Hrsg.): Qualitative Evaluationsforschung. Reinbek 2006

Caspari, Alexandra: (Rigorous) Impact Evaluations – Eine nicht nur für die Entwicklungszusammenarbeit relevante internationale Diskussion. In: Zeitschrift für Evaluation 1/2008, S. 137-141

Caspari, Alexandra: ‚Rigorose' Wirkungsevaluation – methodische und konzeptionelle Ansätze der Wirkungsmessung in der Entwicklungszusammenarbeit. In: Zeitschrift für Evaluation 2/2009, S. 183-213

Caspari, Alexandra: Chancen der Wirkungsorientierung für die entwicklungspolitische Bildungsarbeit. In: Zeitschrift für internationale Bildungsforschung und Entwicklungspädagogik 2/2012, S. 11-17

Caspari, Alexandra; Barbu, Ragnhild: Wirkungsevaluierungen/(Rigorous) Impact Evaluations. Zum aktuellen Stand der Diskussion und dessen Relevanz für deutsche EZ-Evaluierungen (BMZ Evaluation Working Papers). Bonn 2008

CGD – Centre for Global Development: When Will We Ever Learn? Improving Lives through Impact Evaluation. Washington D.C. 2006

Chung, Kimberly: Qualitative data collection techniques. In: Grosh, M.; Glewwe, P. (eds.): Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of Living Standards Measurement Study (vol. 2). Washington D.C. 2000

DeGEval – Gesellschaft für Evaluation e.V.: Standards für Evaluation. Mainz 2008

DG ECHO – Generaldirektion Humanitäre Hilfe und Katastrophenschutz der Europäischen Kommission (Hrsg.): Evaluation of Humanitarian Aid by and for NGOs. A guide with ideas to consider when designing your own evaluation activities. Ohne Ortsangabe 2007

DWHH – Deutsche Welthungerhilfe: Leitfaden Wirkungsorientierung in den Projekten und Programmen der Welthungerhilfe. Teil I: Hintergründe und Definitionen. Bonn 2008

Ezemenari, Kene; Rudqvist, Andres; Subbarao, Kelanidhi: Impact Evaluation: A Note on Concepts and Methods. Washington D.C. 1999

Guijt, Irene: Methodological issues in participatory monitoring and evaluation. In: Estrella, Marisol (ed.): Learning from Change. Issues and Experiences in Participatory Monitoring and Evaluation. London 2000

Kassam, Yusuf: Combining participatory and survey methodologies in evaluation: the case of a rural development project in Bangladesh. In: Jackson, E.T.; Kassam, Y. (eds.): Knowledge Shared: Participatory Evaluation in Development Cooperation. West Hartford 1998

Leeuw, Frans; Vaessen, Jos: Impact Evaluations and Development. NONIE Guidance on Impact Evaluation (Draft Version for Discussion at the Cairo conference March-April, 2009). In: http://siteresources.worldbank.org/EXTOED/Resources/nonie_guidance.pdf (veröffentlicht 2009, abgerufen am 26.4.2016)

Massing, Armin; Rosen, Andreas; Struck, Gabi (Hrsg.): Wirkt so. Handreichung zur Wirkungsorientierung und Antragstellung in der entwicklungspolitischen Inlandsarbeit. Berlin 2010

Mertens, Donna M.: Research methods in education and psychology: Integrating diversity with quantitative and qualitative approaches. Thousand Oaks 1998

Nelson, Jodi: Program Evaluation: Are we ready for RCTs? In: Monday Developments 3/2008, pp. 28-29

OECD; DAC: Principles for Evaluation of Development Assistance. Paris 1991

OECD; DAC: Guidance for Evaluating Humanitarian Assistance in Complex Emergencies (Evaluation and Aid Effectiveness, No. 1). Paris 1999

OECD; DAC: DAC Criteria for Evaluating Development

Assistance. Paris 2000
PHINEO (Hrsg.): Kursbuch Wirkung. Das Praxishandbuch für alle, die Gutes noch besser tun wollen. Berlin 2013
PHINEO (Hrsg.): Wirkungstransparenz bei Spendenorganisationen 2014 (Studie). Berlin 2014
Prowse, Martin: Aid effectiveness: the role of qualitative research in impact evaluation (Background Note December 2007). London 2007 (https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/600.pdf; abgerufen am 26.4.2016)
Ravallion, Martin: Evaluating Anti-Poverty Programs. Washington D.C. 2005
Rossi, Peter H.; Freeman, Howard E.; Hofmann, Gerhard: Programm-Evaluation. Einführung in die Methoden angewandter Sozialforschung. Stuttgart 1988
Schnell, Rainer; Hill, Paul; Esser, Elke: Methoden der empirischen Sozialforschung. München und Wien 1999
Stiftung Zewo (Hrsg.): Wirkungsmessung für NPO. Zewo-Leitfaden für Dienstleistungen und Projekte im Inland. Zürich o.J. (http://www.zewo.ch/wirkunginland/deutsch/Dokumente/Leitfaden_komplett; abgerufen am 26.4.2016)
Stiftung Zewo (Hrsg.): Wirkungsmessung für NPO im Gesundheits-, Sozial- und Umweltbereich. Broschüre zum Zewo-Leitfaden für Dienstleistungen und Projekte im Inland. Zürich 2013 (http://www.zewo.ch/wirkunginland/deutsch/Dokumente/Broschuere; abgerufen am 26.4.2016)
VENRO – Verband Entwicklungspolitik und Humanitäre Hilfe deutscher Nichtregierungsorganisationen (Hrsg.): Humanitarian aid put to the test. Principles, criteria and indicators to ensure and monitor quality in humanitarian aid (VENRO Working Paper No. 14). Bonn 2005
White, Howard: Impact Evaluation. The Experience of the Independent Evaluation Group of the World Bank. Washington D.C. 2006

VON ZAHLEN UND DEM, WAS WIRKLICH ZÄHLT | Ein kritisches Essay

Wilfried Vyslozil

Von Hilfswerken wird immer öfter die Dokumentation ihrer Wirkung verlangt. Der Mensch und seine Werte geraten dabei aus dem Blick.

Zusammenfassung | Der Beitrag setzt sich kritisch mit den jüngsten, wirkungsorientierten Bewertungen von Spendenorganisationen in Ranglisten auseinander. Anhand einiger Beispiele von erfolgreichen Biografien von Kindern aus den SOS-Kinderdörfern wird darüber hinaus gezeigt, dass nicht alle wirksamen Hilfen mit standardisierten Messungen erfasst werden können.

Abstract | The article criticizes recent impact related assesments of charity organisations in rankings. In addition and on the base of some examples of successful biographies of children who grew up in SOS Children's Villages the author explains that not all effektive help can be measured by standard methods.

Schlüsselwörter ▶ SOS Kinderdorf e.V. ▶ Jugendhilfe ▶ Spendenwesen ▶ Evaluation ▶ Vergleich

Eigentlich haben wir gar nicht so schlecht abgeschnitten: In einem jüngst auf Spiegel online veröffentlichten Ranking von großen, in Deutschland ansässigen Hilfswerken erhielten die SOS-Kinderdörfer weltweit 4,17 von fünf möglichen Sternen und landeten damit ziemlich weit vorne. Trotzdem sind solche Listen ärgerlich. Sie reduzieren die unterschiedlichste Arbeit – von denen jede für sich äußerst komplex ist – auf eine simple Kennzahl, die Vergleichbarkeit unterstellt und Empfehlungscharakter hat: Je mehr Sterne, desto besser wird deine Spende eingesetzt.

Der jährlich fast fünf Milliarden Euro schwere deutsche Spendenmarkt ist hart umkämpft und eben deshalb muss jede spendengestützte Organisation Wert darauf legen, in solchen Listen weit oben zu landen. Das erwähnte Fünf-Sterne-Ranking wurde