

10 Thesen zur Evaluation der Wirksamkeit der Entwicklungs- zusammenarbeit

Prof. Dr. Reinhard Stockmann

Stockmann, Prof. Dr. Reinhard:

10 Thesen zur Evaluation der Wirksamkeit der Entwicklungszusammenarbeit

Saarbrücken: Centrum für Evaluation, 2010.

(CEval-Arbeitspapiere; 18)

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>

10 Thesen zur Evaluation der Wirksamkeit der Entwicklungszusammenarbeit

Für die Evaluation der Wirksamkeit der Entwicklungszusammenarbeit lassen sich aus den bisherigen Erfahrungen und Studien einige Lehren für die Zukunft ziehen:

- (1) Jede Wirkungsevaluation sollte auf einem theorie-basierten Konzept aufbauen, in dem klar die verschiedenen Interventionen, die erwarteten Wirkungen und die Rahmenbedingungen analysiert werden. Anders ausgedrückt: Zu Beginn einer Evaluation sind klar Input, Output (die vom Programm erbrachten Leistungen), Outcome (die vom Programm erzielten intendierten Wirkungen) und Impact (alle beobachteten Veränderungen, also auch die, die von anderen Faktoren ausgelöst wurden sowie die nicht-intendierten Interventionswirkungen) analytisch voneinander zu trennen. Dabei kann die Erarbeitung eines LogFrame hilfreich sein. Entscheidend ist die Entwicklung von Ursache-Wirkungs-Hypothesen, die logisch aufeinander aufbauen (Wirkungsketten): wenn A dann B, wenn B, dann C etc.. Dabei besteht allerdings die Gefahr, die Realität allzu sehr zu vereinfachen. Für komplexe Programme reicht deshalb nicht eine simple Wirkungskette aus, sondern es muss vielmehr ein Wirkungsgefüge erstellt werden, das die Kontextbedingungen ausreichend berücksichtigt.
- (2) Die Auswahl des Untersuchungsdesigns hängt nicht nur von der Fragestellung ab, sondern auch von der Art des Programms und den Rahmenbedingungen, also z.B. ob ein experimentelles Design (mit Randomisierung), ein quasi-experimentelles Design oder andere Designs, die ebenfalls robuste Daten liefern können, wie z.B. Pannedesigns und Zeitreihenanalysen, einsetzbar sind. Auf jeden Fall erfordern eine Wirkungsmessung und ihre Ursachenzuschreibung wenigstens den Vergleich zwischen mindestens zwei Zeitpunkten (Vorher/Nachher) und zwischen Ziel- und Kontrollgruppen (randomisiert) bzw. Vergleichsgruppen. Dafür gibt es unterschiedliche Lösungen und keine kann per se für sich behaupten, der Königsweg oder der „Goldstandard“ (Patton) zu sein. Auch die zur Zeit hoch gehandelten randomisierten Kontrollgruppendesigns (Experiment) weisen eine Reihe von Problemen auf (z.B. geringe externe Validität), auf die hier nicht eingegangen werden soll (vgl. statt dessen Stockmann u. Meyer 2010: 191ff.).
- (3) Es herrscht mittlerweile unter den Fachleuten ein weitgehender Konsens, dass für eine Wirkungsevaluation eine breite, möglichst repräsentative Befragung der Zielgruppen (die in den Genuss von Programmmaßnahmen gekommen sind) und Vergleichsgruppen (die davon ausgeschlossen waren) unabdingbar notwendig ist (vgl. White 2006a: 20; Caspari u. Barbu 2008: 30, Böhnke u.a. 2009). Hierfür reichen einige Gespräche mit Zielgruppenvertretern oder die Durchführung einiger Gruppeninterviews oder Fokusgruppen nicht aus. Eine Vielfalt von Studien – auch zur deutschen Entwicklungszusammenarbeit (vgl. Stockmann 1992 u. 1996; Stockmann u.a. 2000; Reade 2008 u. 2009) – zeigen, dass eine standardisierte Befragung (Survey) sowohl von Ziel- und Vergleichsgruppen nicht nur durchführbar, sondern auch finanziell tragbar ist. Da Baseline-Daten häufig fehlen und deshalb mühsam retrospektiv (aus Dokumenten, Statistiken oder über Befragungen) rekonstruiert werden müssen, um den „Vorher“-Zustand beschreiben zu können, liegt hier eine zentrale Herausforderung für anspruchsvolle Wirkungsevaluationen. Es bedeutet nämlich, dass solche Impactstudien schon von Programmbeginn an mitgedacht werden müssen. Idealerweise müssen von Anfang an Daten für Indikatoren gesammelt werden, die dann Monate oder

gar Jahre später als Vergleichsgrößen für die Wirkungsmessung herangezogen werden. Solche Baselinestudien sind in der internationalen wie deutschen Entwicklungszusammenarbeit bisher Mangelware.

- (4) Entsprechend der vorangestellten Ausführungen zum „Methodenstreit“ ist zu konstatieren, dass in den meisten Fällen sowohl quantitative als auch qualitative Datenerhebungsinstrumente notwendig sind. Während quantitative Verfahren in der Regel für eine höhere Repräsentativität sorgen können, dienen die qualitativen Methoden dazu, Sachverhalten die nötige Tiefenschärfe zu verschaffen. Die Frage nach Art und Umfang von Wirkungen benötigt eine gewisse Repräsentativität. Die Frage danach, weshalb und auf welche Weise ein Programm Wirkungen hervorgebracht hat oder nicht, bedarf einer tiefergehenden Analyse, für die sich Intensivinterviews eindeutig besser eignen als standardisierte Fragebögen.
- (5) Wenn eine Programm-/Wirkungs-„theorie“ vorliegt, können bei ausreichender Datenqualität statistisch anspruchsvollere Auswertungsmethoden angewendet werden. Insbesondere mit Hilfe regressionsbasierter Ansätze lassen sich auf der Grundlage des „theoretischen“ Modells Zusammenhänge zwischen Maßnahme und Wirkungen sowie weiterer relevanter Einflussfaktoren statistisch überprüfen.
- (6) Wirkungsevaluationen sind Expertensache. Die theoretischen und methodischen Anforderungen an robuste Wirkungsevaluationen sind hoch und können nur von dafür ausgebildeten Fachleuten geplant und durchgeführt werden. Hinzu kommen Spezialkenntnisse, wie sie z.B. bei der Anwendung von Matching-Verfahren zur Konstruktion adäquater Vergleichsgruppen auf der Basis von Sekundärdaten notwendig sind. Da bei umfassenden Evaluationen, wie es Wirkungsevaluationen sind, nicht nur diese Qualifikationen erforderlich sind, sondern z.B. auch Fach-, Sektor-, Länder-, Sprachkenntnisse etc., ist es oft zweckmäßig, Evaluationsteams zu bilden, die sich komplementär ergänzen und Partnerkräfte vor Ort miteinbeziehen.
- (7) Noch nicht ausreichend thematisiert ist die Frage, wie im Kontext von Impact-Evaluationen die weiteren DAC-Kriterien (Relevanz, Effektivität, Effizienz und Nachhaltigkeit) untersucht werden können. In den vom CEval untersuchten NONIE-Studien wurden nur bei ganz wenigen neben den entwicklungspolitischen Wirkungen noch andere Faktoren (oder DAC-Kriterien) berücksichtigt. Die DAC-Kriterien wurden zwar bereits Anfang der 90er Jahre in den „DAC-Principles“ definiert, ohne jedoch zu erläutern, in welchem Zusammenhang sie zueinander stehen, was in der Regel dazu führt, dass sie in Evaluationsstudien relativ unabhängig voneinander abgehandelt werden. Caspari (2004: 220ff.) macht einen Vorschlag, wie die DAC-Kriterien in ein Pfadmodell für Ursache-Wirkungs-Hypothesen eingebaut werden können. Allerdings sollte dabei nicht außer Acht gelassen werden, dass Wirkungsevaluationen leicht überfrachtet werden können, wenn unter diesem Titel zu viele verschiedene Anforderungen gestellt werden. Hier ist noch weiterer Diskussionsbedarf zu konstatieren.
- (8) Die bisherigen Impact-Studien waren vor allem Programmevaluationen. Mit dem Aufkommen neuer EZ-Instrumente wird jedoch die Wirkungsevaluation vor neue oder zusätzliche Herausforderungen gestellt. Dies gilt bei Gemeinschaftsfinanzierungen (Budgethilfe, General Budget Support) oder vergleichbaren sektorweiten Ansätzen (Sector Wide Approaches = SWAp) zum Einen für die Entwicklung der Programmtheorie und ihren Ursache-Wirkungs-Hypothesen. Da diese neuen Instrumente die Be-

ziehung zwischen Geber- und Empfängerland neu definieren, und da Budgethilfe und sektorweite Ansätze die Partnerländer in die Lage versetzen sollen, ihre eigenen Ziele zu erreichen, ist genauer zu klären, was Gegenstand der Analyse sein soll, z.B. die Wirkungen des Ansatzes oder der Budgethilfe auf die (Sektor-)Politik des Empfängerlandes; die Wirkungen der veränderten (Sektor-)Politik des Partnerlandes bei den Zielgruppen; oder die Wirkungen auf die Haushaltspolitik des Partnerlandes (z.B. Substitutionseffekte).

Zum Anderen erfordert die Evaluation von Wirkungen des „Budget Support“ und SWAp eine besondere Methodik. Während es einen zunehmenden Konsens über Methoden und Techniken für die Wirkungsevaluation von Projekten und Programmen gibt, existieren bisher erst wenige Erfahrungen dazu, wie die Wirkungen dieser neuen Instrumente geprüft werden könnten. Ein grundsätzliches Problem wird darin gesehen, dass sich bei SWAp oder Budget Support schwerlich Vergleichsgruppen zur Abbildung des Kontrafaktischen bilden lassen (vgl. zusammenfassend Caspari 2008: 32f.).

- (9) Nicht jedes Projekt oder Programm muss einer Wirkungsanalyse unterzogen werden. Manchmal ist weniger mehr. Wenn nicht genügend finanzielle Ressourcen zur Verfügung stehen, sollten diese auf eher wenige, dafür qualitativ hochwertige Wirkungsevaluationen konzentriert werden. Entscheidend ist, dass EZ-Organisationen, staatliche wie private, solche Evaluationen in ihrem Evaluationsportfolio haben.

Aufgrund des hohen personellen, finanziellen und zeitlichen Ressourcenaufwands, den robuste Wirkungsevaluationen erforderlich machen, sollten diese sorgfältig ausgewählt werden. Dabei können verschiedene Überlegungen eine Rolle spielen.

- a) Ein zentrales Auswahlkriterium stellt das zu erwartende Lernpotenzial dar, also inwieweit angenommen wird, dass die Evaluationsergebnisse für zukünftige Vorhaben von Bedeutung sein werden. Dies ist dann der Fall, wenn eine „Policy“ oder ein Programm von strategischer Bedeutung für die Erreichung eines Entwicklungsziels (z.B. Armutsreduzierung) ist. Mit solchen Impact-Evaluationen kann man herausfinden, ob bzw. inwieweit ein Programm dazu beiträgt, diese Ziele zu erreichen, wie es verbessert werden kann etc..

Lernpotenzial ist auch dann gegeben, wenn ein neuer vielversprechender innovativer Entwicklungsansatz oder ein neues Förderinstrument auf ihre Wirksamkeit hin geprüft werden sollen, bevor diese in einem größeren Umfang eingesetzt werden. Manchmal interessiert die Frage, ob ein Programm, das sich in einem bestimmten Länder- oder Sektorkontext bewährt hat, auch unter anderen Bedingungen erfolgreich ist. Dies kann ebenfalls ein Ansatzpunkt für eine Wirkungsevaluation sein.

- b) Ein nicht weniger wichtiges Auswahlkriterium ist das Interesse der Öffentlichkeit, des Gebers oder des Partnerlandes, zu erfahren, ob die geförderten Programme oder propagierten Entwicklungsstrategien wirkungsvoll sind, um die dafür aufgewendeten Ressourcen zu legitimieren. In Zeiten knapper Finanzmittel steht die Entwicklungszusammenarbeit immer besonders in der Kritik. Eingangs wurde dargestellt, dass es in zyklischen Abständen immer wieder zu fundamentalen Angriffen auf die Entwicklungszusammenarbeit kommt, denen die EZ jedoch nicht wirklich überzeugende Belege für die Wirksamkeit der EZ entgegensetzen kann. Des-

halb ist es von besonderer Bedeutung, methodisch sauber zu überprüfen, ob bzw. inwieweit die aufgewendeten Mittel die gewünschten Wirkungen zeigen, ohne gleichzeitig so gravierende Nebenwirkungen zu haben, dass diese wieder konterkariert werden.

c) Wirkungsevaluationen erzielen vor allem dann einen „Mehrwert“ gegenüber herkömmlichen Evaluationen, wenn bestimmte Durchführungsvoraussetzungen gegeben sind. Dies ist der Fall, wenn

- der Evaluationsgegenstand klar abgegrenzt werden kann;
- Baselinedaten vorhanden sind, die mit den Daten zum Evaluationszeitpunkt für die ausgewählten Indikatoren verglichen werden können;
- Monitoringdaten und glaubwürdige Sekundärdaten vorliegen, die sich für den Vorher-Nachher-Vergleich verwenden lassen;
- sich Ziel- und Vergleichsgruppen gut voneinander abgrenzen lassen (was z.B. dann schwierig ist, wenn alle von der zu evaluierenden Programmmaßnahme oder von Maßnahmen anderer Geber in diesem Sektor profitieren);
- das Programm, das auf seine Wirkungen hin evaluiert wird, vom Volumen her groß genug war, damit überhaupt messbare Wirkungen entstehen konnten;
- das Programm nur so komplex ist, dass die abgeleiteten Ursache-Wirkungsstränge auch theoretisch wie methodisch isoliert und einzeln auf ihre Wirksamkeit hin untersucht werden können;
- die Projektverantwortlichen des Auftraggebers, aber auch die Partnerorganisationen die Evaluation voll mittragen und unterstützen.

(10) „Geht nicht – gibt's nicht!“ würde man am liebsten den vielen Bedenkenträgern entgegenhalten, die ohne es vorher einmal ausprobiert zu haben, schon wissen, warum Wirkungsanalysen nicht (!) durchgeführt werden können. Ein häufig vorgebrachtes Argument ist, dass sie zu teuer seien. In der Tat, Wirkungsevaluationen mit einem aufwendigen Design benötigen mehr Ressourcen als herkömmliche Evaluationen. Wenn man 5 bis 10% eines Programmbudgets als maximale Obergrenze für eine Wirkungsevaluation ansetzen würde, dann bedeutet dies, dass für Projekte mit einem Programmvolumen von 100.000 bis 200.000 € kaum einer aufwendigen Wirkungsevaluation unterzogen werden können, Projekte von 300.000 € an aufwärts jedoch leicht. Da – wie vorher betont – nicht für jedes Projekt oder Programm eine Wirkungsevaluation durchgeführt werden muss, bedeutet diese Faustregel keineswegs, dass alle Projekte ab einem bestimmten Volumen 5 bis 10% für Wirkungsevaluationen einplanen sollten. Ab einem bestimmten Programmvolumen nimmt der Anteil sowieso degressiv ab. Möchte man Projekte mit geringem Finanzvolumen einer Wirkungsevaluation unterziehen, empfiehlt es sich, mehrere Projekte der gleichen Art zu clustern, um die Evaluationskosten gemeinsam zu tragen. Zudem empfiehlt sich das Anlegen von organisationsinternen Fonds (aus Overheadmitteln oder anderen Quellen) für die Durchführung von ausgewählten Wirkungsevaluationen.

Häufig wird bei dem Kostenargument, das gegen umfassende Wirkungsanalysen vorgebracht wird, übersehen, dass auch die Nicht-Durchführung solcher Evaluationen Geld kostet. Wenn Geld in Programme gesteckt wird, die keine oder kaum Wirkungen

zeigen, wird dieses vergeudet. Niemand würde auf die Idee kommen, ein Medikament einzunehmen, das nicht in mehrfachen klinischen Tests auf seine Wirksamkeit und seine Nebenwirkungen hin ausreichend evaluiert worden ist. Bei EZ-Programmen, für die hohe Finanzbeträge dafür eingesetzt werden, das Verhalten von Menschen zu ändern, ihre Lebensverhältnisse umzugestalten, Gesellschafts-, Unternehmens- oder Familienstrukturen umzubauen, Einfluss auf ihre Kulturen und Wertvorstellungen zu nehmen, wurden (bisher) keine vergleichbaren Standards formuliert.

Dies bringt einen zu dem nächsten, oft gehörten Gegenargument, Wirkungsevaluationen seien unethisch, insbesondere wenn sie ein experimentelles Design verwenden, da die Kontrollgruppe bewusst keine Programmunterstützung erhält. Doch dieses Argument wäre nur dann stichhaltig, wenn alle Personen zu Beginn eines Projekts auch gleichzeitig versorgt werden könnten. Doch dies ist sehr häufig nicht der Fall, da die Ressourcen oder die Logistik gar nicht ausreichen, alle gleichzeitig zu bedienen. Dies kann man sich in einer Wirkungsevaluation zunutze machen, indem frühe Nutzer mit solchen Personen verglichen werden, die erst später von dem Programm profitieren. Selbst ein randomisiertes Experiment ist denkbar, wenn Teilnehmer und Nicht-Teilnehmer z.B. über eine Lotterie ausgewählt werden. Bis jetzt hat jedenfalls noch niemand protestiert, dass das häufig bei der Vergabe von Studienplätzen oder überfüllten Seminaren angewendete Losverfahren unethisch sei. Vielmehr scheint es unethisch, Programme, die zentral in die Lebensbereiche, Wertvorstellungen und Traditionen von Menschen und Völkern eingreifen, nicht auf ihre Wirkungen hin zu untersuchen, wobei den nicht-intendierten Wirkungen keine geringere Bedeutung zukommt als den intendierten.

Die bisherigen Wirkungsevaluationen haben gezeigt, dass derartige Untersuchungen grundsätzlich realisierbar sind und welche administrativen, operativen, theoretischen und methodischen Probleme es dabei zu bewältigen gibt. Auch wie sich Wirkungsevaluationen noch effektiver und effizienter planen und durchführen lassen und was bei der Auswahl von zu evaluierenden Programmen zu beachten ist, um ihren „Mehrwert“ zu steigern, um Lernpotenziale zu erschließen und um den Fundamental-Kritikern, aber auch allen Steuerzahlern und Spendern fundierte Belege für eine wirksame Entwicklungszusammenarbeit zu präsentieren, ist mittlerweile bekannt. Es muss „nur“ noch umgesetzt werden.

Damit hierfür die politischen und institutionellen Voraussetzungen geschaffen werden können, haben sich auf internationaler Ebene staatliche (NONIE) und nichtstaatliche (3IE) Strukturen herausgebildet, die Wirkungsevaluationen unterstützen und finanzieren. In Deutschland lassen sich auf nationaler Ebene bisher keine vergleichbaren Pendanten erkennen. Wie die Untersuchung des Evaluationssystems der deutschen Entwicklungszusammenarbeit ergab, ist eine Reihe von Reformschritten notwendig, um die Unabhängigkeit, Glaubwürdigkeit, Qualität, Nützlichkeit und Partnerschaftlichkeit des Evaluationssystems zu stärken. Diese Befunde können hier im Einzelnen nicht dargestellt werden (vgl. statt dessen Borrmann u. Stockmann 2009).

Im Hinblick auf die Wirksamkeit der Entwicklungszusammenarbeit ist zu konstatieren, dass das Evaluationssystem der deutschen Entwicklungszusammenarbeit bisher wenig profunde, methodisch und statistisch belastbare Ergebnisse vorweisen kann. Dies hat – wie bereits dargestellt – dazu geführt, dass eine Reihe von EZ-Organisationen Wirkungsevaluationen in Auftrag gegeben haben. Die hier in Kapitel 2.4.2 dargestellte Zersplitterung des deutschen

EZ-Systems hat zu einer äquivalenten Parzellierung des Evaluationssystems geführt, die durch das BMZ – wie in der „Systemstudie“ ausführlich dargestellt – trotz erheblicher Anstrengungen nicht ausreichend gesteuert werden kann. Dies führt nun auch bei der Erarbeitung von Konzepten und Instrumenten für rigorose Wirkungsevaluationen dazu, dass jede Organisation an ihren eigenen Entwürfen bastelt. Zwar hat das BMZ schon eine Reihe von gemeinsamen Workshops und Tagungen zu diesem Thema veranstaltet, doch dies reicht natürlich nicht aus, um die Segmentierung der Evaluationsanstrengungen zu überwinden. Was in Deutschland fehlt, ist

- (1.) ein Zusammenschluss der Evaluierungseinheiten auf nationaler Ebene, der die NRO, die an diesem Thema Interesse zeigen, mit einschließt,
- (2.) eine 3IE-vergleichbare Einrichtung, die Forschungs- und Evaluationsaufträge anhand wissenschaftlicher Standards vergibt und
- (3.) einen Finanzfonds, der diese Art von Evaluationen fördert.

Da die wissenschaftlichen Forschungsförderer (wie z.B. DFG oder VW-Stiftung) an dieser Fragestellung bisher kaum Interesse gezeigt haben, die EZ-Ressortforschung schon seit vielen Jahren kaum mehr über Finanzmittel verfügt, und aufwendige Designs für Wirkungsevaluationen nur schwerlich (im Nachhinein) aus den Programmetats finanziert werden können, ist die Bereitstellung von Finanzmitteln und der Aufbau einer an den Interessen und Bedürfnissen der staatlichen wie nicht-staatlichen EZ-Organisationen orientierten Vergabestruktur ein Gebot der Stunde. Auf diese Weise könnten Entwicklungsarbeiten gebündelt, Wirkungsstudien aufeinander abgestimmt, qualitativ hochwertige Studien produziert und die Ergebnisse rasch in den Geberkreis zurückgespiegelt werden. Über die vorgelegten Wirkungsstudien würden nicht nur neue Lernpotenziale für das gemeinsame Lernen aus staatlicher und nichtstaatlicher Entwicklungszusammenarbeit eröffnet, sondern auch die Legitimität und Glaubwürdigkeit der EZ als Ganzes gestärkt.

Literaturverzeichnis

- Böhnke, Jan R.; Koehler, Jan; Zürcher, Christoph (2009): Evaluation von Entwicklungszusammenarbeit zur Stabilisierung in Post-Konflikt-Zonen: Anwendung eines Mixed-Methods-Survey in Nordost-Afghanistan. In: Zeitschrift für Evaluation 2009/2, S. 215-236.
- Borrmann, Axel; Stockmann, Reinhard (Hg.) (2009): Evaluation in der deutschen Entwicklungszusammenarbeit. Band 1: Systemstudie und Band 2: Fallstudien. Studie im Auftrag des Bundesministeriums für Wirtschaftliche Zusammenarbeit und Entwicklung – BMZ. Münster: Waxmann.
- Caspari, Alexandra (2004): Evaluation der Nachhaltigkeit von Entwicklungszusammenarbeit. Zur Notwendigkeit angemessener Konzepte und Methoden. Wiesbaden: VS Verlag.
- Caspari, Alexandra (2009): ‚Rigorese‘ Wirtschaftsevaluation – methodische und konzeptionelle Ansätze der Wirkungsmessung in der Entwicklungszusammenarbeit. In: Zeitschrift für Evaluation 2/2009, S. 183-213.
- Caspari, Alexandra; Barbu, Ragnhild (2008): Wirkungsevaluierungen. Zum Stand der internationalen Diskussion und dessen Relevanz für die Evaluierung der deutschen Entwicklungszusammenarbeit. Bonn: Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung.
- Khandker, Shahidur R.; Koolwal, Gayatri B.; Samad Hussain A. (2009): Handbook on Impact Evaluation: Quantitative Methods and Practices. Washington: The World Bank.
- Reade, Nicolà (2008): Konzept für alltagstaugliche Wirkungsevaluierungen in Anlehnung an Rigorous Impact Evaluations. Erprobung der Durchführung im Rahmen von GTZ Unabhängigen Evaluierungen. CEval Arbeitspapier Nr. 14. Online verfügbar unter http://www.ceval.de/typo3/fileadmin/user_upload/PDFs/workpaper14_01.pdf
- Reade, Nicolà (2009): Durchführung von Pilot-Wirkungsevaluierungen am Beispiel des Wassersektors. In: Zeitschrift für Evaluation 2/2009, S. 237-262.
- Silvestrini, Stefan; Garcia, Melody (2010): Joint ex-post evaluation. Dual Vocational Training, Philippines (Unveröffentlichte Studie im Auftrag der GTZ, KfW, InWent und DED)
- Stockmann, Reinhard (1992): Die Nachhaltigkeit von Entwicklungsprojekten. Eine Methode zur Evaluierung am Beispiel von Berufsbildungsprojekten. Opladen: Westdeutscher Verlag.
- Stockmann, Reinhard (1996): Die Wirksamkeit der Entwicklungshilfe. Eine Evaluation der Nachhaltigkeit von Programmen und Projekten. Opladen: Westdeutscher Verlag.
- Stockmann, Reinhard; Meyer Wolfgang; Krapp Stephanie; Köhne, Gerhard (2000): Wirksamkeit deutscher Berufsbildungszusammenarbeit. Ein Vergleich staatlicher und nicht-staatlicher Programme in der Volksrepublik China. Wiesbaden: Westdeutscher Verlag.
- Stockmann, Reinhard; Meyer, Wolfgang (2010): Evaluation. Eine Einführung. Leverkusen: Barbara Budrich, UTB.
- White, Howard; WB IEG (2006a): Impact Evaluation. The Experience of the Independent Evaluation Group of the World Bank.
- Wolff, Jürgen H. (2005): Entwicklungshilfe: Ein hilfreiches Gewerbe? Versuch einer Bilanz. Münster: Lit Verlag.