Bundesministerium für
wirtschaftliche Zusammenarbeit
und Entwicklung

BMZ Evaluation Division: Evaluation Working Papers

Wirkungsevaluierungen

Zum Stand der internationalen Diskussion und dessen Relevanz für die Evaluierung der deutschen Entwicklungszusammenarbeit

VORWORT

Die vom Referat „Evaluierung der Entwicklungszusammenarbeit, Außenrevision“ des Bundesministeriums für wirtschaftliche Zusammenarbeit (BMZ) herausgegebenen *Evaluation Working Papers* behandeln methodische und konzeptionelle Fragen im Zusammenhang mit Evaluierungen des BMZ oder generelle Fragen der Methodik und Herangehensweisen von Evaluierungen. Sie dienen daher als Referenzdokumente für Evaluierungsberichte und als Diskussionsbeitrag für die Fachöffentlichkeit. Wie auch die Evaluierungsberichte geben Sie die Meinung der Autoren und nicht unbedingt die des BMZ wieder. Die deutschen und englischen Beiträge sind in einer Reihe zusammengefasst.

Der vorliegende Beitrag befasst sich mit „Wirkungsevaluierungen“. Der Begriff ist schillernd, das Ziel ist jedoch klar: durch eine nachvollziehbare Methodik belastbare Aussagen zu Wirkungen der Entwicklungszusammenarbeit treffen zu können. Dies ist nicht selbstverständlich. In aller Regel können im Rahmen von Evaluierungen Wirkungen nur grob abgeschätzt werden, da Daten nicht in ausreichendem Maße vorliegen oder die Art der Entwicklungsmaßnahmen keinen Beleg von Wirkungen im strengen Sinne zulässt. Dies ist in vielen Fällen auch ausreichend, manchmal jedoch nicht gut genug. Der Beitrag von **Alexandra Caspari** und **Ragnhild Barbu**, Centrum für Evaluation (CEval), Universität des Saarlandes, befasst sich mit der neuen Methodendiskussion, die strenge Kriterien an den Nachweis von Wirkungen anlegt, nennt Beispiele und kommt zu Empfehlungen, wann diese aufwändigeren Untersuchungen sinnvoll sein können.

Das Arbeitspapier sollte wie folgt zitiert werden: Caspari, A. u. Barbu, R. (2008): Wirkungsevaluierungen: Zum Stand der internationalen Diskussion und dessen Relevanz für Evaluierungen der deutschen Entwicklungszusammenarbeit. *Evaluation Working Papers*. Bonn: Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung.

BMZ-Referat „Evaluierung der Entwicklungszusammenarbeit, Außenrevision“
Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung

INHALT

ABKÜRZUNGSVERZEICHNIS

ADB	Asian Development Bank
AfDB	African Development Bank
AusAID	Australian Government's overseas aid program
AV	abhängige Variable – zu erklärende Größe
BMZ	Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
CBA	Cost-Benefit Analysis
CGD	Center for Global Development
CEval	Centrum für Evaluation
CIDA	Canadian International Development Agency
Danida	Ministry of Foreign Affairs of Denmark
DFID	UK Department for International Development
DIME	Development Impact Evaluation Initiative
EES	European Evaluation Society
EZ	Entwicklungszusammenarbeit
Finnida/FORMIN	Ministry of Foreign Affairs of Finland
FZ	Finanzielle Zusammenarbeit
GBS	General Budget Support
GIS	Geographisches Informationssystem
IADB/OVE	Inter-American Development Bank/Office of Evaluation and Oversight
IE	Impact Evaluation / Wirkungsevaluierung
IEG	Independent Evaluation Group der Weltbank
IFAD	International Fund for Agricultural Development
IFPRI	International Food Policy Research Institute
IOB	Policy and Operations Evaluation Department, Netherlands Ministry of Foreign Affairs
JBIC	Japan Bank for International Cooperation
JICA	Japan International Cooperation Agency
JPAL	Abdul Latif Jameel Poverty Action Lap
KG	Kontrollgruppe
M&E	Monitoring und Evaluation
NGO, NRO	Nichtregierungsorganisation
NONIE	Network of Networks Impact Evaluation Initiative
NORAD	Norwegian Agency for Development Cooperation
NT	Non-Treated / "Nicht-Teilnehmer/innen"; Vergleichs-/Kontrollgruppe
OECD/DAC	Organisation for Economic Co-Operation and Development/Development Assistance Committee
ODI	Overseas Development Institute
PRA	Participatory Rural/Rapid Appraisal – Participatory Research Approach
PROGRESA	Programa de Educación, Salud y Alimentación / Education, Health, and Nutrition Program of Mexico
PSM	Propensity Score Matching

RCT	Randomized Controlled Trial / (randomisiertes) Kontrollgruppen-Design
RIE	Rigorous Impact Evaluation / rigorose Wirkungsevaluierung
RRA	Rapid Rural Appraisal
PREM	Poverty Reduction and Economic Management
Sida	Swedish International Development Cooperation Agency
SWAp	Sector Wide Approaches / Sektorweite Ansätze
T	Treated / Teilnehmer/innen einer Maßnahme
TBE/TBA	Theory-Based Evaluation/Approaches / theoriebasierte Evaluation/Ansätze
TOC	Theory of Change / Theorie der Veränderung
TZ	Technische Zusammenarbeit
VG	Vergleichsgruppe
UNEG	United Nations Evaluation Group
UV	unabhängige Variable – erklärende Größe
WB	Weltbank
ZG	Zielgruppe

1. ANLASS UND ZIELE DER STUDIE

Die Entwicklungszusammenarbeit (EZ) muss vor dem Hintergrund steigenden Mitteleinsatzes mehr denn je (positive) Wirkungen in den Partnerländern nachweisen. Die in der deutschen EZ-Evaluierung übliche methodische Vorgehensweise (Dokumentenanalyse, Experteninterviews, etc.) scheint dazu nur teilweise geeignet. So kommt etwa der DAC Peer Review zu dem Schluss, dass „Deutschland (…) nur über begrenzte Fähigkeiten zur Beobachtung und aussagekräftigen Berichterstattung über die Wirkung der Entwicklungszusammenarbeit [verfügt]" (DAC 2005: 81). Andere Geberorganisationen und Finanzinstitutionen verwenden (bei ähnlicher Ausgangslage wie in Deutschland) zunehmend anspruchsvollere Methoden, um die Wirkungen ihrer Maßnahmen zu untersuchen. Die deutsche Entwicklungszusammenarbeit hat hier Nachholbedarf, auch um international sprechfähig zu sein.

Vor diesem Hintergrund wurde das Centrum für Evaluation (CEval), Universität des Saarlandes, mit der Forschungsstudie: „Wirkungsevaluierungen – Zum aktuellen internationalen Stand der Diskussion" vom BMZ beauftragt. Übergeordnetes Ziel ist, die in der deutschen EZ Evaluierung angewandten Methoden auf den Stand des sich international herausbildenden, neuen Anspruchsniveaus zu bringen, Wirkungen fundiert zu belegen. Unmittelbares Ziel der vorliegenden Studie ist es, die internationale Diskussion zum Thema Wirkungsevaluierungen / (Rigorous) Impact Evaluation für den deutschen EZ-Gebrauch zu rezipieren.

Der vorliegende Bericht beruht auf einer Literatur- und Dokumentenanalyse: Zum einen wurden relevante Veröffentlichungen hinsichtlich Gemeinsamkeiten und Unterschiede in den Darstellungen zu Impact Evaluation untersucht. Zum anderen wurden diverse als Impact Evaluation bezeichnete Studien internationaler Geberorganisationen systematisch analysiert und ausgewertet. Grundlage waren insbesondere 29 auf der NONIE-Homepage von den Mitgliedern selbst als Impact Evaluation aufgeführte Studien (vgl. http://www.worldbank.org/ieg/nonie/database.html), darüber hinaus wurden weitere relevante IE Berichte erfasst.[1]

[1] Eine Übersicht der berücksichtigten Berichte findet sich im Anhang. Bei den im Text zitierten NONIE-Dokumenten handelt es sich um *Textentwürfe* zu verschiedenen Themenkomplexen aus 2007 und 2008, die zum Zeitpunkt der Berichtserstellung im Internet verfügbar waren http://www.worldbank.org/ieg/nonie/members.html [01/2008].

2. DER HINTERGRUND – THE EVALUATION GAP

Die Millenniums-Erklärung und die Millenniumsentwicklungsziele der Vereinten Nationen sowie die Erklärung von Paris über die Wirksamkeit der Entwicklungszusammenarbeit sind auch für EZ-*Evaluationen* von Relevanz: International richten politische Entscheidungsträger ihre Aufmerksamkeit mehr und mehr auf Wirkungen. In diesem Kontext ist es die Aufgabe von Evaluationen, zuverlässige Ergebnisse über die Wirksamkeit von Projekten und Programmen zur Verfügung zu stellen, die Rechenschaftslegung aber auch Lernen ermöglichen. Vor diesem Hintergrund ist die bereits Ende der 1980er Jahre geführte Diskussion um das damals von Mosley identifizierte Mikro-Makro-Paradox erneut aktuell geworden: Während Evaluationen von Projekten und Programmen tendenziell positive Ergebnisse auf der Mikroebene aufzeigen, sind auf der Makroebene kaum Wirkungen der Zusammenarbeit auf die Entwicklung der Partnerländer i.S.v. Armutsreduzierung, Wachstum des Pro-Kopf-Einkommens etc. nachweisbar. Wurde damals die Ursache der (positiv) verzerrten Ergebnisse von Projektevaluationen auf Seiten der Evaluatoren/innen bzw. der Organisationen vermutet[2], wird die Ursache heute eher bei den Projektevaluationen selbst gesehen: Trotz aller Evaluationsanstrengungen wurden schlichtweg kaum Projektevaluationen durchgeführt, die die Frage nach der *Wirksamkeit* von Maßnahmen, die Frage „Welche Veränderung hat die Maßnahme bewirkt?“ oder vereinfacht „Was funktioniert, unter welchen Bedingungen?“ beantworten konnten. Diese *„Evaluation Gap“* wurde in den letzten Jahren international zunehmend wahrgenommen.

Das Center for Global Development (CGD), ein gemeinnütziger „Think Tank“ mit Sitz in Washington D.C. berief daher 2004 die “Evaluation Gap Working Group” mit dem Auftrag zu untersuchen, warum nur derart wenige (echte) Wirkungsevaluationen/Impact Evaluations[3] existieren, und um praktische Empfehlungen zu erarbeiten, wie dieses Problem gelöst werden kann.[4] Im Mai 2006 legte die Arbeitsgruppe ihren Bericht „When Will We Ever Learn? Improving Lives Through Impact Evaluation“ vor, mit der einleitenden Feststellung: ”Yet after decades in which development 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: 1). Es werden *zwei zentrale Ursachen der Evaluation Gap* ausgemacht – die Quantität und Qualität von Impact Evaluationen:

Die Evaluation Gap – eine *quantitative* und eine *qualitative* Lücke: Es werden zu wenige Evaluationen mit Fokus auf Outcomes oder Impacts durchgeführt und wenn sind diese häufig methodisch nicht ausreichend.

(1) Zwar werden teilweise erhebliche Summen für Monitoring und auch Evaluationen von Maßnahmen ausgegeben, diese untersuchen jedoch meist nur Inputs (Ressourcen) und Outputs (Leistungen). Impact Evaluations in dem hier verstandenen Sinne, die die Outcomes (direkte Wirkungen) und Impacts (entwicklungspolitische Wirkungen) analysieren und dokumentieren, existieren dagegen kaum.

[2] Als Grund für diese Verzerrung ("bias") der Ergebnisse („erwünschte“ Ergebnisse) wurden zum einen die Organisationen selbst identifiziert, die ihre Maßnahmen möglichst positiv darstellen wollten, zum anderen die „abhängigen“ Evaluator/innen, die weitere Aufträge nicht gefährden oder als übermäßig kritisch gelten wollten.

[3] Im Folgenden wird der international gängige Begriff Impact Evaluation (IE), der synonym zu Wirkungsevaluierung bzw. Wirkungsevaluation zu verstehen ist, genutzt.

[4] Die Aufgabenstellung fokussierte auf Programme zur sozialen Entwicklung, insbesondere zu Gesundheit und Bildung.

(2) Die wenigen durchgeführten Impact Evaluations sind häufig methodisch unzureichend und somit irreführend. Die Ergebnisse basieren meist auf Informationen, die ausschließlich bei den Projekt-/Programm-Beteiligten erhoben wurden, so dass Projektwirkungen überschätzt wurden. „Echte“ Wirkungsevaluationen, die analysieren inwieweit Veränderungen auf Seiten der Zielgruppe einer bestimmten Maßnahme eindeutig *zugeschrieben* (attribuiert) werden können, müssen der Frage nachgehen, was ohne Programm geschehen wäre, d.h. das Kontrafaktische/Counterfactual berücksichtigen. Dies erfordert härtere/rigorosere Methoden.

Die Evaluation Gap Working Group definiert Impact Evaluations demnach als “studies that measure the impact directly attributable to a specific program or policy, as distinct from other potential explanatory factors” (CGD 2006: 10).

Impact Evaluationen messen die Wirkungen, die einer bestimmten Maßnahme *direkt zugeschrieben* werden können – in Abgrenzung zu anderen möglichen Erklärungsfaktoren.

Dem Argument, IE seien insbesondere in Relation zu der eigentlichen Maßnahme zu kostspielig oder kompliziert, wird entgegengehalten, dass als Vergleich nicht die Kosten der Maßnahme zugrunde gelegt werden müssten sondern der Wert des generierten Wissens an sich, ein Wert der bemisst, dass Schaden verhindert wird und mehr Menschen mit bewährten Programmen erreicht werden. Denn “ignorance is more expensive than impact evaluations“ – so die Evaluation Gap Working Group in ihrem Bericht (2006: 23ff.). Zwar benötigen IE ein durchaus größeres Budget als übliche Projektevaluationen, sie sind allerdings aufgrund ihrer enormen Wissensgenerierung als öffentliches Gut anzusehen. Denn, wenn auch die Kosten für eine IE von einer einzelnen Organisation zu tragen sind, so können die Ergebnisse nach ihrer Veröffentlichung von allen genutzt werden, um ihre Politiken zu verbessern.

Empfehlung der Evaluation Gap Working Group für eine “Collective Action”: Commitments to Public Good through a New Council.

Entsprechend kommt die Working Group zu ihrer zentralen Empfehlung einer “*Collective Action”*: “(…) the full range of stakeholders – NGOs, foundations, research centers, bilateral agencies, developing country governments, and multilateral development banks – should both reinforce existing initiatives and collaborate on a new set of actions to promote more and better impact evaluations” (CGD 2006: 4). Der Wert der Aktivitäten einer einzelnen Organisation könnte vervielfacht werden, wenn diese durch *kollektive* Maßnahmen ergänzt würden. So können z.B. *gemeinsam* Fragen von allgemeinem Interesse identifiziert und nach dringlichen Themen zusammen gefasst werden; wenn sichergestellt wird, dass IE reliabel und valide durchgeführt, gesammelt und verteilt werden, und so werden auch entsprechende Kompetenzen zur Durchführung von IE in den Partnerländern aufgebaut. Daher wird konkret vorgeschlagen, ein *gemeinsames Gremium oder einen Rat* zu institutionalisieren, wobei jede Organisation sich vertraglich verpflichtet, einen Teil dieser Aufgaben zu schultern, so dass die Kosten für IE verteilt würden und nicht einzelne Organisationen die vollen Kosten für die Wissensgenerierung tragen müssten, die andere nutzen (vgl. CGD 2006: 34ff.).

3. NONIE – NETWORK OF NETWORKS IMPACT EVALUATION INITIATIVE

NONIE – Das Network of Networks Impact Evaluation Initiative: Ein Zusammenschluss von Geberorganisationen und multilateralen Banken über die drei Netzwerke
- DAC Evaluation Network
- UN Evaluation Group
- Evaluation Cooperation Group

Vor diesem Hintergrund schlossen sich im November 2006 die Netzwerke von Evaluierungseinheiten bilateraler EZ-Organisationen, der UN-Organisationen sowie der multilateralen Entwicklungsbanken[5] zu einem „Network of Networks Impact Evaluation Initiative“ (NONIE) zusammen, um die Effektivität der Entwicklungszusammenarbeit gemeinsam zu verbessern, indem nützliche und relevante, qualitativ hochwertige IE vorangebracht werden. Waren zuerst lediglich Vertreter/innen dieser drei Netzwerke beteiligt (ca. 50 Personen aus 31 Organisationen), ist die Mitgliederzahl seither gestiegen; auch Mitglieder regionaler und internationaler Evaluationsgesellschaften, wie auch Vertreter/innen vieler Partnerländer, sind heute einbezogen.

Ziel von NONIE ist, – auf Basis eines gemeinsamen Verständnisses von IE sowie der Vorgehensweise bei der Durchführung – ein Programm für Aktivitäten im Bereich Wirkungsevaluation zu fördern. In der Hauptsache gilt es, ein allgemeines Verständnis von IE voranzutreiben, entsprechende Ansätze zur Durchführung zu fördern und Kooperationen aufzubauen.

Entsprechend wurden drei *zentrale Aufgaben* formuliert:

(1) Erarbeitung von Richtlinien und Leitfäden für IE,

(2) Verständigung auf gemeinsame Maßnahmen zur Durchführung von IE, was das benannte Programm einleiten soll,

(3) Entwicklung einer Ressourcen-Plattform, um IE zu fördern.

Für organisatorische Zwecke wurde in einem ersten Schritt ein Sekretariat eingerichtet, das bei der Independent Evaluation Group (IEG) der Weltbank (WB) angesiedelt ist, die einen Mitarbeiter für die Tätigkeiten zur Verfügung stellt. [6]

Bisher finden sich auf der Internetseite von NONIE (http://www.worldbank.org/ieg/nonie) relevante Ressourcen (zentrale Veröffentlichungen zu IE) sowie Informationen zu den laufenden Aktivitäten. Des Weiteren listet eine Datenbank insgesamt 29 von den Mitgliedern als Impact Evaluation erachtete Studien internationaler Geberorganisationen auf.

Ausgehend von einer detaillierten Analyse dieser Studien[7] und der diversen Arbeitspapiere der Sub-Groups wird im Folgenden der Stand der Diskussion zum Thema IE dargestellt, wobei auch weitere zentrale Veröffentlichungen berücksichtigt werden. Es ist wichtig anzumerken, dass zum einen die Diskussion innerhalb NONIE noch nicht abgeschlossen ist, und zum anderen höchst unterschiedliche Standpunkte vertreten sind. Dies wird in der folgenden Ausführung berücksichtigt.

[5] Im Einzelnen: das DAC Evaluation Network, die UN Evaluation Group (UNEG) sowie die Evaluation Cooperation Group (ECG).
[6] Um die gesetzten Aufgaben zu bearbeiten, wurden sechs „Sub-Groups“ implementiert (vgl. im Detail Anhang 2), drei Sub-Groups sind für einzelne Teile eines zu erarbeitenden Leitfadens (Guidelines) zuständig.
[7] Von den 29 auf der Internetseite aufgeführten Studien waren lediglich 24 online oder als "Hardcopy" erhältlich.

4. ZUM BEGRIFF (RIGOROUS) IMPACT EVALUATION / WIRKUNGSEVALUIERUNG

International wird der Begriff Impact Evaluation vielfältig genutzt mit meist unterschiedlichen Bedeutungen, z.B. synonym für Sektor- oder Länderevaluationen oder auch ex-post Evaluationen (vgl. Abbildung 1).

Abbildung 1: Unterschiedliche Bedeutungen des Begriffs Impact Evaluation im internationalen Kontext

- evaluations that look at the impact of an intervention on *final welfare outcomes*, rather than only at project outputs, or process evaluation that focus on implementation;
- evaluations that are concerned with establishing the *counterfactual*, i.e. the difference projects or programmes made (how indicators behaved with the project compared to how they would have been without it);
- evaluations carried out some time (five to ten years) after the *interventions* have been completed so as to allow time for impact to appear;
- evaluations considering all interventions *within a given sector or geographical area*; and
- studies concerned with environmental or social effects, which are often *ex ante* and so not in fact evaluations.

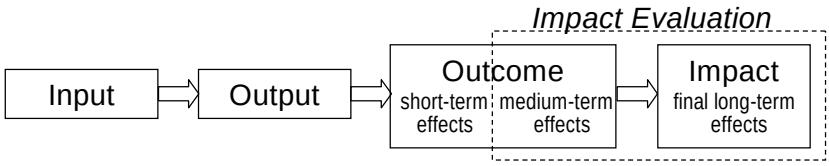
These different forms are not mutually exclusive and may overlap.

Quelle: NONIE Workshop Dokumente 2008: Impact Evaluation Guidance; White 2006a, 2006b

Der Begriff Impact Evaluation beinhaltet für NONIE demgegenüber *zwei zentrale Elemente*:

(1) Die *Zielebene* von Impact Evaluationen: Die Betonung liegt dezidiert auf „*Impact*" gemäß der Definition des OECD/DAC. Demnach sind Impacts "positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (OECD/DAC 2002: 24).[8] Somit untersuchen IE weitaus mehr, als lediglich die Zielerreichung oder die kurzfristigen Effekte einer Maßnahme auf die Zielgruppe. Die interessierenden Ebenen von Impact Evaluationen sind vielmehr die *mittelfristigen, direkten sowie die langfristigen, übergeordneten entwicklungspolitischen Wirkungen* einer Maßnahme:

Abbildung 2: Zielebenen von Impact Evaluationen



In Anlehnung an NONIE SG1 2007: 1; NONIE 2008: 2

(2) Die Berücksichtigung des *Kontrafaktischen/"Counterfactual"*: Ausgehend von dieser Impact-Definition wird die Betonung auf „effects *produced by*" gelegt. D.h. IE analysieren die Effekte einer Maßnahme, die nicht beobachtbar gewesen wären, wenn die Maßnahme nicht durchgeführt worden wäre. Dies impliziert zwangsläufig auch die Frage „Was wäre ohne Maßnahme gewesen?", also die Berücksichtigung des Kontrafaktischen (counterfactual).

[8] Auf diese Definition berufen sich viele Organisationen (gemäß eigener Angaben), z.B. ADB, BMZ, DFID, IEG, IFAD, OED.

Bezüglich Punkt (2) ist anzumerken, dass innerhalb von NONIE bisher ungeklärt ist, inwieweit das Kontrafaktische als notwendiges, zentrales Element in die Definition von IE einfließen soll. Zentraler Diskussionspunkt ist die Frage, ob das Kontrafaktische *explizit* oder *implizit* berücksichtigt werden soll.[9] Diese Diskussion geht einher mit der Definition des Begriffs *„Rigorous Impact Evaluation"* (RIE):

„Rigorous Impact Evaluations" nutzen anspruchsvolle bzw. "harte" (rigorous) Methoden, um systematisch das Problem der Wirkungs-*zuschreibung* anzugehen.

Aufgrund der oben dargestellten vielfältigen Bedeutungen des Begriffs „Impact Evaluation" wurde es notwendig, das hier zugrunde gelegte Verständnis von IE gegenüber diesen abzugrenzen – insbesondere um hervorzuheben, dass IE immer auch spezifische methodologische Ansätze beinhalten müssen, die es erlauben, *systematisch* das Problem der Wirkungs*zuschreibung* ("attribution") anzugehen. Entsprechend sind anspruchsvolle bzw. „harte" („rigorous") Methoden notwendig – solche Ansätze werden zur Abgrenzung häufig „Rigorous Impact Evaluation" genannt. Der Begriff „rigorous" beschränkt sich jedoch *nicht* ausschließlich auf randomisierte Kontrollgruppen-Designs (RCTs, Randomised Controlled Trials): "The term 'rigorous' is not being restricted to the quantitative approaches presented in section 2 of this Guidance, but applies to any approach which systematically tackles the problem of attributing impact"[10] (NONIE 2008: 2).

5. WIE WIRD DAS KONTRAFAKTISCHE BERÜCKSICHTIGT? – RELEVANTE DESIGNS ZUR SYSTEMATISCHEN WIRKUNGSMESSUNG

Bei der Berücksichtigung des Kontrafaktischen werden die tatsächlichen Wirkungen einer Maßnahme isoliert, indem die tatsächlich beobachteten Wirkungen bei der Zielgruppe mit den kontrafaktischen Wirkungen verglichen werden, d.h. den hypothetischen Veränderungen, die auch ohne Maßnahme eingetreten wären. Da Personen entweder Begünstigte oder „Nicht-Begünstigte" einer Maßnahme sein können, aber niemals beides gleichzeitig, kann dieser hypothetische Kontrafakt nicht *direkt* beobachtet werden, sondern wird stattdessen *geschätzt*.

Bisher wurde dies in der Evaluierung der EZ mittels eines *vorher-nachher Vergleichs* bei der Zielgruppe bzw. den Teilnehmer/innen einer Maßnahme gelöst, was auch als reflexiver Vergleich / "reflexive comparison" bezeichnet wird (vgl. (a), Abb. 3): Baselinedaten (erhoben *vor* der Maßnahme) werden mit Daten aus einer Erhebung (Survey) *nach* der Maßnahme verglichen, wobei die aufgefundenen Veränderungen der Maßnahme zugeschrieben werden. Allerdings sind diese beobachteten Veränderungen selten *allein* sondern nur *teilweise* auf die Maßnahme zurückzuführen. Andere, externe Faktoren, wie z.B. Maßnahmen anderer Geber aber auch unerwartete Ereignisse (Naturkatastrophen, Kriege, etc.) oder auch allgemeine Veränderungsprozesse (Wirtschaftswachstum/-krise, Verstädterung, etc.) können die Wirkung der Maßnahme beeinflussen, schwächen oder auch verstärken. Gleichwohl ist anzunehmen, dass sich die Situation der ZG auch ohne Maßnahme verändert hat. Solche Faktoren, die teilweise oder ganz für die beobachtete Veränderung verantwortlich sein können, bleiben bei einem reinen vorher-nachher Vergleich der Zielgruppe unberücksichtigt. In Afrika hat sich z.B. gezeigt,

[9] Eine ausführliche Darstellung der Diskussion findet sich in Kapitel. 8 und 9.
[10] Zur näheren Beschreibung der verschiedenen Designs siehe folgendes Kapitel.

dass die Sterblichkeitsrate der unter 5-Jährigen aufgrund von HIV/AIDS trotz zunehmender Immunisierungsrate und Zugang zu sauberem Wasser gestiegen ist (vgl. White 2006a: 3).

Abbildung 3: In der EZ bisher genutzte Versuchsanordnungen[11]

	DESIGN	Vorher-Daten t_1 (Baseline)	Maß-nahme *X*	Nachher-Daten t_2 (Survey)
Vorexperimentelle Versuchsanordnung:				
(a)	Ein-Gruppen-Vortest-Nachtest-Design	**ZG_{t1}**	**X**	**ZG_{t2}**
(b)	Ein-Gruppen-Nachtest-Design		**X**	**ZG_{t2}**

ZG: Zielgruppe, *t:* Zeitpunkt

Ein vorher-nachher Vergleich zeigt lediglich die Entwicklung der ZG über die Zeit hinweg auf – das Faktische ("factual"), nicht aber das Kontrafaktische – und kann demnach höchst selten eine zuverlässige Antwort auf die Frage nach Wirkungen einer Maßnahme geben.

Ein *Vorher-Nachher Vergleich* kann nur selten die Wirkungen, die allein auf eine Maßnahme zurückzuführen sind, zuverlässig abbilden.
Dies wird als *Zuordnungsproblem* ("attribution problem") bezeichnet.

Die eingeschränkte Aussagekraft eines einfachen vorher-nachher Vergleichs illustriert eine Impact Evaluation des IADB Programms „Social Investment Fund" in Panama dar: Während eine "naïve" vorher-nachher Berechnung zu dem Schluss kam, dass die Armut unter den Begünstigten angestiegen und das Programm somit gescheitert sei, zeigten die Impact Berechnungen das Gegenteil, nämlich eine Reduzierung der Armut (vgl. Ruprah 2008: 25f.).

Häufig wird angemerkt, dass auch Maßnahmen denkbar sind, bei denen ein vorher-nachher Vergleich für eine adäquate Wirkungszuschreibung ausreichend erscheint, z.B. sei es *naheliegend*, dass die Installation von individuellen oder auch öffentlichen Wasseranschlüssen die Zeit für das Wasserholen reduziert (vgl. White 2007: 3; NONIE-SG1 2008: 1; NONIE-SG2 2008: 15). Allerdings ist die Reduzierung der Zeit des Wasserholens auch nicht die letztlich interessierende *Wirkungsebene* (vgl. oben Abb. 2). Die Grenzen der Erkenntnisse von nach dieser Logik durchgeführten "Impact" Evaluationen zeigt folgende Studie des finnischen Außenministeriums (Finnida):

> "One of the greatest benefits from the improved services has been more water closer to people which has meant time saving in fetching water. This undoubtedly has directly benefited women and young girls. (…) As water is fetched several times a day, time saving has in some cases been considerable. The surveys did not establish how the time saved in fetching water is used. Be it in household chores, productive activities, social networking, homework for school, or simply in a child's time to play, the workload of women and young girls has become substantially lighter. Time saving is a good example of how the projects/programmes have been able to respond to practical needs of women. However, it has rarely been used in the WWS project as an indicator to measure progress towards the project goals and overall objectives" (Finnida (2): 52f.).

[11] Die in Abb. 3 aufgezeigte Variante (b) – Betrachtung der Zielgruppe nur zu einem Zeitpunkt, nämlich *nach* der Maßnahme – ermöglicht, auch wenn häufig in der EZ angewandt, keinerlei Aussagen über Wirkungen (*Veränderungen* aufgrund der Maßnahme), da das Kontrafaktische hierbei nicht berücksichtigt wird.

Dieses Beispiel verdeutlicht, dass eine solche „naheliegende“ Wirkungszuschreibung auf Basis eines vorher-nachher Vergleichs für eine IE nicht ausreicht: „Zeitersparnis“ entspricht im vorliegenden Beispiel der "Outcome"-Ebene. Eine IE muss jedoch der Frage nachgehen, inwieweit sich hierdurch die Lebensverhältnisse der Begünstigten verbessert haben.

Zentrale Aufgabe einer IE ist aufzuzeigen, in welchem *Umfang* die beobachteten Veränderungen der Maßnahme *eindeutig zugeschrieben* werden können. Dies wird auch als *Zuordnungsproblem* (“attribution problem“) bezeichnet (vgl. NONIE-SG1 2007: 3; 2008: 1; CGD 2006: 29; Bamberger et al. 2006: 16; ADB 2006: 21). Die Lösung des Zuordnungsproblems bzw. die Berücksichtigung des Kontrafaktischen kann durch das Bilden einer *Kontrollgruppe* (KG) erreicht werden, d.h. Individuen, Haushalten, Firmen, etc., die nicht in den Genuss der Maßnahme (der vermuteten Ursache) kamen, ansonsten aber in allen anderen Aspekten identisch mit der ZG sind. Hierdurch wird ein „mit-ohne Vergleich“ (“with and without“) möglich.

"Single-Difference" (SD) misst Wirkungen über einen einfachen Vergleich zwischen dem Faktischen und dem Kontrafaktischen, d.h. zwischen der Zielgruppe und einer Vergleichsgruppe zu einem Zeitpunkt.

Diese *single-difference Methode* (SD) misst Wirkungen über einen einfachen Vergleich zwischen den Teilnehmer/innen (ZG) und Nicht-Teilnehmer/innen (KG) einer Maßnahme, zwischen dem Faktischen und dem Kontrafaktischen (vgl. Baker 2000, White 2006a).[12] Zentrale Grundannahme des single-difference Ansatzes ist, dass die Ausgangssituation der ZG und der KG vor der Maßnahme identisch sind ("same values of outcome"). Diese Annahme ist in der Realität jedoch selten gegeben, da Maßnahmen häufig gezielt für Personen aufgelegt werden, die entweder besondere Defizite aufweisen oder aber besondere Voraussetzungen erfüllen. Dadurch werden bei der single-diffence Methode die berechneten Wirkungen der Maßnahme je nach Situation über- oder unterbewertet (“biased“).

Diesem Problem kann begegnet werden, in dem eine *vorher Messung* sowohl bei der ZG als auch der KG durchgeführt wird bzw. bereits vorliegende Daten für die ZG und KG genutzt werden (Baselinedaten): Die *double-difference Methode* oder auch *Differenzen-in-Differenzen Schätzungen* (DD) kombiniert den mit-ohne Vergleich mit dem vorher-nachher Vergleich. Die Wirkung einer Maßnahme ergibt sich aus dem Unterschied zwischen ZG und KG nach der Maßnahme (t_2) minus dem Unterschied zwischen ZG und KG vor der Maßnahme (t_1) (vgl. Baker 2000: 56; ADB 2006: 13):[13]

"Double-Difference" (DD) misst Wirkungen über die Kombination eines _mit-ohne_ und eines _vorher-nachher_ Vergleiches.

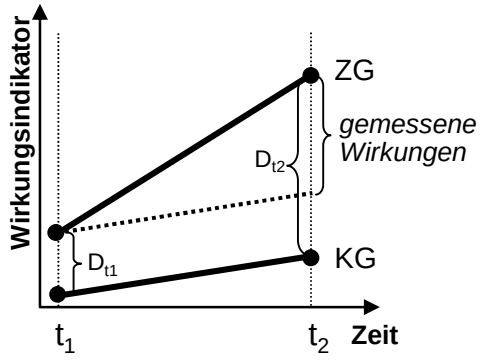
[12] Der single-difference Effekt einer Maßnahme (β_d) bezogen auf die direkte Wirkung (y) entspricht dabei der Differenz der durchschnittlichen Wirkungen (beobachteten Mittelwerten) zwischen der ZG und der KG nach Beendigung der Maßnahme (t_2):

$$\beta_d = \left(\overline{y}_{ZG} - \overline{y}_{KG}\right)$$

[13] „The double difference impact (β_{dd}) of the project is given by the difference between the differences in mean outcomes for control and project areas reported at the end-line and base-line” (White 2006a, 39):

$$\beta_{dd} = \left(\overline{y}_{t_2,ZG} - \overline{y}_{t_2,KG}\right) - \left(\overline{y}_{t_1,ZG} - \overline{y}_{t_1,KG}\right)$$

Abbildung 4: Differenzen-in-Differenzen Schätzung der Wirkungen



ZG: Zielgruppe, KG: Kontrollgruppe,
D: Differenz ZG – KG, *t:* Zeitpunkt

Die Frage ist, *wie* Kontrollgruppen gebildet werden. Hierfür gibt es zwei Möglichkeiten:

(1) Das *Experimentelle Design / "Randomized Controlled Trials" (RCT)*:

Beim experimentellen Design werden Personen (oder Haushalte, Firmen, Gemeinden, etc.) *vor* der Implementation einer Maßnahme nach dem *Zufallsprinzip* (Randomisierung / "randomization") zwei Gruppen zugeordnet. Zum einen eine Gruppe, die an der geplanten Maßnahme teilnimmt also die ZG sein wird (Teilnehmer/innen / "Treated", T), und zum anderen eine *Kontrollgruppe*, die nicht an der Maßnahme teilnehmen wird (Nicht-Teilnehmer/innen / "Non-Treated", NT). Randomisierung bedeutet hierbei nicht, dass die Teilnehmer/innen selbst per Zufall ausgewählt werden, sondern setzt eine Stufe tiefer an: Z.B. wird für eine geplante EZ-Maßnahme nach zuvor festgelegten Kriterien in einem ersten Schritt eine Gruppe potentiell Begünstigter bewusst identifiziert. Diese werden dann zunächst hinsichtlich derjenigen Merkmale oder auch Eigenschaften, die für die Wirksamkeit der Maßnahme als bedeutsam angesehen werden, vorgruppiert, z.B. nach Geschlecht, Alter, Bildung. Danach werden aus diesen Subgruppen per *Randomisierung*, also *per Zufall*, Personen der Gruppe der Teilnehmer/innen bzw. der Kontrollgruppe zugeordnet.

Experimentelle Designs gelten als die rigoroseste Evaluationsmethodologie schlechthin, denn durch den zufallsgesteuerten Auswahlprozess werden systematische Gruppenunterschiede eliminiert, d.h. systematische Auswahlverzerrungen ("selection bias") liegen nicht vor (vgl. ADB 2006: 5f.; Bloom 2006: 1; Baker 2000: 2; NONIE-SG1 2008: 4). Stimmen folglich die Ausgangssituation der ZG und KG vor der Maßnahme (t_1) überein (was bei randomisierter Kontrollgruppe der Fall sein sollte), können alle Unterschiede in den Wirkungen zwischen ZG und der KG zum Zeitpunkt t_2 der Maßnahme zugeschrieben werden (anhand der single-difference Methode). Allerdings trifft dies nur in sogenannten Laborexperimenten zu, die für die EZ nicht relevant sind. In sogenannten *Feldexperimenten*, d.h. die Anwendung einer experimentellen Versuchanordnung in für die Beteiligten authentischer Alltagsumgebung, ist es unmöglich, alle Eigenschaften der Personen vorab zu berücksichtigen, so dass Gruppenunterschiede zwischen ZG und KG nach wie vor gegeben sind. Während die single-difference Methode solche systematischen Gruppenunterschiede nicht berücksichtigt, können diese anhand der *double-difference Methode* erfasst werden: hierbei wird zusätzlich eine vorher Messung sowohl für ZG

als auch KG in das Design aufgenommen. Für die EZ realistische experimentelle Designs (Feldexperimente) haben somit folgende Versuchsanordnungen:

Abbildung 5: Das experimentelle Design

DESIGN		Vorher-Daten t_1 (Baseline)	Maß-nahme *X*	Nachher-Daten t_2 (Survey)
Experimentelle Versuchsanordnung:				
(1)	Kontrollgruppen-Design	ZG_{t1} KG_{t1}	X –	ZG_{t2} KG_{t2}

ZG: Zielgruppe, KG: randomisierte Kontrollgruppe, *t:* Zeitpunkt

Experimentelle Designs werden (auch in den Partnerländern) insbesondere in klinischen Untersuchungen angewandt, z.B. um die Wirkung eines Medikaments oder die Effektivität einer neuen Behandlungsmethode zu prüfen. Im Kontext entwicklungspolitischer Maßnahmen werden experimentelle Designs teilweise aufgrund ethischer und finanzieller Vorbehalte abgelehnt. Bezüglich der Kosten kann entgegnet werden, dass experimentelle Designs nicht teurer als andere Survey-basierte IE sind (vgl. White 2006a: 13).

Aber auch ethische Vorbehalte scheinen nicht grundsätzlich angemessen, da die Mittel für eine Maßnahme oft begrenzt und nicht für alle potentiell Begünstigten verfügbar sind: wenn beispielsweise im Rahmen eines Bildungs-Projektes 5.000 Schulbücher in einer bestimmten Region verteilt werden sollen, in der jedoch 20.000 förderungswürdige Schüler leben; oder bei der Vergabe einer bestimmten Anzahl an Stipendien für Schüler aus Familien mit geringem Einkommen. Verbleiben selbst nach Anwendung bewusster inhaltlicher Kriterien, wie Begabung, Bedürftigkeit, etc. mehr potentielle Begünstigte als Schulbücher oder Stipendien vorhanden sind muss eine Auswahl getroffen werden. Auch in Deutschland wird die Vergabe von Studienplätzen von der Zentralstelle für die Vergabe von Studienplätzen (ZVS) in mehreren Stufen entschieden: bei gleichen Abiturnoten und Wartezeiten entscheidet letztendlich der Zufall (das Los).

Des Weiteren werden manche Maßnahmen derart konzipiert, dass ein experimentelles Design ohne jegliche ethische Bedenken angewandt werden kann. Wenn im Rahmen eines großen Programms mit langer Laufzeit die geplante Maßnahme in *mehreren Phasen* für einzelne Personen (bzw. Regionen, Gemeinden etc.) *zeitversetzt* implementiert wird, und es keine bewusste Entscheidung darüber gibt, warum eine Person an der ersten Phase teilnehmen sollte (andere dagegen an der zweiten oder späteren Phase), dann können die Personen, die erst in einer späteren Phase an der Maßnahme teilnehmen werden, als Kontrollgruppe für die Teilnehmer/innen der ersten Phase herangezogen werden (vgl. NONIE 2007a: 8; Bamberger 2006: 11; White 2006a: 14; Ravaillon 2005: 30; Baker 2000: 73). D.h. wird die Entscheidung, wer an welcher Phase der Maßnahme teilnimmt, per Zufall getroffen, wird bereits bei der *Konzeption* einer Maßnahme ein experimentelles Design angelegt. Später durchgeführte IE können somit auf adäquate Kontrollgruppen zurückgreifen. Diese Vorgehensweise nach dem sogenannten *Pipeline Verfahren* ist nicht nur sehr ökonomisch, sondern liefert äußerst zuverlässige Daten, da keine systematischen Unterschiede in den Charakteristika von Zielgruppe und Kontrollgruppe vorliegen. Ein gutes Beispiel, wie bereits die Konzeption einer EZ-Maßnahme für eine spätere Evaluation eine adäquate Kontrollgruppe ergibt, zeigt die IFPRI-Evaluation „PRO-

GRESA“ (vgl. insb. Behrman/Todd 1999)[14]: In einem ersten Schritt wurden 505 Gemeinden als relevant für die Maßnahme identifiziert.[15] Da nicht alle 505 Gemeinden gleichzeitig an der Maßnahme teilnehmen konnten, wurden die Gemeinden nach dem Zufallsprinzip in zwei Gruppen aufgeteilt: 60% der Gemeinden nahmen sofort an der Maßnahme teil, 40% zwei Jahre später. Diese 40% dienten im Rahmen der IE anschließend als Kontrollgruppe. Weitere Beispiele experimenteller Designs unter Anwendung des Pipeline Verfahrens finden sich bei DFID (2) sowie ADB (2006: 7ff.).

Es ist allerdings zu beachten, dass das Pipeline Verfahren nicht grundsätzlich angewandt werden kann, nur weil eine Maßnahme in verschiedenen Phasen implementiert wurde. Häufig werden die Teilnehmer/innen für die erste Phase der Maßnahme nicht per Zufall sondern *bewusst* aufgrund *bestimmter Kriterien* ausgewählt, z.B. die Gruppe der Ärmsten bzw. Bedürftigsten oder auch die Regionen mit den besten Voraussetzungen. Auch ist es möglich, dass Individuen oder Gruppen *selbst entscheiden*, an einer Maßnahme teilzunehmen (z.B. durch Antragstellung im Rahmen von Kleinkredit-Maßnahmen). In diesen Fällen liegen entsprechend auch entscheidende Gruppenunterschiede vor, so dass die Teilnehmer/innen der weiteren Phasen nicht als Kontrollgruppe herangezogen werden können (vgl. Bamberger 2006: 11).

Es zeigt sich also, dass experimentelle Designs nicht grundsätzlich ethische Probleme aufwerfen. Auch wird deutlich, dass experimentelle Designs bereits durch die Konzeption einer Maßnahme vorgegeben werden, d.h. im Rahmen einer IE ist es nicht möglich, (nachträglich) ein experimentelles Design anzuwenden. Entsprechend sollte bereits bei der Planung von Maßnahmen immer die *Möglichkeit* eines experimentellen Designs überprüft werden (vgl. White 2006a: 13f.). Insbesondere bei Pilotprojekten, Maßnahmen mit begrenzten Ressourcen bei gleichzeitig hohem Bedarf oder auch Maßnahmen, bei denen die Wirksamkeit noch weitestgehend unbekannt ist, liefern experimentelle Designs wichtige, robuste Ergebnisse, so dass auf dieser Erkenntnisbasis größere oder weitere Maßnahmen aufgelegt werden können (vgl. Bamberger 2006: 5). Zwar werden experimentelle Designs im entwicklungspolitischen Kontext durchaus angewandt, die Anzahl in der EZ ist jedoch noch immer sehr gering.

***Experimentelle Designs* liefern wichtige, robuste Ergebnisse. Die *Möglichkeit* eines experimentellen Designs sollte daher immer überprüft werden.**

(2) *Quasi-experimentelle Designs*:[16]

Das Quasi-Experiment orientiert sich an der Experimentallogik, allerdings wird hierbei keine Kontrollgruppe nach dem Zufallsprinzip gebildet, sondern es wird eine *Vergleichsgruppe (VG) (re-) konstruiert*.[17] Insbesondere im Rahmen einer Impact Evaluation ist ein experimentelles Design häufig nicht möglich, da die Evaluatoren/innen meist nicht von

Bei *quasi-experimentellen Designs* wird keine *Kontroll*gruppe gebildet sondern eine *Vergleichs*gruppe *(re-)konstruiert.*

[14] Vgl. auch Evaluationsberichte auf der IFPRI-Homepage: http://www.ifpri.org/themes/progresa/progresa_report.htm

[15] Innerhalb jeder Gemeinde wurden wiederum einzelne Haushalte nach Armutskriterien als Zielgruppe/Begünstigte bestimmt (vgl. Behrman/Todd 199, 1f.).

[16] Quasi-experimentelle Designs werden insbesondere im Englischen teilweise "nonexperimental“ Designs genannt – dies ist jedoch verwirrend, da nicht gleichzusetzen mit vorexperimentellen Designs.

[17] Diese Unterscheidung anhand der Begriffe ist entscheidend: *Kontroll*gruppen (KG) sind immer randomisiert, *Vergleichs*gruppen (VG) dagegen sind nicht-randomisiert.

Beginn an einer Maßnahme in den Prozess eingebunden sind, oder die Maßnahme bereits abgeschlossen wurde – eine Randomisierung folglich nicht mehr möglich ist. Im Unterschied zum experimentellen Design erfolgt die Bestimmung der Vergleichsgruppen daher beim quasi-experimentellen Design im *Nachhinein*, d.h. entweder im Laufe der Maßnahme oder gar erst nach Abschluss der Maßnahme. Abbildung 6 zeigt für die EZ realistische quasi-experimentelle Szenarien auf:

Abbildung 6: Für die EZ realistische quasi-experimentelle Designs

	DESIGN	Vorher-Daten t_1 (Baseline)	Maß-nahme *X*	Nachher-Daten t_2 (Survey)
Quasi-experimentelle Versuchsanordnung:				
(2)	Vortest-Nachtest mit Vergleichsgruppen-Design	**ZG_{t1}** **VG_{t1}**	**X** **–**	**ZG_{t2}** **VG_{t2}**
(3)	„verspätetes" Vergleichsgruppen-Design		**X ZG_{t1}** **– VG_{t2}**	**ZG_{t2}** **VG_{t2}**
(4)	Vortest-Nachtest mit Nachtest Vergleichsgruppen-Design	**ZG_{t1}**	**X** **–**	**ZG_{t2}** **VG_{t2}**
(5)	Survey-Design		**X** **–**	**ZG_{t2}** **VG_{t2}**

ZG: Zielgruppe, VG: nicht-randomisierte Vergleichsgruppe, *t:* Zeitpunkt

Liegen Daten für den Zeitpunkt *vor* Implementierung einer Maßnahme (t_1) vor, z.B. aus einer Baseline-Studie, Feasibility-Studie, Bedarfsanalyse oder auch aus einer allgemeinen Haushaltsbefragung, so kann aus diesen Daten eine VG konstruiert werden.[18] Bei dem „Vortest-Nachtest mit Vergleichsgruppen-Design" (vgl. (2), Abb. 6) kann – bei guter Datenlage und optimaler Umsetzung – die double-difference Methode angewandt werden, was zu ähnlich robusten Ergebnissen über die Wirkungen einer Maßnahme führt wie ein experimentelles Design (vgl. Abb. 5 oben). Aus diesem Grund werden diese zwei Varianten häufig nicht getrennt dargestellt (z.B. bei Bamberger 2006: 5; 8).

Wenn eine Impact Evaluation nicht zu Beginn einer Maßnahme geplant wurde, so ist dennoch denkbar, dass im Kontext einer Zwischenevaluation die Notwendigkeit einer künftigen Impact Evaluation realisiert und beschlossen wird. Sofern die Maßnahme noch nicht allzu lange läuft oder noch nicht richtig begonnen hat, können die zu diesem Zeitpunkt gewonnen Daten als vorher Messung genutzt werden (vgl. (3), Abb. 6). Die gewonnenen Ergebnisse sind nicht wesentlich schlechter als bei Variante (2). Ist die Maßnahme jedoch bereits seit längerer Zeit am Laufen sind die Ergebnisse signifikant schlechter.

Liegen Baselinedaten lediglich für die ZG vor, ist es dennoch sinnvoll, eine Vergleichgruppe für einen nachher Vergleich heranzuziehen (vgl. (3), Abb. 6) und zumindest die single-difference Methode anzuwenden und dem vorher-nachher Vergleich der ZG gegenüber zu stellen. Die Ergebnisse sind hierbei allerdings wesentlich weniger robust als bei den zuvor beschriebenen Designs.

[18] Eine detaillierte Beschreibung zur Konstruktion der Vergleichsgruppen findet sich weiter unten im Text.

Die in der EZ häufig genutzte Variante zeigt Szenario (5): Beim sogenannten "Survey-Design", d.h. Datenerhebungen nur nach Beendigung der Maßnahme, liegen keine Basis-Daten für die ZG vor. Der gefundene Unterschied zwischen der im Nachhinein konstruierten VG und der ZG wird hier allein der Maßnahme zugeschrieben. Da mögliche Gruppenunterschiede zum Zeitpunkt t_1 hierbei völlig unbeachtet bleiben, ist eine äußerst bedachte Auswahl der VG nötig.

Um *quasi-experimentelle Designs* anwenden zu können, existieren unterschiedliche Verfahren zur *Konstruktion von Vergleichsgruppen*:

"Matching on Observables" ist sicherlich die in der EZ am häufigsten angewandte Methode, um für die Anwendung der in Abbildung 6 dargestellten Varianten (4) und (5) eine Vergleichsgruppe für eine nachher Messung (t_2) zu konstruieren*:* Die Evaluatoren/innen identifizieren auf Basis von Gesprächen mit Beteiligten sowie relevanten Unterlagen zentrale Eigenschaften der Zielgruppe, anhand derer eine Vergleichsgruppe gebildet werden kann. Auf Basis von Informationen aus Beobachtungen und Sekundärdaten, wie z.B. Zensus, Haushaltsbefragungen, Schulaufzeichnungen, etc. werden dann Personen (oder eine Region, Gemeinde, Dorf etc.) gesucht, die die höchste Übereinstimmung in den Eigenschaften mit der ZG aufweisen und als Vergleichsgruppe ausgewählt. In der Realität können hierbei meist nur einfache beobachtbare Größen ("observables") berücksichtigt werden. Mögliche unbeobachtbare Unterschiede (z.B. Motivation) sollten daher insbesondere mit qualitativen Methoden untersucht und bei der Auswertung der Ergebnisse berücksichtigt werden.

Robustere Ergebnisse ergibt die Berechnung der double-difference im Kontext der oben dargestellten Designs (2) oder (3), wofür jedoch eine vorher Messung auch der Kontrollgruppe bzw. Vergleichgruppe notwendig ist: liegen keine vorher Daten (t_1) für eine VG vor, so kann auf allgemeine Bevölkerungsumfragen oder sonstige Daten (national surveys etc.), die zu diesem Zeitpunkt erhoben wurden und die interessierenden Fragen bzw. Angaben enthalten, zurückgegriffen werden. Unter Anwendung sogenannter *Matching-Verfahren (Parallelisierung),* insb. dem *"Propensity Score Matching"* (PSM), kann aus solchen Daten im Nachhinein eine adäquate Vergleichsgruppe für den Zeitpunkt t_1 konstruiert werden, die ein sehr genaues Abbild der ZG bildet. Zuerst werden wiederum bestimmte Merkmale bzw. Charakteristika der ZG identifiziert (vgl. "Matching on Observables"). Anschließend wird für jede Person (Haushalt, Gemeinde, Dorf) der ZG eine oder mehrere Personen aus dem vorhandenen Datensatz für die Vergleichsgruppe ausgewählt, die sich bezüglich dieser Merkmale nicht von der ZG-Person unterscheidet („statistischer Zwilling", Matching-Partner).[19] Auf Basis der so ermittelten Vergleichsgruppe kann dann die Wirkung einer Maßnahme über die double-difference Methode bestimmt werden. Wurden die Daten, die zur Bildung der VG herangezogen werden, mit den gleichen bzw. ähnlichen Instrumenten und Fragebögen (Fragen) erhoben wie bei der ZG, so entspricht die Qualität der VG auf Basis des Matching-Verfahrens durchaus der einer Kontrollgruppe aus einem experimentellen Design.

[19] Insbesondere wenn die identifizierten Merkmale, anhand derer die Matching-Partner ausgewählt werden sollen, mehrdimensional sind, ist das *Propensity Score Matching* am besten geeignet und wird daher im Kontext von EZ Impact Evaluationen angewandt (vgl. Bamberger 2006: 11, White 2006a: 14, Baker 2000: 6). Hierbei wird die identische oder ähnlich bedingte Wahrscheinlichkeit (Propensity Score), zur ZG zu gehören, berechnet und als Auswahlkriterium genutzt (vgl. Rosenbaum/Rubin 1983: 1). Wie PSM im Detail durchgeführt wird, findet sich bei Gangel/DiPrete 2004, siehe auch White 2006a: 15, Baker 2000: 50 sowie ausführliche bei Ravaillion 2005, 22ff.

Der entscheidende Vorteil von Matching-Verfahren insbesondere im Kontext der EZ ist eindeutig: bei vielen EZ-Maßnahmen existieren häufig – wenn überhaupt – nur Basisdaten für die ZG, nicht jedoch für eine Vergleichsgruppe. Unter Anwendung von Matching-Verfahren kann jedoch auf nationale Datenquellen (national surveys etc.) zurückgegriffen und eine adäquate Vergleichsgruppe (re)konstruiert werden. So wurde z.B. im Rahmen einer WB/IEG „Impact Evaluation of Intervention to Improve Maternal and Child Health and Nutrition in Bangladesh" eine Vergleichsgruppe mit PSM auf Basis der Daten des „National Representative Nutritional Surveillance Project" konstruiert, nachdem sich gezeigt hatte, das die zuvor ausgewählte Vergleichsgruppe zu klein war (vgl. White 2006a).

Verfahren zur (Re-)Konstruktion von Vergleichsgruppen:
- "Matching on Observables"
- "Matching" Verfahren, insb. "Propensity Score Matching" (PSM)
- "Regression Discontinuity"
- "Multiple Comparison Group"

Darüber hinaus eröffnen spezifische Konzeptionen von EZ-Maßnahmen weitere Möglichkeiten, eine Vergleichsgruppe zu konstruieren. Ist die Teilnahme an einer Maßnahme an eine bestimmte Voraussetzung mit gesetztem *Schwellenwert* gebunden, d.h. im Rahmen eines Stipendienprogramms werden bestimmte Leistungen bzw. Noten vorausgesetzt oder eine Maßnahme ist nur für Familien mit einem Pro-Kopf-Einkommen von weniger als 1US$/Tag vorgesehen, kann anhand der *"Regression Discontinuity"* Methode eine Vergleichsgruppe konstruiert werden. Da bei solchen Maßnahmen die Voraussetzungen überprüft werden, liegen Daten für den Zeitpunkt t_1 für die Personen vor, die letztendlich in die Maßnahme aufgenommen wurden, aber auch für solche, die abgelehnt wurden, da sie den Schwellenwert unter- bzw. überschritten haben. Die Idee ist nun, als Kontrollgruppe diejenigen auszuwählen, die den Schwellenwert nur *knapp* nicht erreicht haben, somit aber sehr ähnliche Charakteristika wie die Teilnehmer/innen aufweisen (vgl. Baker 2000: 103f.; Bamberger 2006: 11). Wird z.B. ein Arbeitsmarktprogramm für Jugendliche bis 24 Jahre aufgelegt, so eignet sich die Gruppe der 25-jährigen gut als Vergleichsgruppe.

Bei Maßnahmen, die wie oben beschrieben zeitversetzt implementiert werden, die Auswahl, wer an welcher Phase teilnimmt, jedoch bewusst bestimmt wird, kann das *"Multiple Comparison Group Design"* angewandt werden: Hierbei werden verschiedene Teilnehmer/innen-Gruppen, die unterschiedliche Eigenschaften aufweisen, *untereinander* als Vergleichsgruppe genutzt und verglichen. Solche Unterschiede können wie im oben beschriebenen Fall in den Eigenschaften der Gruppen begründet sein, z.B. besondere Bedürftigkeit/Armut, besonders gute Voraussetzungen etc.

Aber auch wenn die Maßnahme nicht in Phasen aufgeteilt wird sondern zu einem einzigen Zeitpunkt implementiert wird, kann dieses Verfahren sinnvoll sein: sieht eine größere Maßnahme für verschiedene Gruppen, Regionen oder Dörfer unterschiedliche *Ausgestaltungen/Variationen* oder verschiedene *Kombinationen von Leistungen* vor – d.h. wird z.B. in einer Region ein Produkt oder eine Serviceleistung unentgeltlich angeboten, in einer anderen Region dagegen wird hierfür ein gewisses Entgelt verlangt – können die Gruppen *untereinander* als Vergleichsgruppe dienen und verglichen werden. Hierbei wird allerdings nicht das Kontrafaktische betrachtet. Diese Vorgehensweise eignet sich daher insbesondere, um die Frage zu beantworten, wie eine Maßnahme am besten wirkt, wobei verschiedene Wirkungshypothesen überprüft bzw. verglichen werden.

6. RELEVANTE STÖRFAKTOREN

Die Zuverlässigkeit der aufgefundenen Wirkungen hängt insbesondere von der *internen Validität* des jeweils gewählten Untersuchungsdesigns ab. Interne Gültigkeit ist dann gegeben, wenn eine Maßnahme tatsächlich für die aufgefundenen Veränderungen verantwortlich ist. Veränderungen können aber auch durch einen oder mehrere *Störfaktoren* entstehen, der bzw. die zu Verzerrungen ("bias") führen – dann ist interne Gültigkeit nicht mehr gegeben und die gefundenen Wirkungen sind nicht mehr eindeutig der Maßnahme zuzuordnen.

(1) So muss bezüglich des *Zeitpunkts* der Datenerhebungen (Baseline und nachher Messung) geprüft werden, ob die Situation zu diesem Zeitpunkt „normal" war oder nicht. Wenn z.B. die Baseline-Studie zum Zeitpunkt einer Naturkatastrophe oder eines sonstigen außergewöhnlichen Ereignisses stattfand, dann können die beobachteten Veränderungen stark verzerrt sein. Da im Rahmen von IE die Zeitpunkte der Datenerhebung meist durch die Maßnahme vorgegeben sind und selten von den Evaluatoren/innen bestimmt werden können, müssen Daten immer auf solche möglichen Zeiteinflüsse überprüft werden (vgl. NONIE-SG1 2008: 2).

(2) *Auswahlverzerrung* (Selektionsbias, "selection bias") ist, wie bereits bei der Darstellung zur Bildung der Vergleichsgruppen angesprochen wurde, ein häufiger Störfaktor, der im Rahmen der Auswahl der Vergleichsgruppen immer überprüft werden sollte: Oft wird die ZG einer Maßnahme nach bestimmten Vorgaben ausgesucht, oder aber erfolgt im Sinne einer Selbstselektion, d.h. Individuen oder Gruppen entscheiden selbst, an einer Maßnahme teilzunehmen (melden sich für eine Veranstaltung an, beantragen einen Kredit, etc.). In beiden Fällen ist die Auswahl an bestimmte Eigenschaften, persönliche Charakteristika gebunden. Bei der Bildung der Vergleichsgruppe ist daher darauf zu achten, dass diese die gleichen Eigenschaften aufweist, um eine Auswahlverzerrung zu verhindern und den Einfluss dieser Störgrößen zu eliminieren. Bei Auswahlkriterien (z.B. bestimmte Voraussetzungen für die Teilnahme) kann davon ausgegangen werden, dass diese bekannt und vorab überprüft wurden – also beobachtbar sind ("observables"). In diesem Fall ist es möglich, eine passende Vergleichsgruppe zu definieren und Messabweichungen gering zu halten bzw. zu kontrollieren.[20] Bei Selbstselektion dagegen ist es möglich, dass die Auswahl (bzw. Entscheidung zur Teilnahme) auf unbeobachtbaren Eigenschaften ("unobservables") beruht. Korrelieren diese mit den Wirkungen der Maßnahme, wird eine Schätzung der Wirkungen verzerrt: Es hat sich z.B. im Rahmen einer IE gezeigt, dass Kleinunternehmen, die an einer Mikrokredit-Maßnahme teilgenommen haben, höhere Profite aufweisen als vergleichbare Betriebe, die keine Kredite erhalten. Allerdings haben nur solche Unternehmen einen Kredit erhalten, die im Rahmen eines Auswahlprozesses einen adäquaten Geschäftsplan vorweisen konnten. Es scheint naheliegend, dass diese Unternehmen ohnehin bessere Profite erwirtschaftet hätten, als die Betriebe, deren Geschäftsplan zu schlecht für einen Kredit war. Dieses Beispiel verdeutlicht nochmals die besonderen Vorteile eines experimentellen Designs, denn bei zufallsgesteuerten Auswahlprozessen liegen solche dargestellten, systematischen Auswahlverzerrungen nicht vor (vgl. ADB 2006: 5f.; Bloom 2006: 1; Baker 2000: 2; NONIE-SG1 2008: 4). Alternativ hätte sich hier auch die "Regression Discontinuity" Methode geeignet, indem die Betriebe, deren Ge-

[20] Das Auswahlproblem wird somit endogen und kann anhand der „Instrumental variable" Schätzung gelöst werden (vgl. White 2006, 4)

schäftsplan nur knapp abgelehnt wurde, bei ausreichender Anzahl als Vergleichsgruppe herangezogen worden wären.

(3) *Übertragungseffekte* (“Contamination“/“Contagion“) entstehen aufgrund zweier Möglichkeiten: Zum einen durch die Maßnahme selbst ("spill-over effects"), d.h. die Maßnahme wirkt nicht allein in einer begrenzten/intendierten Zielregion, sondern auch darüber hinaus. Häufig werden Vergleichsgruppen aus der direkten räumlichen Nachbarschaft einer Maßnahme gewählt, da davon ausgegangen wird, dass hier eine vergleichbare Situation vorliegt. Je näher die Vergleichsgruppe jedoch rein räumlich ist, umso größer ist die Wahrscheinlichkeit, dass auch diese Gruppe bzw. Region indirekt von der Maßnahme betroffen ist. Ein Bauprojekt kann z.B. einen kurzfristig größeren Bedarf an Arbeitskräften auslösen, so dass auch Bewohner außerhalb der Maßnahmenregion eine Anstellung finden. Solche Tatsachen sind natürlich als nichtintendierte positive Effekte der Maßnahme zuzuschreiben. Dennoch muss bei der Auswahl der Vergleichsgruppe zwischen erwünschter geografischer Näher (um vergleichbare Eigenschaften zu gewährleisten) und notwendiger Entfernung (um Übertragungseffekte zu vermeiden) sorgfältig ausgewählt werden.

    Eine zweite, weitaus relevantere Form von Übertragungseffekten zeigt sich, wenn die gewählte Vergleichsgruppe oder -region Gegenstand einer Maßnahme mit gleicher oder ähnlicher Zielrichtung einer *anderen* Organisation war. In solchen Fällen kann wiederum lediglich der Unterschied der Maßnahmenkonzepte verglichen werden. Die Vergleichsgruppe kann jedoch nicht zur Untersuchung des Kontrafaktischen genutzt werden. Bei Designs, die nur nachher Messungen beinhalten, kann dies bei der Auswahl der Vergleichsgruppe zuvor überprüft werden. Wurde jedoch ex-ante eine Vergleichsgruppe festgelegt und Baselinedaten erhoben, haben die Evaluatoren/innen natürlich keinen Einfluss darauf, ob diese Region vor der nachher Messung Gegenstand einer anderen Maßnahme wird (vgl. White 2006a: 4.; NONIE-SG1 2008: 3f.). "Spill-over" Effekte können entsprechend auch bei randomisierten Kontrollgruppen auftreten!

Störeffekte haben einen entscheidenden Einfluss auf die Qualität der Ergebnisse. Daher ist es bei allen gewählten Vorgehensweisen äußerst wichtig, dass mögliche Störfaktoren berücksichtigt und kontrolliert werden: Wie dargestellt stellt Randomisierung der Gruppen eine Möglichkeit dar, Störgrößen auszuschalten. Grundsätzlich müssen jedoch sämtliche Eventualitäten bedacht werden, d.h. es muss überprüft werden, ob bzw. welche (vergleichbaren) Maßnahmen anderer Geber im Projektgebiet selbst aber auch im Umfeld der Vergleichsgruppe stattfanden. Auch müssen die beobachtbaren, für die Wirkungen relevanten Eigenschaften (“observables“) der ZG sorgfältig erarbeitet und bei der Bildung der Vergleichsgruppe berücksichtigt werden. Häufig entsprechen die sogenannten unbeobachtbaren Eigenschaften (“unobservabes“) lediglich *unbeobachteten*, d.h. schlichtweg nicht berücksichtigten Eigenschaften! Grundlage, dass sämtliche relevanten Eigenschaften bedacht werden, ist daher ein *theoriegeleiteter Ansatz* der Evaluation.

Störeffekte haben einen entscheidenden Einfluss auf die _Qualität_ der Ergebnisse einer IE. Insbesondere _Auswahlverzerrungen_ sowie _Übertragungseffekte_ müssen daher immer überprüft werden.

7. *WARUM* WIRKT EINE MAßNAHME UND WELCHE UNINTENDIERTEN WIRKUNGEN ZEIGEN SICH? – ZUR NOTWENDIGKEIT THEORIEBASIERTER ANSÄTZE

Die dargestellten experimentellen bzw. quasi-experimentellen Designs, die eine vorher-nachher Messung sowie eine Vergleichsgruppe umfassen und somit eine double-difference Berechnung zulassen, sind zuverlässige Methoden, um Wirkungen *eindeutig* einer Maßnahme zuzuschreiben und auch deren *Umfang* darzustellen. Hierdurch wird zwar die Frage beantwortet „Welche Veränderung hat die Maßnahme bewirkt?“ nicht aber die Frage „Was funktioniert, unter welchen Bedingungen?“, d.h. die Frage „*warum* hat eine Maßnahme Wirkungen entfaltet (oder nicht)?“ bzw. „*wie* hat die Maßnahme gewirkt; unter welchen Bedingungen?“. Evaluationen, die Ergebnisse über die Wirkungen geben, aber keine Angaben machen, *warum* eine Maßnahme die erwarteten Wirkungen gezeigt hat oder nicht, werden auch "*black box*" Evaluationen genannt, denn die fehlenden Informationen der Kausalkette einer Wirkung hinterlässt eine leere "black box" (vgl. Bloom 2006: 18f.; White 2006a: 9; Ravallion 2005: 1, NONIE-SG1 2008: 5).

Double-difference ermöglicht, Wirkungen einer Maßnahme *eindeutig zuzuschreiben* und deren *Umfang* darzustellen.
Die Fragen „*warum* hat eine Maßnahme Wirkungen entfaltet (oder nicht)?“ und „*wie* hat die Maßnahme gewirkt?“ bleibt hierbei jedoch eine „black box“.

Diese Fragen können über die Anwendung *regressionsbasierter Ansätze* beantwortet werden: Auf der Grundlage gängiger (Fach-) Literatur wird ein *theoretisches Modell* über die Zusammenhänge zwischen Maßnahme und Wirkungen sowie weiterer relevanter Einflussfaktoren entwickelt, welches anschließend mit regressionsanalytischen Verfahren statistisch überprüft wird. Bei regressionsbasierten Untersuchungen wird eine 0/1-codierte Variable eingeführt – 1 für Personen der ZG, 0 für Personen der VG –, die als erklärende Variable aufgenommen wird. Hierdurch kann der Unterschied zwischen den Teilnehmer/innen und der Vergleichsgruppe geschätzt werden. Zur Beantwortung der Frage, warum eine Maßnahme gewirkt hat, müssen jedoch weitere Variablen aufgenommen werden, insbesondere muss die *Maßnahme selbst* spezifiziert werden. Während beim double-difference Ansatz die Wirkungsgröße „Maßnahme“ implizit über die ZG und VG „enthalten“ ist, muss bei regressionsbasierten Ansätzen die Maßnahme als Variable dezidiert aufgenommen werden, d.h. die Ressourcen (“Inputs“) und insbesondere die Aktivitäten ("activities") einer Maßnahme müssen operationalisiert werden und mit entsprechenden Indikatoren in die Regressionsgleichung einfließen. Dies bedeutet, dass nicht nur Daten zu den Veränderungen bei der ZG und VG erhoben werden müssen, sondern auch Daten zur Maßnahme selbst. Dies erscheint bei Projekten der Finanziellen Zusammenarbeit (FZ) ggf. über die reine Höhe der finanziellen Zuwendung einfach machbar. Bei Maßnahmen der technischen Zusammenarbeit, die auch Beratungsleistungen umfassen, ist dagegen für jede einzelne Komponente der Aktivitäten ein sinnvoller Indikator zu bilden und die entsprechenden Daten zu sammeln. Soll auch die Frage geklärt werden, *warum* eine Maßnahme gewirkt oder nicht gewirkt hat, so müssen mögliche *intervenierende Einflussgrößen* ebenfalls operationalisiert und entsprechende Daten gesammelt werden.

Regressionsbasierte Ansätze überprüfen demnach zuvor aufgestellte *Hypothesen.* Hierfür ist allerdings die Konstruktion eines expliziten *theoretischen Modells über die erwarteten Ursache-Wirkungs-Zusammenhänge* notwendig. Es wird deutlich, dass ein anspruchsvolles Design – eine vorher-nachher Messung der Wirkungsindikatoren mit Kontroll- oder Vergleichsgruppe

inklusive einer double-difference Berechnung – die Frage nach dem „Warum?“ und „Wie?“ nicht beantworten, d.h. das Problem der „black box“ nicht lösen kann.

Grundlage einer guten IE ist ein *theorie-basierter Ansatz*, der *Hypothesen* über die *Ursache-Wirkungs-Zusammenhänge* konstatiert.

Daher wird in vielen Veröffentlichungen die Notwendigkeit hervorgehoben, bei IE *theorie-basierte Ansätze* (“Theory-Based Approaches“, TBA) zu nutzen (vgl. z.B. White 2006a, Bamberger et al. 2006, Bamberger 2006, Baker 2000): Bei TBA wird mit Hilfe eines logischen Modells (“Logical Framework/ LogFrame“) die einer Maßnahme zugrundeliegende Hypothese über Ursache-Wirkungs-Zusammenhänge (“Theory of Change“, TOC) detailliert ausgearbeitet und tabellarisch (als Matrix) dargestellt. Das LogFrame zeigt dabei auf, welche (impliziten und expliziten) Annahmen über kausale Verknüpfungen (“causal chains“) zwischen der geplanten Maßnahme und den intendierten Wirkungen der Maßnahme zugrunde liegen, d.h. es wird explizit beschrieben, *was* eine Maßnahme *wie* (mit welchen Ressourcen und Aktivitäten), *für wen* und *wozu* (mit welchem übergeordneten Ziel) erreichen soll. Eine LogFrame-Matrix[21] enthält für *jede* Ebene der Wirkungskette eine eigene Wirkungshypothese (TOC) (vgl. Abb. 7). Zusätzlich zur genauen Benennung der einzelnen Wirkungsebenen sowie entsprechender Hypothesen über deren Ursache-Wirkungszusammenhänge, enthält eine LogFrame-Matrix immer auch eine Auflistung geeigneter Indikatoren sowie der Erhebungs- und Auswertungsverfahren (vgl. Baker 2000: 12; Bamberger 2006: 8f.). Insbesondere werden externe Faktoren, die ein mögliches Risiko der konstatierten TOC darstellen, identifiziert und explizit in die LogFrame-Matrix eingetragen. Sind konkurrierende TOC denkbar, wird auch dies festgehalten.

Abbildung 7: Zentrale Elemente einer LogFrame-Matrix

Wirkungs-ebene	Aktivitäten & TOCs	Risiken	Indika-toren	Daten
Impact	*Wirkung* ⬆			
Outcome	*Ursache* *Wirkung* ⬆			
Output	*Ursache* *Wirkung* ⬆			
Input	*Ursache*			

Ursprünglich wurde der LogFrame-Ansatz als Planungs- und Steuerungsinstrument in den frühen 1970er Jahren in der EZ übernommen. Dennoch zeigt sich das Problem, dass insbesondere das zentrale Element, die jeweiligen Ursache-Wirkungs-Hypothesen, bei der Planung von Maßnahmen nicht in dieser Detailliertheit niedergeschrieben werden, was häufig zum Vorwurf der Theorielosigkeit von EZ-Maßnahmen führt. Die Erstellung eines LogFrames – und hierbei insbesondere der Ursache-Wirkungs-Hypothesen – ist keinesfalls einfach. Die Identifikation

[21] Ursprünglich als 4x4-Felder-Tabelle entwickelt existieren heute verschiedene LogFrame-Konzepte. Eine 4x4-Felder-Tabelle wird auch heute noch z.B. von der Weltbank genutzt, um die Zielhierarchie beginnend mit den Aktivitäten, über die Leistungen und Ziele, bis hin zu dem übergeordneten Ziel abzubilden (vgl. White 2006a: 8).

von Hypothesen und Theorien auf jeder Ebene der Wirkungskette ist eine äußerst komplexe Aufgabe. Dies mag der Grund sein, dass Evaluatoren/innen bei vielen der zu evaluierende Maßnahmen kein LogFrame vorfinden oder aber die aufgestellten Hypothesen zu allgemein gehalten sind. Das bedeutet, dass die bei der Planung zugrunde gelegten Hypothesen bzw. teilweise auch die gesamte LogFrame-Matrix im Rahmen einer IE von den Evaluatoren/innen rekonstruiert werden müssen.[22]

Bei der *Erarbeitung* eines LogFrames sollte in einem ersten Schritt auf aktuelle Literatur und wissenschaftliche Erkenntnisse aus dem betreffenden Feld bzw. Sektor zurückgegriffen werden: Bei einem Bildungsprojekt sollten aktuelle Erkenntnisse aus der Bildungsforschung berücksichtigt werden, d.h. die Wirkungshypothesen sollten auf bereits überprüften Theorien aufbauen. In einem zweiten Schritt sind die Stakeholder zu befragen, z.B. über Intensiv-, Experten- und/oder Fokusgruppeninterviews. Des Weiteren wurden für die EZ-Praxis eine Vielzahl vorwiegend partizipativer Tools erarbeitet, wie z.B. SWOT-Analyse, Problem-Baum, "Mind-Map", Venn-Diagramm.

Insbesondere im Rahmen von IE sollten die in der Projektplanungsphase erarbeiteten LogFrames bzw. die aufgestellten Hypothesen vor Beginn der Datenerhebung überprüft werden. Denn: IE untersuchen wie Eingangs dargestellt die *direkten und indirekten, intendierten und nicht intendierten Wirkungen*. Wird einer IE ein bestehender LogFrame zugrunde gelegt, ist die Wahrscheinlichkeit groß, dass indirekte und vor allem nicht intendierte Wirkungen unerkannt bleiben. Denn auch wenn LogFrames Risiken benennen, ist durchaus die Möglichkeit gegeben, dass im Verlauf der Maßnahme andere, zuvor nicht berücksichtigte, Probleme aufgetaucht sind. Wird der LogFrame von den Evaluatoren/innen ausschließlich auf Basis von Literatur erstellt, ist dieses Problem entsprechend wesentlich größer.

Es wird deutlich, dass bei Impact Evaluationen intensive Gespräche und Interviews mit den Stakeholdern ebenfalls entscheidend sind, um adäquate Ursache-Wirkungs-Zusammenhänge zu hypothetisieren und zuverlässige Antworten über die Wirkungen einer Maßnahme zu geben.

In der aktuellen Wirkungsstudie des BMZ: “Assessing the Impact of Development Cooperation in North East Afghanistan” zeigt sich, wie eine theoretisch fundierte Wirkungsmessung auf der Grundlage eines anspruchsvollen Methodenmix das Zuordnungsproblem weitestgehend einschränken und die Fragen nach dem „Wie“ und „Warum eine Maßnahme gewirkt hat?“ beantworten kann. Das gewählte Evaluationsdesign berücksichtig zwei zentrale Ansätze: Erstens wurden einzelne Kooperationsprojekte mit spezifischen direkte Wirkungen verknüpft, sodass bei intensiver Betrachtung des Prozesses (welcher Stimuli eines Projektes führt zu welchem Ergebnis bzw. welcher Wirkung) Zuordnungsmuster sichtbar gemacht werden konnten. Zweitens wurde die ZG einer sorgfältig ausgewählten Vergleichgruppe gegenüber gestellt, um das Kontrafaktische abzubilden. Die ausgewählten Methoden setzen sich zusammen aus qualitativen und quantitativen Methoden sowie GIS, ein computergestütztes Informationssystem mit dem raumbezogene Daten digital erfasst und analysiert werden können (vgl. Zürcher/Köhler 2007). Qualitativ lag der Schwerpunkt auf Fallstudien, die mit unterschiedlichen Stakeholdergruppen realisiert wurden und sich darauf konzentrierten „to generate and to analyze data on causal mechanism, or processes, events, actions, expectations, and other intervening variables that link putative causes to observed effects“ (Zürcher/Köhler 2007: 23). In den quantitati-

[22] Dies kann jedoch nicht als grundsätzliche Aufgabe einer IE gesehen werden!

ven Analysen wurden anschließend statistische Verfahren genutzt, wobei vornehmlich regressionsbasierte Verfahren genutzt wurden, insbesondere multiple Regressionsverfahren, um die Beziehungen zwischen unterschiedlichen unabhängigen oder "predictor" Variablen und den abhängigen Variablen statistisch nachzuweisen (vgl. Zürcher/Köhler 2007: 23).[23]

8. Attribution oder Kontribution – Die Kritik an kontrafaktischen Kausalanalysen

Zwar setzen sich international immer mehr bi- und multilaterale EZ-Organisationen intensiv mit dem Thema IE und den methodischen Implikationen auseinander und führen anspruchsvolle IE durch, dennoch beinhaltet die Diskussion um IE auch kontroverse Standpunkte: Einer der grundlegendsten Diskussionspunkte betrifft die Frage, ob bei Impact Evaluationen das Kontrafaktische *explizit* oder *implizit* berücksichtigt werden sollte.

Die Kritiker des "explizit counterfactual" argumentieren in erster Linie gegen die Verwendung von randomisierten Kontrollgruppen-Designs, wobei die Annahme zugrunde gelegt wird, dass „explizit“ grundsätzlich die Verwendung von RCTs bedingt. Der Begriff RIE wird hierbei *gleichgesetzt* mit der Anwendung von RCTs. So z.B. im aktuellen Statement der Europäischen Evaluationsgesellschaft EES zur IE-Diskussion: ”EES however deplores one perspective currently being strongly advocated: that the best or only rigorous and scientific way of doing so [IE] is through randomised controlled trials (RCTs)“ (EES 2007). Auch Michael Quinn Patton setzt im Rahmen seiner „Debate about Randomized Controls in Evaluation: The Gold Standard Question“ RIE gleich mit der Anwendung von RCTs.[24]

Als Alternative zu *expliziter* Wirkungszuschreibung ("causal attribution"), d.h. der expliziten Berücksichtigung des Kontrafaktischen, wird *"causal contribution"* angeführt: „Analysis of causal contribution aims to demonstrate whether or not the evaluated intervention is one of the causes of observed change. Contribution analysis relies upon chains of logical arguments that are verified through a careful confirmatory analysis” (NONIE-SG2 2008: 16). Es werden zwei zentrale Eigenschaften einer Analyse des kausalen Beitrags angegeben: (1) die Nutzung von Theoriebasierten Evaluationen ("theory-based evaluation", TBE) in einem iterativen Prozess aus Aufstellen, Testen und Verfeinern sowie (2) Berücksichtigung der „Härte“ ("rigour") durch kritische Evidenzanalysen, “not through a particular research design” (NONIE-SG2 2008: 24).[25] Als mögliche Ansätze hierzu werden insbesondere partizipative Ansätze angegeben, die analysieren, inwieweit sich die Situation der Begünstigten aufgrund der Maßnahme verändert hat, wobei eher qualitative Methoden zur Anwendung kommen.[26]

---

[23] In diesem ergänzenden Methodenpapier werden die unabhängigen und abhängigen Variablen aufgezählt und erläutert. Dieser Methodenbericht erscheint hinsichtlich der Transparenz des Vorgehens und der Ergebnisse beispielhaft.

[24] Zwar werden bzw. wurden RCTs aufgrund ihrer hohen internen Validität vereinzelt als zentrales Kriterium von RIE gesehen und als Gold-Standard bezeichnet, dennoch beschränkt sich der Begriff „rigorous“ im Kontext der aktuellen internationalen Diskussion nicht mehr ausschließlich auf RCTs – quasi-experimentelle Designs werden meist als sinnvolle Alternative aufgeführt.

[25] In den Dokumenten wird jedoch nicht erwähnt, dass auch bei TBEs bzw. TOC-Ansätzen die Notwendigkeit des Kontrafaktischen betont wird.

[26] Der Begriff qualitativ ist an dieser Stelle mit Vorsicht zu behandeln, denn die Gleichsetzung partizipativer Methoden mit qualitativer Methoden ist nicht immer gegeben. Denn nicht alle qualitativen Erhebungstechniken sind partizipativ, zudem ergeben partizipative wie auch qualitative Methoden oftmals quantifizierte Daten (vgl. Caspari 2006).

Demgegenüber wird die *explizite* Berücksichtigung des Kontrafaktischen von vielen Autoren und EZ-Organisationen als zentrale Aufgabe einer Impact Evaluation gesehen (vgl. Baker 2000, CGD 2006, Kapoor 2002, NONIE-SG1 2007, White 2006a, b; Ravaillon 2005). Diese Autoren bezweifeln, dass TBE und TOC-Ansätze experimentelle oder quasi-experimentelle Methoden angemessen ersetzen können. Allerdings wird immer wieder die Notwendigkeit dieser Ansätze als zentrales *Element* von RIE betont: „A good evaluation design starts with a theory-based approach which clearly identifies the channels through which the program is expected to operate“ (NONIE 2008: 3; vgl. auch Ravaillon 2005: 56; Baker 2000: 132). TBE und TOC-Ansätze werden demnach als wichtiger *Teil von IE* gesehen, um zu Beginn einer RIE zu wissen, was genau untersucht werden soll, aber auch um die Ergebnisse besser verstehen zu können. Denn TBE und TOC-Ansätze können helfen die Fragen zu beantworten „*Warum* hat eine Maßnahme Wirkungen entfaltet (oder nicht)?“ und „*Wie* hat die Maßnahme gewirkt, für wen, und unter welchen Bedingungen“. In dem Fall, dass eine Maßnahme nicht in der erwartenden Weise verläuft, kann des Weiteren mit zuverlässiger Sicherheit aufgezeigt werden, wo, warum und wie der Misserfolg aufgetreten ist (vgl. Baker 2000: 12).

9. METHODEN-STREIT ODER METHODEN-MIX

Es wird deutlich, dass diese Diskussion nicht losgelöst von dem Schulenstreit „qualitativer“ vs. „quantitativer“ Forschung gesehen werden kann – zwei unterschiedliche Strategien empirischer Forschung, Konstruktivismus und Positivismus: Einerseits wird die Nutzung bestimmter Untersuchungsdesigns zur Untersuchung des Kontrafaktischen als zentrales Element der Bezeichnung "rigorous" eher abgelehnt und der “causal attribution“ eine “causal contribution“ gegenübergestellt, wobei theoriebasierte sowie qualitative bzw. partizipative Ansätze zur Anwendung kommen. Auf der anderen Seite wird die Anwendung „harter“ Methoden bzw. adäquater Designs als notwendige Voraussetzung zur angemessenen Wirkungszuschreibung gesehen, wobei theoriebasierte Ansätze sowie die Kombination von quantitativen und qualitativen Datenerhebungsmethoden ebenfalls als zentrale Elemente einer guten IE angesehen werden. Diese Diskussion ist nicht wirklich neu, eher in einer neuen „Phase“: Spätestens seit Mitte der 1990er Jahre hatten sich partizipative Methoden als vermeintlich bessere Alternative in Abgrenzung zu den lange Zeit genutzten quantitativen Erhebungsverfahren großer Beliebtheit erfreut. Vermehrt auftretende Kritiken an partizipativen Konzepten wurden allerdings lange Zeit ignoriert und fanden keinen ersichtlichen Einfluss auf die Mainstream-Diskussion und die Praxis (vgl. Caspari 2006: 365ff.). Die aktuelle Diskussion um RCTs kann wiederum als Antwort auf diese Entwicklung gesehen werden, nämlich als “reaction against the haphazard use of participatory and qualitative methods in recent years“ (Prowse 2007: 4).

Die Gegenüberstellung quantitativ vs. qualitativ bezeichnet Kromrey als „Ärgernis“: Die Begriffe „hatten vor 50 Jahren ihren Sinn, eignen sich aber heute allenfalls als Kampfbegriffe, um vorgebliche Unterschiede auf den Punkt zu bringen, die so gar nicht mehr existieren“ (Kromrey 2005). Dies scheint auch allgemein für die Diskussion um RIE zutreffend zu sein: “Some argue that the axioms of positivsm and social contstructivism render the two approaches mutually exclusive (…). Such an either/or position is not beneficial to proponents of either standpoint, as 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: 3; vgl. auch Baker 2000: 6f.; NONIE-SG4/5 2008: 6f.).

Entsprechend sind für angemessene IE die Nutzung von beiden Methoden i.S. einer *Triangulation/Methodenmix* notwendig, denn positivistische Methoden beantworten insbesondere die Frage ob Wirkungen eingetreten sind und in welchem Umfang, konstruktivistische Ansätze beantworten dagegen insbesondere die Fragen warum und wie. Die Notwendigkeit der Integration von qualitativen und quantitativen Datenerhebungstechniken auch im Kontext von EZ-Evaluationen ist heute unumstritten (vgl. insbesondere Bamberger 2000; Chung 2000; Guijt 2000; Aplleton/Booth 2001; Ezemenari u.a. 1999; Kassam 1998). Denn "both quantitative and qualitative methods are necessary for a good evaluation. The two approaches strongly complement each other" (Ezemenari u.a. 1999: 28). Diese Notwendigkeit eines Methodenmix in Impact Evaluationen scheint auch innerhalb NONIE unumstritten: In letztendlich allen NONIE-Papieren wird einerseits die Nutzung theoriebasierter Ansätze und andererseits die Notwendigkeit von Triangulation betont: "Good evaluations are almost invariably mixed method evaluations. Qualitative information informs both the design and interpretation of quantitative date. In a theory-based approach, qualitative data provide vital context (…)" (White 2006a: 20; vgl. auch NONIE-SG1 2008: 6).

Angemessene IE bedingen die Nutzung sowohl *qualitativer* als auch *quantitativer Methoden* i.S. einer *Triangulation* / eines *Methodenmix*.

10. ANWENDUNGSBEISPIELE AUS DER PRAXIS

Auf Basis der dargestellten *theoretischen* Diskussion wurden 29 auf der NONIE Homepage als Beispiele für IE internationaler Geberorganisationen eingestellte Studien analysiert. Für die Auswertung der Berichte wurden vorab Kriterien festgelegt: Neben grundlegenden deskriptiven Informationen wie durchführende Institution/Organisation sowie Sektor und Land der Maßnahme sollten die intendierten Wirkungen, die Ebene der Maßnahme (Mikro/Meso/Makro) sowie der Zeitpunkt der Evaluation (Interim-, Abschluss-, ex-post Evaluation) erhoben werden. Weitere Analysekriterien bezogen sich auf die Ebene der IE ("output", "outcome", "impact"), Hypothesenbildung oder Schlüsselfragen, "Log Frame"/"Managment Cycle" o.ä., Triangulation/Methodenmix (Dokumentenanalyse, Literaturanalyse, quantitative sowie qualitative Datenerhebungsmethoden), Berücksichtigung des Kontrafaktischen/Counterfactual implizit oder explizit, wie wurden Kontroll- bzw. Vergleichgruppe gebildet, welche statistischen Test- bzw. Auswertungsverfahren wurden angewandt.

Allerdings hat sich gezeigt, dass eine systematische Auswertung der Studien kaum möglich war, da nur ein geringer Teil der Berichte die der Evaluation zugrundeliegende methodische Vorgehensweise *nachvollziehbar* dokumentiert. Sowohl die ‚DAC Evaluation Quality Standards' (OECD/DAC 2006) als auch die ‚Standards for Evaluation in the UN System' herausgegeben von UNEG (UNEG 2005) aber auch allgemeine internationale Evaluationsstandards (z.B. Sanders/JCS 1994, DeGEval) betonen explizit, dass Evaluationsberichte ausführlich und nachvollziehbar aufführen müssen, welche Methodologie der Evaluation zugrunde lag, wie die Evaluation entwickelt und durchgeführt wurde, welche Indikatoren bzw. Variablen genutzt wurden, wie die Stichprobe ausgewählt wurde, welche Methoden und Techniken für die Datenerhebung und -auswertung genutzt wurden (vgl. UNEG 2005: 20; OECD/DAC 2006: 5ff.). Auch

sollen Validität und Reliabilität der Daten, Einschränkungen und Schwächen der Methodologie angeführt und diskutiert werden: ”The evaluation report describes and explains the evaluation method and process and discuss validity and reliability. (…) Methods for assessment of results are specified. Attribution and contribution/confounding factors should be addressed” (OECD/DAC 2006: 5). Des Weiteren sollen die Ergebnisse nach "input", "output", "outcomes" und "impacts" differenziert dargestellt werden (vgl. UNEG 2005: 21; OECD/DAC 2006: 5). Eine solche Darstellung sollte, wenn nicht explizit im Text, so doch zumindest im Anhang dokumentiert werden. Ein vorbildliches Beispiel einer detaillierten Darstellung der einer IE zugrundeliegenden Methoden und Ansätze ist die oben genannte Wirkungsstudie des BMZ „Assessing the Impact of Development Cooperation in North East Afghanistan“.[27]

Die im Rahmen einer Evaluation *angewandten Methoden* müssen im Bericht *umfassend* und *ausführlich dargestellt* sein – gemäß UNEG- und OECD/DAC-Standards.
Dies ist umso mehr von Bedeutung, wenn Geber untereinander von ihren IE lernen/profitieren wollen.

Die analysierten IE Berichten enthalten derartige Darstellung meist nur rudimentär bzw. nicht vollständig oder ausreichend.[28] Entsprechend mussten für die Auswertung häufig Textpassagen interpretiert werden. Insbesondere die der Evaluation zugrunde gelegten Indikatoren bzw. genutzten Variablen werden selten aufgeführt, auch fehlt meist die Unterscheidung zwischen "output", "outcome" und "impact". Hier besteht *Handlungs- und Verbesserungspotential*, denn die Nachvollziehbarkeit der Methoden von IE ist von grundlegender Voraussetzung, um gegenseitig aus IE zu lernen und somit wiederum die Professionalisierung von IE zu stärken.

Im Folgenden werden die *zentralen Erkenntnisse der Querschnittsanalyse* der in der NONIE-Datenbank aufgeführten Studien dargestellt. Insgesamt werden in der Datenbank 29 Studien aufgeführt, von denen 24 für eine Auswertung zur Verfügung standen.[29] Bei den folgenden Aussagen ist zu beachten, dass bei einigen der Berichte einzelne Analysekriterien nicht eindeutig zu klären waren und somit lediglich *vermutet* sind.[30] Die Ergebnisse werden ergänzt durch eine im Januar 2008 vorgelegte Studie der NONIE-Subgroup 4/5, in der 250 IE-Studien in Hinblick auf Sektor/Thema, geographische Lage sowie angewandte Methoden[31] untersucht wurden, um sektorale, geographische und/oder methodische "Evaluation Gaps“ die anzugehen sind, aufzuzeigen.[32]

[27] Hier wurde zusätzlich zum “Interim Report“ ein eigener Bericht verfasst, der Schritt für Schritt die methodische Vorgehensweise im Detail dokumentiert (vgl. Zürcher/Köhler 2007).
[28] Zu diesem Schluss kommt auch NONIE-SG4/5, die aus diesem Grunde sogar nur 16 der 29 Studien ihrer Auswertung zugrunde legen (vgl. NONIE-SG4/5 2008: 14).
[29] Es fehlen folgende Studien: EC, JICA, FORMIN und zwei Studien von IADB. Die Gründe für das Fehlen der Studien sind sehr unterschiedlich, so ist z.B. die Studie von FORMIN derzeit weder in elektronischer noch in gedruckter Form erhältlich.
[30] So wird z.B. aufgeführt, dass ein mit-ohne Vergleich durchgeführt wurde, in den weiteren Ausführungen findet sich hierzu jedoch kein Beleg. Oder umgekehrt wird die Anwendung der double-difference Methode nicht dezidiert genannt, lässt sich aber anhand der Ausführungen vermuten.
[31] Wird zwar einleitend aufgeführt, dass auch die genutzten Methoden analysiert werden sollen, so findet sich leider im gesamten Papier nur ein kleiner Abschnitt zu dieser Auswertung, der nicht näher auf einzelne Erkenntnisse eingeht.
[32] Die Studien wurden vier zentralen IE-Webseiten entnommen: World Bank DIME (Development Impact Evaluation Initiative), NONIE (Network of Network Impact Evaluation Initiative), PREM (Poverty Reduction and Economic Management) sowie JPAL (Poverty Action Lab). Da auch die NONIE-Studien aufgenommen wurden, könnte es entsprechend zu Überschneidungen kommen. Allerdings wird in einer Ü-

(1) Von den 24 Studien wurden elf von IADB/OVE durchgeführt, drei weitere von IFAD, zwei je von Danida, DFID und FINNIDA, sowie je eine von AfDB, AusAID, CIDA und JBIC. Insgesamt sind es 17 ex-post Evaluationen, die zwei bis neun Jahre nach Beendigung der Maßnahme durchgeführt wurden. Eine weitere als Schlussevaluation bezeichnete Studie untersuchte allerdings ein langjähriges Sektorprogramm und kann somit ebenfalls als ex-post Evaluation gesehen werden. Vier weitere Studien sind *Interimevaluationen*. Des Weiteren sind zwei *Forschungsstudien* enthalten, die spezifische Fragestellungen bearbeiten.

Regional verteilen sich die jeweils untersuchten Maßnahmen auf Afrika (N=7), Südasien (N=3), Südostasien (N=2) sowie Südamerika (N=11), wobei diese elf Studien alle von IADB/OVE durchgeführt wurden.[33] Von den durch NONIE-SG 4/5 untersuchten IE wurde ebenfalls die größte Anzahl an IE in Lateinamerika durchgeführt, wobei auch auf die verstärkten Aktivitäten von IADB im Bereich IE verwiesen wird. Die weitere Verteilung der 250 untersuchten Studien wird wie folgt angegeben: Süd- und Südostasien 27%, Afrika 14%, Osteuropa 6%, China 3%. Hieraus wird ein aktueller Bedarf an IE für Asien und Afrika südlich der Sahara abgeleitet. Warum Asien aufgeführt wird, Osteuropa dagegen nicht, wird hierbei nicht erläutert. Entsprechend scheint insbesondere ein *Bedarf in Ostasien* und *Osteuropa* sowie *Afrika* zu bestehen.

Von den hier untersuchten Maßnahmen sind 35% dem Bereich ländliche Entwicklung (davon 6 Studien aus dem Bereich Landwirtschaft) zuzuordnen, ebenfalls 35% dem Themenbereich Wirtschaft und Beschäftigung (davon 4 Studien zu Mikrofinanzierung), 22% dem Bereich Umwelt und Infrastruktur (davon drei Infrastrukturmaßnahmen) sowie 9% dem Thema Soziale Entwicklung. Zusammenfassend heißt dies, dass im Bereich Ländliche Entwicklung zwar die Themen Infrastrukturmaßnahmen und Landwirtschaft gut vertreten sind, die Themen Wasser/Abwasser sowie Umwelt dagegen eher fehlen. Insbesondere finden sich kaum Studien zu den Themen städtische Entwicklung sowie Gesundheit (nur eine HIV/AIDS-Maßnahme). Demgegenüber identifiziert NONIE-SG 4/5 einen IE Bedarf in folgenden Sektoren: Landwirtschaftliche Entwicklung, Umweltschutz, Gesundheit sowie Gender. Übereinstimmung findet sich demnach für die Bereiche: *Umweltschutz* und *Gesundheit*.

Aktueller Bedarf an IE besteht offensichtlich für *Ostasien*, *Osteuropa* und *Afrika*
sowie in den Sektoren *Umweltschutz* und *Gesundheit*.

(2) Bei insgesamt 18 Studien wird angegeben, dass eine *Kontroll- bzw. Vergleichsgruppe* gebildet wurde. Jedoch lässt sich bei vielen Studien nicht zurückverfolgen, welche Daten dabei für welche Untersuchungseinheit erhoben wurden. Auch werden die Einschränkungen

---

bersicht der evaluierten Studien aufgeführt, dass die NONIE-Datenbank 30 Studien, durchgeführt von DFID, IOB, IEG, BMZ, NORAD and Sida, etc., enthalten würde (vgl. NONIE-SG4/5 2008: 8f.). Dagegen enthält die der hier dargelegten Auswertung zugrundeliegende NONIE-Datenbank, die auch auf der offiziellen Homepage verfügbar ist, keine Studie von IOB, IEG, BMZ, NORAD oder Sida. Auch wird im Text der "Sub-Group"-Studie ausgeführt, dass die NONIE-Datenbank 29 (nicht 30) Studien umfasst. Es wäre daher möglich, dass der Untersuchung der NONIE "Sub-Group" eine andere Datenbank zugrunde liegt (was nicht geklärt werden konnte). Da jedoch ohnehin durch "Sub-Group" 4/5 nur 16 Studien ausgewertet wurden – die übrigen enthielten zu wenig Informationen bzgl. der interessierenden Fragestellungen – und da der Anteil der aus der NONIE-Datenbank entnommenen Studien lediglich 6% der analysierten IE entspricht (vgl. NONIE-SG4/5 2008: 14), können sich die Ergebnisse evtl. eher ergänzen denn überschneiden.

[33] Eine IE – FINNIDA (2) – war eine Sektorstudie, die FINNIDAs Wasser- und Abwassermaßnahmen weltweit untersucht. Eine der Forschungsstudien, IADB (8), basiert u.a. auf der IE IADB (6) und wird daher bei der Länder- bzw. Sektorenauswertung nicht (erneut) berücksichtigt.

bzw. möglichen Verzerrungen nicht immer diskutiert. Die Studie von AusAID, von CIDA sowie eine der zwei FINNIDA Studien nutzen lediglich einen *vorher-nachher Vergleich* (ohne Vergleichsgruppe). Während bei diesen drei IE kein „härteres" Design geplant war zeigt eine Studie von DANIDA (1) wie unerwartete Ereignisse eine Evaluation bzw. das geplante Design beeinflussen kann: Die untersuchte Maßnahme in Mosambik (1985-1999) wurde während des Bürgerkrieges begonnen, wobei der Fokus auf rasche Umsetzung lag – die geplante Baseline-Studie wurde auf später verschoben, letztendlich aber nie durchgeführt. Da somit keine vorher Daten für die ZG (drei Distrikte) vorlagen, wurde auf eine Studie zurückgegriffen, die im Rahmen der Maßnahme ein einziges Dorf über den Zeitraum 1993-1995 mit Hilfe von PRA-Methoden untersuchte. Die Evaluatoren/innen leiteten aus dieser Studie vergleichbare Fragen ab und führten in diesem Dorf eine erneute Befragung durch "in order to quantify the changes" (Danida (1): 22). Darüber hinaus wurde für die nachher Messung ein Dorf „am anderen Ende" eines der Distrikte als Vergleichsgruppe gewählt.

In 18 Studien wurden *Vergleichsgruppen* gebildet.
Bei 11 dieser Studien wurde *PSM* zur Konstruktion der Vergleichsgruppe angewandt.

Interessant ist auch, wie Vergleichsgruppen gebildet wurden: Bei sämtlichen 11 Studien von IADB/OVE wird *"Propensity Score Matching"* zur Konstruktion von Vergleichsgruppen angewandt. Auffallend hierbei ist die große Bandbreite genutzter Daten: Teilweise werden eigene Daten erhoben (Baseline, "follow-up") – die Vergleichsgruppen werden so gut wie immer aus existierenden Census-Daten, Panel-Daten, repräsentativen Haushaltsbefragungen oder sonstigen Datensätze unter Anwendung von PSM gebildet. Um diese existierenden Datensätze „aufzuspüren", ist häufig vor der eigentlichen Evaluationsmission ein gesonderter vor-Ort Besuch vorgesehen. Umfangreiche Erfahrungen, wie an solche Daten zu gelangen ist, liegen offensichtlich bei IADB vor. Es wäre durchaus von Interesse, dieses Wissen in einem „Erfahrungsbericht" mit konkreten evtl. besonders „findigen" Beispielen zusammengefasst öffentlich zugänglich zu haben. Allgemein wäre es grundsätzlich sinnvoll, Erfahrungen und insbesondere „aufgefundene" Daten zusammenzutragen und der Allgemeinheit, unter Berücksichtigung rechtlicher Regelungen, zur Verfügung zu stellen. Eine bessere Koordination und Verknüpfung von IE kann die Expertise in dem Feld deutlich verbessern helfen. „Better coordination of impact evaluation across countries and institutions would make it possible to cluster some studies around common thematic areas and improve the ability to generalize findings" (CGD 2006: 4).

Einschränkend muss allerdings auf eine ebenfalls von IADB durchgeführt Forschungsstudie (8) hingewiesen werden, die aufgrund des steigenden Interesses an *PSM* diese Methode *mit experimentellen Designs* (Pipeline-Verfahren) vergleicht: "This study contributes to the small but growing literature on the performance of the propensity score matching (PSM). (…) Given that the popularity of PSM is spreading rapidly, it is most important to assess whether it provides a suitable substitute for experimental method" (IADB (8): i). Die Untersuchung kommt schließlich zu dem Ergebnis, dass es mit PSM nicht gelingt, die Ergebnisse des Experiments zu wiederholen: "We have been able to corroborate the main – rather pessimistic – conclusion of the existing literature on the performance of PSM as a non-experimental impact estimator. (…) We are not able to replicate the experimental estimates (…)" (IADB (8): 23). Als zentrale Erkenntnis wird aufge-

PSM liefert nur brauchbare Erkenntnisse, wenn die zugrundegelegten Daten mit *ähnlichen Instrumenten* bzw. *Fragen* erhoben wurden wie bei der Zielgruppe

führt, dass PSM nur brauchbare Erkenntnisse liefert, wenn neben der Berücksichtigung verschiedener Störfaktoren bzw. Verzerrungen auch überprüft wird, dass die Daten, die zur Bildung der Vergleichsgruppe herangezogen werden, mit den *gleichen/ähnlichen Instrumenten* und Fragebögen (Fragen) erhoben wurden wie bei der ZG.

Echte *experimentelle Designs* wurden ausschließlich in Kombination mit dem *Pipeline-Verfahren* angewandt: Bei zwei Studien (IADB (6), DFID (2)) zu Finanzierungsmaßnahmen wurde zu Beginn die Gesamtgruppe der Begünstigten bestimmt, anschließend über einen *Zufallsprozess* (Randomisierung) in eine Gruppe eingeteilt, die umgehend bzw. in der ersten Phase Finanzmittel zugewiesen bekam, und eine zweite Gruppe, die erst im Rahmen der zweiten Phase berücksichtigt wurde. Letztere konnte entsprechend als KG der ersten Gruppe (ZG) genutzt werden. Diese zwei Studien sind ein gutes Beispiel, dass echte experimentelle Designs ohne ethische Einwände denkbar sind.

Echte *experimentelle Designs* sind insbesondere über das *Pipeline-Verfahren* ohne ethische Bedenken gut anwendbar.

Während bei den hier ausgewerteten NONIE-Studien nur zwei Studien experimentelle Designs anwenden und hierbei ausschließlich in Kombination mit dem Pipeline-Verfahren, werden bei 32% der 250 von NONIE-SG 4/5 untersuchten IE (angeblich) experimentelle Designs genutzt (vgl. NONIE-SG4/5 2008: 18). Allerdings wird nur in einem kurzen Abschnitt auf die unterschiedlichen genutzten Methoden eingegangen. Daher wird die angegebene Prozentzahl stark in Zweifel gezogen, denn: auch in den hier untersuchten Studien wurden durchaus mehrfach experimentelle *Auswahlverfahren* angewandt, d.h. aus einer zuvor definierten Grundgesamtheit einer angemessenen Vergleichsgruppe (=quasi-experimentelles Design) wurde *per Zufall eine Stichprobe* der zu untersuchenden Einheiten gezogen – dies ist jedoch nicht zu verwechseln mit einem experimentellen *Design.*

Eine andere Form des *Pipeline-Verfahrens* im Rahmen von Finanzierungsmaßnahmen wurde bei einem Projekt zur Mikrofinanzierung von DFID (1) angewandt: Da über die Vergabe der Kredite teilweise innerhalb von 24 Stunden entschieden wurde, war es nicht möglich, die Antragsteller über eine Zufallsauswahl in Gruppen aufzuteilen, die sofort bzw. erst einige Zeit später (in einer zweiten Phase) die Kredite ausgezahlt bekommen. Daher wurden alle Kreditempfänger zum Zeitpunkt der Evaluation in zwei Gruppen geteilt: In eine Gruppe, die innerhalb der letzten sechs Monate einen Kredit erhalten hatte (als Vergleichsgruppe) und eine Gruppe, die vor maximal fünf aber mindesten drei Jahren einen Kredit erhalten hatte (als Zielgruppe). Die Daten wurden zu einem Zeitpunkt erhoben, so dass die Wirkungen unter Berechnung der "single-difference" Methode, d.h. über einen mit-ohne Vergleich abgeleitet wurden.

Eine der DFID-Studien (2) war ebenfalls wie die oben aufgeführte IADB-Studie (8) keine IE im eigentlichen Sinne sondern eher eine Forschungsstudie, die die Wirkungen unterschiedlicher Maßnahme*elemente* miteinander verglich (Reduzierung von Korruption über Partizipation vs. Kontrolle/Audits). In dieser Studie wurden mögliche *"spill-over" Effekte* zuvor bedacht und angemessen berücksichtigt: Es wurde vermutet, dass sich die Nachricht über die Einführung von Audits in den ausgewählten Dörfern über deren Dorfgrenzen hinweg verbreiten könnte. So wurde vermutet, dass auch in Dörfern, bei denen keine Audits eingeführt werden sollten, die Angst vor Audits die Korruption verringern und somit die Ergebnisse verzerren könnte. Da weiterhin davon auszugehen war, dass sich die Entscheidungsträger des Dorfes innerhalb eines Subdistriktes häufig austauschen, die Kom-

munikation zwischen Subdistrikten dagegen begrenzt ist, wurde die Zufallsstichprobe nicht aus der Gruppe der Dörfer sondern aus der Gruppe der *geclusterten Distrikte* ausgewählt. Hierdurch konnten "spill-over" Effekte erfolgreich vermieden werden.

Ein gelungenes und gleichzeitig außergewöhnliches Beispiel für die Konstruktion von Vergleichsgruppen auf Basis von *"matching on observables"* zeigt die JBIC-Studie (1): Im Rahmen einer Infrastrukturmaßnahme sollte durch den Bau einer Brücke der weniger entwickelte Nordwesten mit der besser gestellten östlichen Region verbunden werden. Als Zielgruppe wurden daher fünf Dörfer im Nordwesten gewählt. Als Vergleichsgruppe wurden zwei weitere Dörfer östlich der Brücke gewählt, also der Region, die bereits besser entwickelt war. Ungewöhnlich hier ist, dass als Vergleichsgruppe nicht eine Region gewählt wird, die weniger entwickelt ist (und von der angenommen wird, das sie sich in Zukunft auch nicht entwickeln wird), sondern statt dessen auf eine bereits entwickelte Region zurückgegriffen wird, bei der die Entwicklung – so die Annahme – durch den Brückenbau nicht entscheidend bzw. in weitaus geringerem Umfang beeinflusst wird. Auch die *Auswahl der ZG und VG* ist in dieser Studie sorgfältig gewählt: Auf Basis zweier Haushaltsbefragungen, die im Rahmen anderer Projektmaßnahmen erhoben worden sind, wurden Dörfer im Nordwesten und Osten aussortiert, die sich bezüglich agroökologischer und sozioökonomischer Parameter sehr *ähnlich* waren (observables). Aus diesen Gruppen wurden anschließend per Zufallsauswahl die fünf Dörfer für die ZG und die zwei Dörfer der KG ausgewählt.

„*Matching on observables*“ kann – gut umgesetzt – durchaus zu einer *angemessenen* Vergleichsgruppe führen.

Diese Studie verdeutlicht darüber hinaus auch die Vorteile, die bei IE von *FZ-Maßnahmen* bestehen: Unabhängig davon, ob im Rahmen der Maßnahme vorher Daten erhoben wurden, ist es für Evaluatoren/innen zeitlich dennoch möglich, Daten selbst zu erheben: So wurde im vorliegenden Beispiel 1994 mit dem Brückenbau begonnen und im Juni 1998 beendet. Die Evaluation wurde erst kurz vor Inbetriebnahme der Brücke begonnen. Hierdurch war es möglich, Baselinedaten selbst zu erheben. Entsprechend konnten für die sieben ausgewählten Dörfer Paneldaten[34] erhoben werden (1997/98 und 2003/04).

(3) Bei 16 von den 18 Studien mit Vergleichsgruppe wird angegeben, dass auf Basis der Vergleichsgruppen die *double-difference Methode* angewandt wurde. DD wurde bei allen elf IADB- und allen drei IFAD-Studien angewandt. Des Weiteren wird DD bei der soeben dargestellten Studie von JBIC sowie bei einer der zwei Danida- und einer der zwei DFID-Studien berechnet, um die Wirksamkeit der Maßnahme aufzuzeigen. Bei zwei Studien lagen lediglich zu einem Zeitpunkt Daten für ZG und VG vor, so dass die *single-difference Methode* angewandt wurde: Bei der DIFID-Studie (1) wurden wie oben bereits dargestellt nur zu einem Zeitpunkt Daten erhoben. Bei der zweiten Studie mit VG aber ohne DD (AfDB (1)) lagen keine Basisdaten vor. Ohne die Möglichkeit zu diskutieren, z.B. auf existierende Daten für den Zeitpunkt des Projektstarts zurückzugreifen, wird ein einfacher mit-ohne Vergleich durchgeführt.

Eine Möglichkeit, Informationen über die vorher Situation trotz fehlender Basisdaten zu erhalten, zeigen die drei Studien von IFAD: Bei allen IE wurde aufgrund fehlender – bzw.

[34] Bei Panelerhebungen werden dieselben Personen zu einem späteren Zeitpunkt erneut befragt (nicht eine ähnliche Gruppe) – im hier gegebenen Beispiel wurde des Weiteren eine Vollerhebung (Census) umgesetzt, d.h. es wurden alle Haushalte in den sieben Dörfern befragt.

in einem Fall unbrauchbarer - Baselinedaten die wahrgenommenen Veränderungen mit Hilfe der sogenannten *recall Methode* analysiert, d.h. in den durchgeführten Interviews waren auch Fragen zur vorher Situation enthalten. Diese Vorgehensweise wirft durchaus andere Probleme auf, in den hier dargestellten IFAD-Studien wird dem aber – so die Einschätzung – durch Triangulation, d.h. der Anwendung vielfältigster qualitativer und quantitativer Methoden, entgegen gewirkt.

Wird eine VG gebildet, dann werden Daten zu *zwei Zeitpunkten* erhoben und die Wirkungen der Maßnahmen mit Hilfe der *DD-Methode* kausal zugeschrieben: Bei 16 der 18 mit Vergleichsgruppen durchgeführten Studien wurden DD berechnet.

Es zeigt sich entsprechend – selbst bei Nichtberücksichtigung der überrepräsentativ vertretenen IADB-Studien: Wird eine Vergleichsgruppe gebildet, dann werden auch meist Daten zu zwei Zeitpunkten erhoben (bzw. rekonstruiert) und die Wirkungen der Maßnahmen mit Hilfe der double-difference Methode kausal zugeschrieben.

(4) Bei sechs der Studien wird angegeben, dass der Evaluation ein *LogFrame* zugrunde lag, allerdings werden diese in kaum einer Studie aufgeführt, auch nicht im Anhang. Aus den – teilweise recht kurzen – Darstellungen der LogFrames ist leider nicht ersichtlich, inwieweit die entscheidenden Ursache-Wirkungs-Hypothesen, mögliche Risiken, die Indikatoren sowie die Datenquellen aufgenommen wurden. Die übrigen Studien sind teilweise – so die Aussagen bzw. Interpretationen – *theoriegeleitet* bzw. es werden *Hypothesen* formuliert. In den meisten Studien werden jedoch lediglich Schlüsselfragen formuliert, die eher allgemeiner Art sind.

Die Verwendung von *LogFrames* scheint in der Praxis noch zu gering.

Werden die Berichte, die ohne LogFrame dafür mit – teilweise durchaus elaborierten – Ursache-Wirkungs-Hypothesen arbeiten (wie z.B. alle IADB-Studien), näher betrachtet, zeigt sich Folgendes: Bei der Erarbeitung von LogFrames – sei es in der Planungsphase oder aber im Rahmen der Evaluation – sind die Stakeholder in den Prozess mit eingebunden (bzw. sollten eingebunden sein). Dagegen werden bei Evaluationen, die „lediglich“ Hypothesen erarbeiten, diese meist ausschließlich auf Basis teilweise durchaus umfangreicher Literaturauswertungen aufgestellt. Dass hierdurch nicht-intendierte Wirkungen tatsächlich erfasst werden können ist zu bezweifeln.

(5) In allen Studien wurden *quantitative Methoden* angewandt, d.h. keiner der Studien basiert auf ausschließlich qualitativen Methoden. Bei sieben der Studien wird ausdrücklich angegeben, das ein *Methodenmix* (Triangulation) genutzt wurden. Bei weiteren fünf Studien ist aus dem Bericht erkennbar, dass auch qualitative Methoden angewandt wurden. Als gutes Beispiel kann diesbezüglich auf die Danida-Studie (2) verwiesen werden: Die angewandten Methoden sind derart zahlreich, dass sie hier lediglich im Überblick aufgezählt werden: Dokumentenanalyse, Archivarbeit, Zusammenstellung und Analyse wichtiger Materialen zur Erfassung des Kontextes (Statistiken, Artikel, Berichte, etc.), Interviews mit ehemaligen Verantwortlichen, Kompetenzträgern und anderen zentralen Akteuren des Partnerlandes sowie des Geberlandes Dänemark, Interviews mit aktuellen Stakeholdern sowie zentralen Akteuren, fragebogenbasierte Interviews mit Teilnehmer/innen sowie Nicht-Teilnehmer/innen, Fokusgruppen-Interviews, Tiefeninterviews (themenorientiert und Lebensgeschichten), quanti-

Nicht bei allen Studien wurde ein *Mix* aus *quantitativen und qualitativen Methoden* angewandt.

tative Analyse der Monitoring Daten, repräsentativer Survey der Projektkomponenten, Institutionelles "Mapping", Beobachtungen, Bewertung von Gebäuden, Straßen und Bewässerungskanälen, großflächige Dorfstudien (Surveys).

Demgegenüber zeigt sich bei den IADB-Studien, dass Triangulation offensichtlich nicht grundsätzlich geplant ist, sondern lediglich dann „nachgeschoben“ wird, wenn die Ergebnisse aus selbst durchgeführten standardisierten Interviews bzw. Sekundärdatenauswertungen Fragen aufwerfen. In diesen Fällen wurden Intensivinterviews oder auch Fokusgruppeninterviews nachgeschoben (IADB (3), (5), (11)).

(6) Abschließend sei noch erwähnt, dass eigentlich alle untersuchten Studien eher auf die "Outcome"-Ebene fokussieren – teilweise in der Tat auf mittelfristige Wirkungen, häufig jedoch auf die direkten Ziele (kurzfristige Wirkungen). Jedoch wird meist ausschließlich von Impact gesprochen und nicht zwischen langfristigen Impacts und kurzfristigen oder mittelfristigen Outcomes unterschieden. Lediglich bei den wenigen Studien, bei denen die DAC-Kriterien bearbeitet wurden, wird mit den Begriffen etwas differenzierter umgegangen (AfDB, Finnida, IFAD).

Als zentrale Ergebnisse der Auswertung kann Folgendes festgehalten werden:

- Es liegt kein Beispiel vor, das aufzeigt, wie unter Anwendung *ausschließlich qualitativer Datenerhebungsmethoden* gesicherte Aussagen über die Zuschreibung der Wirkungen zu einer Maßnahme oder auch über den Beitrag einer Maßnahme zu bestimmten Wirkungen möglich sind.
- Andererseits zeigt sich, dass trotz Betonung der Notwendigkeit von *Triangulation/Methodenmix* einige der Studien sich *ausschließlich* auf quantitativ erhobene Daten berufen. Wenn Daten selbst erhoben wurden, dann häufig über standardisierte Interviewmethoden.
- Weiter zeigt sich, dass trotz Betonung der Notwendigkeit eines *theoriebasierten Ansatzes*, ein *LogFrame* mit TOC oder zumindest fundierte Ursache-Wirkungs-Hypothesen, die anschließend systematisch Schritt für Schritt überprüft werden, nur bei einem Teil der untersuchten Studien enthalten sind.
- Bei stark quantitativ fokussierten Studien werden zwar häufig gesicherte Aussagen zur Wirksamkeit gegeben, fundierte Antworten auf die Frage *warum* oder gar ob nicht-intendierte Wirkungen eingetreten sind, werden dagegen selten geliefert.

Aus dieser Bilanz könnte abgeleitet werden, dass die Kritik beider oben dargestellter „Schulen“ an der jeweils anderen berechtigt scheint. Es könnte von Vorteil sein, den jeweils gewinnbringenden Beitrag beider Seiten für anspruchsvolle Impact Evaluationen hervorzuheben. Ein erster Schritt könnte darin bestehen, die folgenden zwei Elemente nicht nur in den diversen Texten zu betonen sondern dezidiert in die *Definition* von RIE aufzunehmen, um somit die zentrale Bedeutung hervorzuheben:

- Die Notwendigkeit eines *theoriebasierten Ansatzes*, um "black box" Evaluationen zu verhindern und die Frage nach dem „warum“ adäquat untersuchen und beantworten zu können.

- Die Verwendung der *gesamten Bandbreite an Datenerhebungsmethoden und Tools* – in Abhängigkeit der jeweiligen Fragestellung. Demnach müssen qualitative Methoden genutzt werden, um Wirkungshypothesen zu erstellen bzw. zu überprüfen und dabei auch mögliche nicht-intendierte Wirkungen zu erfassen, aber auch um gefundene Ergebnisse adäquat interpretieren zu können. Quantitative Methoden sind notwendig, um die hypothetisierten Wirkungen, deren Richtung und Stärke sowie deren Ursachen nachweisen zu können.

Denn auch diese Elemente sind notwendig, um die Wirkungen gemäß der OECD/DAC-Definition aufzeigen zu können.

11. RELEVANZ DER IE-DISKUSSION FÜR DIE EVALUIERUNGS-PRAXIS

Zentrale Erkenntnis der Analyse der relevanten Papiere zum Thema ist, dass IE, die keine *systematische* Befragung der Zielgruppe beinhalten, einheitlich nicht mehr als ausreichend angesehen werden – einige Gespräche mit Zielgruppen*vertretern/innen* im Rahmen von vor-Ort Besuchen genügen nicht mehr: "It is all too common-place to restrict data collection to key informant interviews and perhaps a few focus groups" (White 2006a: 20). Die Wirkungen insbesondere auf Seiten der Begünstigten sind *systematisch* zu erfassen, wobei systematisch die Nutzung angemessener Designs hervorheben soll[35] (vgl. Abb. 9):

Abbildung 9: Übersicht der in der EZ realisierbaren Versuchsanordnungen

QUALITÄT

	DESIGN	Vorher-Daten t_1 (Baseline)	Maß-nahme *X*	Nachher-Daten t_2 (Survey)
	Experimentelle Versuchsanordnung:			
(1)	Kontrollgruppen-Design	**ZG_1** **KG_1**	**X** **–**	**ZG_2** **KG_2**
	Quasi-experimentelle Versuchsanordnung:			
(2)	Vortest-Nachtest mit Vergleichsgruppen-Design	**ZG_1** **VG_1**	**X** **–**	**ZG_2** **VG_2**
(3)	„verspätetes" Vergleichsgruppen-Design		**X ZG_1** **– VG_2**	**ZG_2** **VG_2**
(4)	Vortest-Nachtest mit Nachtest Vergleichsgruppen-Design	**ZG_1**	**X** **–**	**ZG_2** **VG_2**
(5)	nur-Nachtest-Design mit Vergleichsgruppen		**X** **–**	**ZG_2** **VG_2**
	Vorexperimentelle Versuchsanordnung:			
(a)	Ein-Gruppen-Vortest-Nachtest-Design	**ZG_1**	**X**	**ZG_2**
(b)	Ein-Gruppen-Nachtest-Design		**X**	**ZG_2**

ZG: Zielgruppe, KG: randomisierte Kontrollgruppe, VG: nicht-randomisierte Vergleichsgruppe, *t:* Zeitpunkt

[35] Bisher unerwähnt blieb der Hinweis, dass Befragungen der ZG oder VG *repräsentative* Befragungen implizieren.

Eine Befragung im Rahmen einer IE selbst durchzuführen, scheint grundsätzlich auch unter Kostengesichtspunkten realistisch – entscheidend sind die Zeitpunkte: Die Umsetzung eines klassischen Survey-Designs (vgl. Abb. 9 (5)), d.h. eine Befragung sowohl der Ziel- aber auch einer Vergleichsgruppe zum Evaluationszeitpunkt, ist grundsätzlich durchführbar. Von daher sollte dies für EZ-Evaluierungen als Mindestanforderung anerkannt werden. Das heißt in Konsequenz, dass die in Abbildung 9 dargestellten vorexperimentellen Versuchsanordnungen, (a), (b) wobei lediglich die ZG befragt wird, für IE nicht mehr als ausreichend erachtet werden. Grundsätzlich sollte jedoch versucht werden, auch einen vorher-nachher Vergleich umzusetzen, zumindest für die ZG eine vorher Messung im Design zu berücksichtigen (vgl. Abb. 9 (4)). Dies ist jedoch im Rahmen einer Evaluation nicht mehr zu beeinflussen: Baselinedaten für die ZG liegen im Rahmen einer IE entweder vor oder nicht. Alternativ kann hier im Rahmen der Befragungen zum Evaluationszeitpunkt t_2 über retrospektive Fragen (recall Methode) versucht werden, Informationen für den Zeitpunkt t_1 zu erhalten.

Wurden Baselinedaten zumindest für die ZG erhoben, so kann versucht werden über zeitlich entsprechende Sekundärdaten eine Vergleichsgruppe für diesen Zeitpunkt zu konstruieren (vgl. Abb. 9 (3), (2)). Wenn auf Sekundärdaten, d.h. existierende Censen, Surveys, etc., zurückgegriffen werden soll, ist zuvor zu prüfen, ob die interessierenden Variablen enthalten sind und wenn ja, inwieweit die Fragebögen bzw. Fragen mit denen der Baselinedaten der ZG übereinstimmen (vgl. oben die Erkenntnisse aus der IADB-Studie zur Qualität von PSM).

Grundsätzlich stellt sich jedoch die Frage, wie realistisch es ist, *adäquate Sekundärdaten* aufzufinden. Die Erfahrungswerte auf internationaler Ebene scheinen diesbezüglich sehr unterschiedlich. Wie bereits dargestellt, wäre es ein Gewinn, wenn sämtliche Erfahrungen einzelner Organisationen bezüglich Datenbeschaffung in ihrer ganzen Breite und Tiefe zugänglich wären – z.B. in Form von Erfahrungsberichten. Auch könnte überlegt werden, wie diverse Datensätze, die sich im Kontext von IE z.B. für Matching-Verfahren als geeignet erwiesen haben, zusammengetragen und anderen Organisationen zur Verfügung gestellt werden könnten. In diesem Kontext wäre es auch von Vorteil, sich mit Initiativen auszutauschen bzw. abzustimmen, die die Qualität der nationalen Statistikkapazitäten in einzelnen Entwicklungsländern als Basis für eine effektive und zielorientierte Entwicklungspolitik zu verbessern suchen – z.B. „PARIS 21“ (www.paris21.org; vgl. auch Kusek, Rist u. White 2005: 21f.).

Eine weitere Frage, sie sich direkt anschließt, betrifft die *Anforderungen an die Evaluatoren/innen*. Während in allen untersuchten Dokumenten weder die Anforderungen an Daten noch die Beschaffung als problematisch oder gar unlösbar gesehen wird, wird immer wieder auf die Herausforderungen hingewiesen, die die Durchführung von IE an Evaluatoren/innen stellen: “Designing and conducting IEs requires high levels of scientific and professional expertise (…). It has been observed that this degree of technical sophistication is ‘often lacking’ in the field of applied development (…). Rather, the expertise required exists at present in only a limited number of institutions, largely in Northern contexts (…)” (NONIE-SG4/5 2008: 10, vgl. auch Baker 2000, Bamberger et al. 2006, CDG 2006, White 2006a). Allerdings stellt sich die Frage, inwieweit Evaluatoren/innen über den ganzen benötigten Wissensfundus verfügen *können*. Insbesondere für die durchaus sinnvoll erscheinenden Matching-Verfahren zur Konstruktion adäquater Vergleichsgruppen auf Basis von Sekundärdaten sind Kenntnisse nötig, die nicht in einem 3-Tage Fortbildungskurs vermittelt werden können. Hier müssen Lösungen gefunden werden. Eine Möglichkeit wäre, dass Evaluatoren/innen aufgefundene Sekundärdaten inklusive der entscheidenden Merkmale an eine vom Auftraggeber unter Vertrag genommene

Institution, die über entsprechendes Fachwissen verfügt, weiterreichen, die die notwendigen Datensätze konstruieren (berechnet). Evtl. könnten auch die gesamte Datenauswertung an diesen Partner gegeben werden, d.h. auch selbst erhobene Daten inklusive der aufgestellten Wirkungshypothesen (Pfadmodell). Welche Wege auch immer angedacht werden, fest steht: Sollen Matching Verfahren vermehrt Anwendung finden, ist es unumgänglich für das Problem der fehlenden (ökonometrischen) Kompetenz eine Lösung zu finden.[36]

Ein anderer Aspekt bzgl. fehlender Kompetenzen für die Durchführung von IE betrifft die Partnerländer. Hier ist noch weitaus weniger Wissen über und Erfahrungen mit anwendungsbezogener Forschung und Programmevaluationen vorhanden als in den Geberländern. Entsprechend ist "capacity building" der Partner vornehmlich in Grundlagen von IE nötig. Nur dann können bei „gemischten“ Teams die Evaluatoren/innen der Partnerländer gleichberechtigt an der Evaluation mitarbeiten. Capacity Building in den Partnerländern würde nicht nur die Anzahl und Qualität von IE steigern, sondern auch zu einer stärkeren Nutzung der Ergebnisse zur Politikgestaltung in den Ländern selbst führen.

Eine weitere Erkenntnis der Analyse betrifft die zentrale Bedeutung detailliert konstatierter *Ursache-Wirkungs-Hypothesen* als Grundlage einer adäquaten Impact Evaluation. Die Erarbeitung von LogFrames ist keine einfache und insbesondere schnell zu lösende Aufgabe: Wie dargestellt wurde kann es von zentraler Bedeutung sein, wenn diese lediglich auf Basis von (Fach-) Literatur abgeleitet werden ohne die Einbindungen relevanter Stakeholder. Wenn Evaluatoren/innen bereits zu Projektbeginn erstellte LogFrames im Rahmen ihrer Mission überprüfen, um mögliche nicht-intendierte Wirkungen „aufzuspüren“, benötigt dies – unter Beteiligung der Partner umgesetzt – bereits eine beachtliche Zeit. Müssen Evaluatoren/innen aber für die Durchführung einer IE ein gesamtes LogFrame erarbeiten, weil zu Beginn der Maßnahme ein solches nicht niedergeschrieben wurde, so muss dies bei der Auftragsgestaltung bzw. der Formulierung der TORs berücksichtigt werden. Ansonsten werden Evaluatoren/innen weiterhin den wichtigen Aspekt der nachprüfbaren Hypothesenbildung vernachlässigen, was wiederum negative Auswirkungen auf die Qualität der Ergebnisse haben wird. Die Auswertung der Studien zeigte, dass IE auf Basis zuvor formulierter Ursache-Wirkungs-Hypothesen in der Praxis kaum zu finden sind – noch weitaus weniger als z.B. die Anwendung von Matching-Verfahren. Es scheint notwendig, die IE-Diskussion auch auf solche Fragen der praktischen Umsetzung zu fokussieren.

Bei der Evaluation von *neuen Instrumenten* wie *Gemeinschaftsfinanzierungen* oder vergleichbare *sektorweite Ansätze* (SWAps oder Budgethilfe, GBS) kommt den Ursache-Wirkungs-Hypothesen bzw. der Programmtheorie eine weitaus größere Bedeutung zu als bei üblichen IE. Da diese neuen Instrumente die Beziehung zwischen Geber- und Empfängerland neu definieren, verändern sich auch die Hypothesen in der Wirkungskette: GBS und SWAp zielen auf “Alignment“ und verstärkte “Ownership“ ab. Somit ist Ziel von GBS und SWAp, das Partnerland bei der Erreichung *seiner eigenen Ziele* zu unterstützen.[37] Daraus folgt jedoch die Frage,

[36] Denn es scheint auch umgekehrt nicht angemessen, zukünftig Ökonometriker/innen als Evaluatoren/innen einzusetzen, da diesen häufig wiederum grundlegende Kenntnisse in Evaluation sowie Datenerhebungsmethoden insbesondere qualitativen Methoden sowie partizipativer Tools fehlen.

[37] Im NONIE-Papier der Sub-Group 3 wird eine detaillierte Darstellung der „Pawsons’s ‚Simple Principles’ for the Evaluation of Complex Programmes“ aufgelistet (vgl. NONIE-SG3 2008: Anhang V, 42ff.).

was genau Gegenstand der Analyse sein soll: (1) Die Wirkungen der Unterstützung auf die Sektorpolitik des Empfängerlandes, (2) die Wirkungen der veränderten Sektorpolitik des Empfängerlandes, (3) die Wirkungen auf Serviceleistungen oder (4) die Wirkungen auf der Haushaltsebene.

Aber auch methodisch stellt die IE von GBS oder SWAp eine besondere Herausforderung dar. Denn während es einen zunehmenden Konsens über Methoden und Techniken für IE von spezifischen Projekten und Programmen gibt, existieren bisher keine handfesten Empfehlungen, wie die Wirkungen dieser neuen Instrumente überprüft oder gemessen werden können. Grundsätzliches Problem bei Wirkungsuntersuchungen derartiger Ansätze und Instrumente ist, dass es schwerlich möglich sein wird, bei SWAp oder GBS eine Vergleichsgruppe zur Abbildung des Kontrafaktischen zu konstruieren: "While it is definitely possible to give an *assessment* of the effectiveness of budgets support or sector support, it seems almost impossible to measure the (quantitative) impact", denn durchgeführte Ländervergleiche haben gezeigt, dass hierbei zu viele methodologische Probleme zusammenkommen (NONIE-SG3 2008: 3).

Aufgrund der besonderen methodischen Probleme, die IE von GBS oder SWAps mit sich bringen, hat sich innerhalb NONIE eine eigene Sub-Group mit diesem Thema auseinandergesetzt. Für IE auf Sektorebene wurden methodische Vor- als auch Nachteile gegenüber „normalen" IE identifiziert (vgl. NONIE-SG3 2008: 17):

- Während eine Maßnahme meist einen eindeutigen Startpunkt hat, ist dies nicht notwendigerweise auf der Sektorebene gegeben, was einen vorher-nachher Vergleich erschwert.
- Meist sind sektorweite Maßnahmen schwer abgrenzbar und nicht automatisch auf eine spezifische Zielgruppe ausgerichtet, so dass es schwierig bzw. unmöglich wird, eine Zielgruppe und eine Vergleichsgruppe zu unterscheiden.
- Die Heterogenität der Maßnahmen macht es schwierig, einzelne Maßnahmen von anderen zu isolieren, so dass Übertragungseffekte (contagion) überall vorhanden sind.
- Andererseits kann diese Heterogenität genutzt und zum Gegenstand der Analyse gemacht werden.
- Die Nutzung von Sekundärdaten scheint für IE auf Sektorebene erfolgversprechender als auf Projektebene, denn es ist weitaus wahrscheinlicher, für diese Ebene Paneldaten zu erhalten. Allerdings bestimmen solche Daten auch die Analyseebene und schränken die Entscheidungen der Evaluatoren/innen, die beste Analyseebene auszuwählen, stark ein.
- Heterogenität und mögliche Probleme mit der Datenqualität bedingen entsprechend große Datensätze.
- Wenn Sekundärdaten genutzt werden können, liegen meist auch große Datensätze vor. Allerdings sollten diese immer auf Plausibilität und Reliabilität überprüft werden, indem z.B. verschiedenen Datensets verglichen werden.

Wie diese methodischen Vor- und Nachteile letztendlich bei der Analyse der Wirkungen von sektorweiten Maßnahmen berücksichtigt werden, wird zukünftig ein zentraler Diskussionspunkt sein, da diese Instrumente mehr und mehr an Bedeutung gewinnen.

Eine weitere Frage, die im Kontext der IE-Diskussion noch nicht ausreichend behandelt wurde, ist, wie die DAC-Kriterien Relevanz, Effektivität, Effizienz und Nachhaltigkeit im Kontext einer IE sinnvoll Berücksichtigung finden. So wurden nur bei wenigen der hier untersuchten Studien

neben den entwicklungspolitischen Wirkungen die weiteren DAC-Kriterien bearbeitet. Lediglich die Studien von AfDB, Finnida sowie IFAD haben *zusätzlich alle* diese Kriterien untersucht. Allerdings wird aus den Studien nicht eindeutig ersichtlich, wie die einzelnen DAC-Kriterien miteinander verknüpft werden. Dies mag daran liegen, dass es bisher noch keine Diskussion über die sinnvolle Einbindung der DAC-Kriterien in IE gab. Das Problem hierbei ist Folgendes: Die DAC-Kriterien wurden 1991 in den "OECD/DAC-Principles" als im Rahmen von Evaluationen zu erhebende Größen definiert. Allerdings wurde nie erläutert oder diskutiert, in welchen *Zusammenhang* diese einzelnen Kriterien zueinander stehen. Im Ergebnis findet sich häufig ein Nebeneinander der jeweils abgearbeiteten einzelnen Kriterien. Doch die DAC-Kriterien bieten weitaus mehr, sie können in das Pfadmodell der Ursache-Wirkungshypothesen eingebaut werden (vgl. Caspari 2004: 220ff.):

Relevanz ist in erster Linie eine Bewertung des *Zielsystems* in der *Planungsphase* bzw. der eingeführten *Innovation* und somit eine erklärende Variable der Wirkungen. Die *Effektivität* dagegen – im Sinne der Zielerreichung – entspricht den direkten intendierten Wirkungen (Outcomes), die u.a. durch die Ergebnisse des Relevanz-Kriteriums beeinflusst wurden (wenn Maßnahmen nicht mit dem Bedarf der ZG übereinstimmen, wird dies sicherlich die Outcomes schmälern). Das Kriterium *Impact* beinhaltet – wie in diesem Papier ausführlich dargestellt wurde – die mittel- und langfristigen Wirkungen einer Maßnahme. Die Frage der *Effizienz* setzt die Ergebnisse der Kriterien Effektivität und/oder Impact in Relation zu den Kosten. *Nachhaltigkeit* wiederum überprüft die langfristigen entwicklungspolitischen Wirkungen.

Die einzelnen DAC-Kriterien entsprechen demnach einzelnen aufeinander aufbauenden Elementen der Wirkungskette. Es wäre für die weitere Diskussion über IE von Vorteil, wenn die Relevanz der einzelnen Kriterien für die im Rahmen des LogFrame aufzustellenden Ursache- und Wirkungsindikatoren herausgearbeitet würden. In diesem Kontext könnte auch geklärt werden, wie das Problem, dass einige der Kriterien sowohl Fragen bezüglich der erklärenden Variablen (Ursachevariablen, AV) als auch der zu erklärenden Variablen (Wirkungsvariablen, UV) enthalten, gelöst werden kann: Das Kriterium Effektivität enthält z.B. die Frage, ob eine Maßnahme die intendierten direkten Wirkungen vollständig erreicht/nicht erreicht hat – dies entspricht einer Wirkungsmessung auf Ebene der Outcomes (=AV). Dagegen wird auch gefragt, inwieweit die Ziele angemessen waren – dies entspricht wohl eher einer erklärenden Frage, die auf die Zielerreichung einwirkt (UV) denn: Wurden die intendierten Wirkungen (AV) eines Vorhabens nicht erreicht, müssen mögliche Ursachen (UVs) gesucht werden. Eine Möglichkeit ist z.B. eine nicht angemessene Zielformulierung zu Beginn![38]

Im Kontext der IE-Diskussion wird bisher lediglich auf das Kriterium Effizienz eingegangen: Es wird mehrfach betont, dass sich IE bestens als *Grundlage* eignen, um in eine angemessene Kosten-Nutzen- (“cost-benefit analysis“, CBA) bzw. Kosten-Wirksamkeits-Analyse (“cost-effectiveness analysis“) einzufließen: “Incorporating cost-benefit or cost-effectiveness analysis is also strongly recommended. This methodology can enable policymakers to compare alternative interventions on the basis of the cost of producing a given output” (Baker 2000: 15). Diese Möglichkeit wird allerdings bisher bei IE eher selten genutzt (vgl. White 2006a: 12). Dies verdeutlicht, dass sowohl im Rahmen der Diskussion als auch bei der praktischen Durchführung von IE die DAC-Kriterien bisher noch nicht ausreichend eingebunden sind.

[38] Zur grundsätzlichen Problematik der DAC-Kriterien als Teilindikatoren der Projekt*wirkungen* vgl. auch Caspari, 2004.

Die Ausführungen machen deutlichen, dass sich die IE-Diskussion erst am Anfang befindet. Für die Umsetzung in die internationale Evaluierungspraxis sind noch viele insbesondere praktische Fragen zu klären. Von daher scheint es zum jetzigen Zeitpunkt unrealistisch, die aufgeführten Anforderungen an alle Evaluationen zu stellen. Insbesondere da die Durchführung einer IE grundsätzlich einen erheblichen Informations-, Zeit- und Ressourcenbedarf mit sich bringt, sollten Maßnahmen, die einer IE unterzogen werden, sorgfältig ausgewählt werden. Als zentrales Auswahlkriterium eignet sich das erwartete *Lernpotential* einer IE, d.h. inwieweit angenommen wird, dass die Ergebnisse entscheidend für weitere zukünftige Maßnahmen sein werden. „Proposals for rigourous evaluation include calculations showing the study is of large enough size to identify with high probability effects large enough to matter for policy purposes ("power calculation")" (CGD 2006: 73). Verschiedene Möglichkeiten sind denkbar:

- Wenn eine "Policy" oder ein Programm von *strategischer Relevanz* für Armutsreduzierung ist: Bei Maßnahmen, von denen vermutet wird, dass sie einen großen Einfluss auf Armut haben, können IE genutzt werden um sicherzustellen, dass die Bemühungen zur Armutsreduzierung auf dem richtigen Weg sind und ggf. Modifizierungen vorgenommen werden können.
- Wenn ein *neuer vielversprechender, innovativer Ansatz* oder auch eine *neue Policy* eingeführt wurde, können IE genutzt werden, um zu überprüfen, ob diese neuen Ansätze erfolgreich sind und somit ausgeweitet und in einem größeren Umfang eingesetzt werden sollten. Hierbei sollten IE insbesondere auf den Vergleich unterschiedlicher Variationen einer Maßnahme fokussieren, was entsprechend bereits bei der Planung berücksichtigt werden muss.
- Wenn ausreichend Kenntnisse vorhanden sind, dass die untersuchte Maßnahme auch in *anderen Kontexten* wirkt: Der Aufwand einer IE lohnt lediglich, wenn vorab ersichtlich ist, dass die Ergebnisse auf Maßnahmen mit anderen lokalen Bedingungen übertragbar sind. Ist eine Maßnahme dagegen auf spezifische Umstände und Gegebenheiten oder gar Zielgruppen angelegt, kann bezweifelt werden, dass die Ergebnisse einer IE dieser Maßnahme sinnvoll für andere Maßnahmen genutzt werden können.
- Wenn zu erwarten ist, dass intendierte Wirkungen einer Maßnahme *eingetreten* sind: Die grundsätzliche Annahme, Wirkungen könnten prinzipiell erst einige Jahre nach Abschluss einer Maßnahme eintreten und somit überprüft werden erscheint zu pauschal. So muss z.B. bei Ernährungsprogrammen für Schwangere, um das Geburtsgewicht der Babys zu erhöhen, die Wirkung innerhalb von 9 Monaten (d.h. innerhalb der Schwangerschaft bzw. vor der Geburt) eintreten. Andererseits benötigen Maßnahmen zur Verhaltensänderung weitaus mehr Zeit um Wirkungen entfalten zu können (vgl. White 2006a: 23). Entsprechend muss der Zeitpunkt einer IE sorgfältig gewählt werden. Allerdings muss hierbei ein weiterer Punkt beachtet werden: Nicht nur der „richtige" Zeitpunkt zur Wirkungsmessung i.S. von mittel- bis langfristigen Wirkungen ist für IE relevant. Da alle Stufen der Wirkungskette überprüft werden müssen, ist es auch notwendig, Informationen über die *direkten* Wirkungen zu erheben. Von daher scheint es notwendig, für eine IE nicht nur zwei, sondern drei Zeitpunkte zu berücksichtigen: eine vorher Messung (t_1), eine nachher Messung zum Förderende (t_F) sowie eine weitere nachher/ex-post Messung, wenn sich Wirkungen entfalten konnten (t_E) (vgl. Caspari 2004: 52ff.; 135ff.). Diese Notwendigkeit wurde bisher im Rahmen der Diskussion noch nicht angesprochen.

Grundsätzlich sind die notwendigen *Voraussetzungen* für eine IE zu überprüfen: Liegen Baselinedaten vor, gibt es ein M+E-System, existieren in dem Land und/oder Sektor entsprechende Surveys, wurden bereits Evaluationen durchgeführt, etc. Es ist also vorab zu prüfen, ob bei einer Evaluation der notwendige „rigour“ umsetzbar ist. Diesem Problem kann entgegengewirkt werden, wenn eine IE bereits frühzeitig eingeplant wird: „Early and careful planning will, however, provide many more methodological options in designing the evaluation” (Baker 2000, 2).

ANHANG 1: LITERATUR

Appleton, Simon/Booth, David (2001): 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), http://www.q-squared.ca/pdf/Q2_WP14_Appleton&Booth.pdf [Okt 07].

Asian Development Bank (2006): Impact Evaluation. Methodological and Operational Issues.

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

Bamberger (2006): Conducting Quality Impact Evaluation Under Budget, Time and Data Constraints. Independent Evaluation Group, The World Bank.

Bamberger, Michael (Hg.) (2000): Integrating Quantitative and Qualitative Research in Development Projects, Washington, D.C.: The World Bank.

Bamberger, Michael/Rugh, Jim/Mabry, Linda (2006): Real World Evaluation. Working under Budget, Time, Data And Policy Constraints. Overview. Sage Publication.

Behrman, J.R./Todd, P.E. (1999): Randomness in the Experimental Samples of PROGRESA (Education, Health, and Nutrition Program). Research Report of the PROGRESA Evaluation Project of IFPRI.

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

Caspari, Alexandra (2006): Partizipative Evaluationsmethoden – Zur Entmystifizierung eines Begriffs in der Entwicklungszusammenarbeit, in: Flick, U. (Hg.): Qualitative Evaluationsforschung. Reinbek: Rowohlt, S. 365-384.

Caspari, Alexandra (2004): Evaluationen der Nachhaltigkeit von Entwicklungszusammenarbeit. Zur Notwendigkeit angemessener Konzepte und Methoden. Wiesbaden: VS-Verlag

Chung, Kimberly (2000): Qualitative data collection techniques. In: Grosh, Margaret/Glewwe, Paul (Hg.): Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of Living Standards Measurement Study (Bd. 2), S. 337-363, Washington, D.C.: World Bank.

CDG (2006): When Will We Ever Learn? Improving Lives through Impact Evaluation. Washington, D.C, Centre for Global Development.

Deutsche Gesellschaft für Evaluation (2002): Standards für Evaluation. Köln: DeGEval

EES (2007): EES Statement: The Importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and development interventions.

Ezemenari, Kene/Rudqvist, Andres/Subbarao, Kelanidhi (1999): Impact Evaluation: A Note on Concepts and Methods, Washington, D.C.: The World Bank.

Gangel, Markus/DiPrete, Thomas A. (2004): Kausalanalyse durch Matchingverfahren. In: Diekmann, Andreas (Hg.): Methoden der Sozialforschung. Sonderheft 44, 2004 der Kölner Zeitschrift für Soziologie und Sozialpsychologie, S. 396-420.

Guijt, Irene (2000): Methodological issues in participatory monitoring and evaluation. In: Estrella, Marisol (Hg.): Learning form Change. Issues and Experiences in Participatory Monitoring and Evaluation, S. 201-228, London: Intermediate Technology Publications.

Kassam, Yusuf (1998): Combining participatory and survey methodologies in evaluation: the case of a rural development project in Bangladesh. In Jackson, Edward T./Kassam, Yusuf (Hg.): Knowledge Shared: Participatory Evaluation in Development Cooperation, S. 108-121, West Hartford, Connecticut: Kumarian.

Kapoor, Anju G./ OED (2002): Review of Impact Evaluation. Methodologies Used By The Operations Evaluation Department Over Past 25 Years.

Kusek, Jody Z.; Rist, Ray C.; White, Elizabeth M. (2005): How Will We Know the Millennium Development Goal Results When We See Them? Building a Results-based Monitoring and Evaluation System to Give Us the Answers. In: Evaluation. Jg. 11(1), S. 7-26.

Kromrey, Helmut (2005): 'Qualitativ' versus 'quantiativ' – Ideologie oder Realität? Vortrag auf dem 1. Berliner Methodentreffen Qualitative Forschung an der FU Berlin http://www.qualitative-forschung.de/methodentreffen/archiv/texte/texte_2005/kromrey.pdf [Jan 07]

OECD/DAC (2006): DAC Evaluation Quality Standards.

OECD/DAC (2005): Prüfbericht über die Entwicklungszusammenarbeit – Deutschland. Paris http://www.oecd.org/dataoecd/10/22/36770168.pdf [Okt 07]

OECD/DAC (2002): Glossary of Key Terms in Evaluation and Results Based Management

Prowse, Martin (2007): Aid effectiveness: the role of qualitative research in impact evaluation". Background Note December 2007, ODI.

Ravaillon, Martin (2001): The Mystery of the Vanishing Benefits: Ms Speedy Analyst's Introduction to Evaluation. Washington, D.C.: World Bank.

Ravallion, Martin (2005): Evaluating Anti-Poverty Programs. Washington, D.C.: World Bank.

Rosenbaum, Paul R./Rubin, Donald B. (1983): The central role of the propensity score in observational studies for causal effects. In: Biometrica, 70, 1, S. 41-55.

Ruprah, Inder Jit (2008): „You Can Get It If You Really Want": Impact Evaluation Experience of the Office of Evaluation and Oversight of the Inter-American Development Bank. In: IDS-Bulletin (Vol. 39, No. 1, March 2008), S. 23-35.

Sanders, James R./The Joint Committee on Standards for Educational Evaluation (1994): The Program Evaluation Standards (2. Auflage), Thousand Oaks: Sage.

UNEG (2005): Standards for Evaluation in the UN-System.

White, Howard (2007): Evaluating Aid Impact. MRPA Research Paper Nr. 2007/75.

White, Howard/ WB IEG (2006a): Impact Evaluation. The Experience of the Independent Evaluation Group of the World Bank.

White, Howard (2006b): Impact Evaluation: An Overview And Some Issues For Discussion. Room Document 5, 4th meeting 30-31 March 2006. Collaboration IEG and DAC Secretariat.

White, Howard (2005): Maintaining Momentum to 2015? An Impact Evaluation of Intervention to Improve Maternal and Child Health and Nutrition in Bangaladesh. World Bank Operations Evaluation.

Zürcher, Christoph und Jan Köhler (2007)**:** Assessing the Impact of Development Cooperation in North East Afghanistan: Approaches and Methods. BMZ Evaluation Working Papers, Bonn.

Zürcher, Christoph, Jan Köhler und Jan-Rasmus Böhnke (2007): Assessing the Impact of Development Cooperation in North East Afghanistan: Interim Report. BMZ Evaluation Reports 028, Bonn.

ANHANG 2: AUSGEWERTETE NONIE DOKUMENTE[39]

Workshop Dokumente, 17. Mai 2007:

NONIE Sub-Group 1: Experimental and quasi-experimental approaches to impact evaluation (draft 1), http://www.worldbank.org/ieg/nonie/docs/subgroup1.doc [Jan 08]

NONIE Sub-Group 2: Approaches and methods in impact evaluation – An initial concept note, http://www.worldbank.org/ieg/nonie/docs/guidelines%20subgroup%202.doc [Jan 08]

NONIE Sub-Group 3: Impact evaluation of new aid instruments and country programs (draft 1), http://www.worldbank.org/ieg/nonie/docs/NONIE_SG3.pdf [Jan 08]

Room Document: Working Together – NONIE and 3IE, http://www.worldbank.org/ieg/nonie/docs/nonie%20and%203ie.doc [Jan 08]

Workshop Dokumente, 14. Januar 2008:

NONIE: Impact Evaluation Guidance, http://www.worldbank.org/ieg/nonie/docs/NONIEGuidanceIntroduction.doc [Jan 08]

NONIE Sub-Group 1: Impact Evaluation Guidance Section 2 – Experimental and Quasi-Experimental Approaches to Impact Evaluation (draft 2), http://www.worldbank.org/ieg/nonie/docs/SG1GuidanceDraft.doc [Jan 08]

NONIE Sub-Group 2: NONIE Impact Evaluation Guidance, http://www.worldbank.org/ieg/nonie/docs/NONIE_SG2.pdf [Jan 08]

NONIE Sub-Group 3: Impact evaluation of new aid instruments and country programs (draft 2), http://www.worldbank.org/ieg/nonie/docs/NONIE_SG3.pdf [Jan 08]

NONIE Sub-Group 4/5: Improving impact evaluation coordination and uptake A scoping study commissioned by the DFID Evaluation Department on behalf of NONIE, http://www.worldbank.org/ieg/nonie/docs/IEFinalDraftJan6.doc [Jan 08]

NONIE statement on impact evaluation (draft), http://www.worldbank.org/ieg/nonie/docs/DraftNONIEstatement_impact_eval_revised.doc [Jan 08]

[39] Quelle: http://www.worldbank.org/ieg/nonie/members.html [Jan 08]

ANHANG 3: AUSGEWERTETE IMPACT EVALUATION STUDIEN DER NONIE DATENBANK

AfDB:

(1) Yirga-Hall, G./Giorgis, G. (2004): Arab Republic of Egypt El Arish Power Project. Project Performance Evaluation Report. Cairo: AfDB Publications. http://www.oecd.org/dataoecd/46/6/37964745.pdf [Okt 07]

AusAID:

(1) Australian Agency for International Development, Australian Government (2005): Impact evaluation of the Thailand-Australia HIV/AIDS Ambulatory Care Project. Evaluation and Review Series – NO.37 MAY 2005. Canberra: AusAID Publications. http://www.ausaid.gov.au/publications/pdf/qas37_ambulatory.pdf [Okt 07]

CIDA:

(1) Canadian International Development Agency (1998): Basic Human Needs Performance Review. From Service Delivery Towards Governance for Sustainability Report of the Ghana Water Program. Quebec: CIDA Publications.

Danida:

(1) Centre for Development Research Evaluation(2002): The Agricultural Development Project in Tete, Mozambique. Copenhagen: Danida Publications. http://www.um.dk/NR/rdonlyres/5FFDB790-BAC1-45A3-BE5A-A072EED1FB74/0/20023Tete.pdf [Okt 07]

(2) Evaluation Secretariat, Ministry of Foreign Affairs (2002): Noakhali Rural Development Project. Copenhagen: Danida Publications. http://www.um.dk/publikationer/Danida/English/Evaluations [Okt 07]

DFID:

(1) Faisel, A. (2005): KASHF Foundation Impact Assessment. London: DFiD Publications. http://www.kashf.org/administrator/attachment/file/Financials/Impact%20Assessment%202005.pdf [Okt 07]

(2) Olken, B. A. (2007): Monitoring Corruption: Evidence from a Field Experiment in Indonesia. Harvard: Harvard University, National Bureau of Economic Research Publications. http://www.nber.org/~bolken/corruptionexperiments.pdf [Okt 07]

FINNIDA:

(1) Ministry for Foreign Affairs (1996): Road Maintenance Assistance to the Roads Department, Zambia, Phase II. Evaluation Report, Helsinki: FINNIDA Publications.

(2) Skyttä, T/Ojanperä, S./Mutero, J. (2001) Finland´s Support to Water Supply and Sanitation 1968 – 2000. Evaluation of Sector Performance, Helsinki: FINNIDA Publications.

IADB/OVE:

(1) Benavente, J.M./Crespi, G./Maffioli, A. (2006): The Impact of National Research Funds: An Evaluation of the Chilean FONDECYT. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/1192240.pdf [Okt 07]

(2) Brownwyn, H./Maffioli, A. (2006): Evaluating the Impact of Technology Development Funds in Emerging Economies: Evidence from Latin-America. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/1404775.pdf [Apr 08]

(3) Carolyn, J.H./Lopez, Y. (2005): Does Community Participation Produce Dividends in Social Investment Fund Projects? Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/1146714.pdf [Okt 07]

(4) Carolyn, J. H./Cabrol, M. (2005): Programa Nacional de Becas Estudiantiles Impact Evaluation Findings. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/599401.pdf [Okt 07]

(5) Chudnovsky, D./Lopez, A./Rossi, M./Ubfal, D. (2006): Evaluating a Program of Public Funding of Private Innovation Activities. An Econometric Study of FONTAR in Argentina. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/907633.pdf [Okt 07]

(6) Davis, B./Handa, S./Arranz, H.M./Stampini, M./Winters, P. (2005): Agricultural Subsidies, Human Capital Development and Poverty Reduction: Evidence from Rural Mexico. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/599375.pdf [Okt 07]

(7) Diaz, J. J./Jaramillo, M. (2006): An Evaluation of the Peruvian “Youth Labor Training Program” – PROJOVEN. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/907634.pdf [Okt 07]

(8) Diaz, J. J./Handa, S. (2005): An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator: Evidence from Mexico´s PROGRESA Program. Washington: IADB Publications. http://idbdocs.iadb.org/ove/Documents/uploads/cache/862049.pdf [Okt 07]

(9) Galdo, V./Briceno, B. (2005): Evaluating the Impact on Child Mortality of a Water. Supply and Sewerage Expansion in Quito: Is Water Enough? Washington: IADB Publications. http://idbdocs.iadb.org/ove/Documents/uploads/cache/596806.pdf [Okt 07]

(10) Soares, F./Soares, Y. (2005): The Socio-Economic Impact of Favela-Bairro: What do the Data Say? Washington: IADB Publications. http://idbdocs.iadb.org/ove/Documents/uploads/cache/600835.pdf [Okt 07]

(11) Torero,M./Field, E. (2005): Impact of Land Titles over Rural Households. Washington: IADB Publications. http://www.iadb.org/ove/Documents/uploads/cache/2003690.pdf [Okt 07]

IFAD:

(1) International Fund for Agricultural Development (2006a): Republic of Ghana. Upper West Agricultural Development Project. Interim Evaluation, Rome: IFAD Publications. http://www.ifad.int/evaluation/public_html/eksyst/doc/prj/region/pa/ghana/Ghana_uwadep.pdf [Okt 07]

(2) International Fund of Agricultural Development (2006b): Republic of Ghana. Upper East Region Land Conservation and Smallholder Rehabilitation Project (LACOSREP) – Phase II. Interim Evaluation, Rome: IFAD Publications. http://www.ifad.org/evaluation/public_html/eksyst/doc/prj/region/pa/ghana/gh_lacos.pdf [Okt 07]

(3) International Fund of Agricultural Development (2005): Republic of The Gambia. Rural Finance and Community Initiatives Project. Interim Evaluation, Rome: IFAD Publications. http://www.ifad.org/evaluation/public_html/eksyst/doc/prj/region/pa/gambia/rfcip.pdf [Okt 07]

JBIC:

(1) Bayes, A. (2007): Impact Assessment of Jamuna Multipurpose Bridge Project (JMBP) on Poverty Reduction. Tokio: JBIC Publications. http://www.jbic.go.jp/english/oec/post/2006/pdf/te03_full.pdf [Okt 07]

Network of Networks Impact Evaluation Initiative (NONIE)
Proposed sub-groups, March-October, 2007

Sub-Group	Area of work	Description	Members
1. Guidelines 1	Experimental and Quasi-experimental approaches to impact evaluation	This sub-group will develop the section of the guidelines relating to experimental and quasi-experimental approaches. The guidelines should be clear as to when these methods are desirable and when they are not appropriate, and outline the practical implications of adopting them and any pitfalls. The draft submitted by IEG prior to the November 2006 meeting was a preliminary draft of this topic.	IOB, OVE/IADB **IEG**, NORAD
2. Guidelines 2	Developing approaches to impact evaluation	It is widely recognized that many interventions supported by development agencies are not amenable to impact evaluation using the approaches covered by the 'Guidelines 1' subgroup; e.g. policy influence, macro reforms, institutional development activities, and harmonization. This sub-group will develop guidelines on approaches to impact evaluation for these areas.	GEF, IFAD FAO, UNDP **AfrEA**, AFD, DESA
3. Guidelines 3	Impact evaluation of new aid instruments and country programs	Agencies are concerned with being able to demonstrate the impact of the 'new aid modalities' such as budget support and SWAPs. They also want to be able to assess impact of their assistance at the country level. This sub-group will draw on the work of the other two groups to provide guidelines on these issues.	OVE/IADB, **IOB**, IEG, DFID, EBRD, AfREA, DANIDA, **Susan Tamondong (IDEAS), Velayuthan Sivagnanasothy (ENSA)**
4. Consultation Process	Consultation process for guidelines – MERGED WITH SUB-GROUP 5	As agreed at the November 2006 meeting in Paris and subsequent discussions of the Task Team, the guidelines should be subject to a broad process of consultation. This sub-group shall identify stakeholders with whom consultations should be held and identify the means by which those consultations should be held.	WHO, **AfrEA**
5. Common program and priority areas	Moving forward with impact evaluations (+ SG 4 activities)	An ultimate purpose of NONIE is to ensure a more substantial program of impact evaluations is carried out in the future, preferably in a coordinated or joint manner. This subgroup will make proposals for both the means by which this program should be undertaken and the content of the program (or at least means for identifying content).	**DFID**, WHO, FAO, UNDP, Debazou Yantio, DESA, French Treasury, WHO, AfrEA
6. Information-sharing platform	Development of resource platform to support impact evaluation of NONIE members and other agencies	This sub-group will outline ideas for what is to be included on a website to support impact evaluation. Possible areas include: resources for IE, database of IEs, information-sharing on planned activities.	**IEG**, DAC secretariat, Debazou Yantio

01/08/07

Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ)
Referat „Evaluierung der Entwicklungszusammenarbeit, Außenrevision“

Dienstsitz Bonn
Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
Dahlmannstraße 4
53113 Bonn
E-Mail: eval@bmz.bund.de
Tel. + 49 (0) 228 99 535 0
Fax + 49 (0) 228 99 535 35 00

Dienstsitz Berlin
Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
Stresemannstraße 94
10963 Berlin
Tel. + 49 (0) 30 18 535 0
Fax + 49 (0) 30 18 535 25 01

www.bmz.de
http://www.bmz.de/de/erfolg/index.html

Redaktion und Verantwortlich
Michaela Zintl

Stand
Mai 2008