Centrum für Evaluation

Center for Evaluation

Arbeitspapier 14

Centrum für Evaluation – Universität des Saarlandes

Konzept für alltagstaugliche Wirkungsevaluierungen in Anlehnung an Rigorous Impact Evaluations

Erprobung der Durchführung im Rahmen von GTZ Unabhängigen Evaluierungen

Nicolà Reade, MA

Unter der Leitung von:
Dr. Wolfgang Meyer und
Prof. Dr. Reinhard Stockmann

Reade, Nicolà:

Konzept für alltagstaugliche Wirkungsevaluierungen in Anlehnung an Rigorous Impact Evaluations

Saarbrücken: Centrum für Evaluation, 2008.

(CEval-Arbeitspapiere; 14)

NICHT IM BUCHHANDEL ERHÄLTLICH

SCHUTZGEBÜHR: 5 €

BEZUG: Centrum für Evaluation (CEval)
Universität des Saarlandes
Postfach 15 11 50
D-66041 Saarbrücken
info@ceval.de

oder kostenfrei zum Download:
http://www.ceval.de

Inhalt

ABKÜRZUNGSVERZEICHNIS

BMZ	Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
CEval	Centrum für Evaluation
EZ	Entwicklungszusammenarbeit
FZ	Finanzielle Zusammenarbeit
GTZ	Gesellschaft für Technische Zusammenarbeit
IE	Impact Evaluation / Wirkungsevaluierung
KG	Kontrollgruppe
M&E	Monitoring und Evaluation
NGO, NRO	Nichtregierungsorganisation
PFB	Projekt Fortschritts Bericht
PFK	Projekt Fortschritts Kontrolle
PSM	Propensity Score Matching
RCT	Randomized Controlled Trial / (randomisiertes) Kontrollgruppen-Design
RIE	Rigorous Impact Evaluation / rigorose Wirkungsevaluierung
SB	Schlussbericht
TBE/TBA	Theory-Based Evaluation/Approaches / theoriebasierte Evaluation/Ansätze
TZ	Technische Zusammenarbeit
UV	unabhängige Variable – erklärende Größe
VG	Vergleichsgruppe
ZG	Zielgruppe

1. Anlass und Ziel des Konzeptes „Wirkungsevaluation“

Die Millenniums-Erklärung und die Millenniumsentwicklungsziele der Vereinten Nationen sowie die Erklärung von Paris über die Wirksamkeit der Entwicklungszusammenarbeit beeinflussen das internationale Evaluierungsgeschehen erheblich. Vor allem die MDGs stellen erstmals einen gemeinsamen, überprüfbaren Bezugsrahmen für die internationale Entwicklungszusammenarbeit dar, der für die Evaluation relevant ist. Hervorzuheben ist zudem das Prinzip „Orientierung auf Wirkungen“. In diesem Kontext ist es Aufgabe von Evaluationen, zuverlässige Ergebnisse über Wirkungen von Entwicklungsmaßnahmen zur Verfügung zu stellen, welche Rechenschaftslegung aber auch Lernen ermöglichen. Seitdem werden die Instrumente wirkungsorientiertes Monitoring (als Steuerungsinstrument zur Ziel- und Wirkungserreichung) und Evaluation (zum Nachweis von Wirkungen) verstärkt in der internationalen EZ genutzt.

In Deutschland hat insbesondere die GTZ die Wirkungsorientierung ihrer Tätigkeit in den letzten Jahren deutlich verstärkt und zu einer der Leitlinien ihres unternehmerischen Handelns erklärt. Einen wichtigen Impuls dazu löste der neue Auftragsrahmen aus, den die GTZ (unter AURA) und andere deutsche Durchführungsorganisationen 2003 mit dem BMZ vereinbarten. Die GTZ hat seitdem wesentliche Verfahren des Auftragsmanagements (Vorbereitung, Durchführung, Berichterstattung), das M&E-System, das Berichtswesen sowie das Evaluationssystem (Fremdevaluierungen, Schluss- und ex-post-Evaluierungen, PFK) an die Orientierung auf Wirkung angepasst. Ein eigenes Wirkungsmodell als Grundlage des Auftragsmanagements wurde ebenfalls entwickelt (vgl. AURA 2006).

Allerdings entspricht die in der bisherigen deutschen EZ-Evaluierung übliche methodische Vorgehensweise (Dokumentenanalyse, Experteninterviews, etc.) nicht den neuen Ansprüchen an die Wirkungsmessung. Methodisch anspruchsvolle Designs (Längsschnittstudien, Verwendung eines Multimethodenansatzes) werden bisher selten verwendet (vgl. Systemprüfung 1999, 2001 und 2007; Stockmann 2000a). Um Wirkungen tatsächlich nachzuweisen ist ein intensives Zusammenspiel von Evaluation und empirischer Sozialforschung für einen optimalen Methodeneinsatz zwingend notwendig, was auf Seiten der Wissenschaft bereits seit 1990 thematisiert und demonstriert wird (vgl. Stockmann 1992, 1996; Stockmann u.a. 2000; Caspari 2000, 2004). Aktuell wird das Thema „Nachweis von Wirkungen“ in der internationalen und deutschen Entwicklungszusammenarbeit seit Erscheinen der Publikation des Centers for Global Development „When will we ever learn„ unter dem Stichwort der „(Rigorous) Impact Evaluation“ wieder verstärkt diskutiert[1].

Als Reaktion auf die gestiegenen Anforderungen des Wirkungsnachweises greift die GTZ im Kontext des Rahmenvertrags zur Durchführung von Unabhängigen Evaluierungen 2007 bis 2008 methodisch „anspruchsvollere“ Wirkungsevaluierungen auf. 2008 sollen demnach für die drei Entwicklungsmaßnahmen *KV Programm für die Reform des Wassersektors, Sambia* (PN 2005.2125.2 - Zwischenevaluierung), *KV Programm zur Reform des Wassersektors, Kenia* (PN 2007.2039.1 - Zwischenevaluierung) und *Qualifizierung kommunaler Dienste, Türkei* (PN 1998.2179.4 – Ex-post Evaluierung) anspruchsvolle Wirkungsevaluierungen im

[1] Z.B. wurde 2006 ein Zusammenschluss der Netzwerke von EZ-Organisationen sowie multilateralen Entwicklungsbanken gegründet (NONIE – Network of Networks Impact Evaluation Initiative), um die Effektivität der EZ zu verbessern, indem nützliche und relevante, qualitativ hochwertige Wirkungsevaluationen vorangetrieben werden.

Wassersektor unter der Federführung des CEval durchgeführt und erprobt werden.

Das vorliegende Konzept zu Wirkungsevaluationen hat daher die Zielsetzung, für diese drei Evaluierungen ein abgestimmtes und verallgemeinerbares Wirkungsevaluationsdesign unter Berücksichtigung adäquater und anspruchsvoller Methoden aufzustellen. Hierzu erläutert das Konzept zunächst den Begriff „Wirkungen“ (Kapitel 2) und die methodischen Anforderungen an Wirkungsevaluationen (Kapitel 3). Danach wird die bisherige Umsetzungspraxis in der GTZ skizziert (Kapitel 4) und in einem letzten Schritt die allgemeine organisatorische und inhaltliche Umsetzung im Rahmen der vorgesehenen drei Pilot-Wirkungsevaluationen im Wassersektor beschrieben (Kapitel 5)[2].

Ein so abgestimmtes Wirkungsevaluationsdesign soll dazu beitragen, eine einheitliche methodische Vorgehensweise zu ermöglichen sowie eine Basis für die Entwicklung vergleichbarer Fragestellungen für die Datenauswertung zu geben, was wiederum eine Querschnittsauswertung der Wirkungsevaluationen ermöglicht. Die gewonnenen Erfahrungen bei der Umsetzung der Wirkungsevaluationen werden nach der ersten Pilotphase für die Weiterentwicklung des Konzepts in 2009 genutzt.

2. Das Messen von Wirkungen

Wirkungsevaluationen sehen sich vor zwei Herausforderungen gestellt. Zum einen sollen Sie Aufschluss geben über möglichst alle Wirkungen, d.h. der Fragestellung nachgehen: „Welche Veränderungen (Wirkungen) haben sich ergeben?“. Zum anderen sollen diese identifizierten Wirkungen im Rahmen der Wirkungsevaluation kausal Ursachen zugeschrieben werden, d.h. die Fragestellung „Welche Ursache ist für die beobachtete Veränderung verantwortlich?“ steht ebenfalls im Vordergrund. Dabei gebührt den während einer Entwicklungsmaßnahme durchgeführten Maßnahmen besondere Aufmerksamkeit, da sie bestimmte Veränderungen bewirken sollen und nur aus diesem Grund Interventionen vorgenommen werden. Die Betrachtung der eingesetzten Ressourcen (Inputs), Leistungen (Outputs), der Zielerreichung und den Prozess der Leistungserbringung bilden jedoch nicht den Hauptschwerpunkt einer Wirkungsevaluation (sollten als erklärende Faktoren aber mitbetrachtet werden), sondern die Überprüfung der Veränderungen, die in Folge der Leistungen einer Entwicklungsmaßnahme entstanden sind und die Frage nach dem warum und wie die Veränderungen entstanden sind (Prozessbetrachtung). Dabei stellt sich jedoch zunächst die Frage der Definition von Wirkungen, und darauf aufbauend, welche der Wirkungen im Fokus der Wirkungsevaluation stehen. Dementsprechend ist die Abgrenzung des Untersuchungsgegenstands, der Wirkungen und der Wirkungshypothesen für die Durchführung von Wirkungsevaluationen von zentraler Bedeutung.

Nach wissenschaftlichem Verständnis sind Wirkungen alle Veränderungen die nach einer Intervention (Maßnahme) auftreten. Hierzu gehören auch nicht erwartete und unerwünschte Wirkungen. Wirkungen umfassen somit die intendierten, nicht-intendierten, positiv so wie auch negativ, erwarteten oder unerwarteten Veränderungen, die in einer Wirkungsevaluation alle zu erfassen sind.

[2] Die spezifische Umsetzung der vorgesehenen Wirkungsevaluationen im Wassersektor wurde im Rahmen der Aufstellung der Inception Reports für die jeweilige Evaluation spezifiziert.

Zudem kann zwischen internen, bei der Durchführungsorganisation ausgelösten Wirkungen, und externen, in den Politikfeldern der Intervention und bei den Zielgruppen hervorgerufenen Wirkungen, unterschieden werden. Besonders die internen, bei der Durchführungsorganisation ausgelösten Wirkungen sowie auch die organisationsspezifischen Rahmenbedingungen, die das Entstehen von internen und externen Wirkungen maßgeblich beeinflussen, finden im Rahmen von EZ-Evaluationen oft nicht genügend Beachtung. Generell zeigt sich aber dass die Komponenten „Ressourcen“ und „Vernetzung der Durchführungsorganisationen“ nicht nur für die unmittelbare Steuerung der Entwicklungsmaßnahme, sondern auch für die Diffusion der Wirkungen bei den Zielgruppen und die dauerhafte institutionelle Verankerung von zentraler Bedeutung sind. Aus diesem Grund müssen diese organisatorischen Komponenten im Wirkungsmodell mitberücksichtigt werden. Speziell die Betrachtung der Leistungsfähigkeit der Organisationen, die mit der Durchführung der Entwicklungsmaßnahmen betraut sind, sollte unbedingt im Fokus der Untersuchung stehen. Aus EZ-Evaluationen ist bekannt, dass gerade Organisationen als Transmitter von Entwicklungsmaßnahmen und somit deren Kapazitäten, Strukturen und Prozesse als verursachende Variablen eine entscheidende Rolle für den Erfolg und die Wirkungen der Maßnahme spielen. Systemisch angelegte Organisationstheorien können Aufschluss über die Organisations-Umwelt-Beziehung (interne – externe Wirkungen) und über entscheidende Untersuchungsparameter für die interne Prozessbetrachtung innerhalb der Organisation geben. Ein solches Modell und ihre Anwendung werden von Stockmann (2006) thematisiert. Hier wird deutlich, dass Faktoren wie interne Zielakzeptanz, Personalstruktur und Qualifikationen, Organisationsstruktur, finanzielle Ressourcen, und technische Infrastruktur einen maßgeblichen Einfluss auf die Wirkungsentstehung haben. Eine solide Organisationsanalyse die Aufschluss darüber gibt, welche internen, organisationsspezifischen Faktoren zur Wirkungserreichung beigetragen haben bzw. beitragen können, ist daher im Rahmen der Wirkungsmessung durchzuführen und im Wirkungsmodell aufzunehmen (vgl. Stockmann 2006). Die praktische Umsetzung wurde bereits in Evaluationen der deutschen Berufsbildungszusammenarbeit erprobt (vgl. Stockmann 1992; 1996, Stockmann u.a. 2000; Caspari 2000, 2004).

Interne und externe Wirkungen entfalten sich in der Veränderung von Strukturen, Prozessen oder individuellen Verhaltensweisen (vgl. Abb. 1). Ein Strukturwandel im Wassersektor ist z.B. dann gegeben, wenn die Wassergesetzgebung verändert wird um die Wasserversorgung zu verbessern. Prozesswirkungen würden erzielt, wenn etwa Fortschritte bei der Wasserbereitstellung erkennbar sind. Veränderungen individueller Verhaltensweisen wären z.B. bei Einführung von hygienischen Vorsorgemaßnahmen für sauberes Trinkwasser gegeben (vgl. Stockmann 2006 S. 102ff., 2007 S. 65ff.).

Abbildung 1: Wirkungsdimensionen

Wirkungsdimensionen	Geplant	Ungeplant
Struktur	+-	+-
Prozess	+-	+-
Verhalten	+-	+-

Quelle: Stockmann 2008 S. 66

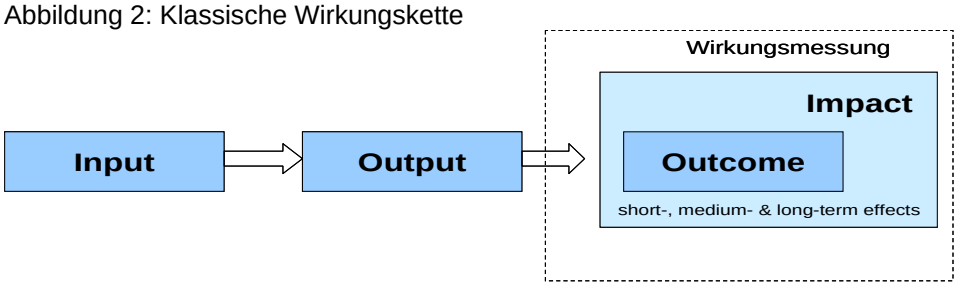
Weitere Differenzierungsmöglichkeiten stellen die Art der Wirkung, ihre Dauer und die Ebene

des Auftretens dar. Es können z.B. ökonomische, soziale, ökologische, kulturelle und politische Wirkungen unterschieden werden. Oder aber kurz-, mittel- oder langfristige Wirkungen die besonders in der internationalen entwicklungspolitischen Diskussion zu Wirkungen eine zentrale Rolle spielten. Schließlich wird noch zwischen Wirkungen auf der gesellschaftlichen Ebene (Makroebene), in Organisationen oder Gruppen (Mesoebene) oder aber bei einzelnen Individuen (Mikroebene) unterschieden (im Detail vgl. Stockmann 2006 S. 101ff., 2007 S. 65ff.). Insbesondere diese Differenzierung hat in den letzten Jahren im Rahmen der Diskussion zur nachhaltigen Entwicklung an Bedeutung gewonnen (vgl. Meyer 2007a).

Welche methodischen Herangehensweisen zur Wirkungsmessung notwendig sind, und welche Wirkungen im Fokus der Wirkungsmessung stehen, wird im Folgenden näher betrachtet. Aktuell zeigen sich zwei Herangehensweisen, die sich durch ein breites und enges Wirkungsverständnis unterscheiden. Das breite, in der wissenschaftlichen Diskussion zur Wirkungsmessung im Rahmen von TZ Evaluationen vorherrschende Wirkungsverständnis, umfasst das Messen aller auf der Mikro-, Meso- und Makroebene auftretenden Wirkungen. Das enge, zurzeit in der aktuellen internationalen Diskussion vorherrschende Wirkungsverständnis hingegen umfasst das Messen von Mittel- und Langzeit-Wirkungen auf der Mikro-, Meso- und Makroebene.

1. Die wissenschaftliche Diskussion zur Wirkungsmessung im Rahmen von TZ Evaluationen bezeichnet alle auftretenden Veränderungen eines Zustands als Wirkungen und subsumiert diese unter dem Begriff „Impact“. Innerhalb des Impacts wird noch zusätzlich der Outcome gekennzeichnet, der Teilbereich der Wirkungen der kausal auf die Interventionsmaßnahme zurückzuführen ist (vgl. Abbildung 2).

Abbildung 2: Klassische Wirkungskette



Idealerweise sollen möglichst alle auftretenden Wirkungen (= Bruttowirkungen/Impact) erfasst werden, die dann um Wirkungen anderer Faktoren und Designeffekte bereinigt werden. Als Ergebnis erhält man alle auf die Entwicklungsmaßnahme zurückzuführenden Wirkungen (= Nettowirkungen/Outcome) (vgl. Stockmann 2006 S. 105, 2007 S. 67). Umfassende Wirkungsevaluationen untersuchen somit weitaus mehr, als lediglich die Zielerreichung oder die kurzfristigen Effekte einer Maßnahme auf die Zielgruppe.

Zur Messung des Impacts, besonders zur strukturierten Suche nach intendierten und nicht-intendierten Wirkungen, ist der Einsatz unterschiedlicher und z.T. aufwendiger sozialwissenschaftlicher Methoden nötig, die in der wissenschaftlichen Auseinandersetzung mit der in der EZ gängigen Evaluationspraxis eingefordert wird. Bereits bei der Evaluation von Wirkungen der deutschen Berufsbildungszusammenarbeit (vgl. Stockmann 1992; Stockmann u.a. 2000) und zur Messung von Wirkungen und Nachhaltigkeit von Maßnahmen in der Entwicklungszusammenarbeit (vgl. Stockmann 1996; Caspari 2000, 2004) gehörten z.B. die heute in der EZ aktuell diskutierten „anspruchsvollen Methoden zur Wirkungsmessung“ zum Standard, wie z.B.

(1) Hypothesengeleitete Ursache-Wirkungsuntersuchung,

(2) Quasi-Experimentelle Forschungsdesigns,

(3) Multi-Methodenansatz,

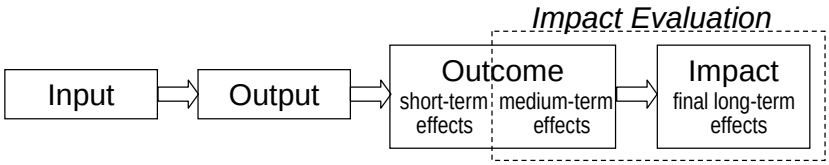
(4) Trichteransatz,

(5) Anwendung fortgeschrittener statistischer Analyseverfahren.

Zusätzlich wurde ein Evaluationsleitfaden zur strukturierten und systematischen Messung von Wirkungen entwickelt. Neue Veröffentlichungen aus der Evaluationsforschung greifen diese Methoden ebenfalls auf (vgl. Stockmann 2006, 2007).

2. Die zurzeit stattfindende internationale Diskussion der EZ-Organisationen zur Wirkungsmessung im Rahmen von EZ-Evaluationen stützt sich auf ein engeres, nur auf das Messen von mittel- und langfristigen Wirkungen bezogenes Begriffsverständnis. Der Begriff Impact umfasst hierbei *ausschließlich* die langfristigen, übergeordneten entwicklungspolitischen Wirkungen. Die der EZ-Logik zugrundeliegende Wirkungskette differenziert demnach die Ebenen Input, Output, Outcome und Impact, wobei Outcome nicht ein Teilbereich des Impacts ist. „Impact-Evaluationen“ sollen zur Messung der Wirkungen allerdings auch den mittelfristigen Outcome berücksichtigen (vgl. Abb. 3). In der internationalen Diskussion wird explizit betont, dass Wirkungsevaluationen darauf achten sollen, im Kontext der Rahmenbedingungen der Vorhaben nicht nur auf der Mikroebene Wirkungen aufzuzeigen, sondern ebenfalls und schwerpunktmäßig Wirkungen auf der Makroebene nachweisbar zu machen, z.B. Armutsreduzierung, Wachstum des Pro-Kopf-Einkommens etc. (allgemein der Beitrag zu den MDGs) (vgl. OECD/DAC 2002 S. 24).

Abbildung 3*:* Zielebenen von Wirkungsevaluationen



Quelle: Caspari/Barbu 2008

Methodische Schwerpunkte der international von den EZ-Organisationen diskutierten Impact-Evaluation sind folgende Aspekte:

(1) einen theoriebasierten Ansatz (d.h. hypothesengeleitete Ursache-Wirkungsuntersuchung unter Berücksichtigung der Kontextfaktoren),

(2) die Berücksichtigung des Kontrafaktischen (durch Anwendung anspruchsvoller Forschungsdesigns),

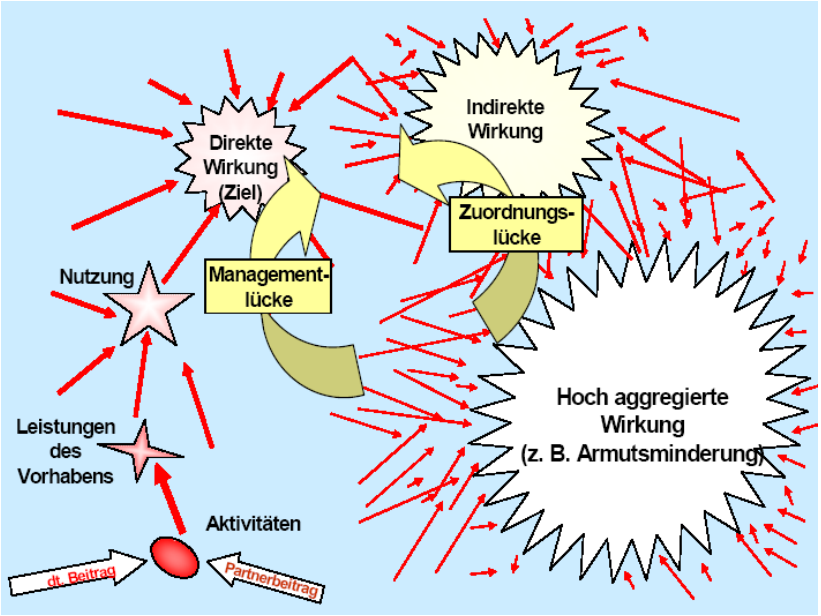
(3) einen Multi-Methodenansatz,

(4) sowie die Anwendung fortgeschrittener statistischer Datenanalyseverfahren (vgl. Caspari/Barbu 2008).

Im Vergleich zur wissenschaftlichen Diskussion bezüglich Wirkungsmessung wird der Schwerpunkt allerdings auf die Messung des Kontrafaktischen durch Anwendung anspruchsvoller Designs und statistischer Auswertungsmethoden gesehen (vgl. NONIE-Netzwerk), der Trichteransatz findet in der Diskussion keine gesonderte Berücksichtigung.

3. Der Wirkungsbetrachtung in der GTZ liegt das nach eigenen Erfahrungswerten ergänzte Wirkungsmodell der internationalen EZ-Diskussion zugrunde. Zum besseren Verständnis der Wirkungsebenen und zur Ausrichtung ihrer Aktivitäten auf Wirkungen wurde eine weitere Ebene hinzugefügt. So lauten die Ebenen des Wirkungsmodells der deutschen EZ (a) Aktivitäten (= Input), (b) Leistungen (= Output), (c) Nutzung der Leistung und (d) direkte Wirkung (c+d = Outcome), (e) indirekte Wirkungen und (f) hoch aggregierte Wirkungen (vgl. Abbildung 4). Dabei bildet die Wirkungskette die zum Beginn der Entwicklungsmaßnahme antizipierten Wirkungen, die bis Ende der Laufzeit erreicht werden sollten, ab.

Abbildung 4: Wirkungskette laut GTZ



Quelle: GTZ 2006

Der Nachweis der direkten Wirkungen (positiv wie negativ) einer Entwicklungsmaßnahme stehen dabei im Vordergrund der durchgeführten Evaluationen. Diese sollen direkt, d.h. ursächlich und kausal der Entwicklungsmaßnahme zugeordnet werden können. Der Nachweis der indirekten Wirkungen und hoch aggregierten Wirkungen kann laut GTZ Verständnis jedoch nur plausibel begründet werden (daher die Zuordnungslücke[3]).

Zur methodischen Vorgehensweise bei der Identifizierung möglicher Wirkungen (besonders der intendierten und nicht-intendierten) werden folgende Aspekte hervorgehoben:

(1) Bewertung des Wirkungsmodells hinsichtlich Qualität (Plausibilität und Anspruchsniveau der Ziel- und Indikatorformulierung unter Berücksichtigung des Kontextes),

(2) der Einsatz mehrerer Methoden (Triangulation) und

(3) der Trichteransatz.

[3] Die GTZ kennzeichnet eine „Zuordnungslücke“ zwischen den Ebenen „direkte Wirkung“ und „indirekte Wirkung“ da, laut GTZ, der Nachweis der indirekten Wirkungen der Entwicklungsmaßnahme nur „plausibel“ erfolgen kann aufgrund zunehmender externer Einflussfaktoren. Das Konzept der Zuordnungslücke wird allerdings in der wissenschaftlichen Diskussion zu Wirkungsmessung umstritten diskutiert. Hier wird davon ausgegangen, dass auch die indirekten Wirkungen durchaus unter Verwendung adäquater Untersuchungsdesigns auf eine Intervention zurückführbar sind.

Spezifikationen zum Evaluationsdesign und den Auswertungsmethoden werden hingegen nicht gegeben, es wird lediglich empfohlen, einen vergleichenden Ansatz zu berücksichtigen. Die bisherige praktische Umsetzung der GTZ Unabhängigen Evaluationen hat jedoch gezeigt, dass ein methodisch sauberer Nachweis von Wirkungen im Rahmen der Unabhängigen Evaluationen noch nicht geleistet wird. Defizite liegen bei der nicht-Anwendung des Multi-Methodenprinzips und eines adäquaten Designs zur Wirkungsmessung (im Detail vgl. Kap. 4).

Das hier entwickelte Konzept für Wirkungsevaluation soll daher für diesen Nachweis eine einheitliche, methodisch adäquate und anspruchsvolle Vorgehensweise skizzieren, die eine fundiertere Betrachtung der indirekten und hoch aggregierten Wirkungen ebenfalls ermöglicht. In den folgenden Kapiteln wird daher die in der wissenschaftlichen Praxis, in der aktuellen internationalen Diskussion der EZ-Organisationen und die durch GTZ angeführten Aspekte der Wirkungsevaluation, die sich überschneiden, näher betrachtet. Hierzu gehören: das Aufstellen theoriebasierter Ursache-Wirkungszusammenhänge, die Anwendung adäquater Forschungsdesigns, das Nutzen eines Multi-Methoden-Ansatzes sowie notwendige Aspekte bei der Datenauswertung.

3. Methodische Besonderheiten der Wirkungsevaluation

3.1 Hypothesengeleitete Ursache- Wirkungsuntersuchung

Wie bereits dargestellt zielen Wirkungsevaluationen darauf ab, mit größtmöglicher Zuverlässigkeit zu messen, ob eine Entwicklungsmaßnahme die intendierten Wirkungen auslöst und welche nicht-intendierten Wirkungen ebenfalls entstanden sind. Die zu beantwortende Fragestellung ist: „Wie hat die Maßnahme unter welchen Bedingungen gewirkt?“. Die Betrachtung des Kontextes und der Rahmenbedingungen als Ursachefaktoren sind somit unumgänglich.

Voraussetzung zur Wirkungsmessung ist, dass differenzierte Ursache-Wirkungszuschreibungen aufgestellt werden (vgl. Stockmann 2006, S. 105; 2007, S. 67) und ein sogenannter *theoriebasierter Ansatzes* (TBA) zur Wirkungsmessung angewendet wird. Hierfür müssen ein explizit *theoretisches Modell über die erwarteten Ursache-Wirkungs-Zusammenhänge mit theoriegestützten Wirkungshypothesen* aufgestellt werden. So kann der Frage nach der Kausalität der Wirkung sowie dem „warum“ adäquat nachgegangen werden (vgl. z.B. White 2006a; Bamberger et al. 2006; Bamberger 2006; Baker 2000; Stockmann 2006 S. 225, 2007). Evaluationen, die zwar Wirkungen nachweisen, aber keine Angaben machen, *warum* eine Maßnahme die erwarteten Wirkungen gezeigt hat oder nicht, hinterlassen eine "*black box*" mit fehlenden Informationen in der Kausalkette einer Wirkung (vgl. Caspari/Barbu 2008 S. 17; Bloom 2006 S. 18f.; White 2006a S. 9; Ravallion 2005 S. 1;).

Bei einem theoriebasierten Ansatz wird mit Hilfe eines Kausalmodells die einer Maßnahme zugrundeliegende Hypothese über Ursache-Wirkungs-Zusammenhänge detailliert ausgearbeitet und tabellarisch oder graphisch dargestellt[4] (= Programmtheorie) (vgl. Caspari/Barbu

[4] Z.B. in der GTZ Wirkungskette mit den Ebenen (a) Leistung, (b) Nutzung der Leistung, (c) Direkte Wirkungen, (d) Indirekte Wirkungen, (e) hoch aggregierte Wirkungen (vgl. GTZ (2006): Handreichung zur Bearbeitung von AURA-Angeboten).

2008, S. 18). Laut Caspari/Barbu zeigt das Wirkungsmodell dabei auf, „welche (impliziten und expliziten) Annahmen über kausale Verknüpfungen zwischen der geplanten Maßnahme und den intendierten direkten und indirekten Wirkungen der Maßnahme zugrunde liegen, d.h. es wird explizit beschrieben, *was* eine Maßnahme *wie* (mit welchen Ressourcen, Aktivitäten und in welchem Kontext), *für wen*, *wozu* (mit welchem übergeordneten Ziel) und *warum* (Wirkungshypothesen) erreichen soll“ (Caspari/Barbu 2008, S. 18). Dadurch soll die Frage der Kausalität von Wirkungen bearbeitet werden. Zu berücksichtigen sind ebenfalls externe Faktoren, die ein mögliches Risiko für die konstatierten Ursache-Wirkungs-Zusammenhänge darstellen (vgl. Caspari/Barbu 2008, S. 18). Wirkungshypothesen müssen für *jede* Ebene des Wirkungsmodells, so wie auch für die Risiken und Rahmenbedingungen formuliert werden. Beachtung finden sollten ebenfalls konkurrierende Ursache-Wirkungs-Zusammenhänge. Klar definierte Ziele und beobachtbare Indikatoren zum Nachweis von Wirkungen und vermuteten Zusammenhängen werden basierend auf dem Wirkungsmodell formuliert und stellen somit ein Referenzrahmen für die Wirkungsevaluation dar (vgl. Stockmann 2006 S. 225).

Laut Caspari/Barbu 2008 ist die Erstellung eines Wirkungsmodells, d.h. der Ursache-Wirkungs-Hypothesen, Risiken und Einflussfaktoren, keinesfalls einfach. Die Identifikation von Hypothesen und Theorien auf jeder Ebene des Wirkungsmodells ist eine äußerst komplexe Aufgabe. Die bei der Planung zugrunde gelegten Hypothesen bzw. teilweise auch das gesamte Wirkungsmodell müssen im Rahmen einer Wirkungsevaluation von den Evaluatoren/innen rekonstruiert und meist überarbeitet werden (vgl. Caspari/Barbu 2008, S. 18f).

Caspari/Barbu 2008 empfehlen, bei der Überprüfung des Wirkungsmodells im Rahmen einer Wirkungsevaluation sollte in einem ersten Schritt auf aktuelle Literatur und wissenschaftliche Erkenntnisse aus dem betreffenden Feld bzw. Sektor zurückgegriffen werden. Bei den vorliegenden Wassermaßnahmen sollten somit aktuelle Erkenntnisse aus der Gesundheitsforschung und dem Wassermanagement berücksichtigt werden, d.h. die Wirkungshypothesen sollten auf bereits überprüften Theorien aufbauen. Zur Reflektion der Richtigkeit des Wirkungsmodells und der Kausalitäten sollten in einem weiteren Schritt die Stakeholder befragt werden (hier eignen sich Intensiv-, Experten- und/oder Fokusgruppeninterviews) (vgl. Caspari/Barbu 2008, S. 19). Ansonsten bleibt die Wahrscheinlichkeit groß, dass indirekte und vor allem nicht intendierte Wirkungen unerkannt bleiben (da dann nur die in dem Wirkungsmodell spezifizierten Wirkungen überprüft werden). Dies macht deutlich, dass bei Wirkungsevaluationen 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.

3.2 Typische Forschungsdesigns für Wirkungsevaluationen

Die Herausforderung von Wirkungsevaluationen ist, die Nettowirkungen, d.h. die Wirkungen die der Entwicklungsmaßnahme ursächlich zugeschrieben werden können, zu identifizieren. Hierbei treten zwei Schwierigkeiten vor dem Hintergrund spezifischer Bedingungen (Zeit, finanzielle Ressourcen, Kooperationsbereitschaft etc.) auf: Zum Einen die möglichst exakte und vollständige Messung geplanter und ungeplanter Wirkungen einer Entwicklungsmaßnahme. Zum Anderen die möglichst eindeutige Bestimmung der Ursachenfaktoren der Wirkungen (vgl. Stockmann 2006 S. 224). Es muss der Frage nachgegangen werden, was ohne

Programm geschehen wäre, d.h. das Kontrafaktische muss bei der Wirkungsmessung berücksicht werden (vgl. Caspari/Barbu 2008).

Hierfür ist die Verwendung spezieller Datenerhebungsdesigns notwendig, die festlegen, wie, wann, wo, und wie oft Daten zu Wirkungen erfasst werden. Das gewählte Design ist dabei entscheidend für den Grad der Gewissheit, mit dem die Frage nach dem Zusammenhang zwischen Ursache und Wirkung beantwortet werden kann (vgl. Schnell, Hill, Esser 2006; Meyer 2006). Aufgrund spezifischer Rahmenbedingungen bei EZ-Evaluationen ist jedoch die Verwendung des optimalen Datenerhebungsdesigns oftmals nicht möglich, trotzdem muss ein wenigstens adäquates Datenerhebungsdesign für Wirkungsmessung Grundlage der Wirkungsevaluation sein. Hierzu gibt es aus der empirischen Sozialforschung und Evaluationsforschung eine Reihe von Empfehlungen, die in Tabelle 1, in Anlehnung an Stockmann 2006 abgebildet sind. Im Folgenden soll auf einige Spezifika der Designs und ihrer Techniken und Verfahren zur Kontrolle von Störfaktoren bei der Wirkungsmessung eingegangen werden.

Tabelle 1: Typische Forschungsdesigns für Wirkungsanalysen

Design		**Auswahl der Untersuchungs-einheiten**	**Art der Kontrollgruppe**	**Datenerhebungszeitpunkte**
I.	‚Echte' Experimente/ Feldexperimente	Randomisierte Auswahl	Randomisierte Kontrollen, oft zusätzlich statistische Kontrollen	Minimum: nur nach der Intervention. Meist vorher und nachher; oft mehrere Messungen während der Intervention
II.	Quasi-Experimente	Unkontrollierte Auswahl	Konstruierte und/oder statistische Kontrollen	Minimum: nur nach der Intervention. Meist vorher und nachher. Oft mehrere Messungen während der Intervention
III.	Querschnitts-analysen	Unkontrollierte Auswahl	Statistische Kontrollen	Nur Nachher-Messungen
IV.	Pretest-Posttest-Untersuchungen	Unkontrollierte Auswahl	Reflexive Kontrollen	Minimum: Vorher- und Nachher-Messung
V.	Retrospektive Vorher-/Nachher-Untersuchungen	Unkontrollierte Auswahl	Retrospektive reflexive Kontrollen	Nachher-Messungen mit retrospektiven Messungen der Ausgangssituation
VI.	Panel- Untersuchungen	Unkontrollierte Auswahl	Reflexive Kontrollen	Mehr als zwei Messungen während der Intervention
VII.	Zeitreihenanalysen	Unkontrollierte Auswahl	Reflexive Kontrollen	Viele Messungen vor und nach der Intervention
VIII.	Gutachtenmodell	Unkontrollierte Auswahl	Generische und/oder Schattenkontrollen	Nur Nachher-Messungen

Quelle: Stockmann 2006 S. 229 in Anlehnung an Rossi u.a. 1988 S. 113; Rossi, Freeman und Lipsey 1999 S. 261.

Zunächst einmal muss festgehalten werden, dass Designs für Wirkungsevaluationen stets auf Vergleichen beruhen (vgl. Stockmann 2006, S. 225). In der Regel werden dazu 2 Gruppen verglichen, eine bei der die geplante Intervention stattfindet (Zielgruppe) und eine zweite, bei der keine Intervention vorgenommen wird (Kontroll/Vergleichsgruppe). Von *Kontrollgruppe* spricht man nur dann, wenn diese per Zufallszuteilung vor Beginn einer Entwicklungsmaßname gebildet wurde, d.h. die Personen, die im Rahmen einer Entwicklungsmaß-

nahme eine Unterstützung erfahren sollen, werden per Zufall der Ziel- & Kontrollgruppe zugeteilt. Von *Vergleichsgruppe* spricht man hingegen dann, wenn die Zuteilung der Vergleichsgruppe ohne Randomisierung z.B. erst nach Start der Entwicklungsmaßnahme gebildet wird. Aus dem Vergleich der Ziel- und Kontroll/Vergleichsgruppe kann dann auf die Wirkung geschlossen werden. Eine Betrachtung nur der Zielgruppe zu nur 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 (vgl. Caspari/Barbu 2008).

Ziel bei der Wahl des adäquaten Datenerhebungsdesigns ist es nun, konkurrierende Erklärungen und Störvariablen auszuschließen, um den Ursache-Wirkungs-Zusammenhang möglichst optimal belegen zu können. Hierzu gibt es folgende Möglichkeiten (im Detail vgl. Stockmann 2006 S. 229ff.; vgl. Tabelle 1):

- Randomisierung: Per Zufall Zuteilung von Personen in Ziel & Kontrollgruppe vor Beginn einer Entwicklungsmaßnahme (verwendet im Experimentellen Design I).
- Konstruierte Kontrollen (Matching on Observables): Personen, die im Hinblick auf bestimmte Merkmalsausprägungen der Zielgruppe gleichen, werden der Vergleichsgruppe zugeordnet (Suche eines äquivalenten Partners) (verwendet im Quasi-Experimentellen Design II).
- Statistische Kontrollen: Zur Überprüfung, ob sich die Ziel- und Kontroll/Vergleichsgruppe tatsächlich in allen wichtigen Merkmalen gleichen. Oder aber zur statistischen Konstruktion einer Vergleichsgruppe basierend auf existierende Census-Daten, Panel-Daten, repräsentativen Haushaltsbefragungen, sonstigen Datensätze allgemeiner Bevölkerungsumfragen oder sonstige Daten[5] (*Propensity Score Matching*) (vgl. Caspari/Barbu 2008) (verwendet im Quasi-Experimentellen Design & Querschnittsdesign II + III).
- Reflexive Kontrollen: Die Zielgruppe wird zur eigenen Vergleichsgruppe indem die Daten zu mehreren Zeiten vor und nach der Intervention gemessen werden (Längsschnittstudie). Falls die Maßnahme noch nicht allzu lange lief, können die zum ersten Zeitpunkt gewonnen Daten wie eine vorher Messung genutzt werden. Eine Variation stellt das *"Multiple Comparison Group Design"* dar. Hier werden bei zeitversetzt implementierten Maßnahmen verschiedene Teilnehmer/innen-Gruppen mit unterschiedlichem Beginn der Intervention und daraus resultierenden unterschiedlichen Eigenschaften *untereinander* als Vergleichsgruppe genutzt und verglichen (vgl. Caspari/Barbu 2008) (verwendet im Design IV, V, VI, VII).
- Generische Kontrollen: Interventionseffekte bei der Zielgruppe werden mit typischen Veränderungen in der Gesamtpopulation verglichen. Kennwerte wie z.B. Sterbe- und Fruchtbarkeitsziffern, Indikatoren zur Charakterisierung der Erwerbsbevölkerung etc. werden herangezogen, um abzuschätzen, was sich ohne die Intervention ereignet hätte. Differenzen bei den Messwerten werden als Wirkung der Entwicklungsmaßnahme zugeschrieben. Solche gesellschaftlichen Kennwerte liegen jedoch nur für wenige soziale Bereiche vor und auch nur in wenig differenzierter Form oder schlech-

[5] Um diese existierenden Datensätze „aufzuspüren“, ist allerdings häufig vor der eigentlichen Evaluationsmission ein gesonderter Vor-Ort Besuch notwendig.

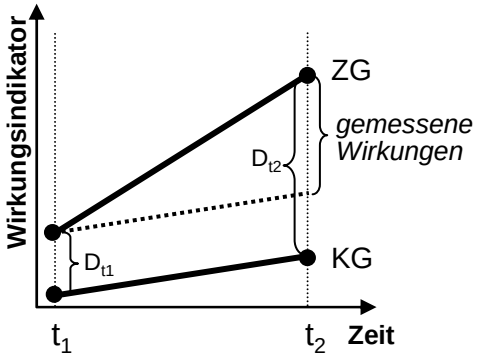
ter Qualität (verwendet im Gutachtenmodell VIII).

- Schattenkontrollen: Die Wirkungen bei der Zielgruppe werden basierend auf Einschätzung von Experten, Programmleiter und/oder Teilnehmer mit dem verglichen, was ‚normalerweise‘, d.h. ohne die Intervention zu erwarten gewesen wäre. Dieses Verfahren führt allerdings zur methodisch schwächsten Bewertung der tatsächlich auf die Entwicklungsmaßnahme zurückzuführenden Wirkungen und ist für wirkliche Wirkungsevaluationen nicht zu empfehlen (verwendet im Gutachtenmodell VIII).

Darüber hinaus eröffnen spezifische Konzeptionen von EZ-Maßnahmen weitere Möglichkeiten, eine Vergleichsgruppe zu konstruieren (siehe Caspari/Barbu 2008, S. 14): „Ist z.B. 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 S. 103f.; Bamberger 2006 S. 11). Wird z.B. ein Arbeitsmarktprogramm für Jugendliche bis 24 Jahre aufgelegt, so eignet sich die Gruppe der 25-jährigen gut als Vergleichsgruppe“.

Eine Vorgehensweise zur Kontrolle konkurrierender Erklärungen und Störvariablen ist das Messen von Wirkungen zu mehreren Erhebungszeitpunkten (Längsschnittuntersuchungen) durch Erhebung gleicher Daten bei Ziel- und Kontroll/Vergleichsgruppe (die sogenannte *double-difference Methode*). Um optimale Aussagen zu bekommen, sollte die Wirkungsmessung möglichst vor der Maßnahme (als Ausgangswert / Referenzwert der Wirkung zum Zeitpunkt t^1 = Baselinedaten) (oder aber kurz nach ihrer Implementation) und spätestens kurz vor (Schlussbetrachtung der Wirkungen) oder nach Beendigung der Maßnahme (Ex-post Betrachtung von Wirkungen) stattfinden (zum Zeitpunkt t^2). Die Identifizierung der Wirkungsveränderung basiert auf dem Vergleich der Werte zum Zeitpunkt t^2 mit den Ausgangswerten gemessen in t^1. Nach Caspari/Barbu 2008 ergibt sich die Wirkung einer Maßnahme somit aus dem Unterschied zwischen Zielgruppe und Vergleichsgruppe nach der Maßnahme (t_2) minus dem Unterschied zwischen Zielgruppe und Vergleichsgruppe vor der Maßnahme (t_1) (siehe Caspari/Barbu 2008 S. 8; vgl. Baker 2000 S. 56; ADB 2006 S. 13; White 2006a) (vgl. Abb. 8).

Abbildung 5: Differenzen-in-Differenzen Schätzung der Wirkungen



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

Quelle: Caspari/Barbu 2008 S. 9

Aus den aufgeführten Möglichkeiten zur Kontrolle von Störfaktoren bei der Wirkungszuschreibung können folgende Schlussfolgerungen für die Anwendung der einzelnen Forschungsdesigns für Wirkungsevaluationen gezogen werden (im Detail vgl. Stockmann 2006 S. 230ff.; vgl. Tabelle 1).

- Experimentelle Designs sind zwar ideal zur Überprüfung und Identifikation von Kausalzusammenhängen (Ursache-Wirkungshypothesen) zwischen den Leistungen einer Entwicklungsmaßnahme (unabhängige Variabel) und den erfassten Wirkungen (abhängige Variable), in der EZ-Evaluationspraxis allerdings selten anwendbar. Eine Kontrollgruppe muss bereits im Rahmen der Konzeption und Durchführung der Entwicklungsmaßnahme per zufallsgesteuertem Auswahlprozess gebildet werden. Dies nachträglich im Rahmen einer Wirkungsevaluation durchzuführen, ist technisch unmöglich. Hinzu kommen oft geäußerte ethische Bedenken hinsichtlich einer randomisierten Zuteilung in Ziel- und Kontrollgruppe[6], sowie soziale, politische und rechtliche Einschränkungen die eine Randomisierung oft unmöglich machen (vgl. Tabelle 1: Experimente I).
- Wenn das experimentelle Design aus methodischen, technischen oder forschungsethischen Gründen nicht eingesetzt werden kann, wird häufig auf Quasi-Experimente ausgewichen, die vereinfacht als „Experimente ohne Randomisierung“ (Diekmann 1995 S. 309) bezeichnet werden können. Ein wesentlicher Unterschied zum Experiment besteht darin, dass die Aufteilung der Ziel- und Vergleichsgruppen nicht per Zufallsauswahl möglich ist, sondern konstruierte und/oder statistische Kontrollen eingesetzt werden. Für die Aufgaben der wirkungsbezogenen Evaluationsforschung ist die quasi-experimentelle Untersuchungsanordnung mit Vergleichsgruppen besonders geeignet (vgl. Kromrey 2002 S: 100; Diekmann 1995 S. 320), auch wenn aufgrund der fehlenden Randomisierung nicht ganz sicher ist, ob eventuelle Drittvariableneffekte neutralisiert werden konnten (vgl. Tabelle 1: Quasi-Experimente II).
- Wenn selbst die Bedingungen für ein Quasi-Experiment nicht gegeben sind, dann kann versucht werden die soziale Realität mit so genannten ‚Ex-post-facto-Designs‘ im Nachhinein zu erfassen. In Querschnittsuntersuchungen (*single-difference Methode)* werden alle zu messenden Variablen nur zu einem Zeitpunkt erhoben (vgl. Schnell, Hill, Esser 1999 S. 218ff. Baker 2000, White 2006a). Mit retrospektiven Fra-

[6] In Caspari/Barbu 2008 werden Möglichkeiten skizziert für die auch ethisch bedenkenlose Einteilung von Personnengruppen in Ziel- & Kontrollgruppe.

gen wird versucht, Informationen über frühere Zeitabschnitte zu erhalten. Bei Querschnittsanalysen können die Variationen der Programmgestaltung in verschiedenen Regionen dazu verwendet werden, Programmeffekte aufzuspüren (vgl. Rossi, Freeman und Lipsey 1999 S. 267). Querschnittsdesigns haben den Vorteil, dass sie in der Regel schnell durchzuführen und relativ kostengünstig und somit das in der EZ bisher am häufigsten vorkommendes Untersuchungsdesign sind. Nach Caspari/Barbu 2008 ist die zentrale Grundannahme, dass die Ausgangssituation der Ziel- und Vergleichsgruppe vor der Maßnahme identisch ist (Caspari/Barbu S. 8). Der gefundene Unterschied zwischen der Ziel- und Vergleichsgruppe wird somit allein der Maßnahme zugeschrieben[7] (Caspari/Barbu S. 13). Die Annahme, dass die Ausgangssituation der Zielgruppe und der Vergleichsgruppe vor der Maßnahme identisch ist, ist in der Realität jedoch selten gegeben. Folglich führt die Verwendung der *single-difference Methode* in der EZ zur Über- oder Unterbewertung der berechneten Wirkungen, da Wirkungen der Maßnahme zugeschrieben werden, die eventuell einen anderen Ursprung haben. Da mögliche Gruppenunterschiede zum Zeitpunkt t_1 völlig unbeachtet bleiben, ist eine äußerst bedachte Auswahl der Vergleichsgruppe nötig (Caspar/Barbu 2008 S. 13) (vgl. Tabelle 1: Querschnittsanalysen III).

- Besonders häufig wird in Evaluationen das Pretest-Posttest-Design (Vorher-Nachher-Vergleich), bei dem die Indikatoren vor und nach der Einführung einer Intervention bei der Zielgruppe gemessen werden, eingesetzt. Die Differenz der Messwerte soll Aufschluss über die Wirkungen der Entwicklungsmaßnahme geben. Dabei wird davon ausgegangen, dass die Messwerte bei Pretest und Posttest gleich ausgefallen wären, wenn es keine Intervention gegeben hätte. Nach Caspari/Barbu 2008 können „andere, externe Faktoren, wie z.B. Maßnahmen anderer Geber aber auch unerwartete Ereignisse (Naturkatastrophen, Kriege, etc.) oder auch allgemeine Veränderungsprozesse (Wirtschaftswachstum/-krise, Verstädterung, etc.) [ ] die Wirkung der Maßnahme beeinflussen, schwächen oder auch verstärken. Gleichwohl ist anzunehmen, dass sich die Situation der Zielgruppe 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“ (Caspari/Barbu 2008 S. 6). Durch die Nichtberücksichtigung des Kontrafaktischen (da keine Vergleichsgruppe) ist dieses Design für Wirkungsmessungen wenig geeignet. Es „zeigt lediglich die Entwicklung der Zielgruppe über die Zeit hinweg auf – das Faktische [ ], nicht aber das Kontrafaktische – und kann demnach höchst selten eine zuverlässige Antwort auf die Frage nach Wirkungen einer Maßnahme geben“ (Caspari/Barbu 2008 S. 7, vgl. Tabelle 1: Pretest-Posttest-Untersuchungen IV).

- Panelanalysen sind den Vorher-Nachher-Designs zuzuordnen, wobei mehrere Messungen bei denselben Personen mit vergleichbaren Fragestellungen und möglichst identischen Datenerhebungsinstrumenten zu verschiedenen Zeitpunkten stattfinden. Durch den Vergleich mehrerer Messungen bei denselben Personen lassen sich nicht

[7] Problematisch – wie bei allen Ex-post-facto-Anordnungen – ist jedoch das Problem der kausalen Reihenfolge, das sich aus der Tatsache der einmaligen, gleichzeitigen Erhebung aller Daten ergibt. Dadurch werden alternative Erklärungen möglich. Beispiel: Es kann sein, dass das Anschauen von Filmen mit aggressivem Inhalt zu aggressivem Verhalten führt, es kann aber auch umgekehrt sein. Zudem ist die Kontrolle von Drittvariablen wesentlich schwieriger zu gewährleisten als in Experimenten. D.h. es können auch andere Variablen als die unabhängige Variable für die beobachteten Veränderungen verantwortlich sein.

nur interindividuelle Veränderung (hinsichtlich einer Fragestellung) in der Gesamtheit der Gruppe (Nettoveränderung), sondern auch intraindividuelle Einflüsse (z.B. Leistungszuwächse, Einkommensveränderungen u.ä.) feststellen (vgl. Schnell/Hill/Esser S. 237ff.). Für Wirkungsanalysen sind Paneldesigns weitaus besser geeignet als z.B. Querschnitts- oder Vorher-Nachher-Untersuchungen, da die Beobachtung individueller Veränderungen eine wesentlich bessere Schätzung der Wirkungsweise einer Intervention gestatten. Der Vorteil von Paneluntersuchungen besteht vor allem darin, dass die unabhängigen und abhängigen Variablen zeitversetzt aufeinander bezogen werden können. Hier bleibt allerdings auch festzuhalten, dass ohne Betrachtung einer Vergleichsgruppe, das Kontrafaktische allein durch ein Paneldesign noch nicht berücksichtigt wird (vgl. Tabelle 1: Panel-Untersuchungen VI).

- Bei Zeitreihenanalysen werden viele Messungen vor und nach der Intervention durchgeführt (empfohlen werden ca. 30 Messungen vor der Intervention). Aus solchen regelmäßig (z.B. monatlich, vierteljährlich, jährlich) erhobenen Daten können Zeitreihen gebildet werden, die sich für ‚Kontrollzwecke‘ nutzen lassen. Die Daten bieten eine relativ sichere Grundlage für Schätzungen, wie die Entwicklung der Zielvariablen ohne Interventionsmaßnahme verlaufen wäre. Dabei wird im Einzelnen so vorgegangen, dass in der Regel an aggregierten Untersuchungseinheiten eine Reihe von Messungen durchgeführt wird, bevor eine Intervention oder bedeutende Programmmodifikation vorgenommen wurde. Aus diesen Daten wird ein ‚Trend‘ berechnet, der eine Vorhersage ermöglicht, was geschehen wäre, wenn es keine Intervention gegeben hätte. Mit diesem ‚Trend‘ werden die Messwerte verglichen, die nach Einführung der Intervention erhoben wurden. Aus der Differenz zwischen dem langfristig errechneten ‚Trend‘ und den Werten nach der Intervention wird auf den Nettoeffekt der Intervention geschlossen. Mit Hilfe inferenzstatistischer Testverfahren lassen sich die auftretenden Zufallsschwankungen kontrollieren (vgl. Tabelle 1: Zeitreihenanalysen VII).
- Das am wenigsten empfehlenswerteste - da am unzuverlässigsten und am ungenauesten - Design zur Wirkungsmessung ist das Gutachtenmodell. Hierfür werden Bewertungen von Experten, Hauptstakeholder und der Zielgruppe zur Schätzung der Netto-Wirkungen herangezogen[8]. Laut Kromrey (2002 S. 103) haben die erhobenen Einschätzungen allerdings „weder den Status von Bewertungen im Sinne ‚technologischer‘ Evaluationen noch von Bewertungen neutraler Experten“. Stattdessen handelt es sich „um individuell parteiische Werturteile von Personen, die in einer besonderen Beziehung – eben als Nutzer, als Betroffene – zum Untersuchungsgegenstand stehen“. Dieses Design kann daher nicht als verlässliches Design zur Wirkungsmessung herangezogen werden (vgl. Tabelle 1: Gutachtenmodell VIII).

Für die Wirkungsmessung ist eine möglichst eindeutige Zuschreibung der Wirkungen zu einer bestimmten Maßnahme anzustreben. Dies kann am ehesten in quasi-experimentellen Designs mit (mindestens) einer Vergleichsgruppe und (mindestens) zwei Messzeitpunkten (Vorher-Nachher-Messung) mit Hilfe einer double-difference Berechnung sichergestellt werden. Anzustreben ist eine möglichst perfekte Zufallszuweisung zu den beiden Untersuchungsgruppen sowie eine möglichst große Zahl an Messzeitpunkten. Da allerdings zeitliche, finanzielle und soziale Restriktionen häufig die Möglichkeiten begrenzen, kann dies nur

[8] Dies ist z.B. bei der Bewertung im e-Val Verfahren der Fall.

als Ideal angesehen werden und es ist im Einzelfall zu prüfen, welches Design (bzw. welche Kombination von Designs) am ehesten umsetzbar und den methodischen Ansprüchen am nächsten kommen.

3.3 Multi-Methodenansatz

Wie die bisherigen Ausführungen gezeigt haben, ist die Nutzung geeigneter Untersuchungsdesigns (z.B. Quasi-Experimentelles Design) ein zentrales Element von Wirkungsevaluationen (vgl. Kap. 3.2). Hinzu kommen theoriebasierte sowie qualitative bzw. partizipative Ansätze zur ersten Identifizierung möglicher Wirkungen (vgl. Kap. 3.1). Für die Attribution identifizierter Wirkungen zu einer Maßnahme müssen sowohl quantitative als auch qualitative Datenerhebungsmethoden verwendet werden (*Triangulation/Methodenmix*) um keine „black box“ in der Kausalkette der Wirkungen zu hinterlassen (vgl. White 2006a S. 20).

Dabei ist darauf zu achten, dass neben den Daten zu Veränderungen bei Zielgruppe und Vergleichsgruppe auch Daten zur Maßnahme selbst erhoben werden müssen. Dies ist bei Entwicklungsmaßnahmen der Finanziellen Zusammenarbeit (FZ) z.B. über die reine Höhe der finanziellen Zuwendung vergleichsweise einfach machbar. Bei Maßnahmen der technischen Zusammenarbeit (TZ), die insbesondere Beratungsleistungen umfassen, ist dagegen für jede einzelne Komponente der Aktivitäten ein sinnvoller Indikator zu bilden und die entsprechenden Daten sind zu sammeln. Nach Caspari/Barbu 2008 müssen zur Klärung der Frage, *warum* eine Maßnahme gewirkt oder nicht gewirkt hat, mögliche *intervenierende Einflussgrößen* ebenfalls operationalisiert und entsprechende Daten gesammelt werden (Caspari/Barbu 2008 S. 17). Auch dies ist im Kontext von Beratungsleistungen deutlich schwieriger zu realisieren und bedarf eines tiefergehenden Verständnisses der sozialen Prozesse, die Wirkungen erst möglich machen oder verhindern können. Zur Untersuchung dieser Aspekte können qualitative Ansätze im besonderen Maße beitragen und quantitative Verfahren sinnvoll ergänzen.

Die angemessene Verwendung eines möglichst breiten Spektrums an Datenerhebungsmethoden und Tools – in Abhängigkeit der jeweiligen Fragestellung – muss im Vordergrund der Datenerhebungsphase im Rahmen der Wirkungsevaluation stehen. Hierdurch wird sichergestellt, dass die methodischen Schwächen eines Instruments durch die Stärken anderer Instrumente ausgeglichen werden. Die ausgewählten Erhebungsverfahren sollten sich ergänzen und ein breites Informationsbild abbilden. Die Datenerhebungsverfahren sollten dabei so gewählt werden, dass die Validität und Reliabilität gewährleistet ist, also die Qualität der erhobenen Daten ausreichend ist. Ferner sollte die Belastung der Beteiligten und Betroffenen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluierung stehen.

Qualitative Methoden können dabei z.B. genutzt werden, um Wirkungshypothesen zu erstellen, diese zu überprüfen, Erfahrungen aus Einzelfällen vertiefend zu untersuchen, die Kausalitätsfrage zu beantworten, dabei auch mögliche nicht-intendierte Wirkungen zu erfassen, und zur Stützung adäquater Interpretationen der gefundenen Ergebnisse beitragen (vgl. Caspari/Barbu 2008 S. 30). Das Methodensprektrum reicht dabei vom Aktenstudium über Leitfadeninterviews, Beobachtung, Gruppenverfahren bis zu den partizipativen Ansätzen (vgl. Flick 2007).

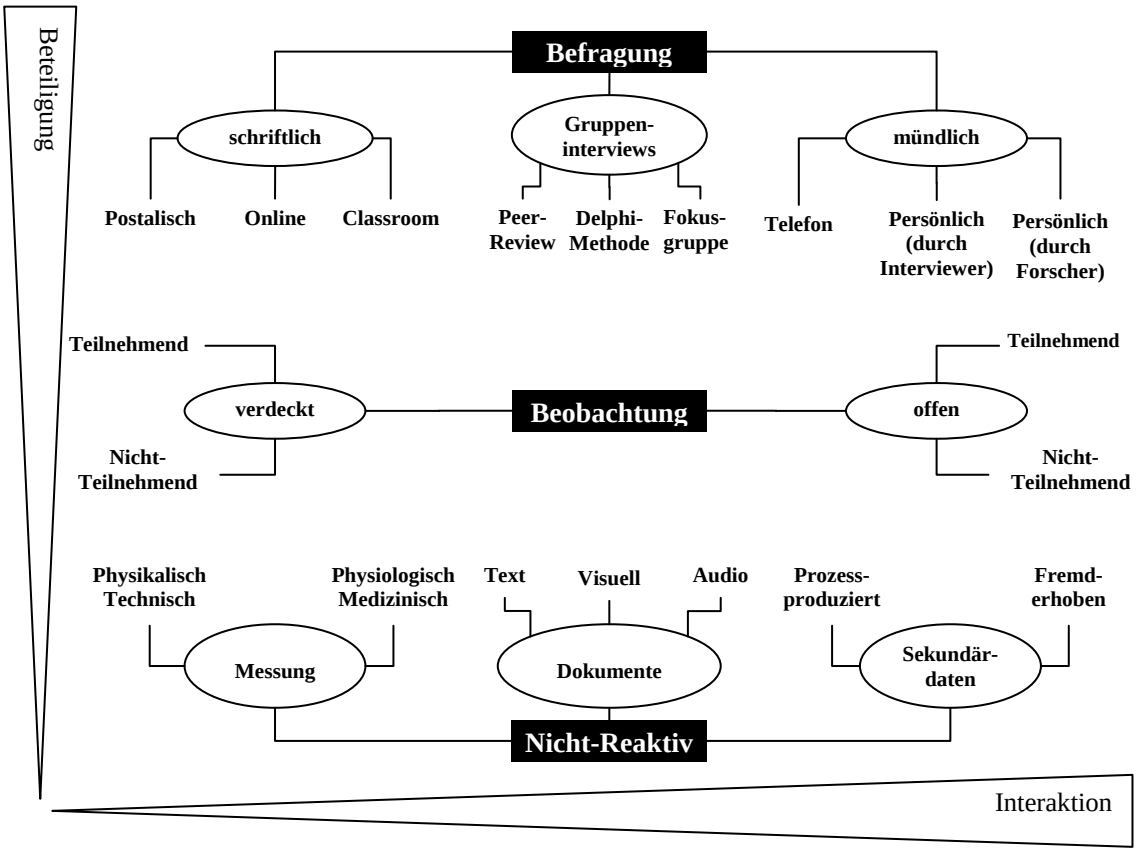
Quantitative Methoden sind notwendig, um die Richtung und Stärke sowie die Ursachen von Wirkungen nachweisen zu können (vgl. Caspari/Barbu S. 30). Hier ist vor allem die Notwen-

digkeit der Datenerhebung mittels standardisierter, repräsentativer und damit eine vorher-nachher Vergleich ermöglichende Befragung der Zielgruppe (und Vergleichsgruppe) hervorzuheben. Nur mittels Vollerhebung oder randomisierter Auswahl durch Anwendung adäquater Samplingmethoden können verallgemeinerbare Aussagen zur Wirkungserreichung ermittelt werden, die für die gesamte Zielgruppe zutreffen (vgl. Schnell/Hill/Esser 2008). Ebenso sollte bei vorhandenen Baseline-Daten darauf geachtet werden, zum Zeitpunkt der Evaluation vergleichbare Daten zu erheben.

Prinzipiell lassen sich drei unterschiedliche Gruppen von Datenerhebungsverfahren unterscheiden, die hinsichtlich der Einfluss- und damit der Manipulationsmöglichkeiten durch die Datenerhebenden wie die Datenliefernden Personen variieren (vgl. Abb. 6). Befragungen kennzeichnen eine mehr oder weniger direkte Beteiligung beider Parteien – Informationsinhaber und Informationsinteressent – an dem Datenerhebungsprozess. Bei der Beobachtung dagegen sollen die Informationsinteressenten nicht aktiv steuernd in den von Ihnen zu beobachtenden sozialen Prozess eingreifen. Nicht-reaktive Verfahren realisieren schließlich die weitgehend personenunabhängige Datenerhebung.

Hinsichtlich des Einflusses der beteiligten Personen lassen sich innerhalb der einzelnen Verfahrensgruppen weitere Differenzierungen vornehmen, die zugleich die Bedeutung anderer Einflussfaktoren (wie z.B. die Beteiligung dritter Personen, die räumlichen Rahmenbedingungen oder die Repräsentativität der Messwerte) in den Fokus der Betrachtung rücken (vgl. Meyer 2007).

Abbildung 6: Datenerhebungsmethoden



Quelle Meyer 2007: 226

Bei der Anwendung der Methoden ist jedoch auf verschiedene Störfaktoren zu achten. Beispielsweise sollte bei mehreren Messzeitpunkten auf Konstanz in der Operationalisierung der Fragestellung geachtet werden, d.h. es sollten die selben Tests, dieselben formulierten Fragen, gleiche Antwortvorgaben etc. verwendet werden. Auch kann ein bei Ziel- und Vergleichsgruppe (möglicherweise unabsichtlich) anderes Verhalten des Evaluators ebenfalls zu einer Beeinträchtigung der Messergebnisse führen (vgl. Stockmann 2006 S. 226).

3.4 Datenauswertung

Bei der Datenauswertung sind sowohl die qualitativ erhobenen Daten als auch die quantitativ erhobenen Daten zusammenfassend unter Berücksichtigung des Gesamtkontextes zu interpretieren. Für qualitative Daten müssen interpretative Verfahren zur Auswertung verwendet werden, bei quantitativen Daten kommen multivariate statistische Verfahren zum Einsatz (z.B. Varianzanalyse, T-Tests, nichtparametrische Testverfahren evtl. auch Regressionsanalyse). Besonders mittels der Anwendung *regressionsbasierter Ansätze* können die auf der Grundlage des aufgestellten *theoretischen Modells* über die Zusammenhänge zwischen Maßnahme und Wirkungen sowie weiterer relevanter Einflussfaktoren entwickelten Wirkungshypothesen statistisch überprüft werden (vgl. Caspari/Barbu S. 17).

Bei der Auswertung der quantitativen und qualitativen Daten geht es darum, eingetretene Veränderungen zu den Messzeitpunkten t^1 (Ausgangssituation) und t^2 (Zeitpunkt der Evaluation) zu identifizieren und diese zu prüfen.

Bei der Datenerhebung und Interpretation, d.h. der Wirkungszuschreibung ist nach Stockmann (2006) ein Trichteransatz zu wählen: Zunächst werden die Veränderungen im Umfeld der Entwicklungsmaßnahme identifiziert. Im nächsten Schritt wird geprüft, welche dieser Veränderungen (Wirkungen) kausal der Entwicklungsmaßnahme zuzuordnen sind. Dies erfolgt durch den Abgleich mit dem Wirkungsmodell der Entwicklungsmaßnahme. Positive und negative Wirkungen sind dabei zu berücksichtigen. Als Ergebnis des Abgleichs wird ersichtlich, welche Wirkungen intendiert waren, welche nicht-intendierten Wirkungen aufgetreten sind – der Entwicklungsmaßnahme aber kausal zuzuordnen sind – und welche negativen Nebenwirkungen ggf. zusätzlich ausgebildet wurden.

4. Bisheriger Schwerpunkt und Umsetzung der GTZ Unabhängigen Evaluationen

Nachdem die Anforderungen an Wirkungsevaluationen skizziert wurden, wird nun nachfolgend beleuchtet, wie die bisherigen, durch die GTZ Stabstelle Evaluierung in Auftrag gegebenen Unabhängigen Evaluationen, umgesetzt werden und wurden und welche speziellen Anforderungen hier greifen.

Die Unabhängigen Evaluationen der GTZ dienen nicht nur der Wirkungsmessung, sondern der allgemeinen Erfolgsbewertung von Entwicklungsmaßnahmen. Der Aspekt der Wirkungsmessung ist dabei ein Aspekt unter vielen. Die Evaluationen sollen generell Auskunft geben über den Erfolg der Maßnahme, über ausgewählte entwicklungspolitisch bedeutsame Themen, die für die weitere Konzept- und Sektorentwicklung von Relevanz sind, sowie über

unternehmenspolitische Fragestellungen, die im Kontext der Weiterentwicklung des Evaluierungssystems, der eingesetzten Instrumente sowie der Produktentwicklung genutzt werden können. Hierzu werden als Orientierungslinie fünf Frageblöcke mit detaillierten Leitfragen zu folgenden Themen vorgegeben (vgl. Muster-TOR GTZ Unabhängigen Evaluationen und Anleitung für die Erfolgsbewertung 2008):

(1) Bewertung nach den international abgestimmten OECD-DAC Evaluierungskriterien (Relevanz, Effektivität, „Impact“, Effizienz und Nachhaltigkeit).

(2) Bewertung der Entwicklungsmaßnahme in Bezug auf Armutsminderung und Millenniumsentwicklungsziele (Einschätzung darüber, inwieweit die Entwicklungsmaßnahme zur Armutsminderung und zur Erreichung der MDG beiträgt).

(3) Bewertung der Entwicklungsmaßnahme in Bezug auf die Förderung der Gleichberechtigung der Geschlechter (Einschätzung darüber, inwieweit die Entwicklungsmaßnahme zur Förderung der Gleichberechtigung der Geschlechter beiträgt).

(4) Bewertung der Entwicklungsmaßnahme in Bezug auf die Förderung nachhaltiger Entwicklung (Einschätzung darüber, inwieweit die Entwicklungsmaßnahme zur Förderung nachhaltiger Entwicklung beiträgt).

(5) Fachbezogene Erfolgsbewertung der Entwicklungsmaßnahme (Fach- und Sektorspezifische Fragestellungen zur Bewertung der Entwicklungsmaßnahme).

Der Wirkungsaspekt wird dabei besonders unter Punkt (1) bei der Bewertung der Kriterien Effektivität und Impact, sowie auch unter Punkt (2) bei der Bewertung hinsichtlich Armutsminderung und Millenniumsentwicklungsziele berücksichtigt.

Bewertungsgrundlage der Evaluation und somit auch Grundlage für die Überprüfung möglicher Wirkungen ist das Wirkungsmodell der jeweiligen Entwicklungsmaßnahme. Dieses Wirkungsmodell wird im Rahmen der Evaluation hinsichtlich Qualität durch die Gutachter überprüft (Plausibilität und Anspruchsniveau der Ziel- und Indikatorformulierung) und ggf. noch ergänzt. In der Praxis wird das Wirkungsmodell basierend auf der Dokumentenanalyse, ersten Gesprächen mit Verantwortlichen der Entwicklungsmaßnahme und der Erfahrung des Evaluierers überprüft und schließlich nach durchgeführter Evaluation bewertet. Eine umfassende theoretische Auseinandersetzung mit den Wirkungshypothesen des Wirkungsmodells unter Einbezug der Hauptstakeholder (nicht nur Projektverantwortliche), wie für Wirkungsevaluationen gefordert (vgl. Kap. 3.1), findet jedoch aufgrund zeitlicher Restriktionen nicht statt. Diese wäre allerdings notwendig, um nicht-intendierte, positive wie negative Wirkungen nahezu vollständig aufzudecken und tatsächlich eine Antwort auf die Fragestellungen „Warum hat eine Maßnahme Wirkungen entfaltet (oder nicht)?“ bzw. „Wie hat die Maßnahme gewirkt, unter welchen Bedingungen?“ zu finden. Beiträge zu höher aggregierten Entwicklungszielen, die gemäß dem Wirkungsmodell der GTZ jenseits der Zuordnungslücke liegen, sollen in der Evaluierung der Entwicklungsmaßnahme jedoch nur durch fundierte Hypothesen belegt werden.

Die Entscheidung, welche Untersuchungsmethoden für die Evaluation angewendet werden sollen, obliegt den Evaluatoren. Die GTZ verlangt jedoch von den Gutachtern der Fremdevaluation als Mindeststandard (1) die Vorlage eines Inception Reports, (2) den Einsatz mehrerer Methoden (Triangulation) und (3) des Trichteransatzes sowie (4) die Nutzung der Befun-

de des obligatorisch vorgeschalteten e-VAL-Verfahrens. Ferner wird darauf hingewiesen, dass ein „Kontrollvergleich“ durchgeführt werden sollte.

Die bisherige praktische Umsetzung der Evaluationen hat jedoch gezeigt, das eine methodisch anspruchsvolle Durchführung der Evaluationen, d.h. besonders ein tatsächlich vorliegender quantitativer und qualitativer Methodenmix (speziell quantitative Zielgruppenbefragungen), die Berücksichtigung des Kontrafaktischen und die Berücksichtigung von Validität und Reliabilität der Datenerhebung und Auswertung aufgrund von zeitlichen Restriktionen kaum verwirklicht werden. Statistische Daten, insbesondere zur Überprüfung von Wirkungsindikatoren, werden über die M&E-Systeme (falls vorhanden), die amtliche und nichtamtliche Statistik sowie aus Beobachtungen erhoben.

Ein einheitlicher Qualitätsmaßstab bei der Durchführung der Evaluationen ist somit nicht gewährt. Dies führt dazu, dass die Vergleichbarkeit der Evaluierungsergebnisse erschwert wird, da jeweils qualitativ unterschiedliche Evaluationsdesigns zur Datengewinnung herangezogen werden.

5. Umsetzungsempfehlungen für Wirkungsevaluationen

Bezugnehmend auf die bisherige Umsetzungspraxis der Unabhängigen Evaluationen in der GTZ und der immer stärker geforderten Schwerpunktsetzung auf Wirkungsmessung wird deutlich, dass bei der Durchführung von Wirkungsevaluationen die hier vorgestellten methodischen Anforderungen stärkere Beachtung finden müssen. Im Folgenden werden daher Umsetzungsmöglichkeiten mit Empfehlungen zu Ablauf und Organisation, gegliedert nach dem Evaluationsprozess 1. Vorbereitung (methodisch und organisatorisch), 2. Durchführung vor Ort und 3. Datenauswertung und Berichterstellung, aufgezeigt. ***Details und länderbezogene Besonderheiten müssen im Rahmen der Inception Reports der Einzelevaluationen noch auf den jeweiligen Kontext und Rahmenbedingungen inhaltlich spezifiziert werden.***

Allgemein sollte zum Ablauf und der Organisation der Evaluation noch festgehalten werden, dass eingängige ‚Standards für Evaluation’ (z.B. der Gesellschaft für Evaluation (DeGEval[9], die DAC Evaluation Quality Standards[10] und die UNEG Standards for Evaluation in the UN System[11]) zu berücksichtigen sind. Darüber hinaus sollte angestrebt werden, die Evaluierung bzw. die einzelnen zu bearbeitenden Fragestellungen im Sinne der Forscher-Triangulation möglichst durch mehrere (mindestens 2) GutachterInnen zu bearbeiten, so dass dem Problem der Subjektivität von Bewertungen und Empfehlungen durch ‚Objektivierung’ der subjektiven Einschätzungen nach dem Konsensualprinzip entgegengewirkt werden kann („Vieraugen-Prinzip“).

[9] Die Standards umfassen die Aspekte Nützlichkeit (N1-8), Durchführbarkeit (D1-3), Fairness (F1-5) und Genauigkeit (G1-9). Siehe: Gesellschaft für Evaluation e.V. (DeGEval) (2002): Standards für Evaluation. Köln: DeGEval.
[10] DAC Evaluation Network (2006): DAC Evaluation Quality Standards. OECD.
[11] UNEG (2005): Standards for Evaluation in the UN System. United Nations.

5.1 Methodische und Organisatorische Vorbereitung

Eine solide methodische und organisatorische Vorbereitung der Wirkungsevaluation ist eine Grundvoraussetzung für eine qualitativ hochwertige Evaluation. Die Vorbereitungsphase setzt den Grundbaustein für die Durchführung der Evaluation, ihr kommt daher eine große Bedeutung zu. Im Rahmen der Vorbereitung wird das tatsächlich mögliche Evaluationsdesign spezifiziert was letztendlich den Datenerhebungsprozess und die Datenanalysemöglichkeiten und somit auch die Berichterstattung maßgeblich beeinflusst.

Die einzelnen Schritte im Rahmen der methodischen und organisatorischen Vorbereitung beinhalten:

(1) Identifikation des Evaluationsgegenstands (Analyse der Konzeption, Interventionslogik und der Durchführung der Entwicklungsmaßnahme).

(2) Konstruktion des Wirkungsmodells.

(3) Festlegung eines Untersuchungsdesigns.

(4) Identifikation der Ressourcepersonen und Datenerhebungsmethoden.

(5) Entwicklung der Datenerhebungsinstrumente.

(6) Planung des Vor-Ort-Aufenthalts.

Die Auflistung deutet bereits an, dass die Vorbereitung faktisch fast ebenso viele zeitliche Ressourcen beansprucht wie die eigentliche Datenerhebung d.h. Durchführung vor Ort. Im Folgenden werden für jeden einzelnen Schritt die notwendigen Aufgaben und die zeitlichen Ressourcen sowie Empfehlungen zur Umsetzung skizziert. Zu beachten ist dabei, dass die Schritte sowie Aufgaben nicht zwangsweise chronologisch erfolgen, sondern in einander übergehen.

(1) Identifikation des Evaluationsgegenstands (Analyse der Konzeption, Interventionslogik und der Durchführung der Entwicklungsmaßnahme)

Die Analyse der Konzeption, Interventionslogik, Kontext- und Rahmenbedingungen, der Durchführung der Entwicklungsmaßnahme sowie der Leistungsfähigkeit der Durchführungsorganisationen ist der erste Schritt im Vorbereitungsprozess der Evaluation. Die Informationen hierzu stehen in den GTZ-spezifischen Dokumenten der Entwicklungsmaßnahme (Prüfberichte, Angebot, PFK, PFB, SB, e-Val-Bericht):

- Die Konzeption beschreibt die zu evaluierende Entwicklungsmaßnahme – somit den Evaluationsgegenstand - deren Umsetzungsstrategien, die Durchführungsinstitution, Mittler, Zielgruppe, Standorte, Rahmenbedingungen etc.
- Die Interventionslogik in Form der Wirkungskette beinhaltet wichtige Leistungen, Ziele, Wirkungen (direkt und indirekt), Systemgrenzen, Kontext- und Rahmenbedingungen und Risiken der Entwicklungsmaßnahme.
- Die Unterlagen zum Durchführungsverlauf der Entwicklungsmaßnahme geben Aufschluss über durchgeführte Steuerungsmaßnahmen, bereits identifizierte Wirkungen, Konfliktbereiche, Risiken, Kontext- und Rahmenbedingungen und die Leistungsfähigkeit der Durchführungsorganisationen.

- Der e-VAL-Bericht gibt Hinweise für weitere Fragestellungen die im Rahmen der Wirkungsbetrachtung notwendig sind.

(1) Identifikation des Evaluationsgegenstands (Analyse der Konzeption, Interventionslogik und der Durchführung der Entwicklungsmaßnahme)		
Aufgaben & Tageverteilung nach Gutachter[12]	**Int.**	**Nat.**
☑ Sichten und Analysieren der GTZ Dokumentation der Entwicklungsmaßnahme	ca.[13] 4	ca. 1
Summe	**ca. 4**	**ca.1**

(2) Konstruktion des Wirkungsmodells

Ausgangslage zur Konstruktion des Wirkungsmodells stellt die Analyse des Evaluationsgegenstandes, der Wirkungskette der jeweiligen Entwicklungsmaßnahme und externer Wirkungsfelder dar. Erste Gespräche mit der Regionalgruppe und Fachgruppe in der GTZ geben dabei zusätzlich Aufschluss über Risiken, intendierte und nicht-intendierte positive wie negative und externe Wirkungsfelder der Entwicklungsmaßnahme. Basierend auf dieser ersten Analyse können die bereits für die Konzeption entwickelte Wirkungskette überprüft und Wirkungshypothesen aufgestellt werden. Dies geschieht unter Hinzuziehung wissenschaftlicher Studien zu Wirkungen im relevanten Sektor (hier der Wassersektor) und weiterer Informationen zum Länderspezifischen Kontext (Länder- und Sektorspezifika, Systemgrenzen) um überprüfbare, theoriegeleitete Wirkungshypothesen (ein Wirkungsmodell mit positiven, negativen, geplanten und ungeplanten Wirkungen) als Grundlage der Wirkungsmessung zu haben, die Hinweise geben können zu den für Wirkungsevaluationen relevanten Fragestellungen: „Warum wirkt eine Maßnahme?“ und „Welche unintendierten Wirkungen zeigen sich wie und warum?“ (vgl. Kap. 3.1).

Ebenfalls im Wirkungsmodell zu berücksichtigen sind Hypothesen zur Organisation (Durchführungsparter und Trägerorganisation), die wie in Kap. 2 dargestellt als Transmitter von Entwicklungsmaßnahmen verursachende Variablen für den Erfolg und die Wirkungen der Maßnahme bilden. Faktoren wie interne Zielakzeptanz, Personalstruktur und Qualifikationen, Organisationsstruktur, finanzielle Ressourcen, und technische Infrastruktur spielen bei der Wirkungserreichung eine maßgebliche Rolle. Zur Wirkungsprüfung müssen deshalb nicht nur fachspezifische Hypothesen formuliert, sondern auch die genannten organisationsinternen Rahmenbedingungen angemessen im Wirkungsmodell berücksichtigt werden.

In der Durchführungsphase sollte das aufgestellte Wirkungsmodell im Rahmen eines Auftaktworkshops ebenfalls mit den Hauptstakeholdern diskutiert werden.

Die einzelnen Aufgaben und die benötigten zeitlichen Ressourcen aufgeteilt nach Verantwortlichkeit[14] sind die Folgenden:

[12] Int. steht für internationaler Gutachter; Nat. für nationaler Gutachter; Messeinheiten sind Tage.
[13] Die angegebenen Tage sind Richtwerte, die je nach Evaluationsgegenstand angepasst werden müssen.
[14] Die hier angegebenen zeitlichen Ressourcen variieren je nach Komplexität der zu evaluierenden Entwicklungsmaßnahme. Hier können lediglich Richtwerte angegeben werden, die im Rahmen der jeweiligen spezifischen Evaluation angepasst werden müssen.

(2) Konstruktion des Wirkungsmodells		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Gespräche mit zuständiger Regionalgruppe und Fachgruppe zum Wirkungsmodell (Rahmenbedingungen, Risiken, Wirkungen)	ca. 1	
☑ Sichten und Analysieren relevanter wissenschaftlicher Studien und Kontextinformationen zu Wirkungen im Wassersektor	ca. 10	
☑ Überarbeitung des Wirkungsmodells der Entwicklungsmaßnahme, Aufstellung theoriegeleiteter, überprüfbarer Wirkungshypothesen	ca. 2	
Summe	**ca. 3** (+10*)	

(3) Festlegung des Untersuchungsdesigns

Im zweiten Schritt ist ein den beschriebenen methodischen Anforderungen entsprechendes Untersuchungsdesign zu entwickeln. Ziel ist es dabei, unter Berücksichtigung der Gegebenheiten vor Ort und den entworfenen Wirkungsmodellen sowie der Interventionslogik (soweit vorhanden bzw. im Vorfeld erarbeitet), ein Untersuchungsdesign zu erstellen, welches die Kontrolle von Störeinflüssen optimal ermöglicht und eine Überprüfung der Wirkungshypothesen bei der Ziel- und mindestens einer Vergleichsgruppe[15] gewährleistet (vgl. Kap. 3.2). Für die drei zu evaluierenden Entwicklungsmaßnahmen im Wassersektor muss im Rahmen des Erstellens der Inception Reports noch geprüft werden, welche Methoden umzusetzen sind. Möglichkeiten die in Erwägung gezogen werden sollten sind:

- Konstruierte Kontrollen (Matching on Observables) und Untersuchung nach der Single-Difference Methode oder aber Double-Difference Methode bei vorhandenen Baselinedaten oder kontinuierlich erhobenen Daten im Rahmen des Monitorings. Ob solche Baseline- oder Monitoringdaten vorhanden sind, und vor allem die Qualität der Daten muss für die drei Wirkungsevaluationen im Wassersektor vorab noch überprüft werden.
- Statistische Kontrollen und Untersuchung nach der Double-Difference Methode, d.h. Bildung einer konstruierten Vergleichsgruppe durch "Propensity Score Matching" (PSM). Diese Methode ist im Rahmen der drei Wirkungsevaluation jedoch nur anwendbar, wenn auf solide Daten allgemeiner Bevölkerungsumfragen oder sonstiger Daten (national surveys, ausführliche Monitoringdaten etc.), die zu Beginn der Maßnahme erhoben wurden und die interessierenden Fragen bzw. Angaben enthalten, zurückgegriffen werden kann.
- Reflexive Kontrollen, speziell das *"Multiple Comparison Group Design"* wobei verschiedene Zielgruppen (Teilnehmer/innen-Gruppen), die unterschiedliche Eigenschaften aufweisen und an z.B. zeitversetzten Maßnahmen teilgenommen haben, *un-*

[15] Ein Kontrollgruppenansatz ist nicht anwendbar, da die Kontrollgruppe bereits bei Start der Entwicklungsmaßnahme gebildet werden muss.

tereinander als Vergleichsgruppe verglichen werden.

Eher zu vernachlässigende Ansätze sind generische Kontrollen und Schattenkontrollen da die Qualität der Wirkungszuschreibung ungenügend ist.

Bei der Bildung der Vergleichsgruppen müssen weitere Störeffekte bedacht werden.

- So muss grundsätzlich überprüft werden, ob bzw. welche (vergleichbaren) Maßnahmen anderer Geber im Entwicklungsmaßnahmengebiet selbst aber auch im Umfeld der Vergleichsgruppe stattfanden.
- Auch müssen die beobachtbaren, für die Wirkungen relevanten Eigenschaften (“observables“) der Zielgruppe sorgfältig erarbeitet und bei der Bildung der Vergleichsgruppe berücksichtigt werden. Häufig entsprechen die sogenannten unbeobachtbaren Eigenschaften (“die unobservabes“) lediglich den unbeobachteten, d.h. den nicht berücksichtigten Eigenschaften.
- Ein weiterer Störfaktor bei der Bildung von Vergleichsgruppen der immer überprüft werden sollte ist die sogenannte Auswahlverzerrung (Selektionsbias)[16].
- Übertragungseffekte sind ebenfalls zu beachten. Diese 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[17].

Durchzuführende Aufgaben und die benötigten zeitlichen Ressourcen sind die Folgenden:

(3) Festlegung eines möglichen Datenerhebungsdesigns		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Festlegung des Datenerhebungsdesigns	ca. 1	
☑ Identifizierung einer Vergleichs(Kontroll)gruppe zur Befragung und für Standortbesuche (Beobachtung)	ca. 2	ca. 1
Summe	**ca. 3**	**ca.1**

(4) Identifikation der Ressourcepersonen und Datenerhebungsmethoden

Um eine reliable und valide Datensammlung zu ermöglichen, müssen die möglichen Datenerhebungsmethoden basierend auf der Fragestellung und den zu befragenden Ressourcepersonen spezifiziert werden. D.h. die anzuwendenden Datenerhebungsmethoden werden durch die zu erhebenden Informationen begründet, die sich wiederum aus der Analyse des

[16] So wird die Zielgruppe einer Maßnahme oft 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, also 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. Soweit die Auswahlkriterien (z.B. bestimmte Voraussetzungen für die Teilnahme) beobachtbar sind (“observables“). 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.

[17] 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.

Wirkungsmodells, den aufgestellten, zu überprüfenden Wirkungshypothesen in Schritt 1 und den Evaluierungskriterien nach DAC (vgl. GTZ Anleitung zur Erfolgsbewertung) sowie den weiteren Evaluationsfragestellungen ergeben (vgl. Kap. 3). Zu beachten ist dabei, dass sich die zu wählende Datenerhebungsmethode aus der Fragestellung ergibt und nicht umgekehrt.

Aufgrund des Multi-Methoden-Anspruchs (vgl. Kap. 2.4), der besonders bei Wirkungsevaluationen gilt, sollten quantitative so wie auch qualitative Methoden zur Datenerhebung genutzt werden. Die ausgewählten Datenerhebungsmethoden sollten die Informationssammlung zu den besonderen unter Wirkungsbetrachtung relevanten Fragestellungen „Welche Wirkungen sind entstanden?“ und „Wie und warum sind diese Wirkungen entstanden?“, ermöglichen.

Folgende Datenerhebungsmethoden sind für die Wirkungsevaluation der drei Entwicklungsmaßnahmen im Wassersektor möglich:

- *Dokumenten- und Aktenanalyse der GTZ Dokumente und weiterer wissenschaftlicher Studien* zur Aufstellung des Wirkungsmodells (bereits unter Schritt 1 dargestellt).
- *Dokumenten- und Aktenanalyse* weiterer verfügbarer Unterlagen zum Sektor, zum Land (Region) und der Entwicklungsmaßnahme hinsichtlich Rahmenbedingungen, Wirkungen und Risiken.
- *Sekundärdatenanalyse* amtlicher statistischer Daten, Daten von Ministerien des Partnerlandes, Regionale Daten etc. insbesondere um die wirtschaftlichen und sozioökonomischen Wirkungen zu ermitteln.
- *Analyse von Baselinedaten der Ziel- & Vergleichs(Kontroll)gruppe* um Informationen zur Situation vor Maßnahmenbeginn zu haben.
- *Gruppendiskussion* mit VertreterInnen verschiedener Stakeholdergruppen zur gemeinsamen Bewertung der Zielerreichung und Wirkungserreichung und weiterer Wirkungshypothesen, besonders zur Frage der Kausalität der Wirkungen.
- *Gruppendiskussion* mit VertreterInnen der Ziel- & Vergleichs(Kontroll)gruppe zur gemeinsamen Bewertung der Zielerreichung und Wirkungserreichung und weiterer Wirkungshypothesen, besonders zur Frage der Kausalität der Wirkungen.
- *Leitfadengestützte Intensivinterviews* mit an der Entwicklungsmaßnahme beteiligten Personen/MitarbeiterInnen auf unterschiedlichen Ebenen, MitarbeiterInnen beim Träger und Zielgruppenvertretern zu Fragen des Verlaufs und Wirkungen der Entwicklungsmaßnahme.
- *Standardisierte Datenerhebung bei Ziel- & Vergleichs(Kontroll)gruppe* zu Fragen des Verlaufs und Wirkungen der Entwicklungsmaßnahme.
- *Besuche von Standorten,* in denen Maßnahmen stattgefunden haben mit *systematischer Beobachtung* und Dokumentation der Nachhaltigkeit der einzelnen Maßnahmen und Identifizierung der Wirkungen.

Aus der geschilderten Vorgehensweise bei der Identifizierung möglicher Ressourcepersonen und Methoden zur Datenerhebung ergeben sich folgende Aufgaben und benötigte zeitliche Ressourcen zur Bearbeitung:

(4) Identifikation der Ressourcepersonen und Datenerhebungsmethoden		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Identifizierung der an der Entwicklungsmaßnahme beteiligten Personen/MitarbeiterInnen/Zielgruppenvertreter und weiterer Stakeholdergruppen	ca. 1	ca. 1
☑ Identifizierung möglicher Datenerhebungsmethoden basierend auf den Evaluationsfragestellungen	ca. 1	
☑ Identifizierung statistischer Sekundärdaten und Überprüfung der Qualität und Beinhaltung interessierender Fragestellungen.	ca. 1-3	ca. 1
☑ Identifizierung möglicher Baselinedaten der Ziel- & Vergleichs(Kontroll)gruppe und Überprüfung der Qualität zur Bestimmung der Messzeitpunkte und der letztendlich möglichen Erhebungsdesigns[18]. Monitoringdaten der Maßnahmen so wie auch landesstatistische Daten sind hier von Relevanz. Hier ist ebenfalls zu prüfen, ob die interessierenden Fragestellungen in den Daten enthalten sind.	ca. 1-4	ca. 1-2
Summe	**ca. 4-9**	**3-4**

(5) Entwicklung der Datenerhebungsinstrumente

Die Evaluationsfragestellungen, denen im Rahmen der Wirkungsevaluation laut GTZ Vorgabe nachgegangen werden muss, beziehen sich auf die Überprüfung:

(1) der OECD-DAC Evaluierungskriterien Relevanz, Effektivität, „Impact“, Effizienz und Nachhaltigkeit anhand der Anleitung zur Erfolgsbewertung (GTZ 2008),

(2) des Beitrags der Entwicklungsmaßnahme zur Armutsminderung und den Millenniumsentwicklungszielen,

(3) der Förderung der Gleichberechtigung der Geschlechter,

(4) der Förderung nachhaltiger Entwicklung,

(5) fachbezogener Fragestellungen (vgl. Kap. 4).

Die Evaluationsfragestellungen überprüfen im Rahmen der Abhandlung der Kriterien „Effektivität“, „Impact“ und „Nachhaltigkeit“ sowie „Beitrag zur Armutsminderung“ und den „MDGs“ intensiv die Wirkungshypothesen.

[18] Bezüglich des *Zeitpunkts* der Datenerhebungen (Baseline und nachher Messung) muss 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. Ferner stellen *Veränderung im Messinstrument* ebenfalls Störfaktoren dar, die sich in der falschen Einschätzung der Wirkung zeigt. Durch die Verwendung von unterschiedlichen Messinstrumenten, z.B. anders formulierte Fragen, ungleiche Antwortvorgaben o.ä. können Effekte möglich werden, die allerdings nichts mit der Wirksamkeit der Entwicklungsmaßnahme zu tun haben.

Informationen zu diesen Evaluations- und Wirkungsfragestellungen müssen mit den unter Punkt 4 dargestellten möglichen Datenerhebungsmethoden gesammelt werden. Hierfür werden adäquate Datenerhebungsinstrumente benötigt, unter anderem:

- ein *Analyseleitfaden* zur Strukturierung der Evaluationsfragestellungen und des gesamten Datenerhebungsprozesses der ebenfalls als Grundlage für die Dokumenten- und Aktenanalyse genutzt werden kann.
- ein *Leitfaden* zur Gruppendiskussion mit VertreterInnen verschiedener Stakeholdergruppen und der Ziel- und Vergleichs(Kontroll)gruppe.
- ein *Leitfadeninterview* für Intensivinterviews mit an der Entwicklungsmaßnahme beteiligten Personen/MitarbeiterInnen auf unterschiedlichen Ebenen, MitarbeiterInnen beim Träger und Zielgruppenvertretern.
- Je ein *standardisierter Fragebogen* zur standardisierten Befragung der Ziel- und Vergleichs(Kontroll)gruppe.
- Ein *Beobachtungsplan* für die Standortbesuche.

Folgende zeitliche Ressourcen sind für die Entwicklung der Erhebungsinstrumente anzusetzen:

(5) Entwicklung der Datenerhebungsinstrumente		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Analyseleitfaden und Analyseraster	ca. 1	
☑ Leitfaden zur Gruppendiskussion	ca. 1	
☑ Leitfadeninterview	ca. 1	
☑ Standardisierter Fragebogen	ca. 2	ca. 1
☑ Beobachtungsplan	ca. 1	ca. 1
Summe	**ca. 6**	**2**

(6) Planung des Vor-Ort Aufenthalts

Als letzter Schritt im Rahmen der Vorbereitung der Evaluation muss der Vor-Ort Aufenthalt und die letztendliche Vor-Ort Durchführung der Evaluation geplant werden. Hierzu ist festzuhalten, dass die gesamte Evaluation gemeinsam von einem internationalen und einen nationalen Gutachter durchgeführt wird. D.h. viele der hier beschriebenen logistischen und organisatorischen Aufgaben müssen schwerpunktmäßig durch den nationalen Gutachter bewerkstelligt werden, der sich mit den Gegebenheiten in der Region der Entwicklungsmaßnahme bestens auskennt. Hierzu muss aber zunächst im Rahmen der Vorbereitungsphase ein geeigneter nationaler Gutachter gefunden werden, der dem Evaluationsteam beitreten kann. Erst wenn dieser vorhanden ist, können die logistische und organisatorische Planung sowie auch bereits oben geschilderte Aufgaben wie die Identifizierung der Zielgruppe und Vergleichs(Kontroll)gruppe und der Standorte für Projektbesuche in Angriff genommen werden.

- Zur logistischen Planung gehören Punkte wie Reiseplanung, Suche einer Unterkunft, Auswählen von Transportmitteln.
- Die organisatorische Planung umfasst die konkrete Kontaktaufnahme mit bereits ausgewählten Ressourcepersonen (für Interviews, Gruppeninterviews, Briefing, Debriefing) und die Vereinbarung von Interviewterminen.
- Ebenso muss die standardisierte Befragung logistisch geplant und organisiert werden. Zur Durchführung der standardisierten Befragung sollte dann eine lokale Institution, die über die notwendigen personellen Ressourcen verfügt (Interviewer) und das Know-How besitzt, engagiert werden.

Die empfohlenen zeitlichen Ressourcen zur Bewerkstelligung dieser Aufgaben sind die folgenden:

(6) Planung des Vor-Ort-Aufenthalts		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Suche und unter Vertragnahme des nationalen Gutachters	ca. 2	
☑ Logistische Reiseplanung (Flug, Hotel, Transport)	ca. 1	ca. 1
☑ Organisatorische Planung (Interviewtermine)		ca. 4
☑ Logistische und organisatorische Planung der standardisierten Befragung (Identifikation einer Institution)	ca. 3	ca. 2
Summe	**ca. 5**	**ca. 7**

Wie verdeutlicht, sind im Rahmen der Vorbereitung einer methodisch anspruchsvollen Wirkungsevaluation zahlreiche Aufgaben zu bewerkstelligen. Daraus resultiert ebenfalls ein dementsprechender zeitlicher Aufwand der veranschlagt werden muss, um eine qualitativ anspruchsvolle, reliable und valide Evaluationsdurchführung zu garantieren. Dies wird im Zeitkontingent der bisherigen Unabhängigen Evaluationen leider nicht entsprechend berücksichtigt. Dort stehen dem internationalen Gutachter lediglich fünf Tage Vorbereitung und dem lokalen Gutachter sieben Tage zu, in der Summe 12 Tage. Wie aufgezeigt, werden zur adäquaten Umsetzung der Vorbereitung jedoch ca. 25 Tage für den internationalen Gutachter und 14 Tage für den lokalen Gutachter benötigt, je nach Komplexität des Evaluandums. Eine deutliche Diskrepanz zum bisherigen Kontingent der Unabhängigen Evaluationen, die sich u.a. in der Qualität der Vorbereitung und Durchführung der Datenerhebung vor Ort niederschlägt.

5.2 Durchführung

Die Durchführungsphase der Evaluation ist primär gekennzeichnet durch die Datenerhebung vor Ort und bereits erste Auswertungen. Die in der Vorbereitungsphase identifizierten benötigten Informationen werden mit den bereits entwickelten Datenerhebungsinstrumenten erhoben.

Insgesamt sollten folgende Datenerhebungen und erste Auswertungen stattfinden, die bereits in Kap. 5.1 unter Punkt 4 thematisiert wurden:

(7a) Datenerhebung & Auswertung vor Ort durch das Gutachterteam

- *Identifikation von Dokumenten und Akten* sowie weiterer verfügbarer Unterlagen zum Sektor, zum Land (Region) und der Entwicklungsmaßnahme und Analyse relevanter Informationen.
- *Identifikation weiterer Sekundärdaten & Baselinedaten* z.B. amtliche Statistiken, Statistiken zuständiger Ministerien, regionale Daten etc..
- *Gruppendiskussion* mit VertreterInnen verschiedener Stakeholdergruppen und mit VertreterInnen der Ziel- & Vergleichs(Kontroll)gruppe zur Wirkungsüberprüfung sowie Auswertung der Ergebnisse. Diese Gruppendiskussion sollte in ein Briefing und/oder Debriefing integriert werden.
- *Leitfadengestützte Intensivinterviews* mit an der Entwicklungsmaßnahme beteiligten Personen/MitarbeiterInnen auf unterschiedlichen Ebenen, MitarbeiterInnen beim Träger und Zielgruppenvertretern zur Überprüfung der Evaluations- und Wirkungsfragestellungen. Verschriftung der Interviews und erste Analyse der Ergebnisse.
- *Beobachtung* an Projektstandorten und Auswertung der Ergebnisse

(7b) Datenerhebung & Datenmanagement durch eine extern beauftragte Institution unter Anleitung des Gutachterteams (oder eines weiteren Gutachters[19]):

- *Standardisierte Datenerhebung bei Ziel- & Vergleichs(Kontroll)gruppe* in Anlehnung an bereits existierende Befragungen um Vergleichbarkeit der Daten und Wirkungsidentifikation zu ermöglichen. (Falls noch keine Zielgruppenbefragungen durchgeführt wurden, sollte bereits in der Vorbereitungsphase eine neue Befragung konzipiert werden mit Retrospektivfragen).
- *Datenmanagement* der standardisierten Befragung, d.h. die Informationen aus den Interviews sollten bereits vor Ort in ein elektronisches Datentableau zur Weiterverarbeitung in Deutschland eingetragen werden.

Hier gilt es darauf zu achten, dass die beauftragte Institution vor Durchführung der standardisierten Befragung eine Schulung erhält um sicherzustellen, dass die Datenerhebung reliabel und valide durchgeführt wird. Ebenfalls ist es empfehlenswert, die Datenerhebung und ggf. den Auswertungsprozess zu Überwachen, denn die Datenqualität ist entscheidend für die Interpretation der Wirkungen. Wenn hier Störeffekte auftreten, wäre im schlimmsten Fall die gesamte standardisierte Befragung zwecklos. In der Schulung muss der Fragebogen besprochen werden, die Vorgehensweise bei der Befragung, sowie auch die Datensammlung und ggf. erste elektronische Auswertungen falls benötigte technische Ressourcen zur Verfügung stehen.

Folgende Aufgaben und zeitliche Ressourcen ergeben sich somit für die Datenerhebung und erste Auswertungen vor Ort:

[19] Ein weiterer Gutachter, der nur für die Koordination der standardisierten Befragung (Schulung, Beobachtung, Analyse) zuständig ist, würde die Effizienz der Evaluation, besonders die Gesamtdauer des Vor-Ort-Aufenthalts erheblich verkürzen.

(7) Datenerhebung & Auswertung vor Ort		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Dokumenten- und Aktenanalyse, Analyse weiterer Sekundärdaten & Baselinedaten der Ziel- & Vergleichs(Kontroll)gruppe	ca. 1-4	ca. 1-4
☑ Durchführung und Auswertung der Gruppendiskussionen	ca. 3	ca. 3
☑ Standortbesuche, Durchführung und Auswertung von Leitfadeninterviews	ca. 15	ca. 15
☑ Koordination der Durchführung der standardisierten Befragung und Datenmanagement	ca. 12	
Summe	**ca. 31 -34**	**ca. 19 -22**

Für die Durchführung der Wirkungsevaluation vor Ort, also konkret die Datensammlungsphase, sind bei Integration einer standardisierten Befragung ebenfalls mehr Tage anzusetzen, als das bei Fremdevaluierungen bisher möglich war. Insgesamt werden für die Durchführung vor Ort ca. 31 Tage veranschlagt. Um die Aufenthaltsdauer vor Ort zu verkürzen, empfiehlt es sich jedoch einen weiteren Gutachter mit der Koordination der Durchführung der standardisierten Befragung zu betrauen. Zusätzliche Kosten müssen noch für die standardisierte Befragung, die an ein externes Institut vergeben werden muss, kalkuliert werden. Bisher waren 19 Tage je Gutachter vorgesehen. Für methodisch anspruchsvolle Evaluationen ist das Zeitkontingent demnach zu erhöhen.

5.3 Datenanalyse und Reporting

Die im Rahmen der Datenerhebung erhobenen Daten werden in der letzten Phase der Evaluation ausgewertet und unter Berücksichtigung des Wirkungsmodells zusammenfassend interpretiert. Dabei sollen insbesondere die für GTZ Evaluationen grundlegenden DAC-Bewertungskriterien nicht nur als Leitfragen zur Erfolgsbewertung betrachtet werden, sondern in das Pfadmodell der Ursache-Wirkungshypothesen eingebaut werden (vgl. Caspari 2004 S. 220ff.):

- *Relevanz* ist in erster Linie eine Bewertung des *Zielsystems* in der *Planungsphase* bzw. der eingeführten *Innovation* in Hinblick auf ihre entwicklungspolitische Bedeutung 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 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.

(8) Datenanalyse & Reporting

Bei der Bewertung und Zuschreibung der Wirkungen zu der Entwicklungsmaßnahme wird der Trichteransatz gewählt (vgl. Kap. 3.4). Zunächst werden die Veränderungen im Umfeld der Entwicklungsmaßnahme bei der Zielgruppe und der Vergleichsgruppe identifiziert. Im nächsten Schritt wird geprüft, welche dieser Veränderungen (Wirkungen) kausal der Entwicklungsmaßnahme zuzuordnen sind. Dies erfolgt durch den Abgleich mit dem Wirkungsmodell der Entwicklungsmaßnahme und durch den Abgleich zwischen Ziel- und Vergleichsgruppe. Positive und negative Wirkungen sind dabei zu berücksichtigen. Als Ergebnis des Abgleichs wird ersichtlich, welche Wirkungen indentiert waren, welche nicht-intendierten Wirkungen aufgetreten sind – der Entwicklungsmaßnahme aber kausal zuzuordnen sind – und somit welche Ziele der Entwicklungsmaßnahme erreicht wurden. Bei dieser Analyse und besonders der Interpretation der Daten sind mögliche Störeffekte, die die Wirkungsmessungen beeinflussen, zu berücksichtigen. Bei Daten die zu mehreren Zeitpunkten erhoben wurden, ist besonders auf das sogenannte „*zwischenzeitliche Geschehen*“ und die sogenannten *Reifungsprozesse* zu achten.

Anzustreben ist eine statistisch anspruchsvollere Datenauswertung, wie bereits in Kap. 3.4 angeführt. Aus diesem Grund sollte das Zeitkontingent für Datenanalyse und Reporting dementsprechend vergrößert werden.

Aufgaben im Rahmen der Datenanalyse und des Reportings sind:

(8) Datenanalyse & Reporting		
Aufgaben & Tageverteilung nach Gutachter	**Int.**	**Nat.**
☑ Analyse der Qualitativen Daten	ca. 4	
☑ Datenmanagement	ca. 3-5	
☑ Analyse der Quantitativen Daten &	ca. 10	
☑ Ggf. Anwendung von Matchingverfahren bei nicht zugrundeliegenden Baselinedaten der Ziel- & Vergleichsgruppe	ca. 0-5	
☑ Berichterstellung	ca. 15-17	ca. 4
Summe	**ca. 32 41**	**ca. 4**

Im Rahmen der bisherigen Unabhängigen Evaluationen wurden dem internationalen Gutachter 12 Tage und dem nationalen Gutachter 4 Tage für Reporting gewährt. Da für Wirkungsevaluationen, bei Verwendung der standardisierten Befragung, ein Mehraufwand bei der

Analyse der quantitativen Daten entsteht, muss dies dementsprechend auch im Zeitkontingent der Berichterstellung beachtet werden.

Schlussbetrachtung

Mit diesem Vorschlag für die Konzeption von Wirkungsevaluationen wird eine erste Grundlage zur Umsetzung der gestiegenen Ansprüche an Wirkungsmessung im Rahmen von Unabhängigen Evaluierungen in der GTZ gelegt. Die vorgestellte Umsetzung versucht sowohl die methodischen Anforderungen als auch die gegebenen Möglichkeiten in Einklang zu bringen. Neben der Bewertung der direkten Wirkungen wird den indirekten und hoch aggregierten Wirkungen ebenfalls die notwendige Aufmerksamkeit gewidmet. Dies wird primär durch die theoretische Ursache-Wirkungsbetrachtung, die Anwendung adäquater Forschungsdesigns, den Multi-Methoden-Ansatz und die anspruchsvolle Datenauswertung erreicht.

Die Ausführungen zu Forschungsdesigns machen jedoch besonders deutlich, dass in Zukunft bereits bei der Planung von Entwicklungsmaßnahmen eine mögliche Wirkungsbetrachtung Berücksichtigung finden sollte. So ist schon in der Planungsphase eine Baselinestudie mit Ziel- und Vergleichsgruppen durchzuführen, auf die dann im Rahmen der Wirkungsevaluation zurückgegriffen werden kann. Ebenfalls sollten die möglichen Untersuchungsdesigns für Wirkungsmessung in der Projektdurchführung und im Rahmen des Monitorings angemessen berücksichtigt werden. Dies trägt dazu bei, dass die Wirkungsbetrachtung und somit auch Wirkungsevaluationen zum integralen Bestandteil von Entwicklungsmaßnahmen werden, und der Frage „Was bewirkt die Entwicklungszusammenarbeit“ besser als bisher nachgegangen werden kann.

6. 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), URL: http://www.odi.org.uk/pppg/publications/papers_reports/gov/ug_ws01/bp.pdf - 10/03.

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, Michael (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.

BMZ (2006): Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit.

Borrmann, Axel u.a. (1999): Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit. Baden Baden: Nomos.

Borrmann, Axel u.a. (2001): Refom der Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit. Baden Baden: Nomos.

Caspari, Alexandra (2004): Evaluation der Nachhaltigkeit von Entwicklungszusammenarbeit – Zur Notwendigkeit angemessener Konzepte und Methoden. Wiesbaden.

Caspari, Alexandra u.a. (2000): Langfristige Wirkungen deutscher Entwicklungszusammenarbeit und ihre Erfolgsbedingungen. Eine Ex-post-Evaluierung von 32 abgeschlossenen Projekten“ (BMZ-Spezial Nr. 19; Bonn.

Caspari, Alexandra. u. Barbu, Ragnhild (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.

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

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

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

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.

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.

Flick, Uwe (2007): Qualitative Sozialforschung. Rowohlt Verlag.

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.

GTZ (2004): Leitfaden zum wirkungsorientierten Monitoring.

GTZ (2006): Handlungsorientierter Auftragsrahmen (AURA).

GTZ (2007): Anleitung zur Erfolgsbewertung.

GTZ (2005): Schluss- und Ex-post Evaluierungen in der GTZ - Konzeptpapier -

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.

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

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.

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

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.

Meyer, Wolfgang (2007): Datenerhebung : Befragungen -Beobachtungen - Nicht-reaktive Verfahren. In: Handbuch zur Evaluation : eine praktische Handlungsanleitung / Reinhard Stockmann (Hrsg.) - Münster [u.a.] : Waxmann.

Meyer, Wolfgang (2007a): Vertical Dimension of Social Integrations. Grass-root Activities for Managing Sustainability. In: Rali, R.N.R: Schwarz-Herion, O. (Ed.): Sustainable Development: Issues and Perspecitves. New Delhi: Printworld, S. 99 – 123.

Michaelowa, Katharina; Borrmann, Alex (2005): Wer evaluiert was, wie und warum? In: H. Ahrens (Hrsg.): Zur Bewertung der Entwicklungszusammenarbeit, Schriften des Vereins für Socialpolitik, Berlin.

Michaelowa, Katharina; Borrmann, Alex (2005): What Determines Evaluation Outcomes? Evidence from Bi- and Multilateral Development Cooperation. HAAW Discussion Paper 310

OECD (2006): Paris Declaration on Aid Effectiveness. (German Translation) http://www.oecd.org/dataoecd/37/39/35023537.pdf

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

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

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

OECD-DAC: DAC Criteria for Evaluating Development Assistance.

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.

Stockmann, Reinhard (1992): Die Nachhaltigkeit von Entwicklungsprojekten – Eine Methode zur Evaluierung am Beispiel von Berufsbildungsprojekten. Opladen.

Stockmann, Reinhard (1996): Die Wirksamkeit der Entwicklungshilfe – Eine Evaluation der Nachhaltigkeit von Programmen und Projekten. Opladen.

Stockmann, Reinhard (2000a): Evaluation staatlicher Entwicklungspolitik. In: Stockmann (Hg.): Evaluationsforschung. Opladen: Leske + Budrich

Stockmann, Reinhard (2006): Evaluation und Qualitätsentwicklung – eine Grundlage für wirkungsorientiertes Qualitätsmanagement. Münster.

Stockmann, Reinhard (Hg.) (2007): Handbuch zur Evaluation: Grundlagen und Praxis. Münster.

Stockmann, Reinhard; Meyer, Wolfgang; Krapp, Stefanie; Koehne, Gerhard (2000b): Wirksamkeit deutscher Berufsbildungszusammenarbeit. Ein Vergleich zwischen staatlicher und nicht-staatlicher Entwicklungszusammenarbeit mit der VR China. Wiesbaden.

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

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

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 (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.

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.