17 | Bamberger Beiträge zur Soziologie

Methoden der Programmevaluation

Ein prozessorientiertes Rahmenmodell für die Begleitung von Evaluationsprojekten

Richard Wolf

University of Bamberg Press

17 Bamberger Beiträge zur Soziologie

Bamberger Beiträge zur Soziologie

Band 17

University
of Bamberg
Press
2017

Methoden der Programmevaluation

Ein prozessorientiertes Rahmenmodell
für die Begleitung von Evaluationsprojekten

von Richard Wolf

University
of Bamberg
Press
2017

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.d-nb.de/abrufbar.

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.
1. Gutachter: Prof. Dr. Friedrich Heckmann
2. Gutachter: Prof. Dr. Hans-Günther Roßbach
Tag der mündlichen Prüfung: 02. März 2016

Herstellung und Druck: docupoint, Magdeburg
Umschlaggestaltung: University of Bamberg Press, Anna Hitthaler

ISSN: 1867-8416
ISBN: 978-3-86309-473-7 (Druckausgabe)
eISBN: 978-3-86309-474-4 (Online-Ausgabe)
URN: urn:nbn:de:urn:bvb:473-opus4-477122

Vorwort

Die vorliegende Dissertationsschrift ist als Resultat meiner mehrjährigen Arbeit als wissenschaftlicher Mitarbeiter am europäischen forum für migrationsstudien (efms) in Bamberg entstanden. Befasst habe ich mich in dieser Zeit hauptsächlich mit der Evaluation von Bildungsmaßnahmen für Migranten. Im Laufe der Beschäftigung mit dem Themenbereich Evaluation hat sich die Grundidee für die Dissertation entwickelt, die Erkenntnisse aus meiner Arbeit für die Weiterentwicklung von Evaluationsmethoden heranzuziehen.

Ich möchte daher allen danken, die mich im Verlauf der langen Entstehungsgeschichte der Dissertation unterstützt haben und mindestens so sehr darunter gelitten haben wie ich. Dafür, dass das Ganze doch noch ein gutes Ende gefunden hat, möchte ich mich zuerst bei meinem Doktorvater Prof. Friedrich Heckmann bedanken. Er hat mich immer ermuntert, mit der Dissertation weiterzumachen, auch wenn mich oft den Eindruck begleitete, es steht ein unüberwindbarer Berg an Arbeit vor mir. Mein Dank gilt auch Maria Matreux, Wolfgang Bosswick, Mihaela Tudose sowie allen weiteren Kolleginnen und Kollegen, die mich während meiner Zeit am efms bei den Evaluationsprojekten unterstützt haben. Bedanken möchte ich mich auch bei der Gemeinnützigen Hertie Stiftung, der Stadt Nürnberg sowie dem Bayerischen Sozialministerium für ihre Zustimmung zur Verwendung des Materials aus den Evaluationsberichten.

Mein Dank geht außerdem an meine Eltern, die mich in der Zeit in Bamberg voll unterstützt haben. Und zu guter Letzt kann ich schwer in Worte ausdrücken, wie sehr ich mich bei Katja bedanken möchte. Katja war während der gesamten Entstehungszeit der Arbeit meine beste Beraterin, meine Lektorin, hat sämtliche Höhen und Tiefen mitbekommen, hat mitgelitten und hat mir auch noch an vielen Wochenenden den Freiraum gegeben, um die Arbeit abschließen zu können. Ohne ihre Unterstützung hätte ich es nicht geschafft.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AEA	American Evaluation Association
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AV	Abhängige Variable
Bamf	Bundesamt für Migration und Flüchtlinge
DeGEval	Deutsche Gesellschaftfür Evaluation e.V.
Efms	europäisches forum für migrationsstudien
MSS	Marburger Sprach-Screening
NPM	New Public Management
OEC	Office of Economic Opportunity
PAT	Parents as Teachers
SOFIS	Sozialwissenschaftliches Forschungsinformationssystem
UV	Unabhängige Variable
WPA	Works Progress Administration

1. Einleitung

Sozialwissenschaftliche Evaluationen werden in Bereichen unternommen, in denen Nachweise über die Leistungsfähigkeit von Programmen, Initiativen, Gesetzen und Maßnahmen gefordert werden. Insbesondere zu nennen sind in diesem Kontext das Bildungswesen (vgl. u. a. Böttcher et al. 2006, Stamm 2003, OECD 2004), die Arbeitsmarkt- und Beschäftigungspolitik (vgl. u. a. Hagen & Spermann 2004, Helmstädter 2009, Caliendo & Steiner 2005, Bussmann et al. 1997), die Verwaltungspolitik (vgl. u. a. Buschhoff 2009, Wollmann & Kuhlmann 2004, 2011), die Integrationspolitik (vgl. u. a. Filsinger 2008; Held et al. 2007) sowie die Entwicklungspolitik (vgl. Stockmann 2006). Evaluation nimmt daher thematisch wie disziplinär ein sehr breites Handlungsfeld ein. Da die Evaluationspraxis durch Methodenvielfalt und eine Fülle an potenziellen Evaluationsgegenständen geprägt ist, stellt sich das Handlungsfeld nicht nur für Laien, sondern sogar für erfahrene Evaluatoren oft als unübersichtlich dar.

Die vorliegende Arbeit verfolgt das Ziel, ein prozessorientiertes Rahmenmodell für die Arbeit von Evaluatoren in der Praxis zu entwickeln. Im Hauptteil der Arbeit werden zunächst in einem historisch-systematischen Vorgehen die Methoden einer Auswahl von Evaluationstheoretikern vorgestellt und hinsichtlich ihrer Relevanz für die Evaluation von sozialen Programmen diskutiert. Neben den Erkenntnissen aus der Evaluationstheorie wird die Anwendung von Evaluationsmethoden in drei Evaluationsstudien des Autors kritisch reflektiert. Die auf diese Weise gewonnen Erkenntnisse aus Theorie und Praxis werden schließlich in einem Prozess für die Durchführung von Evaluationsstudien aufbereitet. Die aus dieser Analyse gewonnenen Erkenntnisse sollen Evaluatoren bei der Gestaltung und Durchführung von Evaluationsstudien zu sozialen Programmen in verschiedenen Etappen und Phasen unterstützen. Diese beginnen bereits mit der Formulierung von adäquaten Zielen der Evaluation, der Auswahl und dem Zusammenführen geeigneter Methoden zu einem Evaluationsdesign, der Auseinandersetzung mit Kontextbedingungen bis hin zur Auswertung und Kommunikation der Ergebnisse.

Bevor auf die methodischen Implikationen von Evaluation eingegangen wird, erscheint es zunächst notwendig, systematisch die wissenschaftlichen Begrifflichkeiten und Konzepte im Feld Evaluation zu bearbeiten. In der Fachliteratur finden sich zahlreiche Definitionen des Begriffs Evaluation, die sich bei näherer Betrachtung voneinander hinsichtlich bestimmter funktionaler Schwerpunktsetzungen (z.B. den Nutzen von Evaluation) unterscheiden. Wenn wissenschaftliche Evaluationen in Betracht gezogen werden, handelt es sich im Idealfall um Bewertungen von klar bestimmbaren Sachverhalten unter Zuhilfenahme von Methoden der empirischen Sozialforschung und unter Wahrung wissenschaftlicher

Standards. Dies bedeutet, dass „genau benannte und empirisch beschreibbare Sachverhalte (Programme, Maßnahmen, Organisationen) und auf diese gerichtete präzise und operationalisierbare Fragestellungen ihr ‚Gegenstand' sind" (Kromrey 2008, S. 117). Kromrey unterscheidet dabei zwischen dem alltäglich verwendeten Evaluationsbegriff, mit dem „Bewertung" gemeint ist, und einem spezifischen wissenschaftlichen Denk- und Verfahrensmodell. Dieses spezifische Denkmodell bezieht sich auf einen Prozess, den Kromrey als zielorientiertes Informationsmanagement beschreibt: „Im allgemeinen Sinne gilt als Evaluation jede methodisch kontrollierte, verwertungs- und bewertungsorientierte Form des Sammelns, Auswertens und Verwertens von Informationen" (Kromrey 2001, S. 1). Abhängig von dem Untersuchungsgegenstand und dem Erkenntnisinteresse können die Methoden variieren.

Eine oft zitierte Definition von Evaluation findet man bei Rossi et al. (1988). In der Vielfalt an Begriffsdefinitionen sticht diese Definition als geeignet heraus, um Evaluationsvorhaben möglichst umfassend zu beschreiben. Es handelt sich dabei um eine einfache, aber eingängige Beschreibung dessen, was Evaluation leisten kann: „Wir definieren sie (Anm. d. A.: Evaluationsforschung) einfach als systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Beurteilung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer Interventionsprogramme " (Rossi, Freeman & Hofmann 1988, S. 3). Die Definition beinhaltet neben der Bewertung von sozialen Programmen anhand sozialwissenschaftlicher Methoden die Berücksichtigung der zeitlichen Komponente von der Planung bis zur praktischen Umsetzung von Interventionen. Dieser Definitionsansatz erscheint daher besonders wertvoll für das Thema dieser Arbeit, bei dem es insbesondere um methodische Überlegungen zu bestimmten Zeitpunkten der Durchführung von Evaluationen geht.

In der Definition von Rossi wurde ein Handlungsfeld von Evaluation – das auch in dieser Arbeit betrachtet wird – bereits genannt: **soziale Programme**. Dieser Gegenstand der Evaluation kann aber auch eine politische Maßnahme, ein Projekt, die Umsetzung eines Gesetzes oder die Arbeit von Organisationen und deren Suborganisationen sein.

In der vorliegenden Arbeit ist der der Gegenstand der Evaluation in allen zu betrachtenden Fällen ein soziales Programm. Als Programme sind dabei systematische und aufeinander aufbauende Handlungen zu verstehen, die aus einem gemeinsamen Zielsystem abgeleitet werden. Programme entstehen aufgrund einer gemeinsamen Zieldefinition von Akteuren mit der Intention, diese in Form von systematischen und strukturierten Handlungen umzusetzen. Für die Evaluation von sozialen Programmen wird auch der Begriff **Programmevaluation** verwendet.

In der Evaluationsforschung wird in Abgrenzung zu Programmevaluationen von Quasi-Evaluationen bzw. Pseudo-Evaluationen gesprochen (vgl. Thoening 2000), wenn die Begleitforschung nicht das primäre Ziel hat, den Wert und den Nutzen bestimmter Programmtätigkeiten zu bewerten. Bei Programmevaluation werden Daten über den Programmverlauf gesammelt, ausgewertet und auf Basis der Ergebnisse (z.B. von Wirkungsmessungen) Empfehlungen zur Fortführung erarbeitet. Neben diesen sozialforschungsnahen Tätigkeiten nehmen Kommunikationsprozesse mit Programmbeteiligten, Auftraggebern von Evaluation und ggf. mit weiteren Partnern einen hohen zeitlichen Stellenwert ein.

Programmevaluationen nehmen gegenwärtig im Bereich der Bildungsförderung eine bedeutsame Rolle ein und beziehen sich auf umfassende Programme oder einzelne Maßnahmen. Einzelmaßnahmen, wie z.B. ein Sprachförderkonzept im Kindergarten, können Bestandteil eines kommunalen Bildungsprogramms sein, das wiederum Teil der kommunalen Strategie zur Verbesserung der Integration von Migranten ist. Cronbach (1963) erläutert den Zweck von Evaluation im Bildungsbereich am Beispiel von Schulevaluationen: „To draw attention to its full range of functions, we may define ‚evaluation' broadly as the collection and use of information to make decisions about an educational program. The program may be a set of institutional materials distributed nationally, the institutional activities of a single school, or the educational experiences of a single pupil" (Cronbach 1963, S. 672).

Genauso wie sich Evaluation auf verschiedene Objekte beziehen kann, zeichnet sich die Verschiedenartigkeit von Evaluationsvorhaben hinsichtlich der beteiligten Personen aus, die ein individuelles Interesse mit dem Programm verbinden (vielfach auch in der deutschen Fachliteratur als *Stakeholder* bezeichnet). Diese vielschichtige Bedeutung von Evaluation in Kombination mit der wachsenden Zahl von Evaluationsstudien macht Evaluation noch zusätzlich zu den oben genannten Punkten zu einem komplexen und nicht einfach zu beschreibenden Handlungsfeld.

Programmevaluationen können sich nach den bisherigen beschriebenen Aspekten – und weiteren hier nicht näher aufgeführten inhaltlichen Aspekten – voneinander unterscheiden. Evaluationsvorhaben bzw. -studien haben dann auch unterschiedliche Bezeichnungen, wie z.B. Effizienzforschung, Wirkungsanalyse, Qualitätsmanagement oder Kosten-Nutzen-Analyse. All dies können Bezeichnungen für Evaluationsvorhaben im Bereich der Programmevaluationen sein und sich doch inhaltlich stark voneinander unterscheiden.

Diese Begriffspluralistik verdeutlicht die aus der Außenbetrachtung vielfach schwer durchschaubare Evaluationsthematik insgesamt. Eine Systematisierung der Begrifflichkeiten kann durch die Einordnung der unterschiedlichen Programmevaluationen nach *ihren Funktionen* erfolgen. Eine konzeptionell einfache

und sehr brauchbare Eingrenzung der Funktionen von Evaluation – die im Laufe der Arbeit mehrmals aufgegriffen wird – nimmt Chelimsky (1997, S. 10ff.) vor. Programmevaluationen können nach dieser idealtypischen Systematisierung grundsätzlich eine von drei Funktionen annehmen:

- Evaluation zu Kontrollzwecken (*evaluation for accountability*)
- Evaluation zu Entwicklungszwecken (*evaluation for development*)
- Evaluation zur Wissenserweiterung (*evaluation for knowledge*)

Bei der Evaluation zu Kontrollzwecken geht es um die **Erfolgskontrolle des Handelns** zu bestimmten Zeitpunkten im Verlauf der Durchführung eines Programms. Die Kontrolle des Verlaufs und der Wirkungsweise einer Maßnahme oder eines Maßnahmenbündels in einem Programm wird beispielsweise im Wissenschaftsmanagement oder in der öffentlichen Verwaltung benötigt. Programme sollen stetig verbessert und weiterentwickelt werden. Die Kriterien für die Erfolgsbeurteilung sind Effektivität (Grad der Zielerreichung), Effizienz (ressourcenschonender Mitteleinsatz) und die Akzeptanz bei den Stakeholdern (mittelbare und unmittelbare Programmbeteiligte). Es werden Evaluationsmethoden herangezogen, anhand derer Output (Ergebnis eines Programms) bzw. Outcome (Ergebnis eines Programms unter Berücksichtigung der Programmziele) von Programmen erfasst wird. Es kann sich aber auch um Wirtschaftlichkeitsprüfungen, Kontrollen der Rechtmäßigkeit im juristischen Sinne oder der Kontrolle der Implementation von sozialen Programmen handeln (vgl. Kromrey 2005, S. 5).

Die zweite Funktion von Evaluation ist die für **Entwicklungszwecke**. Soll ein bestehendes Programm weiterentwickelt oder modifiziert werden, können Programmevaluationen auf das Ziel der Verbesserung hin konzipiert werden. Nach Bussmann (1997) steht im Fokus von Evaluationen zu Entwicklungszwecken auch hier der öffentliche Bereich: „Evaluationen sollen – in Form von Berichten, persönlichen Orientierungen, Datenbasen u.a.m. – Informationen bereitstellen, welche dazu beitragen, öffentliche Politik zu verbessern. Evaluationen sind somit ein Informationsinstrument“ (Bussmann 1997, S. 2).

Eine besonders häufig ausgeübte Form der Programmevaluation für Entwicklungszwecke ist die so genannte **formative Evaluation**. Der Begriff formative Evaluation wurde ursprünglich von Michael Scriven (vgl. 1972, 1980, 1991) entwickelt und bezeichnet eine Rolle, die der Evaluation zukommt. Formativ bezeichnet sie eine Vorgehensweise, bei der parallel zur Programmdurchführung Informationen und Daten mit dem Zweck erhoben werden, die Ergebnisse direkt zur Verbesserung und Weiterentwicklung des Programms zu nutzen. Formative Evaluationen zeichnen sich durch eine „kybernetische“ Funktion aus (Wollmann 2005a, S. 3), indem durch evaluative Tätigkeiten gewonnene Informationen zurück in den Implementierungsprozess des Programms gebracht werden. Es kann

sich dabei u. a. um die wissenschaftliche Begleitung eines Programms, Feedbackerhebungen oder Implementierungsevaluationen von neu gestarteten Programmen handeln. Auch Qualitätsmanagementsysteme beinhalten Evaluationsmethoden, die auf den Zweck der Weiterentwicklung ausgerichtet sind und über reine Kontrollzwecke (z.B. Qualitätssicherung) hinausreichen. Abschließende Evaluationen, die ein zusammenfassendes Urteil (z.B. ein Gutachten) zu einem Sachverhalt beinhalten, werden als **summative Evaluation** bezeichnet.

Die dritte Funktion, **Evaluation zur Wissenserweiterung**, stellt in Bezug auf wissenschaftliche Standards hohe Anforderungen an die Konstruktion des Evaluationsdesigns und die Auswahl der Evaluationsmethoden. In Abgrenzung zu den beiden zuvor genannten Funktionen von Evaluation werden Evaluationsstudien hier als eine Form der angewandten Sozialforschung durchgeführt. Das zu untersuchende Programm kann Bestandteil eines theoretischen Wirkungsmodells sein. Durch die Evaluationsstudie werden die Hypothesen bzw. Annahmen zum Einfluss des Programms auf Veränderungen bei den Teilnehmern überprüft. Die Methoden und Instrumente der Evaluation werden dann auf Basis des theoretischen Modells ausgewählt. Programmevaluation zur Wissensverbreitung konzentriert in diesen Fällen auf die differenzierte **Messung von Programmeffekten** und der möglichst zweifelsfreien Zuordnung der gemessenen Effekte zu den Effekten des untersuchten Programms. Durch die Identifikation von Wirkungen werden Nutzen und Qualität eines Programms beschreibbar. Fragen nach dem Sinn oder Fortführung eines Programms können besser beantwortet werden, wenn eindeutige Wirkungszusammenhänge durch die Evaluation ermittelt wurden. Ernest House (1993) merkt in diesem Zusammenhang an, dass Evaluationen von Programmen im besten Fall etwas zur demokratischen Entscheidungsfindung beitragen, in dem Ergebnisse veröffentlicht werden und dadurch einem öffentlichen Diskurs ausgesetzt sind. Dadurch haben Evaluationen nicht nur für den Auftraggeber einen hohen Nutzwert, sondern auch für die Gesamtgesellschaft.

Feld- und Laborexperimente – wie aus der Medizinforschung und klinischen Psychologie – sind die „Gold Standards“ und stellen nach dem gängigen wissenschaftlichen Verständnis die besten Methoden der Wirkungsmessung dar (Campbell 1966, Cook & Campbell 1979). Klassische Experimente arbeiten mit einem Vergleich von einer Untersuchungsgruppe, deren Teilnehmer Maßnahmen aus einem Programm wahrnehmen, und einer Kontrollgruppe mit Personen, die ähnliche Charakteristika haben, jedoch von den Maßnahmen des Programms ausgeschlossen sind. Die Wirkungen eines sozialen Programms lassen sich durch den Vergleich der Merkmale von Untersuchungs- und Kontrollgruppen im zeitlichen Verlauf identifizieren.

In der Evaluationsforschung besteht weitgehende Einigkeit darüber, dass theoretisch das beste Design für die Wirkungsevaluation von Programmen das Feldexperiment ist. Die Voraussetzungen und Vorgaben für Feldexperimente können jedoch in Realität nahezu nie erfüllt werden; oftmals ist es nicht möglich, alle Einflussfaktoren auf die experimentelle Versuchsanordnung zu kontrollieren bzw. überhaupt zu identifizieren. Zur Messung von Wirkungen wird daher in der Programmevaluation auf alternative Methoden zurückgegriffen, wie z.B. Quasi-Experimente oder Zeitreihenanalysen, um Programmeffekte zu quantifizieren. Damit die Maßnahmen eines Programms überhaupt Wirkungen entfalten können, wird die Durchführung der Wirkungsmessung erst nach Abschluss der Implementation eines Programms empfohlen (Wottawa & Thierau 1998; Chen 1990).

Die Nachfrage nach Evaluationsstudien ist im Verlauf der letzten 20 Jahren in Deutschland gestiegen. Im Bereich der öffentlichen Verwaltung lässt sich dieser Anstieg durch die Einführung von neuen Managementansätzen wie dem New Public Management erklären (vgl. Wollmann 2004, 2005b, Derlien 1997). Evaluation wird beim New Public Management als Instrument zur Generierung und Auswertung von Daten und Informationen verstanden, um basierend auf den Ergebnissen Steuerungsentscheidungen im Bereich der öffentlichen Verwaltung zu treffen. Die Auftraggeber bzw. Initiatoren von Evaluationen erhoffen sich Informationen darüber zu erhalten, ob die zur Verfügung gestellten Ressourcen zielführend und wirkungsvoll (d.h. effizient und effektiv) eingesetzt werden, welchem Qualitätsstandard die Maßnahmen entsprechen sowie eine abschließende Beurteilung darüber, ob und wie die Maßnahmen zu einer Verbesserung der Status-quo-Situation führen und damit nützlich sind. Evaluationsstudien mit den soeben beschriebenen Zielen erfüllen überwiegend die Funktion eines Kontrollinstruments.

Neben dem öffentlichen Sektor äußern insbesondere private Stiftungen – die ihre Förderaktivitäten zunehmend in den Bildungssektor verlagern – eine erhöhte Nachfrage nach Evaluation. Stiftungen initiierten in den vergangenen Jahren – wie noch für die Gruppe der Migranten näher erläutert wird – Programme mit dem Ziel der kompensatorischen Bildungsförderung. Im Vergleich zur öffentlichen Verwaltung werden Evaluationsstudien bei diesen Programmen primär zu Entwicklungszwecken (z.B. das Pilotieren von Maßnahmen) und/oder zum Zweck der Wissenserweiterung durchgeführt.

Den Anstoß für die Beschäftigung mit dem Thema der Arbeit ergab sich während der Tätigkeit des Autors am europäischen forum für migrationsstudien (efms). Der Autor hat am efms mehrere Evaluationsstudien geplant und während der gesamten Laufzeit betreut. Zwangsläufig erfolgte zur Vorbereitung der Studien

eine intensive Auseinandersetzung mit Primärliteratur im Bereich der Evaluationsmethodik. Lehrbücher zur Evaluation (z.B. von Wottawa & Thierau, 1998, Bortz & Döring 2003, Stockmann 2006) geben recht genaue Handlungsanleitungen und Hinweise dazu, wie theoretisch Evaluationsstudien geplant und durchgeführt werden können. Einige Lehrbücher legen den Schwerpunkt auf die Anwendung von Methoden und die Auswahl der richtigen Instrumente und Auswertungsverfahren (z.B. Bortz & Döring 2003, Kromrey 2002). In jüngster Zeit sind Publikationen erschienen, in denen Vorschläge zur Gestaltung des Evaluationsdesigns und den Ablauf einer Evaluationsuntersuchung gegeben werden. Es werden auch Beispiele genannt, wie Evaluationsstudien in verschiedenen Disziplinen durchgeführt werden können (z.B. Flick 2006, Borrmann & Stockmann 2009). In der sehr aktuelle Publikation „Evaluation: Eine Einführung" von Stockmann und Meyer (2014) wird ein Schritt weiter in Richtung einer Unterstützung von Evaluatoren in der Praxis gegangen. So widmet sich ein Großkapitel der Gestaltung des Evaluationsprozesses, der wiederum in verschiedene Phasen unterteilt wird.

Die im Laufe der Auseinandersetzung mit Evaluationsmethoden gereifte Erkenntnis des Autors lautet jedoch, dass es **weiterhin an Anleitungen fehlt**, wie sich Evaluationsstudien als Projekte in der Praxis von der **Planung bis zu der Ergebnisvermittlung** konkret umsetzen lassen. Insbesondere auf Formen des Austausches und der Zusammenarbeit zwischen Evaluatoren und Projektverantwortlichen/-beteiligten im Laufe eines Evaluationsverfahrens wird in Lehrbüchern zur Evaluation nicht im überwiegenden Maß eingegangen. Genauso finden sich selten praxisnahe Ausführungen dazu, welche Schwierigkeiten sich in den verschiedenen Phasen eines Evaluationsprojekts ergeben und wie Evaluatoren diese frühzeitig wahrnehmen und welche Empfehlungen zum Umgang mit Problemen gegeben werden können.

Trotz des Anwachsens der Evaluationsfachliteratur ist Evaluation in Deutschland gegenwärtig ein noch junges Betätigungsfeld. Das Wissen über Methodik, Ansätze und Strategien hat sich zwar aufgrund von Lehrbüchern, dem Engagement zahlreicher Lehrstühle und wissenschaftlicher Institute weiter entfalten können, ein Institutionalisierungs- und Professionalisierungsgrad – im Sinne einer breiten wissenschaftlichen Fundierung von Evaluationstheorien und Evaluationsmethoden sowie deren Anwendung zur Untersuchung gesellschaftlicher Phänomene – ist verglichen mit dem in den USA, in denen Evaluation mit einem hohen Professionalisierungs-grad betrieben wird, bei weitem nicht erreicht. Evaluationsforschung und Evaluationstätigkeit in Deutschland können auch nicht auf einen vergleichbaren Erfahrungsschatz aufbauen; es gibt aber Initiativen, die eine Weiterentwicklung der Evaluationsmethodik vorantreiben. So ist die Gesellschaft für Evaluation (DeGEval e.V.) seit 1997 in Deutschland aktiv und hat sich

zum Ziel gesetzt, den Austausch zwischen verschiedenen Handlungsfeldern voranzutreiben, in denen Evaluation eine Rolle spielt und Ansätze für die Evaluationspraxis zu entwickeln.

Damit Evaluatoren professionell arbeiten können (z.B. in der Evaluation von sozialen Programmen), sind Kompetenzen im Bereich der empirischen Sozialforschung unabdingbar. Neben reinem Methodenwissen sind Kenntnisse bezüglich der Planung und Durchführung von Evaluationsstudien notwendig – gewissermaßen Projektmanagementkenntnisse, die durch Erfahrungswissen angereichert werden. Reine Methodenkenntnisse werden in Studiengängen mit dem Qualifikationsziel Soziologie, Psychologie oder Pädagogik vermittelt, es fehlen jedoch entsprechende Angebote zur Praxis der Evaluationstätigkeit. Evaluatoren sehen sich ab dem Start von Evaluationsstudien mit Fragen konfrontiert, wie z.B.: Welche Arbeitsschritte sind zu Beginn einer Evaluationsstudie wichtig – neben der Konzeption des Evaluationsdesigns? Wie kann festgestellt werden, ob eine Wirkungsevaluation zum vorliegenden Programm tatsächlich Sinn macht? Wie soll mit den Erwartungen der am Programm Beteiligten umgegangen werden? Daher erscheint es zwingend notwendig, Evaluatoren für die Praxis einen Handlungsrahmen zur Orientierung zu geben.

Das Ziel dieser Arbeit ist es, durch eine Aufarbeitung von relevanten Konzepten aus der theoretischen Evaluationsforschung sowie basierend auf den Erkenntnissen der durchgeführten Evaluationsstudien ein Modell für die Programmevaluation ab-zuleiten. Folgende Fragen sollen im Verlauf der Arbeit beantwortet werden:

1. Welche zentralen Diskussionsstränge zu Methoden, Verfahrensweisen und Handlungsanleitungen für die Durchführung von Programmevaluationen haben sich entwickelt?

2. Welche Charakteristika hat ein prozessorientiertes Rahmenmodell und wie lässt sich dieses für Programmevaluationen entwickeln?

3. An welche Stellen in einen Prozessmodell für die Durchführung von Evaluations-projekten können die zuvor identifizierten Methoden, Verfahrensweisen und Handlungsanleitungen positioniert werden?

4. Welche methodischen Erkenntnisse und Folgerungen können aus der Evaluationspraxis herangezogen werden?

5. Welche zentralen Empfehlungen lassen sich aus dem entwickelten Prozessmodell für die Praxis ableiten?

Im zweiten Kapitel folgt – nach einer kurzen Einleitung zur historischen Entwicklung der Evaluationstätigkeit in den USA und Deutschland – eine systema-

tisch-chronologische Erörterung einer Auswahl prominenter Evaluationsansätzen der Programmevaluation. Unter Evaluationsansätzen sind Rahmenkonzepte bis hin zu Methoden zu verstehen, anhand derer die Planung und Durchführung von Programmevaluationen organisiert werden können. Resultierend aus frühen amerikanischen Erfahrungen mit Evaluation im 20. Jahrhundert ist eine Vielzahl von Evaluationsansätzen hervorgegangen. Die Ansätze reichen von erkenntnistheoretischen bis hin zu nutzungsorientierten Vorgehen bei Evaluationen; dementsprechend lassen sich Unterschiede hinsichtlich der Funktion, des Designs und der Methoden festhalten. Eine Systematisierung der Evaluationsansätze hinsichtlich verschiedener Unterscheidungsmerkmale sowie eine erste Abschätzung der Relevanz der Ansätze für das noch zu entwickelnde Prozessmodell schließen das zweite Kapitel ab.

Im den sich anschließenden Kapiteln wird das prozessorientierte Rahmenmodell für eine projektmanagementorientierte Begleitung von Programmevaluationen in der Praxis erarbeitet. Das Rahmenmodell beschreibt zunächst aus der Sicht des Evaluators in generischen Zügen den typischen Ablauf von Programmevaluationen als Prozess. Die einzelnen Prozessschritte (Planung, Konzeption, Umsetzung, Ergebnis) dienen für den sich anschließenden Aufbau der Arbeit als Untergliederungspunkte. In die Darstellung der einzelnen Prozessschritte fließen – spezifisch und inhaltlich passend – die Analyseergebnisse aus dem zweiten Kapitel sowie die Erkenntnisse aus den Evaluationsstudien ein, die der Autor im Verlauf seiner Tätigkeit am efms unternommen hat. Die Bearbeitung der zweiten und dritten Frage erfolgt somit simultan.

Die Analyse von drei Evaluationsstudien – unter Gesichtspunkten wie Entwicklung des Evaluationskonzepts, Effektivität der eingesetzten Methoden, Kommunikation mit Programmbeteiligten sowie Ergebnispräsentation – stellt den zweiten Schwerpunkt der Arbeit dar. Der Autor war in der Zeit seiner Beschäftigung am efms in Bamberg mit der Konzeption, Durchführung, Auswertung und Kommunikation der Evaluationsstudien beauftragt. Alle Programme wurden zwischen 2001 und 2007 evaluiert – einem Zeitraum, in dem in Deutschland auf regionaler Ebene eine große Anzahl von Förderprojekten zur Integration der Bevölkerung mit Migrationshintergrund als Modellvorhaben gestartet wurde. Alle drei untersuchten Programme lassen sich der kompensatorischen Bildungsförderung zurechnen, d.h. durch gezielte Förderung in Form von Projekten oder Programmen sollen Bildungsdefizite ausgeglichen werden. Unabhängig von dem gemeinsamen politischen Ziel der Integration unterscheiden sich die Bildungsprogramme stark hinsichtlich der Programmziele, den Zielgruppen, der Inhalte sowie der Wirkungsannahmen. Alle drei Maßnahmen hatten zum Evaluationszeitpunkt Modellcharakter und stellen daher spezifische Ansprüche an die Evaluation. Das Ziel der Evaluationsstudien war es, aussagekräftige Informa-

tionen darüber zu erhalten, wie Bildungsprogramme arbeiten, in welchem Ausmaß die Fördergrundsätze im Programm verfolgt werden und ob Aussagen zu Wirkungszusammenhängen getroffen werden können. Insbesondere aus der kritischen Analyse der verwendeten Evaluationsmethoden lassen sich Hinweise zur Optimierung der Evaluationspraxis ableiten. Für die Darstellung und Diskussion von ausgesuchten Ergebnissen der Evaluationsstudien wird im Umfang von drei Unterkapiteln bewusst viel Raum gelassen.

Im letzten Kapitel werden die Ergebnisse der Arbeit zu einem Prozessmodell für die Begleitung von Evaluationsprojekten zusammengetragen. Im Prozessmodell wird in schematischer Weise (in Anlehnung an die Darstellung von Qualitätsmanagementprozessen) eine detaillierte Handlungsanleitung für Evaluatoren entworfen. Die Arbeit schließt mit einer Reihe von Empfehlungen für die Praxis sowie mit der Diskussion von möglichen Restriktionen, auf die Evaluatoren bei der Durchführung von Evaluationsprojekten stoßen können.

2. Entstehung und Entwicklung von Evaluationsmethoden für soziale Programme

2.1. Beginn der Evaluationstätigkeit in den USA und in Deutschland

Die Ursprünge der modernen Evaluationsforschung gehen auf Entwicklungen in den USA in den 30er Jahren des letzten Jahrhunderts zurück, Nutzen und Wirkungen umfassender Sozialprogramme zu beurteilen. In dieser Frühphase wurde wissenschaftliche Evaluation für experimentelle Versuchsanordnungen im sozialen Bereich und hauptsächlich für die Bewertung und Verbesserung von Curricula im schulischen Sektor verwendet. Der Einsatz von Evaluation diente der Beurteilung von pädagogischen Maßnahmen sowie der Lernerfolgskontrolle bei Schülern. Die Instrumente für Evaluationen beschränkten sich in den meisten Fällen auf standardisierte Leistungstests für die Beurteilung von Lerneffekten. So arbeitete Ralph W. Tyler (vgl. 1932, 1949), ein renommierter Bildungsforscher in den USA zu der damaligen Zeit, an der Weiterentwicklung von Testtheorie und Testpraxis in der Schulforschung[1].

Die erste nennenswerte zeitliche Hochphase von Evaluationen lässt sich in den 30er Jahren identifizieren. In dieser Zeitphase zwischen der Weltwirtschaftskrise und dem Eintritt der USA in den zweiten Weltkrieg wurde unter Präsident Roosevelt durch massive staatliche Investitionen versucht, die amerikanische Binnenkonjunktur anzukurbeln. Die sozialpolitischen Maßnahmen in dieser Zeit werden seither mit dem Namen „New Deal“[2] in Verbindung gebracht. Die regierungsgeleiteten Maßnahmen zielten auf die Überwindung der Großen Depression („Great Depression“), ausgelöst durch die Weltwirtschaftskrise Ende der 20er Jahre infolge einer „Laissez-faire“-Politik und eines unregulierten Börsenkapitalismus (Stockmann 2006, S. 23).

Die „New Deal“-Maßnahmen reichten von sozialpolitischen Innovationen, Bildungsmaßnahmen bis hin zu technologischen, agrarpolitischen und vor allem baulichen Großprojekten. Die eingeführten Maßnahmen basierten nicht auf einem vordefinierten Konzept oder Leitgedanken, sie waren stark pragmatischer Natur: an den Stellen, bei denen Unterstützung benötigt wurde, leistete die Re-

[1] Eine relevante Veröffentlichung von Ralph W. Tyler aus 1932 trägt den Titel „Service Studies in Higher Education“ und kann als prototypische Evaluationsarbeit im Bereich Schul- und Hochschulevaluation betrachtet werden.

[2] Präsident Franklin D. Roosevelt hat am Tag seiner Nominierung zum Präsidentschaftskandidaten der Demokratischen Partei einen „New Deal for the American people“ versprochen. Alle Maßnahmen unter der Agenda des Reformprogramms von Roosevelt wurden in der Folge als „New Deal“ zusammengefasst.

gierung entsprechende organisatorische und finanzielle Hilfestellung. Es wurden zudem Behörden gegründet, die für die Durchführung der Projekte zuständig waren. Mit der Works Progress Administration (WPA) wurde seinerzeit die größte Behörde der USA geschaffen, die hauptsächlich für den Straßen- und Brückenbau zuständig war. In der Sozialpolitik wurden zahlreiche Innovationen durch die Einführung des „Social Security Act" umgesetzt. Neben der Einführung einer staatlichen Rente und einer Arbeitslosenversicherung wurde Kinderarbeit verboten und ein progressives Steuersystem eingeführt.

Infolge der Einführung der „New Deal"-Maßnahmen nahm die Anzahl wissenschaftlicher Auftrags-Evaluationen in den 30er Jahren kontinuierlich zu (Stockmann 2006, S. 23). Geprägt wurde die weitere Entwicklung vor allem durch historische und sozialpolitisch hervorgerufene Veränderungen in den USA nach dem zweiten Weltkrieg[3]. Als Konsequenz aus dem „Sputnikschock" wurde beginnend mit Ende der 50er Jahre mehr in den Bildungssektor investiert, da die Angst weit verbreitet war, im Kalten Krieg mit der Sowjetunion im technologischen Wettbewerb abgehängt zu werden. So wurden im Bildungsbereich die Curricula in Schulen grundlegend überarbeitet. Im Zuge dieser Reformbewegungen im Bildungsbereich gewann die Evaluationstätigkeit an Bedeutung. Im Gegensatz zur Arbeit von Sozialforschern wie Ralph Tyler, der lokal begrenzt in einzelnen Schulen gearbeitet hatte, wurde nun staaten- und gar bundesweit in Hunderten von Schulen gleichzeitig evaluiert. In der Folge wurden an Universitäten Evaluationslehrstühle eingerichtet und erste Publikationsorgane (z.B. die Fachzeitschrift „Evaluation Review") geschaffen[4].

In den 60er Jahren wurde mit den Reformprogrammen unter Kennedy und Johnson die Verbesserung der Lebens- und Arbeitsbedingungen benachteiligter Bevölkerungsgruppen angestrebt. Präsident Johnson rief den „War on Poverty" aus, gefolgt von milliardenschweren Reformprogrammen zur Verbesserung der Chancengleichheit[5]. Kernstück des „War on Poverty" war der Economic Opprotunity Act von 1964. In dieser Gesetzesinitiative von Präsident Johnson waren die

[3] Nachdem die damalige UdSSR erfolgreich den ersten Satelliten in eine geostationäre Umlaufbahn um die Erde gebracht hatte, starteten die Vereinigten Staaten ein ganzes Initiativen-Paket (viele umgesetzt durch das amerikanische Verteidigungsministerium), um die Vormachtstellung bezüglich der technologischen Innovation der damaligen Zeit zurück zu gewinnen.

[4] Bemerkenswert ist dabei der schon von Beginn an feststellbare hohe Professionalisierungsgrad: die *American Evaluation Association (AEA)* ist die größte professionelle Vereinigung von Evaluationsforschern in der Welt.

[5] Das Programm entstand eher als Schlusspunkt einer Kette von Ereignissen, die sich in den 50er Jahren aufstauten: die Entscheidung des Supreme Court gegen eine Desintegrationspolitik 1954 und „The other America" von M. Harrington, in dem Schattenseite des amerikanischen Wachstums beschrieben wurde (Hellstern u. Wollmann 1984).

Hauptinstrumente zum "War on Poverty" enthalten, darunter Bildungsmaßnahmen für den amerikanischen Mittelstand und Vergünstigungen für kleine und mittelständische Unternehmen. Die Maßnahmen sollten Massenarbeitslosigkeit und Armut von Grund auf bekämpfen.

Durch diese Bündelung von Maßnahmen im sozialen Bereich eröffnete sich für Evaluationsaufträge und für die Evaluationsforschung ein riesiges Betätigungsfeld. Verbunden mit Evaluation war das vordergründige Interesse der Verwaltung, genaue Informationen über die Ausgaben und Mittelverwendung der „War on Poverty"-Projekte zu erhalten. Zudem war von Bedeutung, inwieweit die Projekte auch einen Nutzwert entfalten. Das populärste Programm dieser Zeit und im Fokus evaluativer Untersuchungen war „Head Start"[6]. Bei „Head Start" handelt sich um ein kompensatorisches Vorschulerziehungsprogramm mit intensiver Elternbeteiligung. Psychologen und Sozialforscher der 60er Jahre wiesen in ihren Studien auf den positiven Zusammenhang zwischen Frühförderprogrammen und der Entwicklung von kognitiven und sozial-emotionalen Fähigkeiten bei Kindern aus Haushalten mit geringem Einkommen hin. Nach „Head Start" hat in den USA die Evaluationstätigkeit im Bildungsbereich stark zugenommen. In den folgenden Jahrzehnten sind mehrere umfangreiche Programme zur Förderung der sozial benachteiligten Bevölkerungsgruppen entstanden (z.B. „Even Start"), die seitdem zusätzlich zum weiterentwickelten „Head Start"-Programm angeboten werden.

Zunächst auf die Evaluation weniger Programme beschränkt, haben Evaluationsvorhaben in den USA seitdem stark in Quantität und Komplexität zugenommen. Die Evaluationsforschung, als eigenständiger Zweig in den Sozial- und Geisteswissenschaften, der sich mit der Erforschung der Evaluationstätigkeit befasst, konnte sich in den USA als eigenständige Disziplin etablieren und blickt auf eine lange Tradition zurück.

Derlien (1997) unterscheidet drei Phasen der Institutionalisierung von Evaluation im internationalen Kontext. Die Zunahme von Evaluationstätigkeit ist nach Derlien auch mit einer wellenartigen Entwicklung vergleichbar. Zu der ersten Phase der Entwicklung zählt Derlien Länder wie die USA, Deutschland, Schweden und Großbritannien. Diese Länder zeichnen sich durch einen zeitlichen Vorsprung bei der Einführung von Evaluationen aus. Die zweite Phase umfasst eine Zunahme der Evaluationstätigkeit in den Ländern Dänemark, Frankreich, die Niederlande, Norwegen und der Schweiz, in denen sich Evaluation in den 80er Jahren etablieren konnte. In die dritte und letzte Phase der Zunahme lassen sich hauptsächlich die übrigen europäischen Länder zählen.

[6] Auf das Programm „Head Start" wird im folgenden Kapitel im Zusammenhang mit Evaluationsansätzen detailliert eingegangen.

Die frühe Erfahrung mit umfangreichen Evaluationen zur Unterstützung von Entscheidungsprozessen auf Regierungsebene hat einen entscheidenden Teil dazu beigetragen, dass Evaluation innerhalb der Verwaltung – und hier vor allem in den Rechnungshöfen – stark institutionalisiert ist (Meyer 2002). Zwar kann keine direkte inhaltliche Verbindung zwischen der amerikanischen Entwicklung und der in anderen Ländern hergestellt werden, aufgrund des großen Vorsprungs durch Erfahrungswissen ist aber davon auszugehen, dass andere Länder dieses Know-how für sich nutzen konnten.

Derlien sieht Deutschland zusammen mit Schweden und Kanada in einer Gruppe von Ländern, in denen es seit mehreren Jahrzehnten Erfahrungen mit Evaluationen gibt, aber eine vergleichbar hohe Institutionalisierung wie in den USA noch nicht erreicht wurde. Die Entwicklung von Evaluation in den genannten Ländern dürfte nach Derlien (1990) mit wirtschaftlichen, politischen und verfassungsrechtlichen Situationen zusammenhängen, unter denen die Regierungen in den 60er und 70er Jahren zu wirken hatten: Die „...Gruppe von Ländern erhielt den Impuls zur Programmevaluation in einer Situation wirtschaftlichen Wachstums und wachsender Budgets, die die Regierungen in die Lage versetzten, kostspielige Sozialprogramme sowie Reformen in Bildungs- und Gesundheitswesen in Angriff zu nehmen“ (Derlien 1990, S. 6). Die Unsicherheit, wie die innovativen Programme operieren und welcher Output zu erwarten ist, führte zu einer erhöhten Nachfrage nach Wirkungskontrollen. Die Rolle der Evaluation bestand darin, die Funktion der Programme zu überprüfen, um diese effektiver zu gestalten (vgl. Stockmann 2006, S. 28). Das positive Klima gegenüber Evaluationen wurde zudem durch Initiativen der regierenden Parteien begünstigt. Die sozialdemokratischen Parteien in Deutschland und Schweden förderten die Durchführung von Evaluationen. Der neu definierte Informationsbedarf „fand zum einen in der Installierung von verwaltungsinternen Informations-, Indikatoren- und Berichtssystemen seinen Ausdruck, worin sich vor allem die kommunale Ebene als Schrittmacher erwies“ (Hellstern & Wollmann 1984, S. 9). Wollmann zufolge äußerte sich der Bedarf nach Evaluationen bereits durch die Reformwelle der 60er Jahre (Wollmann 2005a, S. 3f.). Das Verwaltungshandeln orientierte sich zunehmend an neuen Managementmodellen und ihren zugrunde liegenden Bewertungszyklen: Zielsetzung, Umsetzung und Ergebnis (vgl. Wollmann 2005a, S. 2). Im verwaltungspolitischen Handeln eingesetzte Evaluationen erhielten dadurch eine umfassende Steuerungsfunktion. Evaluation wurde Anfang der 70er Jahre im Rahmen von Haushaltskonsolidierungen erstmals im politisch-administrativen Bereich vorgeschrieben (Lachenmann 1977, S. 61).

Die Evaluationsdiskussion der 60er und 70er Jahre war an Prozesse der Planung und Programmentwicklung im Kontext staatlicher Steuerungsfunktionen orien-

tiert, Ende der 70er standen Fragen der Legitimation kostenintensiver Reformprogramme und der Verwendung staatlicher oder kommunaler Budgets stärker im Mittelpunkt (Haubrich und Lüders 2003, S. 9f.; Wollmann 2005a, S.3f.).

Im Bildungsbereich kamen in den 70er Jahren vor allem in zwei Bereichen Evaluationen zum Einsatz: in der Schul- und in der Curriculumforschung (Höhne 2005, S. 14)[7]. In der Schulforschung betraf dies den Bereich Gesamtschulen. Die Diskussion über eine radikale Reform des Bildungssystems entwickelte sich bereits in den 60er Jahren und erreichte schnell eine Ebene der starken konzeptionellen und ideologischen Auseinandersetzung. Die grundsätzliche Fragestellung lautete in diesem Zusammenhang, ob das traditionelle „geschichtete" Schulsystem (Grundschule und Gymnasium) beibehalten, oder ob dieses durch Gesamtschulen ersetzt werden sollte. Infolge dieser Überlegungen startete in den 70er Jahren eine Reihe von Evaluationen mit experimentellem Anspruch, die das Ziel hatten, die Ausbildungseffektivität beider Systeme miteinander zu vergleichen. Die Evaluationen lieferten sehr heterogene Ergebnisse, so dass „hinsichtlich des methodischen Ansatzes, der Vorgehensweise und der eingesetzten Datenerhebungsmethoden (...) kein abschließendes Urteil über die Effektivität der neuen Schulform gegenüber dem traditionell gegliederten Schulsystem möglich war" (Stockmann 2006, S. 26). In politischer Hinsicht hatte das Gesamtschulenkonzept jedoch Erfolg. Die starke öffentliche Wahrnehmung der Debatte führte zur Einführung der Gesamtschulen in den meisten Bundesländern als eine Alternative zum traditionellen Schulmodell.

Im Bereich der Bildungspolitik kam es dazu noch zu einer anderen Entwicklung: Modellprogramme, die wissenschaftlich begleitet wurden, gewannen zunehmend an Bedeutung. Als „Ausgangspunkt für sozial- und fachpolitische Innovationen" (Haubrich und Lüders 2003, S. 9) bieten sich Modellprogramme nahezu an, politische und theoretische Überlegungen hinsichtlich der Praxistauglichkeit zu testen.

In den späten 70er Jahren bot sich eine völlig veränderte wirtschaftliche Situation. Noch vor der Einführung des New Public Management (NPM) wurden Methoden der Haushaltskürzung eingeführt. Der Fokus von Evaluationen wurde umgelenkt: Es wurde nicht mehr der Versuch unternommen, bestehende Programme effektiver zu gestalten, sondern im Sinne einer Kontrollfunktion von Evaluation ineffektive zu erkennen, um die Legitimation für Budgetkürzungen zu erhalten (Derlien 1997, S. 6). Die zunehmende Finanzknappheit der öffentlichen Haus-

[7] Höhne unterscheidet im historischen Sinne vier Phasen der Evaluationsforschung entlang thematischer Schwerpunktsetzungen: 1) Verdichtung des Kontrollwissens im 19. Jahrhundert, 2) Taylorismus und Testpsychologie, 3) Bildungsreform, Humankapital und Curriculum, und 4) Die Durchsetzung des Neoliberalismus (vgl. Höhne 2005, S. 13ff.).

halte führte dazu, dass Evaluation auf nahezu alle politikrelevanten Bereiche ausgedehnt wurde. In den 80er und 90er Jahren wurde durch die Einführung des NPM der Fokus verstärkt auf Performance-Messungen und die Verwendung von Leistungsindikatoren gesetzt. Ziel war hierbei die Steuerung des Verwaltungshandelns über Ergebnisse: „Management by Results" (Kuhlmann 2005, S. 7).

In den 90er Jahren setzte sich diese Orientierung durch neue Konzepte politischer Steuerung (New Public Management, Neue Steuerungsmodelle) weiter durch (Leeuw et al. 1999). Von Kardorff (2006) spricht von einem Paradigmenwechsel weg von der zentralisierten Steuerungskonzeption hin zu dezentralen und indirekten Anreizen („Fördern und Fordern"). Evaluation wurde vermehrt zum gesetzlichen Pflichtbestandteil in der Strukturpolitik, im Gesundheitswesen, in der Schul- und Hochschulpolitik und der Kinder und Jugendhilfe. Evaluation wurde immer mehr zur „Selbstverständlichkeit" (Haubrich & Lüders 2003).

Seit den 90er Jahren wurde auch in Deutschland im regionalen und kommunalen Bereich vermehrt in die schulische und vorschulische Förderung von benachteiligten Personengruppen investiert. Die Folge der kommunalpolitischen, von NGOs initiierten, gemeinnützigen oder privaten Anstrengungen ist eine seit der damaligen Zeit stark wachsende Zahl von Fördermaßnahmen[8]. Mit der steigenden Zahl an Fördermaßnahmen hat die Evaluation dieser Programme an Bedeutung gewonnen. Da es sich bei vielen Programmen um Drittmittel-Projekte handelt, besteht ein natürliches Interesse der Geldgeber, Details über die Verwendung der Mittel zu erfahren.

Neben Auftragsevaluationen sind groß angelegte Studien im Bildungsbereich zu nennen, die gemäß der in der Einleitung getroffenen Definition sowie den verschiedenen Funktionen auch als Evaluationsstudien im großen Maßstab bezeichnet werden können. Bekannte Beispiele für groß angelegte Evaluationen im Bildungsbereich, die eine stärkere öffentliche Wahrnehmung erreicht haben, sind die internationale PISA-Studie (Baumert et. al. 2000, Prenzel et. al. 2006, OECD 2001a, 2001b), das Nationale Bildungspanel (NEPS) (z.B. Blossfeld, Roßbach & von Maurice 2011) sowie das aktuelle Integrationspanel zur Überprüfung der Wirksamkeit von Integrationskursen (Rother 2008, 2009). Hier muss jedoch einschränkend gesagt werden, dass es sich bei der PISA-Studie sowie dem Nationalen Bildungspanel um breit angelegte Forschungsvorhaben handelt, die nicht auf die Analyse von einzelnen Programmen abzielen. Anhand der Ergebnisse lassen sich Rückschlüsse über die Effektivität von nationalen bzw. bundesweiten (Aus-) Bildungssystemen erarbeiten.

[8] Ein Beispiel ist die Zusammenstellung „Projektjahrbuch 2011. Potenziale Nutzen – Integration Fördern" des Bundesamtes für Migration und Flüchtlinge (Bamf): http://www.bamf.de/SharedDocs/Meldungen/DE/2012/20120706-projektjahrbuch-2011.html (letzter Zugriff: 06.12.16).

2.2. Ausmaß der Evaluationstätigkeit sowie der Professionalisierung in Deutschland

Eine objektive und verlässliche Quelle mit Zahlenmaterial zur Abschätzung des Entwicklungsstandes der Evaluationstätigkeit in Deutschland existiert zwar nicht, es können aber Informationen aus recherchierbaren Datenbanken zu Modellvorhaben und Fördermaßnahmen im Bildungsbereich zusammentragen werden. Recherchen in Projektdatenbanken geben Anlass zur Vermutung, dass Evaluationsstudien in Deutschland oft als Auftragsevaluationen vergeben werden. Der Arbeitsauftrag kann z.B. lauten, Wirkungen von Fördermaßnahmen zu identifizieren und die Ergebnisse als Entscheidungsgrundlage dafür heranzuziehen, ob die evaluierten Programme gestoppt, fortgeführt oder weiter ausgebaut werden sollen. Weitere Ziele, die mit Auftragsevaluationen verknüpft werden können, sind – bei nachweislich effektiven Programmen – die Prüfung der Ausweitung bzw. Übertragbarkeit auf andere Standorte. Zudem besteht der Wunsch besonders auf politischer Ebene, die Programme hinsichtlich ihrer nachhaltigen Wirkungsweise zu untersuchen.

Mit einer bestimmten Plausibilität kann angenommen werden, dass sich ein Großteil der existierenden Evaluationsstudien einem engen und sehr spezifischen Untersuchungsgebiet widmet bzw. gewidmet hat. Eine Internetrecherche unterstützt diese Vermutung: So ergab eine Suchabfrage auf der Internetseite des Deutschen Bildungsservers unter den Schlagwörtern „Evaluation“ und „Empirische Untersuchung“ im Jahr 2013 insgesamt 99 Treffer[9]. Dabei handelte es sich größtenteils um kleine Studien im Bereich der Bildung und Bildungsförderung aus den letzten 15 Jahren. Die Studien wurden meist an universitären Instituten oder Abteilungen durchgeführt (vielfach im Rahmen von Dissertationsvorhaben oder Auftragsprojekten). Eine grobe inhaltliche Auswertung der Projektbeschreibung hinsichtlich der Charakteristika der Evaluationsprojekte ergibt folgendes Bild: Es handelt sich überwiegend um kleine Evaluationsprojekte, die für programmspezifische Zwecke konzipiert wurden, und es werden hauptsächlich qualitative Evaluationsmethoden und Auswertungstechniken genannt. Weiterhin konnten aus den Suchergebnissen keine Evaluationsstudien identifiziert werden, die für das Messen von Programmwirkungen einschlägige Evaluationsmethoden (z.B. quasi-experimentelle oder sogar experimentelle Methoden) nennen. Häufig

[9] Die Treffer wurden über den Bildungsserver (http://www.bildungsserver.de) von der FIS Bildung Literaturdatenbank abgerufen. In FIS Bildung werden Beschreibungen empirischer Forschungsvorhaben archiviert. Wissenschaftliche Einrichtungen werden regelmäßig aufgefordert neue Projekte zu benennen und den Stand von laufenden Projekten zu aktualisieren. Die thematische und disziplinäre Bandbreite der beschriebenen Projekte ist groß: Ausgehend von psychologischen und soziologischen Untersuchungen beziehen sich die Projekte auf Bildungsmaßnahmen in jedem Lebensalter.

findet sich der Hinweis, dass die Evaluationsergebnisse nicht der Öffentlichkeit zur Verfügung stehen. Neben der Tatsache, dass viele Evaluationsprojekte zum Recherchezeitpunkt noch nicht abgeschlossen waren, kann dies auch als Hinweis dafür interpretiert werden, dass es sich um Auftragsvorhaben handelt bzw. gehandelt hat.

Zu einem vergleichbaren Schluss kommt die differenzierter erfolgte Recherche von Widmer, Beywl und Fabian aus dem Jahr 2009. Hier ergibt die Suche unter dem Schlagwort „Evaluation" in den Datenbankarchiven der Länder Deutschland, Österreich und Schweiz insgesamt 2.370 Einträge (Bezugsjahre 2000-2008)[10]. Bei näherem Hinschauen kommen die Autoren zu dem Schluss: „In der Konsequenz bestehen für die Abbildung des Evaluationsgeschehens in den drei Ländern erhebliche Lücken. So gibt es für den besonders wichtigen Bereich Schule, namentlich berufliches Schulwesen, keinerlei bundesweiten Überblick, weder für Deutschland noch für Österreich, über Evaluationen auf Bundes- und Länderebene (geschweige denn kommunaler Ebene)" (Widmer, Beywl & Fabian 2009, S. 18).

Nach Brandt gibt es zudem nur wenige Metaanalysen (d.h. Evaluationsstudien über Evaluationen) in Deutschland im Bereich Evaluation (Brandt 2009, S. 89). Zum Thema Entwicklungszusammenarbeit gibt es demnach eine Analyse von Evaluationsstudien von Stockmann und Caspari (vgl. Stockmann & Caspari 1998). Des Weiteren lässt sich noch die Analyse von Becher und Kuhlmann (1995) zu Evaluationen in der Forschungs- und Technologiepolitik nennen. Beide Meta-Analysen deuten darauf hin, dass die meisten untersuchten Evaluationsstudien auf einem methodisch überschaubaren Niveau bleiben (Brandt 2009, S. 90). In einer Metaanalyse von Bergmann et al. (1998) zur Ausbildungsqualität von Evaluatoren in der Schweiz wurden problematische Aspekte hinsichtlich der verwendeten Methoden diskutiert. In der Studie wurden von den befragten Evaluatoren auch Schwierigkeiten in der Kommunikation zwischen Auftraggebern und Evaluatoren thematisiert. Insbesondere die Problematik der Abhängigkeit der Evaluatoren von den Auftraggebern wurde diskutiert, etwa bei der Vergabe von Folgeprojekten (zitiert nach Brandt 2009, S. 90, Bergmann et al. 1998, S. 22 ff.).

Die Ergebnisse der wenigen Meta-Analysen lassen die Schlussfolgerung zu, dass zwar auf der einen Seite viel über die Evaluationsmethoden und wissenschaftlichen Grundsätze bekannt ist, es aber auf der anderen Seite kein gesichertes Wissen über die Erfahrungen aus der Praxis der Evaluationstätigkeit in Deutschland gibt. Zwar lassen sich in den oben genannten Meta-Analysen einzelne Hinweise

[10] Es wurde eine Abfrage von zentralen sozialwissenschaftlichen Datenbanken mit Angaben zu Projektinformationen in den drei Ländern durchgeführt: für die Schweiz SIDOS, für Österreich FODOS und in Deutschland SOFIS (vgl. Widmer, Beywl & Fabian 2009, S. 18).

für die Evaluationspraxis finden, die Mehrheit der Hinweise wird beispielhaft genannt und sind nicht explizit als Anleitungsmodell für Evaluatoren zu verstehen.

Die Evaluationstätigkeit ist zwar in den letzten Jahren in das Bewusstsein von Kommunal- und Länderverwaltungen sowie Stiftungen gerückt, die zu dem typischen Kreis der Auftraggeber für Evaluationsstudien zählen. Will man die Ergebnisse der SOFIS-Abfrage zum Anlass nehmen, scheint die Entwicklung jedoch kaum über lokal begrenzte Evaluationsstudien spezifischer Maßnahmen hinauszugehen. Die Rechercheergebnisse können die Schlussfolgerung zulassen, dass Evaluationstätigkeit in Deutschland kein weit verbreitetes Phänomen ist, sondern eine sich noch im Wachstum und in der Weiterentwicklung befindliche Branche der Sozial- und Geisteswissenschaften. Eine stärkere Institutionalisierung der Evaluationstätigkeit (z.B. in Forschungsbereichen der Universitäten) würde für eine stärkere Transparenz von Evaluation im gesellschaftlichen und politischen Gebiet sorgen und dadurch Evaluation auch als Forschungsinstrument im Bereich der Geistes- und Sozialwissenschaften einsetzbar machen.

Trotz des konstatierten Wachstums der Evaluationstätigkeit in Deutschland liegt jedoch aufgrund der Intransparenz der gewonnenen Erkenntnisse die Annahme der Hypothese nahe, dass das Know-how über Methoden und deren Anwendung innerhalb der Sozialwissenschaften wenig Verbreitung findet. Der schwache Institutionalisierungs- und Professionalisierungsgrad in Deutschland zeichnet sich überwiegend durch das Fehlen von Formen einer strukturierten Weitergabe von Kenntnissen und Fähigkeiten und der kontinuierlichen Weiterentwicklung dieser aus. Nach Brandt (2009) ist unter Professionalisierung „ein multidimensionaler, richtungsoffener Prozess der Herausbildung einer bestimmten Organisationsform eines Tätigkeitsfeldes“ (Brandt 2009, S. 33) gemeint. Brandt unterscheidet vier Dimensionen der Professionalisierung:

1. „eine spezifische, abgrenzbare Wissensbasis;
2. der organisatorische Zusammenschluss der Akteure im Tätigkeitsbereich;
3. Standards der Evaluation;
4. vorhandene Ausbildungs- und Weiterbildungsangebote“ (vgl. Brandt 2009, S. 37).

Die erste Dimension der Wissensbasis wird durch das Know-how bestimmt, d.h. der Kompetenzen zur Anwendung von geeigneten Evaluationsmethoden und der Kompetenz, das gewonnene Datenmaterial auszuwerten und zu interpretieren. Hier wurden in Deutschland Evaluationstheorien und Methoden aus der amerikanischen Evaluationsforschung übernommen und auf den spezifischen Kontext

angewendet[11]. Im zweiten Kapitel dieser Arbeit wurden dazu eingehend zentrale Evaluationsansätze im Bereich der Sozial- und Geisteswissenschaften behandelt, die sich im Verlauf der Entwicklung der amerikanischen Evaluationsforschung herauskristallisiert haben.

Eine frühe Organisationsform von Evaluation, die Aufschluss über die Entwicklung der zweiten und dritten Dimension von Professionalisierung gibt, ist die 1997 gegründete Deutsche Gesellschaft für Evaluation (DeGEval). Ein erstes Arbeitsergebnis der Gesellschaft war die Veröffentlichung der DeGEval-Standards für die Durchführung von Evaluationsverfahren[12]. Die Standards sollen zur Professionalisierung von Evaluation beitragen und den Austausch zwischen Evaluatoren, Auftraggebern, also zwischen der Wissenschaft und der Praxis, fördern. Dadurch stellen die Standards gewissermaßen ein Rahmenwerk für die Planung und Durchführung von Evaluationen dar. Im engeren Sinne handelt es sich um Kriterien, die Evaluatoren eine Orientierungshilfe geben sollen, um in Abstimmung mit den Auftraggebern eine objektive, methodisch korrekte und hinsichtlich der Evaluationsziele nützliche Evaluation durchzuführen[13]. Ohne Professionalisierung kann Evaluation nach Haubrich und Lüders (2004) schnell als „Zauberwort" enttarnt werden. In der Konsequenz führt dies zu einer Diskrepanz zwischen den Erwartungen an Evaluation und dem tatsächlichen Leistungspotential (Haubrich & Lüders 2004, S. 316f.). Während die Wissensbasis in Deutschland vorhanden ist, fehlt es somit an Formen der Institutionalisierung und der Wissensvermittlung.

Die Entwicklung der Evaluationstätigkeit in Deutschland lässt sich kurz folgendermaßen zusammenfassen: Die wachsende Anzahl von Evaluationen ist Ausdruck dafür, dass die Bedeutung von sowie die Nachfrage nach Programmevaluation im Bildungs- und Wirtschaftssektor zugenommen hat. Trotz der Bedeutungszunahme der Evaluationstätigkeit in Deutschland beruhen die angewendeten Methoden und Verfahren zum größten Teil auf dem jahrzehntelang in den USA gereiften Erfahrungs- und Professionalisierungsgrad.

Dieser kurze Überblick zur Entwicklung der Evaluationstätigkeit verdeutlicht einen Erfahrungsvorsprung der USA vor Deutschland. Aufgrund der vergleichbar jungen Geschichte der Evaluation in Deutschland kann festgestellt werden, dass

[11] Ein prominentes Beispiel dafür ist das Lehrbuch zur Programmevaluation von Rossi, Freeman und Hoffmann aus dem Jahr 1988.

[12] Die Standrads der DeGEval sind im Internet abrufbar: http://www.degeval.de/degeval-standards. (Zuletzt geprüft: 06.12.2016)

[13] Eine der Masterstudiengang Evaluation an der Universität des Saarlandes dar (http://www.master-evaluation.de/; letzter Zugriff: 06.12.2016). Des Weiteren wird in Form einzelner Kurse der Weiterbildungsstudiengang Evaluation von der Universität Bern an der Universität Köln angeboten.

nur bei einer kleinen Anzahl der existierenden Maßnahmen im Bereich der Bildungsförderung Programmevaluationen durchgeführt wurden. Für die Analyse von Evaluationsmethoden ist es daher entscheidender, die historische Entwicklung der Evaluationsmethodik in den USA systematischer zu untersuchen. Aus der Fülle an Verfahren der Evaluation, wie sie in den USA seit Jahrzehnten im Bereich der Bildungsförderung breit eingesetzt werden, lassen sich – wie in den folgenden Unterkapiteln dargelegt wird – differenziert Erkenntnisse für die Gestaltung von Evaluationsstudien ableiten. Aus dem noch entwicklungswürdigen Professionalisierungsgrad der Evaluationstätigkeit in Deutschland lässt sich die Forderung nach praxisnaher Unterstützung von Evaluatoren bei der Umsetzung ihrer Evaluationsprojekte ableiten.

2.3. Entwicklung von Ansätzen für die Evaluation von sozialen Programmen

Die Entwicklung der Evaluationstätigkeit in den USA ist in der Fachliteratur detailliert dokumentiert. In Publikationen zur Evaluationshistorie werden mehrere voneinander zeitlich abgrenzbare Entwicklungsphasen im 20. Jahrhundert unterschieden. Eine in der Fachliteratur oft zitierte chronologische Einordnung der Entwicklung der Evaluationstätigkeit nennt drei Zeitphasen (vgl. Cook und Matt 1990), die nicht als inhaltlich voneinander trennscharf abzugrenzende Einteilungen zu verstehen sind, sondern als symbolische Darstellungsform dafür, dass die Entwicklung durch Phasen der Konzentration auf verschiedene methodische Schwerpunkte geprägt ist[14]. Angelehnt an die zeitliche Unterteilung nach Cook und Matt werden in dieser Arbeit drei Phasen mit folgenden methodischen Schwerpunktsetzungen unterschieden:

1955-1975: Evaluationsansätze für Wirkungsmessungen

1975-1982: Prozess- und nutzenorientierte Evaluationsansätze

Seit 1982: Theoriegeleitete Evaluationsansätze

Wenn man die Entwicklung der Evaluationsforschung in den USA näher betrachtet, dann wird in diesem Zusammenhang häufig von der Weiterentwicklung einer Theorie bzw. von mehreren Theorien der Evaluation gesprochen. „Die traditionelle Theorieauffassung besagt, dass Theorien logisch verknüpfte Satzsysteme bzw. Aussagensysteme sind, die aus universellen Gesetzeshypothesen bestehen und in axiomatisierter Form als vollendete Theorien dargestellt werden können“ (Lenk 1999, S. 171).

[14] Eine vergleichbare Einteilung der Evaluationshistorie in drei Zeitphasen nehmen auch andere Autoren wie beispielsweise Shadish, Cook und Leviton (1991) vor.

In der amerikanischen Fachliteratur wird ausschließlich von Evaluationstheorien gesprochen (Alkin 1994), auch wenn dies einem Vergleich mit streng wissenschaftlichen Theoriedefinitionen wie der obigen nicht standhalten würde. Bei näherer Betrachtung der „Evaluationstheorien" handelt es sich – wie es in den folgenden Unterkapiteln deutlich wird – im eigentlichen Sinn um methodische Durchführungsbeschreibungen von Evaluationen, die zumeist auf wissenschaftstheoretischen Annahmen beruhen. Auch Beschreibungen von klassischen Experimenten, wie sie in der Psychologie angewendet werden, werden in der amerikanischen Literatur häufig als Evaluationstheorien bezeichnet (vgl. Shadish, Cook & Levinton 1991, Cook & Matt 1990).

Anstatt von Theorien zu sprechen, wird im Rahmen dieser Arbeit vorgeschlagen, den Begriff **Evaluationsansatz** zu verwenden, da es sich oft um Rahmenkonzepte der Evaluation von sozialen Programmen handelt, bei denen nicht ein System von aufeinander bezogenen Aussagen oder Hypothesen zu einem bestimmten Ausschnitt der Realität im Zentrum der Untersuchung steht. Evaluationsansätze stellen dagegen ein Rahmenkonzept für die Konzeption und Durchführung von Evaluationsstudien sowie für die Verwertung der Ergebnisse dar.

Beywl, Speer und Kehr bieten folgende Definition für den Begriff Evaluationsmodell an, der synonym zum Begriff des Evaluationsansatzes verwendet werden kann: „Unter einem Modell der Evaluation wird eine ausformulierte, theoretisch begründete und durch praktische Evaluationserfahrungen gesättigte Anleitung verstanden, wie praktische Evaluationen geplant und durchgeführt werden sollen. Modelle sind oft mit Namen bestimmter Evaluationstheoretiker/innen verbunden, welche diese erstmals konkretisiert und weiter entwickelt haben" (Beywl, Speer & Kehr 2003, S. 68). Evaluationsmodelle bzw. Evaluationsansätze können auf verschiedenen wissenschaftstheoretischen Hintergrundüberlegungen beruhen und lassen sich, bildlich gesprochen, wie bei einem Container mit Evaluationsmethoden bis hin zu konkreten Evaluationsdesigns befüllen. In dieser Arbeit wird daher der Begriff Evaluationsansatz durchgängig verwendet, um zwischen den verschiedenen theoretischen Herangehensweisen an Evaluation zu differenzieren.

Wie aus den folgenden Unterkapiteln noch deutlich hervorgehen soll, hat sich die Entwicklung der Evaluationstätigkeit als ein kontinuierlich voranschreitender Prozess dargestellt, bei dem die einzelnen Entwicklungsschritte auf den zuvor gewonnenen Erkenntnissen und Erfahrungen beruhen und darauf aufbauen. Alle noch zu beschreibenden Evaluationsansätze haben gemeinsam, dass sie nicht auf einer durch empirische Überprüfung allgemein anerkannten Theorie beruhen: „Thus, in the strictest sense, what we will refer to as evaluation theories do not fully qualify for that status" (Alkin 2004, S. 5). Die Betrachtung der histo-

rischen Entwicklung der Evaluationsforschung hat vielmehr gezeigt, dass zahlreiche Evaluationsansätze entstanden sind, die sich hinsichtlich ihrer Zielsetzung, den Funktionen und der methodischen Schwerpunktsetzung unterscheiden lassen.

Wissenschaftliche Publikationen zur Evaluationstätigkeit in den USA in den 60er und 70er Jahren sind stark durch die Arbeit von wenigen, prominenten Evaluationstheoretikern geprägt. Im Folgenden werden die Hauptcharakteristika und Grundzüge der Ansätze einer Auswahl von Evaluationstheoretikern mit dem Ziel beschrieben, den *Entwicklungspfad der Evaluationsmethode* aufzuzeigen. Die Erörterung der vorgestellten Evaluationsansätze orientiert sich an den folgenden Unterscheidungsmerkmalen:

- Abdeckungsgrad der Evaluationsstudien (Fokussierung auf einzelne Elemente im Evaluationsprozess vs. umfassender Ansatz); Art und Weise der Berücksichtigung von Rahmenbedingungen;
- Ansatztypen (abhängig vom Rollenverständnis, Methoden nach Alkin): methodenzentriert, nutzer- und verwertungsorientiert;
- Funktion der Evaluation (Typisierung nach Chelimsky): Erkenntnisgewinnung, Kontrolle, Entwicklung und Weiterentwicklung;
- Rollenverständnis des Evaluators;
- Ergebnisverwertung und Nutzen der Evaluation.

Die Vorstellung der Evaluationsansätze folgt grob einer chronologischen Entwicklung und wird sich in den folgenden Kapiteln auf die Darstellung der Verfahren und Methoden konzentrieren, um in einem Abschlusskapitel die Schwerpunktsetzungen und Unterschiede zwischen den diskutierten Ansätzen zu skizzieren. Geeignete Konzepte und bewähre Elemente der diskutierten Evaluationsansätze sollen im darauf folgenden Kapitel für die Konstruktion des Prozesses für die Begleitung von Evaluationsprojekten herangezogen werden – das Hauptprodukt der vorliegenden Arbeit.

2.3.1. Erste Phase: Evaluationsansätze für Wirkungsmessungen

Bei der Beschäftigung mit der Arbeit von prominenten Evaluatoren der 50er und 60er Jahre wird zunächst ersichtlich, dass die Konzepte und Verfahren für Programmevaluationen der damaligen Zeit durch die wissenschaftsphilosophische Richtung des logischen Positivismus (oder auch logischer Empirismus) geprägt sind. Eine Annahme des logischen Positivismus lautet, dass sich durch Erhebungsverfahren – wie z.B. der systematischen Beobachtung – objektive Erkenntnisse über die Realität gewinnen und zu Erfahrungswissen zusammenführen lassen. Eine Maßnahme oder ein soziales Programm, dass sich nach Durchlaufen

eines Evaluationsverfahrens und strengen empirisch-methodischen Bewertungskriterien als „erfolgreich“ herausstellen würde, sollte in der Breite umgesetzt und Praktikern in der öffentlichen Verwaltung sowie Projektmanagern als Instrument der Problemlösung angeboten werden.

Zwei Vertreter der Evaluationsforschung, die sich in ihrem methodologischen Verständnis dem logischen Positivismus nahe fühlen, sind Michael Scriven und Donald T. Campbell. Sie vertreten strenge methodologische Standards von Evaluation. Beide – sowohl Evaluationspraktiker als auch Evaluationstheoretiker in den 60er Jahren – prägten die Entwicklung der Evaluationsmethodik insbesondere in der Psychologie. Sowohl Scriven als auch Campbell setzen in ihren Evaluationsansätzen den Schwerpunkt auf die Generierung von Erkenntnissen durch den Einsatz geeigneter Evaluationsverfahren. Evaluationsverfahren müssten den höchsten Standards der empirischen Sozialforschung entsprechen, wobei die stärksten Methoden zur Identifikation von Programmeffekten gerade gut genug seien. Im Gegensatz zu Campbells starker Schwerpunktsetzung auf die Verwendung der richtigen Methoden in Evaluationsvorhaben ist Scrivens Hauptbeitrag zur Evaluationsmethodik die differenzierte Analyse unterschiedlicher Formen und Zielsetzungen von Evaluation.

Nach Scriven besteht das Wesen eines Evaluationsvorhabens in der Bewertung eines Gegenstandes oder eines Sachverhaltes. Scrivens Definition von Evaluation lautet: „’evaluation’ refers to the process of determining the merit, worth or value of something, or the product of the process” (Scriven 1991, S. 139). Dieser Definition von Evaluation entspricht es demnach, die Bewertung eines Evaluationsgegenstandes mit objektiven Verfahren durchzuführen. Daher seien allen Evaluationsmethoden Vorzug zu geben, die nach strikten wissenschaftstheoretischen und logischen Regeln Informationen und Daten generieren. Teilnehmerorientierte Beurteilungsverfahren (z. B. subjektive Urteile von Programmteilnehmern) seien dagegen in den Hintergrund zu stellen. Scriven trennt in seiner Definition zwischen der Funktion, die Evaluationsstudien zu erfüllen haben und dem Aspekt, für was die Evaluationsergebnisse verwendet werden sollen: „Evaluation is what it is, the determination of merit or worth, and what it is used for is another matter“ (Scriven 19866, S. 7).

Zu dem Beitrag von Scriven zählt auch, zahlreiche Konzepte und Begrifflichkeiten im Bereich der wissenschaftlichen Evaluationstätigkeit eingeführt zu haben. In seinem bereits in den 60er Jahren verfassten Artikel „Die Methodologie der Evaluation“, der später 1972 auch in deutscher Sprache erschienen ist (Wulf 1972), geht Scriven auf zwei Grundkonzepte von Evaluation ein, die bis in die heutige Zeit in der Praxis Anwendung finden und die Frage der Verwendung von Evaluationsergebnissen thematisieren. Es handelt sich dabei um die **formative und summative Art der Evaluationsdurchführung**.

Im Gegensatz zu den Zielen von Evaluation kann die Rolle der Evaluation je nach Untersuchungsgegenstand und Anforderung der Stakeholder oder Auftraggeber sehr unterschiedlich sein. Evaluationsergebnisse können zur direkten und unmittelbaren Verbesserung von sozialen Programmen verwendet werden (vgl. Evaluation mit der Funktion der Entwicklung (Chelimsky 1997)). Diese Form wird von Scriven als **formative Evaluation** bezeichnet (Scriven 1972, S. 61). Die Rolle des Evaluators ist es dabei, Programmverantwortliche bzw. Auftraggeber von Evaluationen durch geeignete evaluative Verfahren kontinuierlich mit Evaluationsergebnissen zu versorgen. Die zusammengetragenen Erkenntnisse bleiben nach Scriven innerhalb der Programmentwicklung und dienen zur Verbesserung des Programms. Die formative Evaluation eignet sich daher vor allem zur stetigen Verbesserung von Programmen oder Projekten, die sich erst in der Modellprojektphase befinden und daher weiterer konzeptioneller Optimierung bedürfen (Scriven 1972, S. 62)[15].

Eine komplett andere Rolle spielt nach Scriven die **summative Evaluation**. Sie soll Entscheidungsträger mit den nötigen Informationen versorgen, ob das Programm überhaupt Wirkungen zeigt, die eine Fortführung rechtfertigen würde (Scriven 1972, S. 62 f.). Die Ergebnisse summativer Evaluationen bieten eine Gesamtbetrachtung der Wirkungsweise von Projekten zu einem festgelegten Zeitpunkt. Sie sind dazu gedacht, „alle Effekte eines Programms zu fassen" (Cook & Matt 1990, S. 17)[16]. Projektkoordinatoren und -planer nutzen beispielsweise die positiven Erkenntnisse summativer Evaluationen, um für die Fortsetzung des Projekts auf politisch-administrativer Ebene zu werben. Scriven kritisiert in seinem Artikel „Die Methodologie der Evaluation" die Position von Cronbach, der aussagt, dass „der größte Beitrag, den Evaluation leisten kann, darin liegt, die Aspekte des Curriculums eines Programms herauszuarbeiten, für die eine Neubearbeitung erforderlich ist" (Cronbach 1963, S. 41). Cronbachs pragmatische Position gibt dadurch der formativen Rolle von Evaluation einen höheren Stellenwert als der summativen. Scriven betont jedoch die summativen Rolle von Evaluation, da nur diese im Endergebnis Aussagen über den tatsächlichen Nutzen eines Programms und damit über die Legitimation der Weiterführung ermöglicht (vgl. Stufflebeam & Shinkfield 2007, S. 374; Cook & Matt 1990, S. 18 f.). Angelehnt an die Begriffe der formativen und summativen Evaluation unterscheidet

[15] Das im Laufe dieser Arbeit vorgestellte Modellprojekt „In Deutschland zu Hause" (ein Integrationskursangebot für Migranten) stellt ein Beispiel für eine stark formativ verlaufene Evaluationsstudie dar. Die Ergebnisse wurden nahezu ausschließlich zur Optimierung des Konzepts verwendet.

[16] Die noch im folgenden Kapitel näher charakterisierte Evaluationsstudie zum Projekt *frühstart* beinhaltete eine Sprachstandsmessung bei Kindern im Kindergartenalter und kann als Beispiel für summative Evaluation genannt werden.

Scriven des Weiteren die **intrinsische und die Ergebnisevaluation**. Bei der **intrinsischen Evaluation** sind es die Inhalte und die Instrumente eines Programms, die der Evaluation unterzogen werden. Die **Ergebnisevaluation** beschränkt sich auf die Erfassung von Effekten bzw. Wirkungen eines Programms[17]. Mit der Unterscheidung beider Begriffe soll nochmal betont werden, dass in Abhängigkeit von den Evaluationszielen unterschiedliche Verfahren der Evaluation anzuwenden sind.

Neben verschiedenen Rollen und Formen von Evaluation unterscheidet Scriven zwischen Evaluationsuntersuchungen, die sich auf die Erfassung des Nutzens und der Leistung eines Programms konzentrieren, und Prozessuntersuchungen (Scriven 1972, S. 69 ff.). Von der formativen Evaluation unterscheiden sich Prozessuntersuchungen grundsätzlich dadurch, dass sich Letztere auf die Erfassung des Ablaufs eines Programms beschränken. Während der Evaluator durch die formative Evaluation eine aktive und gestaltende Rolle im Evaluationsverfahren einnimmt, werden in Prozessuntersuchungen nur die Programmabläufe systematisch erfasst und einem Bewertungsprozess unterzogen. Daher geben Prozessuntersuchungen direkt keine Auskunft über den Nutzen oder die Leistung eines Programms insgesamt (Scriven 1972, S. 70)[18].

Beywl (1988) stellt fest, dass Scriven eine generelle Frage aufgeworfen hat, die von Suchman (1967), einem Vertreter der frühen Evaluationsforschung, bereits zu einem früheren Zeitpunkt thematisiert wurde: Wenn Evaluation von Nutzensabwägungen bei sozialen Programmen ausgeht, die den Zielen von Aktivitäten zugewiesen werden, soll sie dann auch zur Weiterentwicklung von auf gesellschaftlicher Ebene geführten Nutzendiskussionen beitragen? Suchman verstand Ende der 60er Jahre Evaluation als eine Form der angewandten Forschung und prägte, um diesem Zusammenhang besondere Bedeutung zu verleihen, den Begriff *evaluative research*. Seine Definition von Evaluation war lange Zeit prägend für die gesamte Forschungslandschaft: „Evaluation is the social process of making judgements of worth" (Suchman 1967, S. 7). Die ursächliche Erklärung von Phänomenen sowie das Testen von Theorien durch die Methoden der Evaluation sei gemäß Scriven nicht die Aufgabe von Evaluationen sondern der Sozialwissen-

[17] In vielen Evaluationsansätzen werden vergleichbare Differenzierungen zwischen der Evaluation von Wirkungen und der Evaluation der Prozesse oder der Inhalte geleistet, die zum Ziel des Projekts führen sollen. Donald Stufflebeam unterscheidet beispielsweise in seinem Evaluationsansatz „CIPP" prinzipiell zwischen vier verschiedenen Gegenständen der Evaluation: concept, input, process und product (vgl. Stufflebeam 1983, S. 117-141).

[18] Prozessforschung ist nach Scriven aus den genannten Gründen nicht immer Teil von Evaluationsforschung. Scriven sieht die Evaluationsforschung als eine Methode zur Identifikation des Nutzens eines Programms. Dabei ist die Anwendung von Methoden der empirischen Sozialforschung fundamental (vgl. Scriven 1991).

schaft. Somit sei es auch nicht das Ziel von Evaluationen, Einfluss auf gesellschaftliche Diskussionen zu nehmen. Der Evaluator hat nach Scriven die Aufgabe, auf Grundlage von empirischen Verfahren Bewertungen zum Programm vorzunehmen und diese nicht den Programmverantwortlichen zu überlassen. Scriven geht es daher ausschließlich um die Bewertung des Nutzens des evaluierten Programms (vgl. Alkin 2004, S. 32).

Einen weiteren relevanten Beitrag hat Scriven betreffend der grundsätzlichen Herangehensweise an Evaluationsstudien geleistet, in dem er den Ausdruck der **zielfreien Evaluation** (goal free evaluation) in die Diskussion eingebracht hat (Scriven 1991). Das Ziel von Evaluationen sollte es im Idealfall nach Scriven sein, alle mit einem Programm verbundenen Effekte zu identifizieren und zwar unabhängig davon, ob die Wirkungen intendiert oder nicht intendiert sind. Bei der Suche nach den Effekten eines Programms können die strategischen und operativen Programmziele den Evaluator in seinem Handeln beeinflussen und von den tatsächlichen, mit dem Evaluationsverfahren zu erfassenden Effekten des Programms ablenken. Ziele werden von Programminitiatoren häufig pauschal formuliert und sind durch politische Bestrebungen beeinflusst (vgl. Cook & Matt 1990, S. 19; Youker & Ingraham 2014, S. 54).

Die Objektivität von Evaluationen ist für Scriven die wichtigste Voraussetzung für wissenschaftlich fundierte und zuverlässige Ergebnisse. Bei der Entwicklung des Evaluationsdesigns und bei der Datenerhebung sollen mit den Möglichkeiten der empirischen Sozialforschung nur diejenigen Methoden ausgesucht werden, die den kleinstmöglichen Spielraum für Verzerrungen mit sich bringen. Experimentelle Designs mit randomisierten Versuchsanordnungen und quasi-experimentelle Designs mit Kontrollgruppen gehören für ihn zu den besten Methoden, um zuverlässig valide Aussagen über die Leistung und den Wert eines Programms zu erhalten (vgl. Scriven 1972, S. 85 ff.).

Neben Scriven war Donald T. Campbell ein wichtiger Vertreter der methodisch ausgerichteten Evaluationsforschung in den 60er Jahren. In den folgenden Jahrzehnten hat Campbell einen substantiellen Beitrag zur Weiterentwicklung der Methodik von Experimenten bei Programmevaluationen geliefert. Mehrere Standardwerke zur Methodenlehre sind aus dieser Arbeit hervorgegangen, die bis in die heutige Zeit an Aktualität nichts eingebüßt haben, allen voran das methodische Werk zur Planung und Umsetzung von **experimentellen und quasi-experimentellen Untersuchungen** unter dem Titel: *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Campbell, Cook, Shadish 2001), das seit 1963 in zahlreichen Neuauflagen unter wechselnden Ko-Autoren erschienen ist. In den deutschen Evaluationsansätzen von Lange (1983) oder Tulodziecki (1982) findet dieser Ansatz starke Anwendung (vgl. auch Hinweise in Hellstern & Wollmann 1984). In Campbells und Stanleys Ur-Ausgabe aus dem Jahr 1963

beschreiben die Autoren die notwendigen Voraussetzungen, um eine experimentelle Untersuchung mit zufälliger Auswahl der Teilnehmer im sozialen Bereich durchzuführen. Bis in die 60er Jahre wurden in den Vereinigten Staaten wenige Untersuchungen mit strengen experimentellen Anordnungen durchgeführt (vgl. Oakley 1998). Bis zu dieser Zeit fanden Experimente hauptsächlich in der Medizin und in der Psychologie Anwendung. Campbell und Stanley haben durch ihre Arbeit an der Weiterentwicklung der Evaluationstheorie einen großen Beitrag zur Einführung von experimentellen Versuchsanordnungen im Sozialplanungsbereich geleistet. Die Autoren entwickeln Regeln und Verfahrensabläufe für Experimente, um die Wirkungen von sozialen Programmen in den Bereichen Psychologie, Soziologie und Pädagogik festzustellen.

In den USA finden seitdem experimentelle und quasi-experimentelle Evaluationsmethoden häufig Anwendung, beispielsweise in der Bildungsforschung zur Untersuchung von Förderprogrammen[19]. Nach dem Vorbild der klinischen Psychologie wird die Methodik von Laborexperimente als methodischer „Goldstandard" auf die Untersuchung sozialer Phänomene übertragen. Die objektiven Erkenntnisse können dann – nach bestimmten Maßgaben – verallgemeinert werden und als empirische Grundlage für die Weiterentwicklung des sozialwissenschaftlichen Verständnisses von der Gesellschaft verwendet werden. Als Vertreter der evolutionären Erkenntnistheorie vertrat Campbell die Einsicht, dass die Lösung von manifesten sozialen Problemen erst durch das Testen verschiedener alternativer Lösungswege und die Auswahl der effektivsten Maßnahme möglich ist (Campbell 1974). Campbell beschreibt den Zweck von Evaluationen sowie die Umsetzung dieses Ansatzes in dem Artikel „Reform as Experiments" (vgl. Campbell 1969).

Trotz der Bewertung der randomisierten experimentellen Versuchsanordnungen als so genannte „Gold Standards" in der Evaluationsforschung bieten Campbell und Stanley eine Lösung für Situationen an, in denen richtige Experimente weder durchführbar noch gewünscht sind. Als Alternative zum randomisierten Experiment schlagen beide die Anwendung von quasi-experimentellen Designs in der Sozialforschung vor. Diese unterscheiden sich im besten Fall von echten Experimenten nur durch die nicht-zufallsbedingte Auswahl der Einheiten für die Untersuchungs- und Kontrollgruppe. Während die Qualität von randomisierten Ex-

[19] Einen exemplarischen Überblick zu evaluierten Programmen im Bereich der vorschulischen Bildungsförderung in den USA bietet die Zusammenstellung von Chambers et al. (2010). Die darin genannten Programme wurden hinsichtlich Wirkungen mit verschiedenen methodischen Designs evaluiert; darunter befinden sich auch zahlreiche experimentelle und quasi-experimentelle Studien.

perimenten von der Kontrolle der Parameter abhängt, ist die Qualität quasi-experimenteller Designs von der Genauigkeit und Kontrolle des Forschungsdesigns abhängig.

Der Evaluationstheoretiker Suchman sah daraufhin durch die Einführung von quasi-experimentellen Designs in die „evaluative Forschung" eine Möglichkeit, dass „die selbstsüchtigen Tendenzen der Programmadministratoren durch sozialwissenschaftliche Methoden der Prüfung kausaler Beziehungen diszipliniert werden könnten" (Cook & Matt 1991, S. 19; vgl. auch Campbell & Stanley 1966). Den Zuspruch für quasi-experimentelle Designs, der von Suchman erbracht wurde, konnte jedoch Campbell in den Folgejahren weniger teilen. Es zeigte sich, dass in bestimmten Forschungsbereichen der Begriff des quasi-experimentellen Designs inflationär für jede Studie verwendet wurde, die zwar kein echtes Experiment darstellt, aber die Überprüfung kausaler Hypothesen mit einer mehr oder weniger äquivalenten Kontrollgruppe beinhaltet (vgl. insbesondere dazu den Fachartikel von Cook (Cook 2003)). Campbell und Stanley (1966) und später auch Cook und Campbell (1979) fordern seitdem einen selbstkritischen Umgang mit den Methoden in quasi-experimentellen Untersuchungen: „All quasi-experimental designs and analyses are not equal" (Cook 2003, S. 142). Aufgrund der Heterogenität der quasi-experimentellen Designs seien Ergebnisse der in der Vergangenheit durchgeführten Untersuchungen kaum miteinander vergleichbar. Zudem könnte niemals ausgeschlossen werden, dass so genannte intervenierende Drittvariablen – wie Intelligenz oder Motivation der Teilnehmer an der Untersuchung – einen Einfluss hatten. Die Einflussgröße von Drittvariablen könne nach dem Messzeitpunkt nicht mehr quantifizieren werden. Aufgrund dieser Implikationen generieren quasi-experimentelle Untersuchungen häufig Ergebnisse, die im Vergleich mit Ergebnissen von randomisierten Experimenten unbefriedigend ausfallen würden (vgl. Cook & Matt 1990, Campbell & Erlebacher 1970, Campbell & Boruch 1975).

Campbell ist vor allem an objektiven, empirischen Methoden der Sozialforschung interessiert, mit deren Hilfe sich kausale Zusammenhänge beschreiben lassen. Die Arbeit Campbells an der Weiterentwicklung der sozialwissenschaftlichen Methodik und vor allem an Experimenten hat ihn zur Einsicht geführt, dass es nicht die perfekte Evaluationsmethode für die Analyse von Wirkungszusammenhängen gibt. Campbell schlägt daher vor, die Güte von Evaluationsverfahren, in denen quantitative Methoden zur Messung von Programmwirkungen angewendet werden, durch Aussagen zur **internen Validität** der Ergebnisse zu überprüfen (Campbell & Stanley 1966). Eine experimentelle oder quasi-experimentelle Untersuchung zeichnet sich durch geringe interne Validität aus, wenn externe Einflussfaktoren auf das zu untersuchende Programm sowie das Messverfahren nicht durch den Evaluator kontrolliert wurden. Externe Einflussfaktoren (von Campbell auch „Threats to internal validity" genannt (Campbell & Stanley

1966) können sich bei Experimenten z.B. durch Reifungsprozesse bei den Teilnehmern mehreren Messzeitpunkten, unzuverlässige Messinstrumente oder durch begangene Fehler bei der Bildung von Untersuchungs- und Kontrollgruppen ausdrücken[20].

Neben interner Validität nennt Campbell das Konzept der **externen Validität**. Die externe Validität ist dann hoch, wenn mit statistischer Sicherheit die Erkenntnisse aus einer experimentellen oder quasi-experimentellen Untersuchung generalisiert, d.h. auf größere Gruppen als die Untersuchungs- und Kontrollgruppen übertragen werden können. Campbell nennt auch bei der externen Validität Einflussfaktoren, durch die eine Übertragbarkeit der Ergebnisse auf andere Kontexte gefährdet sein kann. Zu diesen zählt beispielsweise das reaktive Verhalten der Teilnehmer an einem Programm (Campbell & Stanley 1966).

Die Konzepte der internen und externen Validität, die auch als Indikatoren zur Abschätzung der Güte einer Evaluationsstudie herangezogen werden können, waren zudem Gegenstand der wissenschaftlichen Auseinandersetzung und Diskussion um die Ergebnisse der ersten Wirkungsevaluationen des Head Start-Programms in den USA gegen Ende der 60er Jahre. Das 1964 vorbereitete Programm zur kompensatorischen Förderung von Kindern aus sozial benachteiligten Familien wurde 1965 in der Regierung unter Präsident Johnson umgesetzt. Wenngleich Head Start ursprünglich als Programm zur Gesundheitserziehung entwickelt wurde, setzte es sich auch aus den Förder-Komponenten Bildung, Sozialberatung und Elternbeteiligung zusammen. Das Programm stieß bei der Einführung auf große Resonanz: so nahmen 561.000 Kinder im ersten Jahr an dem Förderprogramm teil. Infolge dessen wurde Head Start schnell ausgeweitet: aus vielen Sommerprogrammen wurden im Verlauf der ersten Jahre Programme mit einer Länge von bis zu neun Monaten. Zu den Zielen des Programms zählen die Verbesserung der kognitiven und sozial-emotionalen Fähigkeiten bei teilnehmenden Kindern, die Verbesserung der gesundheitlichen Situation, eine stärkere Einbeziehung der Eltern in die Förderung ihrer Kinder, Förderung der Arbeitsmarktintegration und Nachbarschaftshilfe[21].

Einer der Auftragnehmer der Evaluation von Maßnahmen im Zusammenhang mit dem Head Start-Programm war Ende der 60er Jahre die Westinghouse Learning Corporation (Evaluationsstudie von Cicirelli et al. 1969). Das Design der Evaluationsstudie war eine retrospektive Querschnittstudie mit einer Kontrollgruppe

[20] Die einzelnen Einflussfaktoren der internen und externen Validität werden detailliert in Kapitel drei behandelt. Zudem werden sie auch bei der Ergebnisinterpretation zu den diskutierten Evaluationsstudien im dritten Kapitel aufgegriffen.

[21] Zur Geschichte von Head Start siehe: http://www.acf.hhs.gov/programs/ohs/about/history-of-head-start. (Letzter Zugriff: 06.12.2016).

(vgl. Wu & Campbell 1996). Dazu wurden Head Start-geförderte Kinder mit Kindern in einer Kontrollgruppe verglichen, die zwar ähnliche sozio-demographische Charakteristika aufwiesen, jedoch nicht gefördert wurden. Das Ziel der Studie war es, den Einfluss des Head Start-Programms auf die kognitive Entwicklung der geförderten Kinder zu untersuchen.

Statt der erhofften durchschlagenden Wirkung zeigte die Studie der Westinghouse Learning Corporation, dass die einzelnen Maßnahmen mittel- und langfristig wenig effektiv sind, um die kognitiven Fähigkeiten der geförderten Kinder zu steigern (Zigler & Styfco 1993, S. 9). In Folge der geringen Effektstärke hatten die Maßnahmen auch keine nachhaltige Wirkung (Zigler & Styfco 1993, S. 9). Das Design der Studie sah sich in der Folge einiger Kritik aus der Bildungsforschung ausgesetzt. Vielfach wurde bemängelt, dass aufgrund der Zusammensetzung die Merkmale der Kontrollgruppen-Kinder nicht vergleichbar mit denen der Untersuchungsgruppe waren. Kinder in der Kontrollgruppe wiesen bei der genannten Studie einen durchschnittlich höheren sozioökonomischen Status aus. Zudem wurden ihre Testergebnisse erst drei Jahre nach dem Ende der Head Start-Projekte in der Evaluation für einen Vergleich mit der Untersuchungsgruppe herangezogen (vgl. Campbell & Erlenbacher 1970; Datta 1976). Die von Nixon vorzeitig an die Öffentlichkeit gebrachten, eher negativen Ergebnisse lösten eine auf wissenschaftlicher Ebene intensiv geführte Diskussion über Forschungsmethoden und Evaluationskriterien aus und führten auf Regierungsseite zu der Überlegung, die Umsetzung von Head Start auf unbestimmte Zeit auszusetzen.

Ziegler und Muenchow (1992) nennen zudem weitere methodologische Schwächen der Evaluation von Frühförderprogrammen in den USA in den 60er Jahren. Demnach wurden viele Evaluationen der frühen Head Start-Programme von Forschungseinrichtungen durchgeführt, die, streng nach den Vorgaben der Auftraggeber, den Gesundheitszustand und die Ernährung der Kinder untersuchen sollten (deren Verbesserung ein Hauptziel von Head Start war). Neben Merkmalen, die in den Bereich der Gesundheitsforschung fallen, sollte das Engagement der Kommunalverwaltung zur Lösung stadtteilspezifischer Probleme mit Familien erhoben werden, die unter der Armutsgrenze leben. Im Gegensatz zu diesen Hauptzielen von Head Start wurde der Programmerfolg hingegen anhand von kognitiven Tests, darunter hauptsächlich Intelligenztests, gemessen.

In den Jahren nach der Veröffentlichung der ersten Evaluationsergebnisse wurden die Daten der frühen Head Start-Phase mehrfach neu untersucht. Bronfenbrenners (1975) Re-Analyse kam bei fast allen untersuchten Projekten zu dem Schluss, dass die geförderten Kinder sich nach der Teilnahme durch eine signifikante, jedoch schwach ausgeprägte Steigerung der kognitiven Kompetenzen auszeichneten. Der identifizierte kognitive Effekt hatte keine nachhaltige Wirkung

und verschwand nahezu komplett nach zwei bis drei Jahren Grundschulbesuch. Auch Campbell nahm sich in den 90er Jahren einer Re-Analyse der Westinghouse-Studie an und kam im Großen und Ganzen zu einen ähnlichen Ergebnis wie Bronfenbrenner Jahre zuvor (vgl. Wu & Campbell 1996, auch Bentler & Woodward 1978).

Die Ergebnisse der ersten Evaluationen von Head Start blieben zudem politisch nicht ohne Wirkung. Entscheidungsträger überdachten zunehmend die Konzepte der Frühförderung im Verlauf der 70er Jahre. Datta bezeichnet die Zeit von Anfang bis Mitte der 70er Jahre als „winter of disillusion and some despair about education and the Great Society in general" (Datta 1979, S. 405). Die Kritik an der Methode früherer Head Start-Evaluationen führte jedoch zu einer Weiterentwicklung der Evaluationsmethodik für soziale Programme insgesamt. Ebenfalls wurden differenzierte Designs von Wirkungsevaluationen entwickelt. Campbell lieferte dafür mit den Kriterien der internen und externen Validität einen wichtigen substantiellen Beitrag, um die bis dato verwendeten Methoden der Wirkungsmessung auf den Prüfstand zu stellen. Zwar stellen sich die experimentellen und quasi-experimentellen Designs als der im Prinzip richtige Weg für Wirkungsmessungen heraus. Die bis dato verwendeten Designs weisen jedoch methodische Lücken auf, die bei näherer Betrachtung der verwendeten Evaluationsdesigns deutlich werden.

So werden in einer Meta-Analyse aktueller Evaluationsstudien im Bereich der Frühförderung von Horton (2006) Faktoren genannt, die nicht mit dem Konzept der Frühförderung zusammenhängen, aber einen nachweisbaren positiven Einfluss auf die kognitive, soziale und emotionale Entwicklung von Kindern haben. Zu diesen zählen nach Horton die Programmstruktur, die Qualität des Förderprogramms (z.B. bezüglich der Auswahl der Inhalte und der didaktischen Aufbereitung), Charakteristika der geförderten Kinder sowie ihrer Familien sowie die Auswahl des Messinstruments und der Umgang damit in der Evaluationsstudie (Horton 2006, S. 51). Dazu verdeutlichen White und Phillips (2001) am Beispiel der frühen Head Start-Untersuchungen, dass die Ziele und Verfahren der Evaluationsstudien auf die Messung von Intelligenzquotienten reduziert und vereinfacht wurden (White und Phillips 2001, S. 98)[22].

Das Head Start-Programm ist in den USA weiterhin ein fester Bestandteil der Bildungsförderung. Head Start gliedert sich in mehrere zielgruppenspezifische

[22] Die ersten Evaluationen von Head Start wurden mehrheitlich von Entwicklungspsychologen durchgeführt. Psychologische Studien zur kognitiven Entwicklung von Kleinkindern in den 60er Jahren zeigten, dass Interventionen in der Lebenswelt von Kindern zu einem frühen Zeitpunkt in der Entwicklung zur Steigerung des Intelligenzquotienten führen (vgl. Bloom 1964, McVicker Hunt 1961, Susan Gray's Early Training Project 1974).

Unterprogramme, die auf staatlicher und lokaler Ebene angepasst an die Bedingungen vor Ort durchgeführt werden. In den einzelnen Programmen von Head Start wurden seit den 60er Jahren mehrfach Wirkungsevaluationen unternommen. Aufgrund der jahrzehntelangen Laufzeit liegen zudem Ergebnisse von Längsschnittanalysen vor, bei denen Teilnehmer im Vorschulalter bis zum Abschluss der Schullaufbahn begleitet wurden. Barnett (2008; siehe auch Horton 2006) kommt in seiner Auswertung von Evaluationsstudien zu mehreren umfangreichen Frühförderprogrammen (darunter auch Head Start) zu dem Schluss, dass auf Staaten-Ebene und lokal angebotene Head Start-Programme mit einem hohen qualitativem Standard und guter Ausstattung zu signifikanten Bildungserfolgen sowie positiven sozialen und ökonomischen Effekten in der Bildungsbiographie der Teilnehmer führen können (vgl. Barnett 2008, S. 20; Ludwig & Phillips 2007). Ob Head Start-Maßnahmen die intendierten Wirkungen zeigen, hängt gemäß den Meta-Analysen von mehreren Faktoren ab. Neben dem Curriculum, das den Fördermaßnahmen zugrunde liegt, beeinflussen demnach die Bedingungen der Umsetzung vor Ort sowie die Charakteristika der Programmteilnehmer den Fördererfolg (Ludwig & Phillips 2007, S. 11 ff.). Aus den Meta-Analysen geht hervor, welche Faktoren bei der Konzeption von Wirkungsevaluationen berücksichtigt werden müssen. Um die interne Validität der Ergebnisse zu maximieren, müssen Daten zu den genannten Faktoren im Evaluationsverfahren zusätzlich erhoben werden und in die Auswertungsverfahren integriert werden. Neben quantitativen Methoden der Datenerhebung kommen prinzipiell auch qualitative Verfahren in Betracht, um z.B. den Kontext eines Förderprogramms differenziert zu erfassen.

Neben der Weiterentwicklung der quantitativen Evaluationsmethode hat Campbell in den 70er Jahren auch Beiträge über die Anwendung von qualitativen Verfahren veröffentlicht (Campbell 1974 und 1975). Dabei sieht Campbell die Funktion von qualitativen Untersuchungen als eine wertvolle Ergänzung zu quantitativen Erhebungen. Für die Bereiche, in denen experimentelle Designs keine Antwort auf Fragestellungen der Stakeholder bieten, eignen sich qualitative Methoden. Neben seiner ausdrücklichen Empfehlung, quantitative Evaluationsverfahren auf Programme anzuwenden, sprach sich Campbell für qualitative Einzelfallstudien als Zusatzinstrumente der Evaluation aus (LeVine & Campbell 1972). Zu diesen zählt nach Campbell der Bereich der Durchführung von Programmen, in dem detaillierte Informationen dazu erhoben werden, wie das Programm in der Praxis umgesetzt wird (Shadish et. al 1991, S. 135). Campbells Hauptbeitrag ist jedoch die Weiterentwicklung der quantitativen Evaluationsmethoden (hier insbesondere die Arbeiten zu den randomisierten Experimentanordnungen) und die Einführung von quasi-experimentellen Designs in die Evaluationsforschung von sozialen Programmen.

Fasst man die Entwicklung der Evaluationstätigkeit im Bereich der Programmevaluation in dieser Frühphase zusammen, dann kann zunächst besonders die Dominanz einer Evaluationskultur festgestellt werden, deren Evaluationsansätze sich hauptsächlich auf die **Wirkungsmessung von sozialen Programmen** konzentriert haben. Dieses Verständnis von Evaluation drückt sich deutlich in den vorgestellten Evaluationsansätzen von Scriven und Campbell aus.

Durch Evaluationen sollte kontrolliert werden, ob Sozialprogramme die vorgesehenen Wirkungen auf die Teilnehmer entfalten. Für große Sozialprojekte wie Head Start in den USA wurden entsprechende Evaluationen in Auftrag gegeben, die die Programme hinsichtlich ihrer Wirkungen in der Sozial- und Bildungsförderung analysieren sollten.

Wirkungsevaluationen hatten in den 60er Jahren aus der Perspektive des Projektmanagements primär eine **Kontrollfunktion** zu erfüllen. Dazu zählt, dass die Ergebnisse Programminitiatoren eine Hilfestellung bei Kosten-Nutzen-Abwägungen bieten sollten (Cook & Matt 1990, S. 18 f.). Außerdem erhofften sich Auftraggeber empirische Evidenz auf Grundlage der Ergebnisse, um über den Stopp oder die Fortführung von Programmen entscheiden zu können (vgl. Campbell 1969, 1974). Mit der Evaluation der Programme wurden u.a. wissenschaftliche Einrichtungen an Universitäten beauftragt, die für die Datenerhebung meist standardisierte, quantitative Instrumente aus der psychologischen Forschung und Schulforschung anwendeten.

Des Weiteren führte in der ersten Phase der Evaluationstheoretiker Scriven Definitionen und Konzepte von Evaluation ein, die bis in die heutige Zeit Gültigkeit besitzen. Dazu lässt sich auch die grundsätzliche Herangehensweise bei der Planung von Evaluationsstudien, d.h. mit den Zieldefinitionen und der Benennung der Funktionen der Studie zu beginnen. Scriven hat des Weiteren die Bezeichnungen *formative und summative Evaluation eingeführt*, die heute noch unter Evaluatoren und im Projektmanagement für die Bezeichnung der Art und Weise der Durchführung von Evaluationsprojekten Anwendung finden.

Hauptsächlich durch die Arbeiten von Campbell ausgelöst, wurde – angelehnt an Forschungsinstrumente der Psychologie und Medizin – die Methode des Laborexperiments zur Methode des Feldexperiments weiterentwickelt und in die Verfahren der Programmevaluation integriert. Neben reinen **Experimenten** hat in der Evaluationspraxis die Methode des **Quasi-Experiments** Einzug gehalten, das mit ungleichen Teilnehmergruppen operiert. Das Tätigkeitsfeld eines Evaluators zeichnete sich durch eine hohe Methodenkompetenz in Kombination mit fachlichen Kenntnissen aus. Die grundlegenden, empirischen Prinzipien und Methoden von Campbell et al. sowie die Arbeiten von Scriven zu der Konzeption und Durchführung von Evaluationsstudien werden bei der Gestaltung des Leitfadens für die Begleitung von Evaluationsprozessen im dritten Kapitel Eingang finden.

Insbesondere die Verfahren zur Konstruktion von experimentellen und quasi-experimentellen Designs sowie die dazugehörigen Auswertungsverfahren sind für die Gestaltung von Wirkungsevaluationen sind bis in die heutige Zeit wegweisend geblieben.

Durch die Diskussion der ersten Wirkungsevaluationen bei Programmen der Bildungsförderung (darunter insbesondere zum Head Start-Programm) konnten hingegen auch methodische Schwachstellen in den bis dato angewendeten Designs identifiziert werden. Zu diesen zählen bei experimentellen und quasi-experimentellen Untersuchungen die bis zu diesem Zeitpunkt mangelnde Berücksichtigung der Kontextbedingungen und potentiellen externen Einflüssen, die das Programm und seine Teilnehmer beeinflussen können, die Verfahren der Bildung von Untersuchungsgruppen im Evaluationsprozess sowie der Umgang mit den Ergebnissen der Evaluation. Eine tiefergehende Auseinandersetzung mit den Schwachstellen erfolgte in der zweiten Phase.

2.3.2. Zweite Phase: Prozess- und nutzenorientierte Evaluationsansätze

Die zweite Phase der Weiterentwicklung der Evaluationsmethode lässt sich zeitlich auf die gesamte Spanne der 70er Jahre verorten und ist durch die Reaktion auf die Erfahrungen mit Evaluationsstudien in den 60er Jahren und der damit zusammenhängenden kritischen Hinterfragung der bis dato verwendeten Evaluationsdesigns für Wirkungsevaluationen geprägt. Wenngleich durch die „frühen“ Wegbereiter der Evaluation in den Vereinigten Staaten wie Campbell und Scriven die Entwicklung der Evaluationsforschung eine gewisse Struktur und Schwerpunktsetzung erhalten hat, ist der Beginn der 70er Jahre in der Fachliteratur mit einem **„Paradigmenwechsel“** in der Evaluationsforschung verbunden (vgl. Fend 1977, S. 67). Um 1970 erreichte der Boom von Evaluationsstudien in den Vereinigten Staaten seinen Höhepunkt. Besonders nationale Behörden investierten in groß angelegte Evaluationsprojekte. Mit dem Ende der Administration Nixon wurden dagegen die Ausgaben für Evaluation reduziert. In der Evaluationsforschung wird argumentiert, dass ein ausschlaggebender Grund für die Ausgabenreduktion die mangelhafte Verwertung von Wirkungsevaluationen zur Verbesserung der Programme im Bereich der Bildungsförderung war (vgl. Guba & Lincoln 1989).

Der Beginn der Wirkungsevaluation von Bildungsprogrammen in den 60er Jahren war mit der Vorstellung verbunden, die generierten objektiven Evaluationsergebnisse würden eine angemessene Informationsgrundlage für rationale Entscheidungen zur Verbesserung der Programme bilden. Die Kritik an den Methoden – beispielsweise der Head Start-Studien – führte jedoch zu einer stärkeren Untersuchung der Evaluationsdesigns nach ihren Schwachstellen. Kritik an der Methode der Evaluation wurde u. a. von Weiss, Rossi, Guba und später Stufflebeam (vgl. Stufflebeam et al. 2000, Stufflebeam & Shinkfield 2007) geäußert –

Evaluationsforscher, auf deren Hauptkritikpunkte im Folgenden eingegangen wird. Die Auseinandersetzung der Evaluationsforscher mit dem Themenfeld erfolgte zu Beginn der 70er Jahre bezogen auf folgende Diskussionspunkte:

1. Nutzung der Ergebnisse von quantitativen Evaluationsstudien,
2. Einbindung von Stakeholdern (Programminitiatoren und -beteiligte) bei Evaluationsverfahren,
3. Diskussion von alternativen Evaluationsansätzen.

Anfang der 70er Jahre beschäftigte sich Carol Weiss mit der Fragestellung, wie Evaluation zusammen mit sozialwissenschaftlicher Forschung für den politischen Entscheidungsprozess genutzt werden kann. Experimentelle und quasi-experimentelle Untersuchungen sind für Weiss geeignete Verfahren der Evaluation, um Wirkungen von sozialen Programmen differenziert zu erfassen (vgl. Weiss 1974). Damit unterstreicht Weiss die Bedeutung der Arbeit von Scriven und Campbell bei der Weiterentwicklung der Evaluationsmethodologie. Weiss befasste sich zu Beginn ihrer Arbeit mit experimentellen Designs für Evaluationsstudien, wurde dann aber auf Probleme aufmerksam, die sie für die Notwendigkeit einer Weiterentwicklung der Evaluationsverfahren sensibilisierte (vgl. Shadish et. al 1991, S. 183).

Der eigentliche Kritikpunkt von Weiss lautete Anfang der 70er Jahre aber, dass die bis dato in der Praxis angewandten Evaluationsmethoden auf die spezifischen Ziele und Arbeitsweisen der zu evaluierenden Programme nicht im ausreichenden Maße abgestimmt seien (vgl. Weiss 1972). Nach Weiss sind gerade die Erkenntnisse aus Evaluationsstudien mit experimentellen Designs insbesondere für Steuerungs- und Entscheidungsprozesse bei Programmen wertvoll, weil die Ergebnisse im Idealfall Programmverantwortliche und Initiatoren mit genauen Informationen darüber versorgen, wie ein Programm im Detail funktioniert und ob die intendierten Programmziele erreicht wurden. Weiss sah hier den Handlungsbedarf bei der Weiterentwicklung der Evaluationsmethodologie und erkannte als Evaluationsforscherin die Notwendigkeit der Berücksichtigung der **Bedarfe und Interessen von Stakeholdern** bei der Planung und Umsetzung von Evaluationsverfahren.

Weiss sieht die mangelhafte bzw. nicht-adäquate Nutzung von Evaluationsergebnissen auf einer irrtümlichen Grundannahme von Evaluatoren begründet, dass die politische Entscheidungsstruktur ähnlich rational aufgebaut sei wie die Funktionsweise von sozialen Programmen. Auf die Evaluation von sozialen Programmen übertragen würde ausgehend von dieser Annahme folgen, dass Programme mit positiven Evaluationsergebnissen weitergeführt und Programme mit nicht-positiven Evaluationsergebnissen eingestellt würden. Weiss vertritt dagegen die These, dass die mit dem Programm in Beziehung stehenden Interessensgruppen

(u.a. Politiker, Entscheidungsträger, Evaluatoren) zumeist nicht diese Art von „Stop-go"-Entscheidungen treffen. Entscheidungen über die Fortführung von sozialen Programmen seien in der Realität von längeren Verhandlungsprozessen der Interessensgruppen abhängig oder begleitet. Cook und Matt (1990) stellen bezugnehmend auf Weiss fest, dass Entscheidungen „... überhaupt nur selten ‚gefällt' werden. Es ist eher der Fall, dass man in Entscheidungen hineinrutscht oder dass sie eine Konsequenz von Entscheidungen der Vergangenheit sind, welche die kompromissfähigen Entscheidungen der Gegenwart stark beschränken" (ebd. S. 22). Den politischen Entscheidungsprozess nennt Weiss decision accretion: „... the build-up of small choices, the closing of small options and the gradual narrowing of available alternatives" (Weiss 1976, S. 226). Politische oder administrative Entscheidungen sind daher im konkreten Betrachtungsfall keine Handlungen, die ausschließlich durch Erkenntnisse aus Evaluationsstudien geprägt sind sondern werden zudem durch die Interessen der am Programm beteiligten Akteure beeinflusst.

Cohen und Garet (1975) nennen ein weiteres Argument dafür, warum Erkenntnisse aus Evaluationen nicht unmittelbar zu politischen oder administrativen Entscheidungen führen müssen. Differierende Strategien von Entscheidungsträgern zu einem Programm, Allianzen von Akteuren auf politischer Ebene und Spielräume in den finanziellen Budgets hätten nach Cook einen deutlich stärkeren Einfluss auf operative Entscheidungen, was mit einer Maßnahme passieren soll, als Ergebnisse wissenschaftlicher Untersuchungen. Cook (1978) äußerte in diesem Zusammenhang die Vermutung, dass die wissensgenerierende (Wissenschaftler)-Kultur mit Evaluationsstudien an den Bedarfen der wissensnutzenden (Praktiker-)Kultur vorbei produziert hat. Evaluatoren hätten demnach lange Zeit zu viel Wert auf die Messung von Effekten einzelner Programme oder Projekte gelegt und dabei auf der qualitativen Ebene die Details der Programmdurchführung sowie der Einbeziehung der am Programm Beteiligten weitgehend außer Acht gelassen (Cook & Matt 1990, S. 22). „Dem setzt die nutzenfokussierte Evaluation den Anspruch entgegen, die Lücke zwischen der Gewinnung von Evaluationsergebnissen und deren Nutzung für Entscheidungen und Programmverbesserungen zu schließen" (Giel 2013, S. 73).

Ein weiterer Kritikpunkt von Weiss zielt auf die bis dato mangelhafte systematische Erfassung der Arbeitsweise eines Programms. Bei einem Experiment werden die Daten vor und nach einem Treatment erhoben, wodurch nur Aussagen zum Zielerreichungsgrad eines Programms möglich sind: „...the evaluator takes ‚before' measurements on factors relevant to program goals, the subjects are then exposed to the program (an unexamined entity like a black box), and then he records ‚after' measurements" (Weiss 1974, S. 331). Wenn jedoch systematisch keine Informationen zur Programmdurchführung gesammelt werden, sind nach

Abschluss der Auswertungen der Evaluationsdaten keine Aussagen darüber möglich, wie das untersuchte Programm Wirkungen bzw. keine Wirkungen bei den Teilnehmern erzeugt hat. Weiss nennt dieses methodische Manko bei Experimenten auch **Blackbox Studies**. Weiss schlägt daher vor, der Blackbox-Problematik mit zwei Maßnahmen zu begegnen: Die Entwicklung einer Programmtheorie vor den Erhebungen sowie die Kontrolle von potentiellen Störfaktoren während der Durchführung des Programms. Auf die Entwicklung von Programmtheorien wird im nächstfolgenden Kapitel eingegangen.

Bei experimentellen Untersuchungen unter Laborbedingungen (z.B. mit Probanden in der klinischen Psychologie oder in der Medizin) lassen sich die Wirkungen von Störfaktoren bzw. externen Einflussfaktoren auf das Programm durch die „sterilen" Untersuchungsbedingungen minimieren. Der Ausschluss von Störfaktoren ist bei „Feldexperimenten" nicht möglich, im Gegenteil, nach Weiss müssten studienspezifische Strategien entwickelt werden, wie im Evaluationsverfahren mit Störfaktoren umzugehen ist. Neben von außen auf das Programm wirkende Faktoren können auch Veränderungen im Programm während der Durchführung die Interpretation der Ergebnisse einer Pretest-Posttest-Messung verzerren. Dies kann z.B. bei einem Sprachförderprogramm für Kinder ein vom Curriculum abweichendes Instruktionsverhalten seitens der Erzieherinnen bzw. Pädagogen sein. Nach Weiss kann die Black-Box-Problematik nur dadurch gelöst werden, dass Evaluatoren bei der Konzeption des Evaluationsdesigns mögliche interne wie externe Einflussfaktoren auf die Untersuchung antizipieren und vor der Untersuchung entsprechende Maßnahmen ergreifen, um diese zu beseitigen bzw. deren Einfluss auf das Programm während der Evaluationsdurchführung zu kontrollieren[23]. Weiss wurde dadurch zur Ideenbereiterin für die Entwicklung von späteren Evaluationsansätzen, die seitdem Lösungsansätze für die Behandlung von Blackbox-Problematiken bei der Evaluation von sozialen Programmen bieten. Zwei dieser Evaluationsansätze werden im anschließenden Kapitel genauer beschrieben.

Ein weiterer Vertreter der Evaluationsforscher in den 70er Jahren war Lee Cronbach, der ebenfalls Vorschläge entwickelte, wie Evaluationsergebnisse besser für die **Optimierung von Programmen** genutzt werden können. Ein zentraler Arbeitsbereich von Cronbach war die Bildungsforschung. In Cronbachs Beschäftigungskontext hat Evaluation die Funktion, als Instrument zur Überprüfung von bildungspolitischen Maßnahmen eingesetzt zu werden: „We see evaluation as an integral part of policy research, and we therefore blend ideas appearing in the policy-research literature with ideas of evaluation itself" (Cronbach 1980, S. 19).

[23] Auf die verschiedenen Einflussfaktoren nach Campbell sowie den Umgang des Evaluators mit den sich daraus ergebenden Herausforderungen im Evaluationsprozess wird detaillierter in Kapitel 5.3.4. eingegangen.

In seinem Hauptwerk „Toward Reform of Program Evaluation“ (Cronbach 1980) entwickelt Cronbach eine Vorstellung von Evaluation, die für die praktische Anwendung gedacht ist und die Entscheidungsfindung unterstützen soll.

Cronbach bereitete in seinem 1963 verfassten Artikel „Evaluation zur Verbesserung von Curricula“ die Wende in der empirischen Bildungsforschung in Schulkontexten von der ergebnisorientierten Lernzielevaluation zur entscheidungsorientierten Curriculumsevaluation vor. Dies bedeutet, dass der Fokus der Betrachtung nicht nur darauf gelegt wurde, ob das Lernziel erreicht wurde. Cronbachs Anliegen ist es, Evaluationen nicht nur zur abschließenden Bewertung konkurrierender Programme oder Lehrverfahren zu nutzen, sondern sie darüber hinaus zum festen Bestandteil der **formativen Curriculumsentwicklung in der Schule** zu machen und ihnen somit einen ganz konkreten Nutzen zu geben, um z.B. durch Maßnahmen eine Verbesserung des Status quo zu erreichen. Um den Nutzenaspekt zu betonen, unterscheidet Cronbach für den schulischen Kontext drei Arten von Konsequenzen, die sich aufgrund von Evaluationen ergeben können: „1) Curriculumsverbesserung [...] 2) Entscheidungen über Individuen [...] 3) Administrative Regelungen [...]“ (Cronbach 1972, S. 42). Erkenntnisse aus Programmevaluationen können daher nutzenorientiert in einem kontinuierlichen Verbesserungsprozess – vergleichbar mit einem Qualitätsmanagementverfahren – zur Weiterentwicklung einzelner Programmelemente eingesetzt werden. Cronbach greift damit den Ansatz der formativen Evaluation von Scriven auf und führt ihn am Beispiel der Evaluation von bildungspolitischen Maßnahmen ein. Cronbach war es in diesem Zusammenhang wichtig zu betonen, dass Evaluationsergebnisse nach Möglichkeit unmittelbar (durch formative Evaluationsverfahren) zur Verbesserung der drei genannten Bereiche eingesetzt und nicht bis zu einer bestimmten summativen Ergebnisfeststellung gewartet werden sollte.

Die Betonung der Nutzen- und Verbesserungsfunktion von Evaluationsverfahren findet sich bei Cronbach auch bei näherer Betrachtung seiner Bewertung der Konzepte der **internen und externen Validität** von Evaluationsergebnissen. Campbell und Cronbach nehmen unterschiedliche Positionen ein. Während Campbell betonte, dass Wirkungsmessungen primär dahingegen überprüft werden sollten, ob die Ergebnisse für den Untersuchungskontext und damit intern valide und objektiv sind, geht der Anspruch Cronbachs einen Schritt weiter. Nach Cronbach sollte das Ziel von Evaluationen sein, externe Validität der Ergebnisse herzustellen, d.h. im Idealfall liefern Experimente nicht nur Einsichten über die Wirkungsweise einer Intervention bei den Teilnehmern einer Stichprobe, sondern die gewonnenen Erkenntnisse lassen sich für große Personen- bzw. Bevölkerungsgruppen verallgemeinern und/oder auf andere Kontexte übertragen. Die Wahrscheinlichkeit der Nutzung von Evaluationsergebnissen wird maximiert, wenn die Ergebnisse den Ansprüchen sowohl der internen als auch der externen Validität entsprechen.

Neben den Arbeiten von Cronbach und Weiss, die sich kritisch mit den Evaluationsmethoden auseinandersetzten und eigene Beiträge zur Weiterentwicklung der Evaluationsverfahren lieferten, etablierte sich in den 70er Jahren ein Kreis von Evaluatoren, die konkrete Vorschläge für die Evaluation von sozialen Programmen entwickelten. Die so entstandenen Evaluationsansätze tragen zumeist programmatische Namen (z.B. *Responsive Evaluation*) und bieten zum Teil ähnliche Lösungsstrategien für die zuvor beschriebene Nutzungsproblematik von Evaluationsergebnissen. Während es sich bei den Evaluationsansätzen der 60er Jahre im Grunde um die Beschreibung der Anwendung von quantitativen empirischen Methoden handelte, werden bei den neuen Evaluationsansätzen zusätzlich der Einsatz von qualitative Methoden, Kommunikationsprozessen und Projektmanagementaktivitäten sowie Verfahren für die kontinuierliche Verbesserung der Programme thematisiert. Aus dem Kreis der Evaluatoren, die sich der Entwicklung neuer Evaluationsansätze gewidmet haben, sind hier stellvertretend die Ansätze von Michael Patton *Creative Evaluation* (Patton 1978), Ernest Stake mit dem Ansatz *Responsive Evaluation* (Stake & Easley 1978) und Ernst Guba und Lincoln (Guba & Lincoln 1989) mit einem konstruktivistisch geprägten Evaluationsansatz zu nennen. Die Hauptmerkmale dieser Ansätze werden im Folgenden kurz beschrieben.

Patton (1978) beschreibt den Evaluator als Künstler und Wissenschaftler zugleich. In der Publikation *Creative Evaluation* beschreibt Patton die Weiterentwicklung der Evaluationstheorie als einen Prozess weg von einheitlichen rein deduktiv-hypothesenprüfenden Verfahren hin zur **Modellierung von individuellen Evaluationsstrategien**, die über die Verwendung von herkömmlichen quantitativen Messverfahren hinausgehen. „In the infancy of the discipline, methods decisions dominated the evaluation decision-making process. Evaluators were primarily concerned about technical adequacy, validity, reliability, and measurability. [...] Evaluation research was dominated by the largely unquestioned, natural science paradigm of hypothetico-deductive methodology. The paradigm defines the epitome of 'good' science as quantitative measurement, experimental design, and multivariate, parametric statistical analysis" (Patton 1981, S. 20). Patton spricht sich für einen kreativen Umgang des Evaluators mit dem Evaluationsinstrumentarium aus, ohne dabei jedoch Kriterien der Wissenschaftlichkeit und die Ziele der Evaluationsstudie außer Acht zu lassen. Patton empfiehlt für die Auswahl der geeigneten Methoden das „Werkzeugkastenprinzip": Abhängig vom Evaluationsauftrag sucht sich der Evaluator aus einer Sammlung von Evaluationsinstrumenten die – für den jeweiligen Kontext – passenden Instrumente aus. Dies können dann quantitative oder qualitative Methoden sein. Diese Vorgehensweise ist als eine Abkehr von dem bis dato vorherrschenden Methodenzentrismus zu sehen. Pawson und Tilley nennen Pattons Evaluationsansatz deshalb

auch „pragmatism incarnate in their presentation of research method in the classic ‚toolbox' manner" (Pawson & Tilley 1997, S. 13).

Ein weiterer Vertreter, der den Fokus auf den Prozess der Durchführung von Evaluationsstudien legt, ist Ernest Stake. Stake (1975) unterscheidet zunächst zwischen den Evaluationsansätzen *Responsive Evaluation* und *Preordinative Evaluation*. Bei *Preordinativen Evaluationen* sind im Evaluationskonzept die Methoden und zum größten Teil auch die strategische Vorgehensweise zur Durchführung der Evaluationsstudie schon vor Beginn des Vorhabens festgelegt. So werden zu Beginn einer Evaluationsstudie detaillierte Angaben zu den beabsichtigten evaluativen Verfahrensweisen und den anvisierten Produkten der Evaluationstätigkeit definiert. Das Evaluationskonzept umfasst „(1) eine Beschreibung der Ziele der Studie, (2) eine Angabe über die geplanten (objektiven) Erhebungsinstrumente, (3) welche Standards als Bewertungsgrundlage genutzt werden und (4), dass die Ergebnisse in wissenschaftlichen Berichten festgehalten werden" (Stake 1975, S. 15 [Anm.: Übersetzung aus dem Englischen durch den Autor]).

Der von Stake favorisierte und von ihm ausgearbeitete Gegenvorschlag dazu ist die Responsive Evaluation. Das wesentliche Merkmal von Responsiven Evaluationen ist die **Konzentration des Ansatzes auf die Kommunikationsstrukturen zwischen Evaluatoren und den Stakeholdern** im Vorfeld und während der Durchführung von Evaluationsstudien. Bei Responsiven Evaluationen „antworten" Evaluatoren mit Vorschlägen zur Gestaltung und Durchführung von Evaluationsstudien auf Anforderungen und Bedarfe seitens der Programmbeteiligten und Programmverantwortlichen. Stake beschreibt den Charakter responsiver Evaluationen folgendermaßen: „An educational evaluation is responsive evaluation (1) if it orients more directly to program activities than to program intents, (2) if it responds to audience requirements for information, and (3) if the different value perspectives of the people at hand are referred to in reporting the success and failure of the program" (Stake 1975, S. 10). Stake kommt zu dem Schluss, dass die Wahl und der Aufbau einer Evaluationsstrategie abhängig ist von den Kontextbedingungen, in denen das Programm durchgeführt wird, und den Anforderungen der Stakeholder: „Different styles of evaluation will serve different purposes" (Stake 1975, S. 28). Der Ansatz basiert auf dem Prinzip, dass Programme keinen instrumentellen Wert haben müssen, um für eine Evaluationsstudie in Frage zu kommen. Der „Payoff" eines Programms kann sich in der Anfangsphase einer Evaluation als diffus herausstellen und erst nach einer langen Laufzeit des Programms eintreten. Im Unterschied zur Preordinativen Evaluation bedeutet dies, dass sich ein **ursprünglich entwickeltes Evaluationskonzept im Verlauf einer Evaluationsstudie ändern muss**, wenn einerseits Änderungen im Programm eintreten oder sich die Interessen der Stakeholder verlagern. Durch die formative Vorgehensweise und den iterativen Charakter des Austausches zwischen Evaluator und Programmbeteiligten würden nach Stake Erkenntnisse aus

Evaluationsstudien, die nach den Vorstellungen der *Responsiven Evaluation* geplant und durchgeführt werden, einen hohen Nutzen für die Stakeholder erzeugen und die Weiterentwicklung des Programms erleichtern.

Bei *Responsiven Evaluationen* wird daher das Methodenportfolio für die Untersuchung erst als Resultat einer intensiven Programmbegleitung zu Beginn des Evaluationsprojekts (z.B. durch Ist-Aufnahmen, Analyse von zugänglichem Informationsmaterial, Begehungen und Gespräche mit Programmbeteiligten) zusammengestellt. Anstatt für eine Evaluationsstudie ein festes Methodenkonzept vorzugeben, sei es für alle Programmbeteiligten bezüglich der zu erwartenden Ergebnisqualität gewinnbringender, die Planung einer Evaluationsstudie explorativ zu gestalten. Im Kontext von sozialen Programmen betrifft dies beispielsweise Fragen der Beteiligung der Zielgruppe am Angebot, der Berücksichtigung von unterschiedlichen Lerntypen sowie der Umgang mit Hindernissen bei der Implementierung eines Programms. Indem der Evaluator diese Aspekte bei der Entwicklung des Evaluationsdesigns antizipiert, gelänge es ihm, den Nutzen von Programmen zielgruppenspezifisch differenzierter herauszuarbeiten.

Die Evaluationstheoretiker Guba und Lincoln (1989) gingen mit dem von ihnen vorgestellten Evaluationsansatz einen Schritt weiter und formulierten einen **methodischen Gegenvorschlag** zu den bis dato angewendeten Evaluationsansätzen. Der von Guba und Lincoln für ihren Evaluationsansatz genutzten Begriff *Fourth Generation Evaluation* soll verdeutlichen, dass diese neue Generation von Evaluationsansätzen mit einem veränderten Verständnis verbunden ist, wie Evaluationsstudien hinsichtlich der Methoden und hinsichtlich des Prozesses der Evaluation durchzuführen sind. Die ersten drei Generationen beziehen sich auf Evaluationsverfahren (zur dritten Generation zählen z.B. nach Guba die im Kapitel zuvor beschriebenen Ansätze von Scriven und Campbell) mit einer starken methodischen und messtheoretischen Schwerpunktsetzung. Guba führte in den USA eine Metaevaluation zahlreicher Evaluationsstudien der 60er Jahre hinsichtlich Kriterien wie Ziele, Durchführung und Folgemaßnahmen durch und kam in der Auswertung zu dem Schluss, dass der Nutzen vieler Evaluationsstudien ungenügend sei. Nach Guba wurden Evaluationsergebnisse häufig nicht genutzt – weder für Haushaltsentscheidungen, noch für politische Entscheidungen oder für die die Programmweiterentwicklung (vgl. Guba & Lincoln 1989).

Waren der Zweck, Ziel und Methoden früher Evaluationsansätze durch die wissenschaftsphilosophischen Annahmen des logischen Positivismus geprägt, wurde der Ansatz von Guba und Lincoln mit **Einflüssen aus dem Konstruktivismus** entwickelt. Guba und Lincoln definieren das Produkt eines Evaluationsprozesses folgendermaßen: „Evaluation is a process whereby evaluators and stakeholders jointly and collaboratively create (or move toward) a consensual valuing construction of some evaluand“ (Guba & Lincoln, 1989, S. 263). Demzufolge

haben Evaluationsstudien die Funktion, Erkenntnisse zu einem sozialen Programm in einem iterativen Prozess zwischen Evaluatoren und Programmbeteiligten und Programmverantwortlichen in einem gemeinschaftlich, nach vorne gerichteten und auf Konsensfindung gestalteten Prozess zu sammeln (vgl. auch Guba & Lincoln 1989, S. 41). Stufflebeam erläutert die Annahmen des Konstruktivismus im Kontext von Evaluationstudien: „Constructivism rejects the existence of any ultimate reality and employs a subjectivist epistemology. It sees knowledge gained as one or more human constructions, uncertifiable, and constantly problematic and changing. It places the evaluators and program stakeholder at the centre of the inquiry process, employing all of them as the evaluator's 'human instruments'" (Stufflebeam, Madaus & Kellaghan 2000, S. 71).

Die Ausgangshypothese von Guba und Lincoln lautet, dass für Evaluatoren Realität nicht als messbarer Zustand vorliegt, der sich durch sozialwissenschaftliche Methoden objektiv erfassen lässt. Sinnvolle Evaluationsverfahren können nicht die Realität erfassen, sondern würden in ihren Ergebnissen lediglich ein Bild der Realität beschreiben, dass hauptsächlich durch die im Evaluationsprozess beteiligten Akteure konstruiert sei. Realität würde sich demnach als eine ständig sich verändernde soziale Konstruktion aus Handlungen, Beziehungen und Interaktionen von Akteuren darstellen. Guba und Lincoln (1989) beziehen Konflikte als Bestandteil des Evaluationsprozesses ein: „Conflict, rather than consensus, must be the expected condition in any evaluation taking account of value differences" (S. 210). Der Prozess von Evaluation ist von Konflikt geprägt, mit dem Ziel, einen Konsens zwischen den beteiligten Akteuren (Stakeholder) zu erreichen.

Guba und Lincoln zufolge verlaufen Evaluationsverfahren nach dem **hermeneutischen Prinzip**. Der Evaluator tritt mit den Programmbeteiligten in einen Kommunikationsprozess, bei dem Informationen zu den Zielen, der Planung, zur Durchführung sowie zu den Ergebnissen der Evaluation eines sozialen Programms in einem schleifenartigen Verfahren zunächst gesammelt und kritisch diskutiert werden. Die Auseinandersetzung mit den gesammelten Informationen hat zum Ziel, mit den am Diskurs beteiligten Personen eine gemeinsam getroffene Vereinbarung zu resultierenden Handlungen zu erreichen. Sobald Änderungen im Programm und seiner Bestandteile eintreten, werden diese wiederum in den Kommunikationsprozess eingebracht. Eine fortlaufende Interpretation und Anpassung des Programms soll bei Evaluatoren und Programmverantwortlichen zu einem zunehmend festen und detaillierten Verständnis über die Funktionsweise des Programms inklusive des Kontextes führen, in dem es durchgeführt wird (vgl. Hitzler & Honer 1997). Dies stellt zugleich das Produkt der Evaluationsstudie dar.

Die Aufgabe des Evaluators sei es, diese unterschiedlichen Erfahrungen im Evaluationsprozess zu interpretieren und zu bewerten. Dabei würden zum Teil divergierende Vorstellungen und Bewertungen der Akteure im Evaluierungsprozess fortlaufend gesammelt und durch die Projektbeteiligten neu interpretiert und bewertet werden. Evaluatoren würden im Idealfall eine unabhängige und neutrale Position einnehmen und agieren als Moderatoren. „Dem Evaluator fällt daher die besondere Rolle zu, ‚hermeneutische Zirkel' zu schaffen, indem er unterschiedliche Sichtweisen zusammenführt und durch ihre Interpretation und Synthese allen Beteiligten ermöglicht, sie wechselseitig zu ergründen" (Descy & Tessaring 2006, S. 33). In der praktischen Umsetzung bedeutet dies, dass Evaluatoren für die Ansprüche und Probleme der Stakeholder und Zielgruppen des Programms stets offen sein und diese systematisch und konstruktiv in den Evaluationsprozess einbinden sollten. „[...] seeking the views of those present on why (if at all) the implicit ideas behind a scheme have crossed their paths and changed their reasoning" (Pawson & Tilley 1997, S. 18).

Die neueren Evaluationsansätze unterscheiden sich von den Ansätzen davor durch ihre **Schwerpunktsetzung auf den Prozess der Evaluationsdurchführung** und heben die Bedeutung einer **Eingangsphase von Evaluationsstudien** hervor. Zu dieser zählen die Konzeption eines Evaluationsplans, die Beteiligung der Akteure des Programms bei Festlegung der Ziele der Evaluation sowie die stärkere Einbindung der Sichtweisen der Programmbeteiligten während des gesamten Evaluationsprozesses. Stake, Guba und Patton betonen, wie wichtig es sei, schon in der Eingangsphase die Erfahrungen und Sichtweisen der Projektbeteiligten zu erfassen (vgl. Stake 1975, Guba & Lincoln 1981). Die Auswahl der Erhebungsinstrumente für die Studie folgt dann als zweiter Schritt, basierend auf den Erkenntnissen aus der Eingangsphase. Ein Evaluationsprojekt sei ein handwerklich gut vorbereitetes und durchgeführtes Vorhaben mit dem vorrangigen Ziel, Entscheidungen über die Fortführung eines Programms zu beschleunigen (Stake 1975).

Aufgrund der engen Verknüpfung von Evaluation mit Aspekten wie Moderation und Projektmanagement verschiebt sich die Schwerpunktsetzung weg von der nahezu ausschließlichen Anwendung von quantitativen Verfahren und hin auf Formen des Umgangs des Evaluators mit Bewertungen und Interessen der Akteure im Evaluationsprozess. Dies wird beispielsweise dann bei der Durchführung einer Evaluationsstudie relevant, wenn Programminitiatoren andere Interessen verfolgen als diejenigen, die unmittelbar mit der Programmdurchführung befasst sind.

Auf den Evaluator kommen bei Evaluationsstudien, wie sie durch Guba und Lincoln, Patton und Stake nahe gelegt werden, mehrere Kompetenzanforderungen zu. Der Evaluator sollte zum einen als Programmmanager tätig sein, in dem er

mit den Programminitiatoren die Ziele der Evaluation festlegt und mit den verschiedenen Akteuren und Interessensvertretern in Kontakt tritt. Außerdem sollte eine gewisse Sensibilität im Umgang mit divergierenden Interessen der Beteiligten zeigen und ggf. als Vermittler und Moderator zwischen den Interessensgruppen auftreten. Dieses Anforderungsprofil für Evaluatoren reicht über die reine Methodenkompetenz hinaus und birgt die Gefahr der Überforderung von Evaluatoren, indem der ursprünglich mit den Auftraggebern der Evaluationsstudie vereinbarte Zweck der Evaluation (z.B. Ermittlung des Zielerreichungsgrades, Generierung von Informationen zur Wirkungsweise des Programms) in den Hintergrund tritt und sich die Tätigkeit in einer endlos erscheinenden Moderation und Vermittlung zwischen den Akteuren verliert.

Die Ergebnisse der zweiten Phase der Entwicklung der Evaluationstätigkeit lassen sich wie folgt zusammenfassen: Die kritische Auseinandersetzung der Evaluatoren mit den Ergebnissen von Evaluationsstudien der 60er Jahre führte insgesamt zu einer **methodischen Weiterentwicklung der Evaluationsverfahren**. An den empirischen Methoden und Grundmodellen für Wirkungsevaluationen haben Evaluatoren in den 70er Jahren im Prinzip festgehalten, jedoch haben einige Evaluationstheoretiker ihrer Kritik an den bis dato praktizierten Verfahren der Wirkungsevaluation von sozialen Programmen deutlich Ausdruck verliehen. Dazu lässt sich die in diesem Kapitel beschriebene Kritik von Carol Weiss nennen. Eine Erkenntnis, die in Folge der ersten Phase der Evaluationstätigkeit unter Evaluationsforschern gereift ist, lautet, dass die **Übertragung von quantitativen Methoden** aus der klinischen Psychologie oder Medizin auf die Evaluation von sozialen Programmen **alleine nicht ausreichend** sei. In der Folge wurden von Evaluatoren wie Weiss und Cronbach Schwachstellen bisheriger quantitativer Designs aufgezeigt und Vorschläge für alternative Herangehensweisen erarbeitet. So entwickelte Weiss grundsätzliche Überlegungen zu einer Konzeptionsphase von Evaluationsstudien, in der zunächst Annahmen zur Funktionsweise des zu evaluierenden Programms in Form einer **Programmtheorie** im Detail erarbeitet werden sollte. Auf Basis der Programmtheorie ließen sich anschließend die Evaluationsmethoden gezielter auswählen sowie die Blackbox-Problematik bei Wirkungsevaluationen behandeln.

Als eine weitere Schwachstelle der frühen Evaluationsansätze identifizierten Evaluationsforscher die mangelhafte **Nutzungsorientierung und Verwertungsmöglichkeit der Ergebnisse** (vgl. Guba und Lincoln 1989). Der **Prozess der Durchführung von Evaluationsstudien** rückte in den Vordergrund der Überlegungen und insbesondere die Beantwortung der Frage: Wie können Evaluationsstudien zielgerichtet konzipiert werden, damit die Adressaten der Studien (v.a. Programmverantwortliche) sinnvolle Rückschlüsse aus den Evaluationsergebnissen ziehen können? Neue Impulse hat die Evaluationsforschung durch Vertreter wie Patton und Stake erhalten, die mit Modellen der *Creative Evaluation* (Patton) und der

Responsive Evaluation (Stake) gerade die Problematik der Ergebnisverwertung in den Vordergrund der Diskussion setzten. So sind in der zweiten Phase Evaluationsansätze entstanden, die den Raum für die Anwendung von quantitativen Methoden zur Wirkungsmessung ermöglichen, jedoch das methodische Instrumentarium um qualitative Methoden erweitern. In der sich anschließenden dritten Phase wurden von Evaluationsforschern Vorschläge erarbeitet, wie methodisch mit der von Weiss geschilderten Blackbox-Problematik bei der Evaluation von Programmen umgegangen werden kann.

2.3.3. Dritte Phase: Entwicklung von komplexen, theoriegeleiteten Evaluationsansätzen

Mit der Weiterentwicklung der Evaluationstätigkeit in den 80er Jahren wurde wie beim Übergang zwischen der ersten und zweiten Phase der Fokus auf die Lösung von Defiziten bei bis dato praktizierten Evaluationsverfahren gelegt. Im Fokus der Weiterentwicklung stand die Optimierung der Designs für Wirkungsevaluationen. Ausgehend von den Designs quasi-experimenteller und experimenteller Verfahren sind in der dritten Phase differenzierte und zum Teil sehr komplexe Evaluationsansätze entstanden. Im Folgenden werden zwei Ansätze vorgestellt, die in der Fachliteratur repräsentativ für diese Phase der Evaluationstätigkeit genannt werden können: Der so bezeichnete *theoriegeleitete Evaluationsansatz* von Rossi und Chen (1983) (Theory-driven Evaluation) und der sich *realistische Evaluationsansatz* nennende von Pawson und Tilley (1997) (Realistic Evaluation). Beide Evaluationsansätze widmen sich drei zentralen Themen:

- Der Entwicklung von komplexen Designs für Wirkungsevaluationen, die auf Programme angewendet werden können, die an mehreren Standorten und ggf. unter verschiedenen Rahmenbedingungen durchgeführt werden.
- Der Erarbeitung von spezifischen Lösungsansätzen für die Blackbox-Problematik.
- Der Einarbeitung von bewährten Erkenntnissen aus den vorangegangenen Phasen der Entwicklung der Evaluationstätigkeit (z. B. zu der Einbindung von Programmbeteiligten in den Evaluationsprozess).

Rossi und dann zu einem späteren Zeitpunkt Chen (ab Ende der 80er Jahre) entwickelten ihre Evaluationsansätze u.a. als Reaktion auf die methodische Diskussion über die Relevanz der internen und/oder externen Validität von Evaluationsergebnissen sowie deren praktischen Verwertungsmöglichkeiten (Alkin 2004, S. 26). Chen vertritt die Ansicht, dass anhand eines geeigneten methodischen Instrumentariums in Evaluationsstudien prinzipiell beide Validitätstypen adressiert werden sollten und stellt als mögliche Lösung den theoriegeleiteten Ansatz für die Evaluation von sozialen Programmen vor. Zum Umgang des Evaluators mit den Konzepten der internen und externen Validität schloss Chen: „We are

not convinced that the trade-off problem between internal and external validity, which is so sharply portrayed by Campbell, Cronbach, or others, that dealing with one type of validity must seriously sacrifice the other types of validity“ (Chen & Rossi 1987, S. 97). Auch für die zuvor bereits erörterte Blackbox-Problematik könnte nach Chen und Rossi durch die Anwendung der Prinzipien von theoriegeleiteten Evaluationen eine strukturierte Vorgehensweise erarbeitet werden.

Theoriegeleitete Evaluationen sind sehr detaillierte, analytische Untersuchungen der Wirkungsweise eines Programms (vgl. Chen & Rossi 1983, Chen 1990, Chen 2005, Weiss 1972, Weiss 1998). Nach Chen und Rossi ist es noch vor der Festlegung der Evaluationsziele, der Entwicklung des Evaluationsdesigns und der Auswahl der Methoden unverzichtbar, die **Maßnahmen hinsichtlich ihrer Funktionsweise genau zu erfassen und zu analysieren**. Ein Jahrzehnt zuvor war es bereits Carol Weiss (1972), die eine systematische und detaillierte Untersuchung der Mechanismen eines Programms bei Evaluationen forderte, zum damaligen Zeitpunkt jedoch ihre Vorschläge nicht in dem Grad an Strukturierung und Detail ausgearbeitet hatte, wie dies ein Jahrzehnt später durch Chen und Rossi (1983) erfolgte.

Insbesondere war es dann Chen, der mit seinem Werk „Theory-driven evaluations“ 1990 den Fokus der Betrachtung auf die Programmtheorie gelegt hat. In seinem Evaluationsansatz konzentriert sich Chen auf die Identifikation von Sekundäreffekten, so genannte intendierte und nicht-intendierte Effekte, die sich bei der Programmdurchführung ergeben können. Chen verleugnet nicht die potentiellen gewünschten Effekte von sozialen Programmen, die auch den definierten Programmzielen entsprechen können. Er geht jedoch in diesem Zusammenhang einen Schritt weiter und weist auf mögliche Interaktionen des Programms mit seiner Umwelt hin, wodurch die Ergebnisse von Evaluationen beeinflussen sein können. Als s**ystematische Vorgehensweise zur Erfassung und Kontrolle dieser Einflussfaktoren** schlägt er den Ansatz der theoriegeleiteten Evaluation vor.

In experimentellen und quasi-experimentellen Untersuchungen werden Daten zum Programm und zu den Teilnehmern in den Untersuchungs- und Kontrollgruppen zu bestimmten, im Design festgehaltenen Zeitpunkten mit dem Ziel erhoben, Wirkungen des Programms auf die Teilnehmer zu erfassen. Diese Vorgehensweise kann bei theoriegeleiteten Evaluationen beibehalten werden, vor der Auswahl und Anwendung von Methoden erfolgt jedoch zuerst die Ausarbeitung der so genannten **Programmtheorie**. Chen bezeichnet die Programmtheorie auch als **conceptional framework** (Chen 2012, S. 17). Evaluatoren, Programmbeteiligte und Programmverantwortliche entwickeln zunächst eine gemeinsame Vorstellung davon, wie das zu evaluierende Programm theoretisch wirken soll. Diese Vorstellung von der Funktionsweise eines Programms wird als Programmtheorie bezeichnet und beinhaltet eine sehr detaillierte Vorstellung davon, wie einzelne

Arbeitsschritte ablaufen und zu den intendierten Wirkungen führen sollen. Annahmen über die intendierten Wirkungen der einzelnen Maßnahmen sind vorab zu formulieren – inklusive der Art und Weise, wie die Wirkungen erreicht werden sollen (Chen 1990, S. 43). „Descriptive assumptions, called change model, deal with what causal processes are expected to happen to attain program goals. Prescriptive assumptions, called action model, deal with what actions must be taken in a program in order to produce desirable changes. Theory-driven evaluation uses the action model and change model to address contextual factors and planning and implementation issues that are greatly interested to stakeholders." (Chen 2012, S. 18).

Eine Vorlage für die Ausgestaltung der Programmtheorie lässt sich nach Chen als Schema darstellen (Siehe Abbildung 1). Grob unterteilt sich das Modell in zwei Bereiche: **Action Model** und **Change Model**. Das Action Model beschreibt die Beziehung zwischen Organisationen und Akteuren, die an der Durchführung eines Programms beteiligt sind. Dazu zählen die Programminitiatoren, -verantwortlichen und -mitwirkenden auf der einen Seite sowie auf der anderen Seite die Programmteilnehmer und weitere beteiligte Institutionen oder Akteure. Das Action Model gibt Auskunft darüber, wie das Programm tatsächlich im Detail durchgeführt wird. Im Change Model, dem zweiten Bereich der Programmtheorie, werden die Wirkungsannahmen formuliert. Hier sind – wie in den Absätzen zuvor beschrieben – die Annahmen den Wirkungen des Programms nach erfolgter *Intervention* (z. B. mit einem Förderkonzept) und unter Berücksichtigung von externen und internen Faktoren (*Determinants*) aufgeführt. Mit *Outcomes* werden die unmittelbaren intendierten und nicht-intendierten Wirkungen des Programms nach der Intervention bezeichnet.

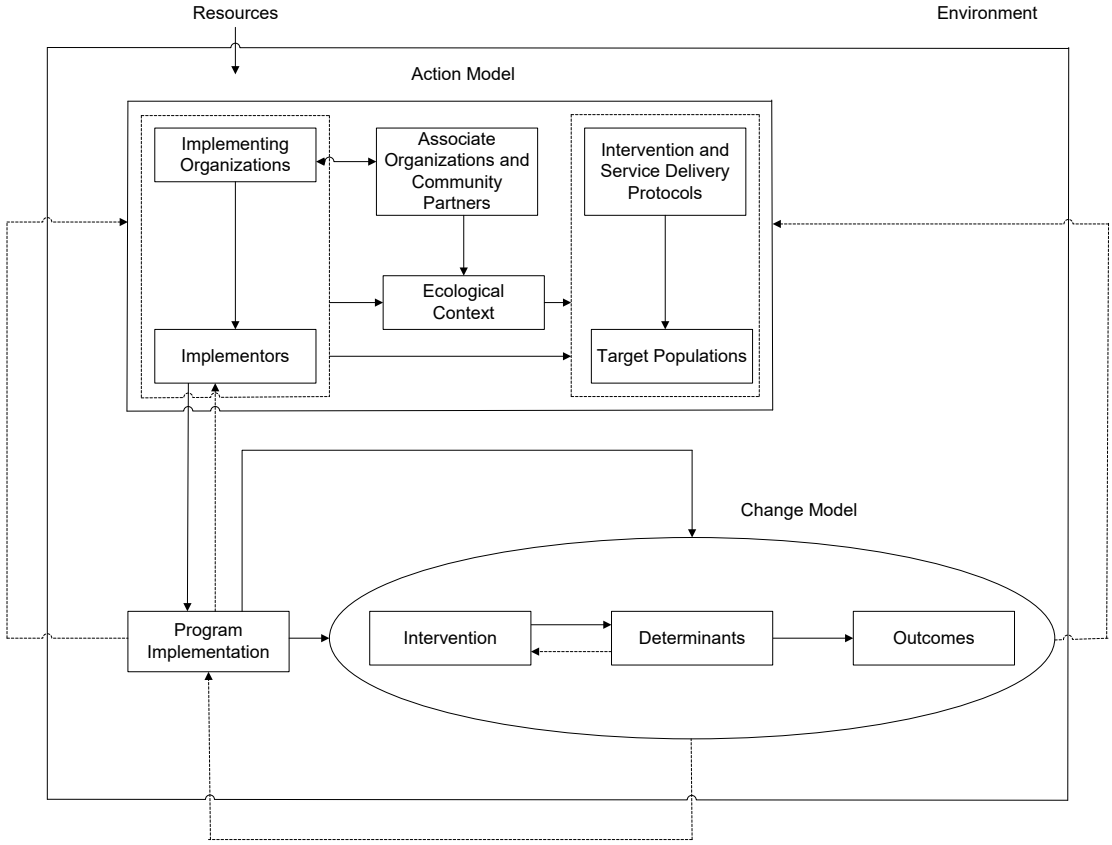


Abbildung 1: Generisches Modell der Bestandteile einer Programmtheorie nach Chen (Chen 2005, S. 31)

Im Vergleich zu den Evaluationsansätzen der zeitlich davorliegenden Phasen wird durch das generische Programmtheorie-Modell eine komplexe, systemische Herangehensweise an die Evaluation von Programmen deutlich, die durch die Einbeziehung von Organisationen, Akteuren sowie den Kontextbedingungen weitere Differenzierungskomplexität gewinnt (vgl. Coryn et al. 2011, S. 202; auch Patton 2010).

Nachdem eine erste Version der Programmtheorie beschrieben ist, wird diese mit den Programmverantwortlichen abgestimmt, so dass Programmverantwortliche und Evaluatoren von den gleichen Annahmen über die Funktionsweise des zu untersuchenden Programms ausgehen. Dieser Arbeitsschritt sollte noch vor der Entwicklung des eigentlichen Evaluationsdesigns vollzogen sein.

Die Ausgestaltung einer Programmtheorie muss in der Praxis nicht zwingend mit der anschließenden Durchführung einer Wirkungsevaluation verbunden sein. Es kann mit dem Ansatz der theoriegeleiteten Evaluation auch z.B. eine Evaluation der Programmprozesse durchgeführt werden. Chen unterscheidet zwischen dem **normativen Typ** und dem **kausalen Typ** von theoriegeleiteten Evaluationsstudien. Der normative Typ umfasst die Entwicklung der Programmtheorie und die Festlegung des Evaluationsdesigns und des Evaluationsplans. Daran

schließt sich eine Analyse der *Prozesse und Mechanismen* im Programm an. Durch die Anwendung von quantitativen und qualitativen Methoden werden Informationen und Daten gesammelt, deren Interpretation zu einer **Modifikation und Anpassung der Programmtheorie** führen kann.

Beim kausalen Typ werden Gültigkeit der in der skizzierten Programmtheorie formulierten Annahmen sowie deren Interdependenzen mithilfe von **Wirkungsmessungen** überprüft. Dies kann dann beispielsweise durch experimentelle und quasi-experimentelle Methoden erfolgen. Bei beiden Typen von theoriegeleiteten Evaluationen steht die Generierung von neuen Erkenntnissen über das Programm sowie dessen Durchführung im Vordergrund. Nach Chelimsky (1997) hat Evaluation hier die Funktion der Wissenserweiterung.

Evaluationsansätze, die sich der Gruppe der theoriegeleiteten Evaluationen zuschreiben lassen, wurden seither in der Praxis beispielsweise von der W. K. Kellogg Foundation für die Evaluation ihrer Programme in Kommunalverwaltungen in den USA angewendet. Ein weiteres Beispiel ist die Weltbank, die für die Evaluation von humanitären Hilfsprogrammen auch theoriegeleitete Evaluationsansätze verwenden (für diese und weitere Beispiele vgl. Coryn et al. 2011, S. 200f.).

Der Ansatz der theoriegeleiteten Evaluation wurde seit Ende der 80er Jahre weiterentwickelt und es sind in der Folge eine Reihe an Evaluationsansätzen entstanden, die sich an dem Grundansatz von Rossi und Chen orientieren. Wie Donaldson und Lipsey (2006) schreiben, ist jedoch unklar, inwieweit es sich um inhaltlich vergleichbare Ansätze handelt in Anbetracht des häufig synonym verwendeten Vokabulars mit Begriffen wie „... program-theory evaluation, theory-based evaluation, theory-guided evaluation, ...“ (Coryn et al. 2011, S. 200).

Ein weiterer Ansatz der „Familie“ der theoriegeleiteten Evaluationsansätze trägt den Titel **Realistic Evaluation** und wird hier kurz skizziert. Der Evaluationsansatz orientiert sich an den Grundprinzipien des von Chen und Rossi entwickelten theoriegeleiteten Evaluationsansatzes und wurde von den Briten Pawson und Tilley in den 90er Jahren vorgestellt. Realistische Evaluationsansätze bauen auf den Postulaten des wissenschaftlichen Realismus auf, in dem der Schwerpunkt der Evaluationstätigkeit auf die empirische Erhebung und Überprüfung von kausalen Zusammenhängen innerhalb der Wirkmechanismen eines Programms sowie in der Beziehung zur Umwelt (Kontextbedingungen) gelegt wird. Wenn es um die Beurteilung der Zielerreichung und damit um die Effektivität von Programmen geht, konzentrieren sich realistische Evaluationsansätze zunächst immer auf die Untersuchung der zugrunde liegenden Programmtheorie – wie schon beschrieben bei den theoriegeleiteten Evaluationen. Evaluationsansätze, die sich unter dem Ansatz Realistic Evaluation subsumieren lassen, entsprechen nach Chen dem kausalen Typ von theoriegeleiteten Evaluationsansätzen.

Ausgangspunkt war für Pawson und Tilley bei der Erarbeitung ihres Lösungsvorschlages die **Unzufriedenheit mit der Behandlung der Blackbox-Problematik** durch die zum damaligen Zeitpunkt in der Praxis verwendeten Evaluationsansätze. Alleine durch die Anwendung von experimentellen oder quasi-experimentellen Designs ist es schon aus designtechnischen Gründen nach Pawson und Tilley nicht möglich, die Mechanismen im Programm sowie die Wirkzusammenhänge differenziert zu untersuchen (siehe auch Weiss 1972). Als Protagonisten des wissenschaftlichen Realismus haben Pawson und Tilley (1997) verstärkt darauf hingewiesen, dass die Maßnahmen eines Programms nicht nur Schritte in einer Kausalkette zur beabsichtigen Wirkung darstellen, sondern dass **Kontextbedingungen** entscheidend auf die Maßnahmen und dadurch auf die Programmwirkungen Einfluss nehmen können.

Die zentrale Frage, die zugleich die Ausgestaltung und Durchführung einer realistischen Evaluation leiten soll, lautet: Was wirkt für welche Teilnehmergruppen unter Berücksichtigung der Kontextbedingungen wie? Anhand der Fragestellung wird deutlich, dass durch realistische Evaluationen im wissenschaftstheoretischen Sinn Erklärungen für komplexe soziale Phänomene gesucht werden. Bei der Entwicklung des Evaluationsvorhabens können Programmtheorien als komplexe Gebilde entstehen. Der Evaluator muss sich anschließend entscheiden, **ob die gesamte oder nur Bestandteile der Programmtheorie im Rahmen der Evaluationsstudie auf kausale Zusammenhänge überprüft werden sollen**. Dies trifft insbesondere für die Evaluation von sozialen Programmen mit einer hinsichtlich des Umfangs großen sowie heterogenen Teilnehmerstruktur zu als auch für Programme, die an verschiedenen Standorten durchgeführt werden und daher in unterschiedlichen Kontexten „wirken".

Pawson und Tilley haben dafür den so genannten CMO-Ansatz entwickelt. CMO ist die Abkürzung für Context, Mechanisms und Outcomes: „Programs work (have successful ''outcomes'') only in so far as they introduce appropriate ideas and opportunities (''mechanisms'') to groups in the appropriate social and cultural conditions (''contexts")" (Pawson & Tilley 1997, S. 51). In einer formelhaften Schreibweise drückt sich der CMO-Ansatz wie folgt aus:

$$\text{Mechanisms} + \text{Context} = \text{Outcomes}$$

In der Welt der realistischen Evaluation zeichnen sich soziale Programme durch eine komplexe Struktur sozialer Interaktion aus: „[...] realists regard programmes as rather sophisticated social interactions set amidst a complex social reality" (Pawson & Tilley 1997, S. 5). Realistische Evaluationsansätze suchen nach Erklärungsmustern auf der Ebene der Programme und ihrer Umwelt. Die Untersuchung von Kontext, Mechanismus und Ergebnis bildet die Grundlage für die Bewertung, ob ein Programm zielgerichtet, nützlich, wirkungsvoll und somit bestimmten Erfolgskriterien entspricht. Somit bietet die konsequente Befolgung

des CMO-Ansatzes die Möglichkeit, in die Blackbox von Programmen zu schauen (vgl. Kazi 2001, S. 1). Der Ansatz von Pawson Tilley lässt sich außerdem graphisch darstellen und anhand eines Beispiels erläutern.

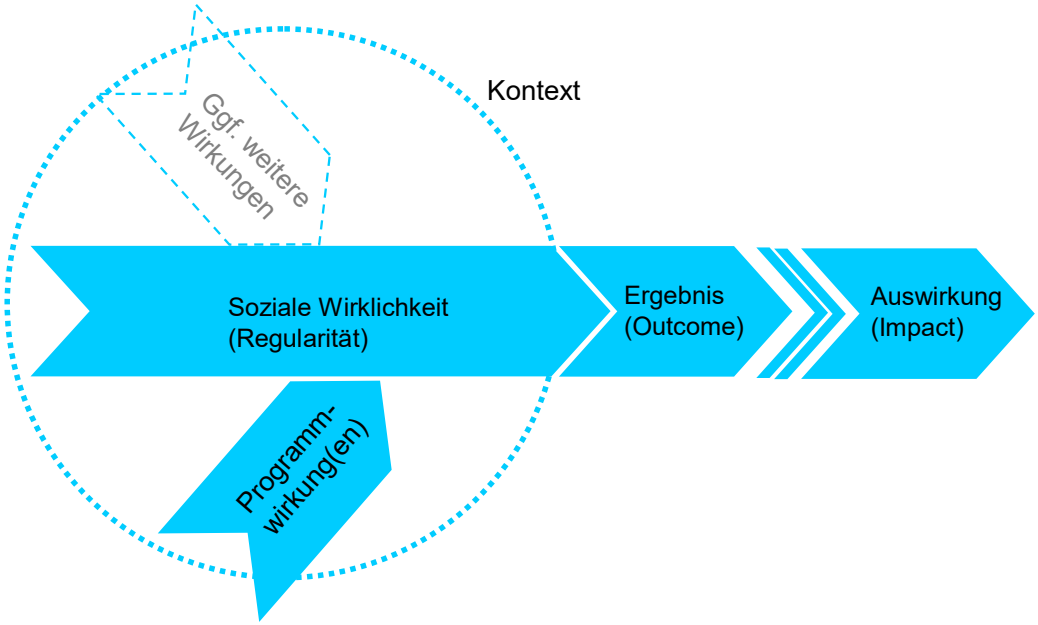


Abbildung 2: Schematische Darstellung der Wirkung von Programmen

In der Grafik ist die Auswirkung einer Intervention durch ein Programm auf bestimmte Strukturen in schematischer Form als Flussdiagramm dargestellt. Am Beispiel der Sprachförderung lässt sich der Wirkungszusammenhang illustrieren. Unter sozialer Wirklichkeit soll beispielsweise die Betreuung von Kindern in einer Kindertagestätte verstanden werden. Bleibt man bei diesem Beispiel, kann die Wirkung einer Interventionsmaßnahme folgendermaßen beschreiben werden: In der Kindertagestätte wurden Defizite in der Sprachentwicklung bei einem Teil der Kinder festgestellt (Zeitpunkt t1). Daraufhin wurde beschlossen, die Kinder mit einem speziellen Sprachförderkonzept für einen festgelegten Zeitraum zu fördern und anschließend noch einmal den Sprachstand der Kinder zu messen (Zeitpunkt t2). Das Sprachförderkonzept ist daher das intervenierende Programm, dessen Wirkung für den Zeitraum t1-t2 bestimmt werden soll. Die Annahme lautet: Da die Sprachförderung bei den Kindern der Kindertagesstätte im Zeitraum t1-t2 durchgeführt wurde, verfügen die geförderten Kinder in der Zeit danach über eine bessere Sprachkompetenz. Im Ergebnis der Sprachförderung drückt sich die durch die Intervention veränderte soziale Wirklichkeit aus. Ver-

änderungen lassen sich beispielsweise durch Sprachtests bei den Kindern erfassen: z.B. den Anteil der geförderten Kinder, die nach der Intervention über einen dem Alter entsprechenden Sprachentwicklungsstand verfügen. Die Nachhaltigkeit der Förderung drückt sich durch die langfristigen Wirkungen (Impact) aus, z.B. bei der Entwicklung der sprachlichen Leistungen von geförderten Kindern in der Grundschule.

Zwei Aspekte erschweren die Wirkungsevaluation: Zum einen müssen für den gesamten Messzeitraum eventuelle Störfaktoren bestimmt und kontrolliert werden. Dies bedarf abhängig von den Programmspezifika einer umfangreichen vorherigen Planung der Evaluationsdurchführung. Zum anderen muss der Kontext des Programms berücksichtigt werden. Auf diesen Umstand haben Pawson und Tilley (1997) in ihrem Konzept der realistischen Evaluation im besonderen Maße hingewiesen, da hiervon die Aussagekraft und damit verbunden die praktische Relevanz der Ergebnisse zusammenhängt. Beide Aspekte sind bei der Konzeption des Evaluationsdesigns zu berücksichtigen.

Pawson und Tilley (1997) betonen zudem, dass Programme keine statischen Konstrukte darstellen, sondern von **fortlaufenden Änderungsprozessen** geprägt sein können. Wenn das Konzept im Verlauf der Weiterentwicklung eines Förderprogramms beispielsweise angepasst wird, kann dies auch zu Änderungen bei der Zielgruppe führen. In der erfassten Programmtheorie müssen dann solche Änderungen nachvollzogen werden. Neben der Konzeption und Durchführung von Wirkungsmessungen kommt auf den Evaluator die weitere Anforderung zu, Veränderungen am Programm zu erfassen und darauf ggf. mit einer Anpassung des Evaluationsdesigns zu reagieren. Ein Beispiel für eine detailliert erfasste Programmtheorie ist im Kapitel 4.4.

In der gegenwärtigen Entwicklung wird der realistische Evaluationsansatz von einigen Evaluationsforschern auf den Prüfstand gestellt. So sieht Rogers (2008) in theoriegeleiteten Evaluationsansätzen mit komplexen und komplizierten Programmtheorien **die Gefahr einer Überschätzung der Möglichkeiten im Rahmen einer Evaluationsstudie** zu durchweg objektiven und validen Ergebnissen zu gelangen. Rogers verweist bei seiner Kritik insbesondere auf die sensible Auswertungs- und Interpretationsphase in einem Evaluationsprojekt, wenn die gesammelten Informationen mit den vor der Datenerhebung festgehaltenen Annahmen der Programmtheorie in Beziehung gesetzt werden. Ein weiterer Hinweis von Rogers ist, dass Programme Veränderungsprozesse durchlaufen. Programme können mit anderen Programmen interagieren oder selbst Teil von umfassenden Programmen sein. So kann ein Sprachförderprogramm für Vorschulkinder eine von mehreren Maßnahmen eines größeren Bildungsprogramms in einer Stadt darstellen.

Die Berücksichtigung der ablaufenden Prozesse bei gleichzeitiger Betrachtung der Kontextbedingungen sei ressourcenintensiv und stelle hohe Ansprüche an die Evaluationsforscher (siehe auch Pedersen & Pieper 2005, Gill & Turbin, 1999). So weist Patton darauf hin, dass eine einmal aufgestellte Programmtheorie sich im Verlauf der Evaluationsstudie nicht stabil verhalten muss: „Once a program is in operation, the relationships between links in the causal hierarchy are likely to be recursive rather than unidirectional. The implementation and attainment of higher-level objectives interact with the implementation of lower-order objectives through feedback mechanisms [and] interactive configurations [...] In short, the cause-effect relationships may be mutual, multidirectional and multilateral" (Patton 1998, S. 232; zitiert nach Rogers 2008, S. 38). Mit steigendem Komplexitätsgrad von sozialen Programmen erscheint es zunehmend schwieriger zu werden, mithilfe von theoriegeleiteten oder realistischen Evaluationsansätzen Wirkungsevaluationen durchzuführen (vgl. Rogers 2008, Coryn et al. 2011, Astbury & Leeuw 2010). In Evaluationsverfahren, die im Auftragsverfahren vergeben werden, könne dies mit restriktiven Zeit- und Ressourcenvorgaben in den seltensten Fällen realisieren. Zu empfehlen sei in diesen Fällen anstatt dessen, nicht zu versuchen, alles mit einer Wirkungsevaluation zu erfassen, sondern sich auf die **Überprüfung von abgrenzbaren Bereichen einer Programmtheorie** zu konzentrieren.

Schließlich konzentriert sich die Kritik auf einen heiklen Punkt von theoriegeleiteten Evaluationen: „In many of the cases reviewed, the explication of a program theory unmistakably was unnecessary, or almost an afterthought in some instances, and was not visibly used in any meaningful way for formulating or prioritizing evaluation questions nor for conceptualizing, designing, conducting, interpreting, or applying the evaluation reported. In these cases, from a methodological perspective, such evaluations very likely would have produced the same results and conclusions even in the absence of articulating or expressing an underlying theory. In other cases, however, the explication of a plausible program theory noticeably was essential to the planning, design, and execution of the evaluation" (Coryn et al. 2011, S. 216). Zusammengefasst bedeutet das Statement von Coryn, dass Programmtheorien sich zwar für den Aufbau eines gemeinsamen Verständnisses über die Funktionsweise eines Programms zwischen Evaluatoren und Programmbeteiligten eignen, für den eigentlichen Bestimmungszweck – die Überprüfung von Wirkungsannahmen innerhalb eines Programms – nur selten verwendet werden.

Auch in der dritten Phase kann eine Weiterentwicklung der Evaluationsmethodik festgestellt werden. In dieser Phase wurden Evaluationsansätze vorgestellt, die umfassende bis hin zu komplexen Evaluationsvorhaben adressieren. Evaluationsstudien können als komplex bezeichnet werden, wenn das Programm an mehreren Standorten und mit unterschiedlichen Teilnehmergruppen umgesetzt wird.

Die Leistung von Rossi und Chen ist es, Ansatzpunkte für eine Zusammenführung der bewährten Evaluationsansätze herzustellen, die sich seit dem Beginn der ersten Phase der Evaluationsforschung in den USA entwickelt haben, um auf diese Weise Lösungsansätze für die Evaluation von komplexen Programmen anzubieten.

Evaluation kann nach Rossi und Chen *mehrere Funktionen* erfüllen: Der Ansatz könne zur Überprüfung der Handlungsziele von Entscheidungsträgern dienen, zur Informationsaufbereitung für Entscheidungsprozesse genutzt werden, zur inhaltlichen Weiterentwicklung des Programms führen oder zur Messung von Programmwirkungen eingesetzt werden. Nach Chen (1990, S. 43) ist es für die Konzeption des Evaluationsdesigns von entscheidender Bedeutung, dass **die Maßnahmen in einem ersten Schritt hinsichtlich ihrer Zielverfolgung genau operationalisiert werden**. Als zweiter Schritt erfolgt die **Ausarbeitung einer Programmtheorie**, d.h. in der Untersuchung bzw. Überprüfung der einem Programm innewohnenden Prozesse sowie der Ergebnisse.

Der realistische Evaluationsansatz nach Pawson und Tilley (1997) geht einen Schritt weiter und bezieht den Kontext von Programmen in die Programmtheorie sowie in das Evaluationsdesign mit ein. In methodischer Hinsicht werden bei realistischen Evaluationen Wirkungsmessungen so geplant und durchgeführt, so dass **Kausalaussagen zu den Wirkungsmechanismen von Programmen** empirisch überprüft werden können. Im Kapitel zur Ausarbeitung des Prozessmodells in der Phase der Entwicklung und Auswahl der Evaluationsinstrumente wird eine Programmtheorie konkret am Beispiel eines zu evaluierenden Programms ausgearbeitet. In der Evaluationsbranche wird der Nutzen einer Programmtheorie für die Behandlung von Blackbox-Problematiken jedoch auch kritisch diskutiert. Die Ergebnisse dieser jüngsten Metaanalysen deuten darauf hin, dass in einem Prozessmodell für die Begleitung von Evaluationsprojekten die Programmtheorie eher für planerische und organisatorische Zwecke als für die Durchführung von Wirkungsevaluationen vorgesehen werden kann (Rogers 2008, Coryn et al. 2011, Astbury und Leeuw 2010).

Für die Gestaltung des Prozessmodells können weitere Elemente herauskristallisiert werden: Dazu für zählen die *projektmanagementorientierte Herangehensweise* im Form einer starken ziel- und verwertungsorientierten Herangehensweise an die Planung und Durchführung von Evaluationsprojekten sowie die ausgeprägte Stakeholderorientierung und das damit verbundene Verständnis von Evaluationstätigkeit als Dienstleistung. Die Weiterentwicklung der Evaluationsmethodik in der dritten Phase charakterisiert sich zudem als eine Kombination von bewährten, quantitativen Methoden mit neuen, prozessmanagementorientierten Vorgehensweisen.

2.4. Typen und Phasen der Evaluationsforschung

Seit den Arbeiten von Campbell haben sich die Ansätze hinsichtlich Ziele, Vorgehensweisen und Methoden der Evaluationsforschung verändert. Die Entwicklung der Evaluationsansätze ist vergleichbar zu dem eines Pfades, von dem in seinem Fortgang Seitenwege abzweigen, die sich noch mal weiter unterteilen. Bildhaft benutzt Alkin (2004) die Metapher eines Baumes mit den ersten als solchen definierten Evaluationsansätzen als dessen Wurzeln. Im Wurzelbereich verortet Alkin die Beiträge von Scriven und Campbell, die gewissermaßen die theoretischen Grundlagen und den Beginn der modernen Evaluationstätigkeit darstellen sollen. Das Wachsen des Baumes hat dazu geführt, dass in der Gegenwart die Baumkrone aus zahlreichen Evaluationsansätzen besteht.

Tatsächlich sind Evaluationsansätze entstanden, die oft gleiche oder zumindest sehr ähnliche Charakteristika aufweisen. Unter Titeln wie „partizipative Evaluation“, „kreative Evaluation“ oder „theoriegeleitete Evaluation“ kann der Anwender im Regelfall kaum differenzieren. Zwar ist für die Praktiker der Evaluation umfangreiches Sondierungsmaterial vorhanden, für die Auswahl und Anwendung der Evaluationsmethoden, besteht jedoch noch Entwicklungsbedarf. Die folgende Tabelle bietet eine Zusammenfassung der drei Zeitphasen der Evaluationsforschung hinsichtlich der in der Einführung zum Kapitel erläuterten Unterscheidungsmerkmale.

Zeitphasen	1955-75	1975-80	Ab 1980
Phasen der Evaluation	Ansätze für Wirkungsmessungen	Prozess- und Nutzenorientierte Ansätze	Theoriegeleitete Ansätze
Vorgestellte Theoretiker	Campbell, Scriven,	Guba, Lincoln, Stake, Patton, Weiss, Cronbach	Rossi, Chen, Pawson & Tilley
Vorherrschende Typen von Evaluation	Wirkungsanalysen	Programmbegleitende Evaluation	Wirkungsanalysen, Programmbegleitende Evaluation
Funktionen von Evaluation	Überwiegend Kontrollfunktion und wachsende Bedeutung der Erkenntnisfunktion	Wachsende Bedeutung der Entwicklungs- und Beratungsfunktion	Alle drei Funktionen möglich: Kontrollfunktion, Entwicklungsfunktion, Erkenntnisfunktion
Methodologische Fokussierung	Methodenzentrierte Vorgehensweise	Konstruktivistischer Ansatz, Methodenmix	Realismus, Pragmatismus, Lösungsansätze für die

			Evaluation von komplexen Programmen
Ideale Rolle des Evaluators	Methodenexperte mit fachlicher Kompetenz	Zusätzlich zu ersten Phase: Vermittler, Berater und Moderator	Zusätzlich Experte im Fachgebiet
Vorherrschende Forschungsmethoden	Experimentelle und quasi-experimentelle Ansätze, standardisierte Tests	Betonung qualitativer Methoden: Fokusgruppen, Fallstudien, kreative Evaluationsverfahren	Methodenmix: Auswahl der Methoden in Abhängigkeit von dem Ziel der Evaluationsstudie
Schwerpunkt der Programmanalyse	Programmeffekte, Output-Untersuchungen, Kosten-Nutzen-Analysen	Prozesse, Kontextbedingungen	Analyse von Programmtheorien: Input, Prozess, Kontext, Outcome,

Tabelle 1: Die drei Phasen der Entwicklung der Evaluationsforschung nach Unterscheidungsmerkmalen

Bei der Betrachtung der drei Phasen wird deutlich, dass die Evaluationsforschung eine kontinuierliche Entwicklung durchlaufen hat. Die Schwerpunktsetzung hat sich im Laufe der Zeit verlagert: Insbesondere sind unter dem Stichwort theoriegeleitete Evaluation in der dritten Phase Evaluationsansätze für die Untersuchung von komplexen Programmen entstanden. Während in der ersten Phase Wirkungsmessungen dominierten, wurde das Instrumentarium der Evaluation im Laufe der Jahre um qualitative Methoden erweitert. Diese Entwicklung hatte zwei Konsequenzen: Evaluationsansätze wurden zwar komplexer, andererseits wurden Studien fokussierter auf die spezifischen Erkenntnisinteressen der Programminitiatoren konzipiert.

Die Weiterentwicklung des Evaluationsinstrumentariums hatte zudem Konsequenzen auf die Gestaltung der Rolle, die Evaluatoren bei der Durchführung von Evaluationsprojekten einnehmen. So kann eine Entwicklung vom reinen Methodenexperten mit fachlicher Kompetenz hin zu einem Typ von Evaluator konstatiert werden, der idealerweise zusätzlich über eine Reihe von individuellen und sozialen Schlüsselkompetenzen verfügen sollte (z.B. Moderationsfähigkeiten).

Die in diesem Kapitel zusammengetragenen Erkenntnisse aus der Betrachtung einer Auswahl von Evaluationsstudien lassen sich für die Gestaltung eines Prozessmodells für die Begleitung von Programmevaluationen systematisch heranziehen. Aus jeder Phase wurden Elemente herausgearbeitet, die sich im Sinne der Weiterentwicklung der Evaluationsmethodik bewährt haben. Aus der ersten Phase kommen insbesondere die methodischen Instrumente für die Durchführung von Wirkungsevaluationen und die damit verbundene Arbeit von Campbell

und Scriven in Betracht. Hier spielen die experimentellen und quasi-experimentellen Untersuchungsdesigns eine hervorgehobene Rolle. Die Produkte der zweiten Phase können für die Konzeption der Grundstruktur des Prozesses sowie für die Festlegung der Prozessteilschritte verwendet werden. Aus dem Evaluationsansatz der partizipativen Evaluation nach Stake lassen sich Aspekte herausarbeiten, die für die Zusammenarbeit zwischen Evaluatoren, Programmverantwortlichen und Programmbeteiligten wichtig sind. Die Evaluationsansätze der dritten Phase bieten schließlich Lösungen für die systematische Untersuchung von Programmmechanismen (theoriegeleitete Evaluation nach Rossi und Chen) sowie für Evaluationsstudien zu Programmen, die an mehreren Standorten durchgeführt werden. Die genaue Systematisierung und Platzierung der Elemente aus den Evaluationsansätzen zu einzelnen Prozessschritten erfolgt im folgenden Kapitel.

3. Ein Prozessmodell für Programmevaluationen

In den vergangenen Jahrzehnten hat eine Entwicklung hin zu Evaluationsansätzen stattgefunden, deren methodisches Instrumentarium mehrere Funktionen von Evaluation erfüllen können. Im Verlauf dieser Entwicklung hat eine Differenzierung der Evaluationsansätze nach wissenschaftstheoretischen Grundüberlegungen in Kombination mit der Methodenvielfalt der Sozial- und Geisteswissenschaften stattgefunden. In Orientierung an einer Auswahl der diskutierten Evaluationsansätze wird in diesem Kapitel zunächst die Grundstruktur eines Prozesses für die Planung und Durchführung von Evaluationsstudien entworfen. Daran anschließend werden die einzelnen Phasen im Evaluationsprozess beschrieben und detailliert.

In die Darstellung fließen Erkenntnisse aus den Ergebnissen von drei Evaluationsstudien ein, die der Autor in einem Zeitraum von sechs Jahren (2001 bis 2007) durchgeführt hat. Die Einsichten, die der Autor aus der Konzeption, Planung und Durchführung der Studien gewonnen hat, werden als Beispiele situativ in den folgenden Unterkapiteln entlang der einzelnen Phasen im Evaluationsprozess eingebracht und diskutiert.

3.1. Hintergrundinformationen zu den drei für die Modellentwicklung analysierten Evaluationsstudien

Alle drei evaluierten Programme beziehen sich auf Sprachförderung für Migranten, jedoch in unterschiedlicher Intensität und mit unterschiedlichen Konzepten. Es erscheint zur Erläuterung des Hintergrunds der Programme sinnvoll, einleitend und kurz auf einige zentrale und allgemeine wissenschaftliche Erkenntnisse zum Stand der sprachlichen Integration von Migranten in Deutschland einzugehen, bevor die Evaluationsstudien hinsichtlich Zielsetzung, Zielgruppen und Rahmenbedingungen im Detail beschrieben werden[24].

Die Ergebnisse der ersten PISA-Studie zu Beginn der 2000er zeigten, dass ein signifikant höherer Anteil von Migrantenkindern über schlechtere Deutschkenntnisse verfügen als vergleichbare deutsche Referenzgruppen (z.B. Prenzel et al. 2006; Artelt & Stanat 2010). Die Integrationsforschung bietet seit langer Zeit empirische Belege dafür, dass das Beherrschen der Sprache des Einwanderungslandes der Schlüssel für die weitere Integration in die Aufnahmegesellschaft bedeutet (vgl. Esser 2006, 2008, 2010; Nauck et al. 1997; Kalter, Granato & Kristen

[24] Innerhalb der Integrationsforschung stellen die Bereiche Spracherwerb und Sprachförderung eigenständige Forschungsfelder dar. Im Rahmen dieser Arbeit kann lediglich der Forschungsstand in diesen Bereichen sehr kurz umrissen werden.

2011). Eine Reihe von Folgeuntersuchungen nach den ersten PISA-Ergebnissen im Jahr 2000 können diese frühen Ergebnisse bestätigten und zeigen außerdem, dass Defizite beim Erwerb und Gebrauch der deutschen Sprache einen negativen Einfluss auf die Bildungsentwicklung im späteren Leben haben – insbesondere was die Schullaufbahn und die Integration auf dem Arbeitsmarkt betrifft (vgl. Naumann et al. 2010; Esser 2006; Kristen & Granato 2004; Prenzel et al. 2006; Haug 2005, 2008). Als Folge der deutlich für sich sprechenden empirischen Ergebnisse zur Sprachkompetenz als Türöffner zur gelungenen Integration in die Aufnahmegesellschaft wurde die Umsetzung von Maßnahmen zur Sprachförderung von Migranten auf breiter Linie in Deutschland vorangetrieben (vgl. Dollmann & Kristen 2010).

Neben der Verpflichtung für Neuzuwanderer, bei fehlenden Deutsch-Sprachkenntnissen an Integrationskursen teilzunehmen, rückten Migrantenkinder in den Fokus der Förderung. Stiftungen, Kommunen, Landesverwaltungen und der Bund stellten Finanzmittel für die Sprachförderung von Kindern und deren Familien zur Verfügung. So findet sich unter dem Stichwort „Frühförderung“ eine Vielzahl an spezifischen Sprachförderprogrammen, die in elementarpädagogischen Einrichtungen durchgeführt werden und deren Zahl stark zugenommen hat (Lisker 2010, 2011). Über die Wirksamkeit der unterschiedlichen Sprachfördermaßnahmen ist nicht viel bekannt (Wolf, Stanat & Wendt 2011, S. 36 f.). Kiziak, Kreuter & Klingholz (2011) vom Berliner Institut für Bevölkerung und Entwicklung machen in ihrer Metaanalyse zur Sprachförderung vom Migrantenkindern deutlich, dass der Sprachstand von Kindern im Vorschulalter zwar mit standardisierten Verfahren regelmäßig erfasst wird, diese Verfahren jedoch in Ihrer Art sehr unterschiedlich und die Ergebnisse daher nicht miteinander vergleichbar sind. Die Ergebnisse der wenigen zugänglichen wissenschaftlichen Meta-Evaluationen zeigen zudem keine signifikanten Aufholprozesse in der Sprachentwicklung bei Migrantenkindern, nachdem diese an einem Förderprogramm teilgenommen haben (vgl. Kiziak, Kreuter & Klingholz 2011, S. 14ff.). Das Ergebnis der Meta-Analyse bedeutet jedoch im Umkehrschluss nicht, dass Sprachförderung in der Einzelfallbetrachtung nicht zu den gewünschten Effekten führt. In den Studien wird darauf hingewiesen, dass die Sprachentwicklung von vielen Faktoren abhängig sein kann, die mit einer Wirksamkeitsüberprüfung alleine nicht erfasst und kontrolliert werden können. Eine Evaluation von drei Sprachförderkonzepten von Hoffmann et al. nennt für den Kindertagesstättenkontext eine Vielzahl an Faktoren, die Effekte von Sprachförderung bewirken (Hofmann et al. 2008, S. 297 ff.). Dazu zählen die Dauer der Förderung, die Motivation der Einrichtungsleistung oder die Einführung von altersdifferenzierten Fördermaßnahmen. Die Autoren empfehlen eine intensivere wissenschaftliche Untersuchung der Effekte von Sprachförderkonzepten und weisen zugleich auf den aktuell unbefriedigenden Erkenntnisstand in dieser Hinsicht hin.

In zwei von den drei am efms durchgeführten Studien wurden Programme evaluiert, in denen mit festen Sprachförderkonzepten gearbeitet wurde. Dies betrifft das Programm *frühstart* und das Programm „Spielend lernen". In der dritten vorgestellten Studie – dem Integrationskurs „In Deutschland zu Hause" – war das primäre Ziel, erwachsenen Migranten Orientierungswissen und sozialkundliche Kenntnisse näher zu bringen. Sprachförderung erfolgte dagegen situativ im Kurszusammenhang.

Im Folgenden wird der Entstehungskontext der Studien kurz erläutert sowie die Ziele der Evaluation beschrieben. Die Gliederung der Vorstellung der Studien orientiert sich an dem Kategorienschema nach Cronbach (1982), mit dessen Hilfe die grundlegenden Informationen zu Evaluationsstudien aufbereitet werden können. Das Schema gliedert sich in die Aspekte *Units* (Teilnehmer am Programm), *Treatments (Programmmaßnahmen)*, *Observations* (erhobene Daten und Informationen im Programm) und *Situations* (Rahmenbedingungen)[25]. Ergänzt werden die Beschreibungen mit Ausführungen zu den Zielen und Hintergründen zum Programm.

3.1.1. Das Programm „In Deutschland zu Hause" – Integrationskurse für Migranten

Ziele und Hintergründe zum Programm

Gemeinsam mit dem Bildungszentrum Nürnberg (BZ) erarbeitete das efms eine Integrationsmaßnahme für seit längerer Zeit in Deutschland lebende Migranten. Dabei handelte es sich um die konzeptionelle Entwicklung, Durchführung und Evaluation von sozialkundlichen Integrationskursen – einer Kursform, die hinsichtlich Inhalt und der Zielgruppe der Vorbereitung der Einbürgerung dienen sollte und zum ersten Mal in Deutschland erprobt wurde. Die Kursreihe trug den Titel „In Deutschland zu Hause – Politik, Geschichte und Alltagswissen für Zuwanderer und Einbürgerungswillige" und wurde im Zeitraum Oktober 2001 bis Mai 2003 durchgeführt[26].

Ziel des Modellprojekts war es, eine Bildungsmaßnahme „konzeptuell zu entwickeln, zu testen und zu evaluieren" (Heckmann & Wunderlich 2001, S. 1). Das Ergebnis und Produkt der Evaluation aus der Projektphase sollte „[...] ein ausgearbeitetes Curriculum, das auch anderen Institutionen zur Verfügung gestellt werden kann" (Heckmann & Wunderlich 2001, S. 1) sein. Für die wissenschaft-

[25] Die Bedeutung der vier Aspekte nach Cronbach wird ausführlich im Kapitel 4 zur Eingangsphase der Evaluation erläutert.

[26] Das Projekt wurde durch das Bayerische Staatsministerium für Arbeit und Sozialordnung, Familie, Frauen und Gesundheit und das Bildungszentrum Nürnberg gefördert.

liche Begleitung der Kurse bedeutete dies, dass das Kurskonzept ein Schwerpunkt der Evaluation während der Modellprojektphase war. Einzelne Kriterien dieser Programmevaluation waren die Qualität des Curriculums, die Eignung der verschiedenen Veranstaltungsformen und die generelle Akzeptanz des Kursangebots seitens der Teilnehmer. Neben der Evaluation des Kursangebotes war die Erfassung von Wirkungen der Integrationskurse auf die Teilnehmer ein weiteres Vorhaben der Evaluationsstudie[27].

Die Konkretisierung der Projektziele im Rahmen einer inhaltlich/didaktischen Ausgestaltung in Grob- und Feinziele und die daraus resultierenden Lehrhandlungen erfolgten in der konzeptionellen Entwicklungsphase des Curriculums, die von dem späteren Leiter der Kurse übernommen wurde. Anknüpfend an den möglichen Nutzen der Integrationskurse für die Teilnehmer, wurden folgende Ziele definiert:

- Die Zielgruppe soll motiviert werden, sich mit dem gesellschaftlichen und politischen System, der Kultur und der Geschichte Deutschlands auseinanderzusetzen.
- Durch den Besuch der Integrationskurse sollen Migranten Möglichkeiten zur weiteren Partizipation in der Aufnahmegesellschaft aufgezeigt werden.
- Parallel zur Vermittlung von sozialkundlichem Wissen sollen die Sprachkenntnisse vertieft werden.
- Die Teilnahme an den Kursen sollte die Zielgruppe auf die Einbürgerung vorbereiten.

Die Kurse zielten zusätzlich darauf ab, „zur Vermittlung von Gefühlen der Zugehörigkeit und Loyalität beizutragen“ (Heckmann et al. 2001, S. 4) und dadurch die identifikatorische Integration zu stärken. Integrationskurse mit kulturellen, gesellschaftlichen und sozialkundlichen Inhalten sollten als Angebot seitens der Aufnahmegesellschaft zur kulturellen Annäherung verstanden werden. Neben einer Stärkung der kulturellen und identifikatorischen Integration der Kursteilnehmer sollten die Integrationskurse Einbürgerungskandidaten auf die Einbürgerung vorbereiten. Die Kursteilnehmer sollten nicht nur bei der Aneignung der notwendigen sozialkundlichen Kenntnisse unterstützt werden, sondern durch ein entsprechendes didaktisches Konzept auch zur Anerkennung und Auseinan-

[27] Seitens der Auftraggeber bestand vor allem hinsichtlich der Wirkungsevaluation ein besonderes Interesse. Hintergrund war hier ein vorangegangenes Gutachten des efms für die Bayerische Staatsregierung zur Integrationssituation von Migranten in Deutschland, indem die damalige theoretische Erkenntnis sowie Ergebnisse der Integrationsforschung verarbeitet wurden. Eine Empfehlung des Gutachtens lautete, die Entwicklung einer Einbürgerungskultur durch geeignete Angebote an die Migranten anzustoßen. Einbürgerungskurse bzw. Integrationskurse sind demgemäß Beispiele für verpflichtende Maßnahmen.

dersetzung mit zentralen demokratischen Grundüberzeugungen angeregt werden. Sicherlich erschien es nicht plausibel, von einem Modellprojekt mit einer zweijährigen Laufzeit zu erwarten, dass es bei den teilnehmenden Migranten zur nachhaltigen Identifikation mit der deutschen Gesellschaft kommt.

Rahmenbedingungen (Settings)

Die Kurse richteten sich allgemein an erwachsene Migranten, die über grundlegende Deutschkenntnisse verfügen. Zu der ersten Zielgruppe, den Einbürgerungskandidaten, konnte eine grobe Schätzung des Teilnehmerpotentials vorgenommen werden. Die Zahl der Einbürgerungen betrug im Durchschnitt der vorhergehenden 10 Jahre etwa 2000 Personen pro Jahr. Die größte Gruppe der Einbürgerungsbewerber stellten durchgehend mit etwa 60% türkische Staatsangehörige. Auf dem zweiten Platz folgten Bewerber aus den Nachfolgestaaten Jugoslawiens. Mit 25% gehören die meisten Antragssteller der Altersgruppe der 30- bis 40-Jährigen an, gefolgt von den 20- bis 30-Jährigen und den unter 20-Jährigen (je etwa 20%). Den Angaben des Nürnberger Einwohneramtes zufolge unterschied sich die Altersstruktur der Antragssteller nicht erheblich im Vergleich der verschiedenen Herkunftsländer[28].

Teilnehmer am Programm (Units)

Das Modellprojekt „Integrationskurse" richtete sich in der ursprünglichen Planung an die Zielgruppe der Einbürgerungskandidaten und Einbürgerungswilligen. Die Zielgruppen für die Teilnahme an dem freiwilligen Kursangebot wurden in den Werbetexten (vorwiegend Broschüren und Anzeigen) folgendermaßen definiert:

- Personen, die sich im Einbürgerungsprozess befinden
- Ausländer mit langjährigen und gesicherten Aufenthaltsstatus, d.h. Personen mit Aufenthaltsberechtigung, befristeter und unbefristeter Aufenthaltserlaubnis
- Aussiedler

Es haben insgesamt während einer Laufzeit von 1,5 Jahren (Oktober 2001 – Mai 2003) 8 Integrationskurse stattgefunden – 5 Abendkurse und 3 Wochenendkurse.

[28] Die Zahlen basieren auf inoffiziellem Zahlenmaterial und Statistiken des Einwohnermeldeamtes, die dem efms zur damaligen Zeit zur Verfügung gestellt wurden. Eine quantitative Aufstellung der Bildungs- und/oder Qualifikationsniveaus konnte durch das Einwohneramt Nürnberg – aufgrund rechtlicher Restriktionen bei der statistischen Erfassung dieser Merkmale – nicht geleistet werden. Aus den Erfahrungen des Amtes mit Einbürgerungsbewerbern ließen sich Qualifikationsniveau und soziale Stellung grob beschreiben. Laut Einwohneramt *„haben viele Bewerber keine Berufsausbildung und teilweise auch nur sehr geringe Schulausbildung"*. Das Einwohneramt regte daraufhin an, bei der Konzeption der Integrationskurse das niedrige Bildungsniveau der potentiellen Teilnehmer besonders zu berücksichtigen.

Bei allen Kurseinheiten war der Frauenanteil überproportional hoch. Insgesamt 38 Frauen (70%) nahmen an dem Kurs teil. An den Wochenendkursen war das Verhältnis mit 13 Frauen und 9 Männern ausgeglichener. Das Durchschnittsalter der Teilnehmer lag bei 37,5 Jahren. Die Teilnehmer der Wochenendkurse waren verglichen zu denen der Abendkurse etwas jünger, jedoch ist vor allem die erzielte Altersspannweite von 32 Jahren bei beiden Kursformen besonders zu betonen.

Die Zusammensetzung der Teilnehmer im Kurs (bezogen auf ihre Herkunftsländer) kann nur als sehr heterogen bezeichnet werden. Insgesamt nahmen Teilnehmer aus 30 unterschiedlichen Ländern an den Kursen teil. Die Herkunftsländer aus dem geographischen Raum Ost- und Südosteuropa (inklusive Russland) stellten mit ca. 40% die größte Gruppe der Kursbesucher. Die zweitgrößte Gruppe war die der Südamerikaner und Afrikaner mit insgesamt gut 23%. Der überwiegende Teil der Teilnehmer lebte seit den 90er Jahren in Deutschland, d.h. zum Zeitpunkt der Kursteilnahme seit 6 bis 7 Jahren. Bemerkenswert war, dass die anteilsmäßig am stärksten in Nürnberg vertretene Gruppe der türkischen Migranten, die auch die stärkste Gruppe der Eingebürgerten stellt, kaum für den Kurs interessiert werden konnte[29].

Trotz intensiver Rekrutierungsbemühungen und breiter Streuung der Kursinformationen sind zum Start der Kursreihen weniger Teilnehmer erschienen als zuvor erhofft. In der Vorbereitungsphase der Kurse wurde damit gerechnet, dass etwa 100 Personen an den Kursen bis zum Schluss teilnehmen werden. Nach Abschluss der Kurse im Sommersemester 2003 haben nur 47 Personen bis zum letzten Kursabend teilgenommen. Der Grund für die wenigen Kursinteressenten wurde von den Organisatoren und dem Evaluator gleichermaßen in den fehlenden Anreizen zum Kursbesuch gesehen. Da es sich bei den Integrationskursen um ein freiwilliges Angebot zur sozialkundlichen Bildung handelte, konnten während der Projektlaufzeit nur gebildete soziale Schichten, also politisch interessierte Personen unter der Migrantenbevölkerung Nürnbergs, angesprochen werden. Um das Kursangebot auf Dauer in Nürnberg zu etablieren, hätten sich (nach Angaben der Organisatoren am Bildungszentrum) mindestens 10 Personen pro Kursreihe anmelden müssen.

[29] Informationen über das Bildungs- und Qualifikationsniveau der Teilnehmer konnten durch die Vorstellungsrunden zu Beginn jeder Kursreihe gewonnen werden. Bei etwa der Hälfte der Kursteilnehmer handelte es sich um Akademiker aus dem Herkunftsland. Teilnehmer ohne einen qualifizierten Bildungsabschluss stellten die Ausnahme dar. Bei den Teilnehmern der Wochenendkurse war das Bildungsniveau höher als bei den Abendveranstaltungen.

Programmmaßnahmen (Treatments)

Die Teilnehmer der Kursreihe erhielten einen Überblick zu Geschichte, Kultur und Rechts- und Wahlsystem der Bundesrepublik Deutschland und zu den Möglichkeiten politischer Mitwirkung. Die Kursreihe wurde in zwei Formen angeboten: als zehnteiliger Abendkurs und als Wochenendkurs. Mit den Integrationskursen am Bildungszentrum wurden im Raum Nürnberg Migranten angesprochen, die schon seit längerer Zeit in Deutschland leben. Als Voraussetzung für die Teilnahme an dem Kurs wurden grundlegende Kenntnisse der deutschen Sprache seitens der Teilnehmer erwartet. Nach der Bewilligung des Antrags im Dezember 2000 übernahm das Bildungszentrum in Nürnberg die Durchführung der Kursreihe und der Autor der Arbeit für das efms die wissenschaftliche Begleitung und Erstellung von Evaluationsberichten.

Erhobene Daten und Informationen in der Evaluationsstudie (Observations)

Die Evaluation des Modellprojekts „In Deutschland zu Hause" wurde als programmbegleitende Evaluation konzipiert, d.h. der Evaluator hat die Umsetzung des Projekts während der gesamten Laufzeit begleitet. Da es sich um einen Integrationskurs für Migranten handelte, gingen zudem Überlegungen in die Evaluationskonzeption ein, wie der Erfolg der Kursreihe gemessen werden kann. In der Erprobungsphase der Kursreihe wurde stark mit verschiedenen Unterrichtsmethoden und didaktischen Hilfsmitteln experimentiert. Daher musste eine Evaluationsstrategie angewendet werden, mit Hilfe derer sowohl der Kursverlauf als auch der Effekt der Kurse bei den Teilnehmern erfasst werden konnte. Die Evaluation von „In Deutschland zu Hause" steht beispielgebend für eine programmbegleitende Evaluationsform, die nicht die Erfassung von Wirkungen im Fokus hat sondern die kontinuierliche Verbesserung des Curriculums.

3.1.2. Das Programm „Spielend lernen in Familie und Stadtteil"

Ziele und Hintergründe zum Programm

„Spielend lernen in Familie und Stadtteil"[30] war ein Integrationsprogramm der Stadt Nürnberg, dass zwischen 2004 und 2008 in den zwei Nürnberger Stadtteilen St. Leonhard/Schweinau und Langwasser durchgeführt wurde. „Spielend lernen" und richtete sich an sozial benachteiligte Familien mit Kindern im Alter von 0 bis 11 Jahren. Größtenteils handelt es sich dabei um Familien mit Migrationshintergrund. Das efms wurde 2005 von der Stadt Nürnberg mit der Evaluation des Programms beauftragt. Die folgenden Angaben beziehen sich daher auf den Evaluierungszeitraum 2004 bis 2007.

[30] Im Folgenden wird das Modellvorhaben mit „Spielend lernen" abgekürzt. „Spielend lernen" wurde nach 2008 durch Folgeprojekte auf Stadtteilebene abgelöst, bei denen weiterhin die Arbeit von so genannten Stadtteilkoordinator/innen im Mittelpunkt standen.

Teilnehmer am Programm (Units)

„Spielend lernen“ strebte eine stärkere Koordination und Vernetzung von Einrichtungen auf Stadtteilebene an, um eine zielgerichtete Unterstützung von benachteiligten Familien zu ermöglichen. Sozial benachteiligten Familien (mehrheitlich mit Migrationshintergrund) wurde bis zum Zeitpunkt der Einschulung ihrer Kinder eine Reihe von Fördermaßnahmen in beiden Projekt-Stadtteilen angeboten. Diese Maßnahmen konzentrierten sich größtenteils auf den Familienkontext und waren zielgruppenspezifisch an Familien mit Kindern verschiedener Altersgruppen adressiert. Die relevanten Zeitpunkte in „Spielend lernen“ waren das erste Lebensjahr eines Kindes, die Kindergartenzeit (4-6 Jahre), der Schuleintritt und der Übergang zu weiterführenden Schulen nach Abschluss der Grundschule. Darüber hinaus standen Familien Angebote zu Verfügung, die entweder direkt in der Familie ansetzten oder in Form sprachlicher Frühförderprogramme in Kindergärten und Kindertageseinrichtungen und in Grundschulen umgesetzt wurden. Bei den Teilnehmern am Programm handelte es sich entweder um Kinder mit Migrationshintergrund nach verschiedenen Altersstufen und/oder deren Eltern. Bei dem hier untersuchten Programm „Phono-logisch: Hand in Hand“ wurde die Phonologie-Fähigkeit bei Kindern im Alter zwischen 5 und 7 Jahren in Kitas der Spielend lernen-Stadtteile untersucht.

Programmmaßnahmen (Treatments)

Das Programm umfasste zum Zeitpunkt der Umsetzung bis 2007 eine Fülle an Einzelmaßnahmen. Es wurde angestrebt, diese Maßnahmen hinsichtlich ihrer Wirkungen zu evaluieren und miteinander konzeptionell zu verknüpfen. „Spielend lernen“ beabsichtigte des Weiteren durch den dauerhaften Einsatz von Stadtteilkoordinator/innen vorhandene Ressourcen im Stadtteil zu bündeln, um Einzelmaßnahmen besser miteinander in Einklang zu bringen.

Wie berichtet, wurden in den Stadtteilen zum Teil unterschiedliche Einzelmaßnahmen angeboten. Das Programm „Phono-logisch: Hand in Hand“ wurde in beiden Stadtteilen durchgeführt. Dieser Umstand sowie die Tatsache, dass es sich um ein standardisiertes Förderprogramm handelte, waren ausschlaggebend für die Entscheidung, Evaluationsmaßnahmen auf dieses Programm zu fokussieren.

Rahmenbedingungen (Settings)

Im Gegensatz zu der Integrationskursreihe handelt es sich bei „Spielend lernen“ um ein komplexes Programm mit vielen Einzelmaßnahmen und Vernetzungsaktivitäten. Die Evaluationsstrategie in „Spielend lernen“ entspricht der einer beratenden, responsiven Evaluation. Der Autor des Evaluationsprojekts war beispielsweise während der Modellphase eng in die Arbeit der Gesamtkoordination des Programms eingebunden. Die Evaluationsergebnisse wurden außerdem in den Gremien der Stadt Nürnberg, die mit der Bearbeitung der Themen Migration

und Integration beauftragt sind, vorgestellt und hinsichtlich möglicher Handlungsempfehlungen diskutiert. Soweit möglich, sollten aufgrund der erhobenen Daten auch Aussagen zu Wirkungen der Maßnahmen von „Spielend lernen“ in den Stadtteilen gemacht werden.

Erhobene Daten und Informationen in der Evaluationsstudie (Observations)

Der Autor hatte Anfang 2005 mit der Arbeit an der Evaluation des gesamten Projekts begonnen. Der Hauptauftrag an das efms lautete, durch Erkenntnisse aus einer geeigneten Programmevaluation die Arbeit der Gesamtkoordination sowie der Stadtteilkoordinator/innen zu unterstützen. Die Evaluationsform von „Spielend lernen“ hatte daher einen **starken formativen Charakter**. Auf diese Weise sollten zu bestimmten Einzelprojekten in „Spielend lernen“ begleitend **Feedback- und Zufriedenheitsbefragungen** durchgeführt werden. Außerdem sollten erste Aussagen zu den Wirkungen von einzelnen Maßnahmen auf Basis der Evaluationsergebnisse ermöglicht werden. In Fokus dieser Betrachtung stand dann die Einzelmaßnahme „Phono logisch: Hand in Hand“. Vor und nach der Durchführung der Förderung mit „Phono logisch“ wurden die phonologischen Fähigkeiten bei Kindern im Vorschulalter erhoben, so dass Aussagen zur Sprachentwicklung in diesem spezifischen Bereich zwischen den Messzeitpunkten möglich wurden. Die Ergebnisse dieser Untersuchung werden Im Kapitel 6 zu der Ergebnisphase von Evaluationsstudien vorgestellt.

3.1.3. Das Programm *frühstart*

Ziele und Hintergründe zum Projekt

frühstart wurde als Pilotprojekt in einem Zeitraum von zwei Jahren in 12 hessischen Kindertagesstätten in Frankfurt am Main, Gießen und Wetzlar durchgeführt. Die Projektpartner im *frühstart*-Projekt waren zum damaligen Zeitpunkt die Gemeinnützige Hertie-Stiftung, die Türkisch Deutsche Gesundheitsstiftung, die Herbert Quandt-Stiftung und das Hessische Sozialministerium. Bis zum Projektabschluss der Pilotphase Ende 2006 kombinierte *frühstart* die Elemente Sprachförderung, interkulturelle Erziehung und Elternarbeit in einem ganzheitlichen Projekt[31].

Das efms hat im August 2004 mit der Arbeit an der wissenschaftlichen Evaluation des gesamten Projekts begonnen. Der Hauptauftrag an das efms lautete, durch eine geeignete Wirkungs- und Programmevaluation herauszufinden, ob die Maßnahmen im Zusammenhang mit *frühstart* zur Verbesserung der Deutschkenntnisse bei den Kindern, zur Förderung der interkulturellen Beziehungen

[31] Informationen zu den Bausteinen der *frühstart*-Förderung: http://www.projekt-fruehstart.de (Letzter Zugriff am 06.12.16).

und zu einem stärkeren Engagement der Eltern durch die Arbeit von Elternbegleitern führen. Vor dem Start der Evaluation wurde – wie in den zuvor diskutierten Evaluationsstudien – ein Antrag entwickelt, in dem entsprechend das Evaluationsdesign spezifiziert wurde und ein Vorschlag für den Ablauf der Evaluation unterbreitet wurde.

Teilnehmer am Programm (Units)

Am Programm beteiligten sich 12 Kindertageseinrichtungen aus Wetzlar, Frankfurt und Gießen. Alle teilnehmenden Kitas zeichneten sich durch einen hohen Anteil von Kindern mit Migrationshintergrund aus sowie einer anderen Erstsprache als Deutsch. Diese Kinder wurden in den Kitas zu *frühstart*-Fördergruppen zusammengefasst und erhielten eine regelmäßige Förderung durch die ErzieherInnen, die im Umgang mit dem pädagogischen Konzept nach Elke Schlösser „Wir verstehen uns gut – Spielerisch Deutsch lernen" begleitend geschult wurden[32].

Rahmenbedingungen (Settings)

Bei *frühstart* wurde eine Wirkungsevaluation der Sprachförderung an drei Standorten in Hessen durchgeführt. Zu allen drei Standorten konnten Kontrollgruppen-Kitas angeworben werden, die sich bereit erklärten, ohne selbst das Förderkonzept „Wir verstehen uns gut" anzuwenden, die Sprachentwicklung der Kinder mit einer anderen Erstsprache als Deutsch mithilfe des selben Screeningverfahrens zu überprüfen wie die *frühstart*-Kitas. Auf diese Weise konnten ein quasiexperimentelles Evaluationsdesign realisiert werden.

Programmmaßnahmen (Treatments)

ErzieherInnen der *frühstart*-Einrichtungen und ehrenamtliche Elternbegleiter wurden in regelmäßigen Fortbildungen für die Förderung ausgebildet. Die Sprachförderung in *frühstart* wurde durch das Konzept „Wir verstehen uns gut – Spielerisch Deutsch lernen" umgesetzt. Das Konzept der Elternarbeit wurde von der Türkisch-Deutschen-Gesundheitsstiftung für *frühstart* entwickelt und zeitgleich zur Sprachförderung in den Kitas angeboten.

Erhobene Daten und Informationen (observations)

In der ersten Erhebungswelle wurden grundlegende Daten und Informationen zum Sprachstand der *frühstart*-Kinder und zum Verlauf der Fortbildungen für ErzieherInnen und Elternbegleiter gesammelt und ausgewertet. Der Schwerpunkt der Evaluation wurde auf die detaillierte Erfassung des Sprachstands der

[32] Elke Schlösser (Dipl. Sozialarbeiterin) ist Expertin im Fachbereich Interkulturelle Pädagogik im Elementarbereich und hat zahlreiche Publikationen zu diesem Thema veröffentlicht. Das Förderkonzept „Wir verstehen uns gut – Spielerisch Deutsch lernen" legt den Schwerpunkt in den pädagogischen Bereich.

Kinder gelegt. In der zweiten Erhebungswelle wurde der Sprachstand der Kinder ein zweites Mal erhoben und die Bereiche Interkulturelle Erziehung und Elternarbeit differenziert evaluiert.

Das Projekt *frühstart* gliedert sich in drei inhaltliche Bereiche: Sprachförderung, Interkulturelle Erziehung und Elternarbeit. Das Ziel der Evaluation von *frühstart* war es, die Wirkungen des Projekts in diesen drei Bereichen zu erfassen und hinsichtlich der intendierten Ziele von *frühstart* zu bewerten. Folgende Fragen sollten im Rahmen der Evaluation beantwortet werden[33]:

Sprachförderung:

- Führt die Sprachförderung durch die ErzieherInnen im Projekt *frühstart* bei den Kindern zu einer Verbesserung der Deutschkenntnisse?
- Haben die ErzieherInnen in den Fortbildungen neue Methoden zur sprachlichen Förderung gelernt?

Interkulturelle Erziehung:

- Konnte die interkulturelle Handlungskompetenz der ErzieherInnen und Eltern-begleiter durch die Teilnahme an den Fortbildungen gestärkt werden?
- Haben ErzieherInnen und Elternbegleiter durch die Teilnahme an den Fortbildungen ihre Kenntnisse im interkulturellen Bereich verbessern können?

Elternarbeit:

- Haben Elternbegleiter durch die Teilnahme an den Fortbildungen Kenntnisse über Methoden der Elternarbeit erhalten?
- Hat die Arbeit der Elternbegleiter zu einer stärkeren Beteiligung der Eltern in der Einrichtung geführt?

Für die Interpretation der Wirkungen der Sprachförderung wurde – wie beschrieben – eine Kontrollgruppe mit Kindern aus Kindertagesstätten gebildet, die nicht am *frühstart*-Projekt teilnehmen. Die Bereiche Interkulturelle Erziehung und Elternarbeit wurden durch standardisierte Befragungen der ErzieherInnen, Eltern-

[33] Die hier dargestellten Fragen entsprechen nicht im Wortlaut den ursprünglichen Fragestellungen. Da das *frühstart*-Konzept im Modellzeitraum (2004 bis 2006) kontinuierlich weiterentwickelt wurde, sind die Fragen entsprechend inhaltlich umformuliert worden. Im Rahmen dieser Arbeit wird im Folgenden nur auf die Methoden, das Verfahren und die Ergebnisse der Wirkungsevaluation eingegangen. Das Evaluationskonzept für das Projekt frühstart beinhaltete darüber hinaus Befragungen von ErzieherInnen und Elternbegleitern.

begleitern und Eltern evaluiert. Im Kapitel zur Umsetzung von Evaluationsstudien wird ausführlich auf die Ergebnisse der *frühstart*-Untersuchung ausschließlich im Bereich der Sprachförderung anhand von zwei Beispielen eingegangen.

3.2. Elemente eines allgemeinen Prozessmodells

Im Vordergrund dieser Arbeit steht die Frage nach den Optimierungspotentialen von Evaluationsprozessen zum Zwecke der Unterstützung von Evaluatoren bei ihrer Tätigkeit. Wie bei jedem Projekt, kann auch für Evaluationsstudien konstatiert werden, dass es formale Ähnlichkeiten im Prozessablauf gibt. Dies bedeutet, dass bestimmte Phasen im Prozess einer Evaluationsstudie in stets der gleichen Reihenfolge einsetzen und zwar unabhängig von den inhaltlichen Zielen und Kontextbedingungen der Evaluation. Die Analyse der Evaluationsansätze auf prozessartige Strukturen zeigt zwar eine hohe Differenzierungskomplexität wenn es um die Messung der Programmwirkungen geht, die Planung und Umsetzung von Evaluationsstudien geschieht jedoch im herkömmlichen Sinne. Bei allen untersuchten Evaluationsansätzen lässt sich eine prozessartige Vorgehensweise identifizieren, die vergleichbare Elemente der Planung und Umsetzung von Evaluationsstudien beinhaltet. Verfahren des Qualitätsmanagements und des Projektmanagements arbeiten mit einer prozessartigen Betrachtung von Abläufen. Für Evaluationsstudien bietet die Systematik dieser Verfahren eine gute Ausgangsbasis, um den Evaluationsprozess transparenter zu gestalten.

Zu Beginn einer Evaluationsstudie werden den Bewerbern um einen Evaluationsauftrag grundsätzliche Informationen über die Ziele, Inhalte und organisatorischen Abläufe des Programms zur Verfügung gestellt. Dabei kann es sich um den inhaltlichen Input seitens des Auftraggebers handeln, der in die Ausarbeitung und Durchführung der Evaluation einfließen soll. Der Auftraggeber von Evaluationsstudien unterstützt den Evaluator, da er an gesicherten Ergebnissen der Evaluationsstudie interessiert ist, die ihm wiederum bei seinen Überlegungen unterstützen sollen, wie mit dem Programm oder der Maßnahme weiter verfahren werden soll. Im betriebswirtschaftlichen Sinn ist die Evaluation von Programmen als spezifischer Unterstützungsprozess, der Teil eines klassischen Managementprozesses zur Optimierung der Unternehmensperformance ist, zu verstehen. Als Prozess wird eine Kette von zeitlich miteinander verbundenen Aktivitäten bezeichnet, die ausgelöst durch ein Ereignis bis zu einem Endzeitpunkt ablaufen. Bei Berücksichtigung des Input- und Output-Gedankens von Prozessen kann ein Prozess allgemein wie folgt definiert werden: „Any activity or group of activities that takes an input, adds value to it, and provides an output to an internal or external customer. Processes use an organisation´s resources to provide definitive results." (Harrington 1991, S. 9).

Nach Osterloh und Frost (2006) kann zwischen Kern- und Supportprozessen unterschieden werden (vgl. Daniel 2008, S. 52). Abhängig von Typus und Form der Evaluation (z.B. internes Verfahren oder extern beauftragte Organisation) kann der Evaluationsprozess als Kernprozess oder als Unterstützungsprozess definiert werden. In den meisten Fällen ist Evaluation als Unterstützungsprozess definiert. Managementprozesse sind klar den Kernprozessen zuzuordnen: „Ein Managementprozess besteht aus einer auf das Entstehen von Performance ausgerichteten, strukturierten Sammlung von Planungs-, Entscheidungs-, Durchsetzungs- und Kontrollaktivitäten, die so auf die betriebliche Leitungssphäre einzuwirken haben, dass die Unternehmensziele nachhaltig erreicht werden“ (Daniel 2008, S. 66). Nach Wild (1982) besteht die Managementfunktion aus Zielfindung und -durchsetzung sowie aus Problemerkenntnis und -analyse. Unterstützungsprozesse sind Prozesse, deren Ergebnisse einen Beitrag zur Entscheidungsfindung in den Kernprozessen leisten (Daniel 2008, S. 64).

Durch Management induzierte Veränderungen vollziehen sich aufgrund von Steuerungs- und Regelungseingriffen. An dieser Stelle können Evaluationen durch aufbereitete empirische Daten als Entscheidungsgrundlage eine entscheidende Rolle für Managementprozesse spielen. Im Fall von Auftragsevaluationen wird die Evaluation nicht intern, sondern durch einen Auftragnehmer durchgeführt. Im Bereich der Bildungsförderung sind typische Prozessverantwortliche private Stiftungen, Regierungsorganisationen (auf Bundes- oder Länderebene), Kommunalverwaltungen, Verbände sowie zu einem kleineren Teil NGOs.

Der Evaluationsprozess kann aus der unternehmerischen Perspektive zur Optimierung von Organisationsabläufen betrachtet werden. Evaluation nimmt dann die Funktion der Bewertung der Funktionsbereiche von Organisationen ein. Auf den Bereich der Bildungsförderung übertragen bedeutet dies, dass Bildungsförderer (in Gestalt von Auftraggebern der Evaluation) Evaluationsergebnisse nutzen, um das Erreichen ihrer mit den initiierten Bildungsprogrammen verbundenen Ziele zu bestimmen sowie daraus Maßnahmen für deren Optimierung abzuleiten. Ein typischer Projektmanagementprozess kann so in der Evaluationspraxis nützlich umgesetzt werden. Im Grunde ergibt sich für die Planung und Durchführung von Evaluationsstudien ein *in seinen Grundzügen generischer und in seinen Einzelphasen logisch aufeinander bauender Prozess*, der sich in seiner relativ abstrakten Form als Rahmenwerk in Evaluationsstudien anwenden lässt. Der in Abbildung 3 dargestellte Prozess stellt den allgemeinen Ablauf der Evaluation von sozialen Programmen dar. Im Folgenden wird dieser allgemeine Prozess in den jeweiligen Einzelphasen differenziert betrachtet. Als Erörterungsgrundlage dazu dienen die im zweiten Kapitel schwerpunktmäßig analysierten Evaluationsansätze sowie die Erkenntnisse des Autors aus selbst konzipierten und umgesetzten Evaluationsstudien.

Ein Hauptergebnis des zweiten Kapitels ist die Illustration der Komplexität existierender Evaluationsansätze für soziale Programme hinsichtlich der Zielsetzungen, den wissenschaftsphilosophischen Grundannahmen, der Berücksichtigung der Programmumwelt (Kontextbedingungen) und der Art und Weise der Anwendung der vorgeschlagenen Methoden der Evaluation. Es wurde gezeigt, dass die besprochenen Evaluationsansätze sich hinsichtlich bestimmter Hauptcharakteristika wie Methodenbezug oder Nutzungszentrierung bereits in der Vergangenheit typisiert wurden (vgl. Cook & Matt 1990, Shadish et al. 1991). Diese Typisierungen helfen bei der Zuordnung der verschiedenen Evaluationsansätze zu den einzelnen Phasen eines Evaluationsprojekts. In Anbetracht der dargestellten Vielschichtigkeit der Evaluationsansätze ist für die Evaluationspraxis die Information wertvoll, welche Erkenntnisse aus den Ansätzen zu welchen Phasen im Evaluationsprozess Berücksichtigung finden können. Dadurch geht der prozessorientierte Ansatz über eine reine Typisierung von Evaluationsmethoden hinaus und bietet vielmehr Evaluatoren eine Orientierungshilfe in ihrer Arbeit.

Abbildung 3: Nach Phasen unterteilter Prozess der Evaluation von sozialen Programmen

Grob kann zunächst der Evaluationsprozess folgendermaßen beschreiben werden: Evaluationsprozesse beginnen mit einer Start- bzw. Eingangsphase, in der programmrelevante Informationen gesichtet und für die spätere Verwendung aufbereitet werden. Die Auseinandersetzung mit den verschiedenen Zielen, die mit dem zu evaluierenden Programm verbunden sind, sollte von Auftragnehmern als erster Schritt in der Konzeption einer Evaluationsstudie wahrgenommen werden. Neben den Zielen lassen sich die Programmcharakteristika (z. B. das Curriculum eines Förderprogramms) dazu zählen. In der Eingangsphase verschaffen sich Evaluatoren typischerweise einen umfassenden Überblich zu dem

Evaluationsobjekt sowie den Kontextbedingungen, bevor in der sich daran anschließenden zweiten Phase das Evaluationsdesign entwickelt wird. Auf die Entwicklung des Evaluationsdesigns folgt die Umsetzungsphase, in der die Methoden angewendet und Daten erhoben werden. In der Ergebnisphase werden die Daten ausgewertet und im Fall einer Auftragsevaluation an den Auftraggeber weitergeleitet.

Die im zweiten Kapitel dargestellten Ansätze wurden in Abbildung 3 den vier Phasen eines Evaluationsprozesses zugeordnet. Die Verteilung der Ansätze auf die Evaluationsphasen folgt der grundsätzlichen Logik, dass Ansätze mit einer methodischen Schwerpunktsetzung in der Entwicklung und Umsetzungsphase relevant sind. So kann z.B. der Evaluationsansatz von Campbell insbesondere in der Entwicklungs- und Umsetzungsphase für Wirkungsevaluationen herangezogen werden. Aus den ziel- und nutzenorientierten Überlegungen von Scriven können ebenso wie aus dem responsiven, kollaborativen Ansatz von Stake können Erkenntnisse für die Gestaltung einer Handlungsanleitung für die Eingangsphase von Evaluationsstudien verwendet werden. Diese nutzungs- bzw. kommunikationsorientierte Ansätze können zudem für die Gestaltung der Endphase in einer Evaluationsstudien Anwendung finden. In ihren Publikationen thematisiert z.B. Weiss den Aspekt, wie die Evaluationsergebnisse mit Programminitiatoren und Auftraggebern von Evaluationsstudien diskutiert werden können. Der vorgestellte generische Evaluationsprozess dient daher als Ausgangsbasis für die Entwicklung eines differenzierten Vorgehens bei der Evaluation von Programmen in der Praxis. Der individuelle Beitrag der einzelnen Evaluationsansätze für die Gestaltung des Prozessmodells wird in den folgenden Kapiteln im Detail dargelegt und diskutiert.

4. Eingangsphase der Evaluation

In der Literatur zu Evaluation gibt es eine Reihe von Anleitungen und Vorschlägen dazu, wie der Einstieg in die Evaluation eines sozialen Programms erfolgen sollte (vgl. Wottawa & Thierau 1998, Bortz & Döring 2003). Die Vorschläge unterscheiden sich insbesondere dadurch, ob die Ziele der Evaluationsstudie vor dem Start des Verfahrens feststehen bzw. durch die Programminitiatoren als solche festgelegt wurden, oder ob vergleichsweise wenig über den Evaluationsgegenstand bekannt ist. Daher ist es das primäre Ziel des Evaluators, sich eine gute Informationslage über den zu evaluierenden Gegenstand vor dem Einsatz von Evaluationsmethoden zu verschaffen. Die Entscheidung für oder gegen bestimmte Evaluationsinstrumente, die Fokussierung auf bestimmte Elemente des Programms, die es sich zu evaluieren lohnt sowie die Datenerhebung und Datenauswertung sollten sorgfältig geplante Schritte im Evaluationsprozess sein.

Neben der **Informationssammlung** ist die Analyse des Programms von zentraler Bedeutung, ob eine Evaluation überhaupt in der Praxis durchgeführt werden kann und ob Ergebnisse erwartbar sind, anhand derer die Erkenntnisinteressen der Auftraggeber der Evaluation und der Programminitiatoren befriedigt werden können. Das zu diesem Zweck durchführbare Verfahren wird **Evaluierbarkeitsprüfung** genannt.

Bei diesem Vorgehen darf sich das Evaluationsteam nicht davor scheuen, das „Pferd von hinten aufzuzäumen“ sondern im Gegenteil diese Vorgehensweise als Strategie für sich entdecken. Dies bedeutet, dass zuerst die intendierten Programmeffekte in den Fokus der Betrachtungen gerückt werden sollten. Lassen sich die Programmwirkungen quantifizieren? Bei Förderprogrammen ist eine konkrete Beschreibung dessen wichtig, was die Teilnehmer nach dem Abschluss des Programms in der Lage sein sollten. Derartige Informationen zur Outcome-Orientierung sind aus dem Curriculum oder dem inhaltlichen Konzept des Programms zu entnehmen. Diese intendierten Programmergebnisse dienen zu einem späteren Zeitpunkt der Evaluationskonzeptausarbeitung als Kriterien für die Festlegung von Evaluationsmethoden und Datenauswertungsstrategien.

Der Evaluationsprozess lässt sich in dieser Eingangsphase in mehrere aufeinander aufbauende Arbeitsschritte unterteilen. Grob setzen sich diese Schritte zunächst aus der Informationsbeschaffung und der Analyse dieser Programminformationen zusammen. Erst nach dieser Eingangsphase befasst sich das Evaluationsteam mit der Ausarbeitung der Evaluationsmethoden. In den nächsten Absätzen werden die Einzelschritte in der Eingangsphase der Evaluation beschrieben und mit Praxisbeispielen und Theoriebezügen unterfüttert. Die folgenden Ausführungen und Hinweise lassen sich gut bei **Auftragsevaluationen** anwenden –

wenn ein Evaluator, mehrere Evaluatoren oder eine Institution mit der Durchführung einer Evaluationsstudie zu einem sozialen Programm beauftragt wurden.

4.1. Informationssammlung

Die Entscheidung für bestimmte Evaluationsinstrumente, die Auswahl des Evaluationsgegenstandes und die Datenerhebung müssen sorgfältig geplante Schritte im Evaluationsprozess sein. Die Planung einer Evaluationsstudie in die Praxis ist des Weiteren immer in Anbetracht der Rahmenbedingungen zu betrachten. Eine sorgfältige Aufbereitung der zur Verfügung stehenden Informationen stellt den ersten Schritt bei der Konzeption einer Evaluationsstudie dar. Bei der **Informationssammlung** trägt der Evaluator alle zugänglichen Informationen zu dem Programm zusammen, strukturiert und bewertet diese.

Dabei empfiehlt es sich, **die gesammelten Informationen nach bestimmten Kategorien und Bereichen zu gruppieren**. Eine auf diese Weise aufbereitete Informationslage zum Programm erleichtert die sich daran anschließenden analytischen und konzeptionellen Arbeitsschritte bis hin der zu einem späteren Zeitpunkt zu erfolgenden Ausarbeitung der Evaluationsmethoden. Eine Zusammenstellung von Informationen und Daten zu einem Programm kann beispielsweise entlang der Auseinandersetzung der Evaluatoren mit den folgenden **Leitfragen** erfolgen:

- Liegen Informationen in Form von Konzepten, Programmbeschreibungen, Internetauftritten, Berichten, Datensätzen etc. vor, aus denen Aussagen zu den Zielen, Inhalten, Ressourcen, Rahmenbedingungen und der Arbeitsweise des Programms entnommen werden können?
- Zeichnet sich das Konzept des zu evaluierenden Programms durch wissenschaftliche Fundierung aus (z.B. das Konzept eines Förderprogramms beruht auf einer spezifischen Lerntheorie)?
- Wurden Ergebnisse des Programms chronologisch festgehalten (z.B. in Publikationen, im Internet) und sind diese zugänglich?
- Sind Informationen zu den Zielen der beabsichtigten Evaluation des Programms zugänglich?
- Gibt es Ansprechpartner oder Wissensträger auf den Seiten der Programmbeteiligten oder Programmverantwortlichen, die zu dem Programm und zu den Zielen der Evaluation detaillierte Auskünfte geben können?
- Wurden zu einem früheren Zeitpunkt bereits Evaluationen durchgeführt? Sind das Evaluationskonzept, Daten sowie Berichte zu den Ergebnissen der Evaluationen zugänglich?

In den meisten Fällen stellt der Auftraggeber Informationen über die Charakteristika des Programms, dessen Entstehungsgeschichte sowie aktuelle Daten zur Programmdurchführung (z.B. Ansprechpersonen, Teilnehmerstatistiken) zur Verfügung. Die gesammelten Informationen sollten als nächstes nach bestimmten Merkmalen geordnet werden, die eine spätere Verwendung – z.B. die Erstellung des Evaluationskonzepts – erleichtern. Eine empfehlenswerte Kategorisierung sind die Evaluationsmerkmale nach Cronbach (1982), wie sie bereits bei der Beschreibung der Studien im Kapitel zuvor beispielgebend skizziert wurden. Cronbach (1982) nennt unter dem Akronym **UTOS** spezifische Merkmale für die Charakterisierung Evaluationsstudien. UTOS lässt sich nicht nur auf Evaluationsverfahren sondern auch für die Beschreibung von sozialen Programmen und – wie noch zu zeigen sein wird – für die weitere Ausarbeitung des Evaluationskonzepts verwenden. Cronbach unterscheidet zwischen folgenden Merkmalen:

- **Units** (Teilnehmer am Programm),
- **Treatments** (Programmmaßnahmen),
- **Observing operations** (erhobene Daten und Informationen) und
- **Settings** (Rahmenbedingungen) (Cronbach 1982, S. 82-84).

Diese vier Merkmale lassen sich noch sinnvoll mit einem fünften Merkmal zu den **Zielen des Programms** erweitern, die im nächsten Kapitel behandelt werden. Cronbach verwendete ursprünglich das UTOS-Konzept, um die Ergebnisse von Evaluationen bei großen, komplexen Bildungsprogrammen hinsichtlich der Möglichkeiten ihrer Generalisierbarkeit zu erläutern. So sei UTOS (großgeschrieben) die Grundgesamtheit der Teilnehmer im Programm, der Programmmaßnahmen, der erhobenen Daten und Informationen sowie der Rahmenbedingungen. Bei utoS (kleingeschrieben) handelt es sich um die Stichproben aus der Grundgesamtheit, wobei S großgeschrieben bleibt, da aus den Rahmenbedingungen keine Stichprobe gezogen werden kann. Evaluationsstudien können direkt die Grundgesamtheit UTOS behandeln oder sich auf Stichproben und somit utoS beschränken. Nach Cronbach sollte es das Ziel von Evaluationen sein, Ergebnisse mit validen Aussagen zu UTOS zu produzieren, d.h. die Generalisierbarkeit der Ergebnisse von utoS zu UTOS herzustellen (Cronbach 1982, S. 84). Am Beispiel von *frühstart* wurde das UTOS-Konzept bereits in Kapitel 3.1.3. angewendet. Die Evaluationsstudie zu *frühstart* befasste sich mit der Grundgesamtheit (UTOS), indem der Versuch unternommen wurde, alle Teilnehmer an allen Standorten sowie alle Maßnahmen in einem Evaluationskonzept zu integrieren.

Wie gründlich sich der Evaluator schon zu Beginn mit der Informationslage auseinandersetzen muss, kann davon abhängig sein, in welcher Entwicklungsphase sich das Programm befindet und welchen Komplexitätsgrad es aufweist (z.B.

mehrere Standorte, mehrere Programmmaßnahmen). Zum Merkmal *observing operations* liegen beispielsweise mehr Informationen vor, wenn das Programm bereits einmal evaluiert wurde. Je komplexer und weiterentwickelter ein Programm ist, desto mehr Informationen müssen seitens der Evaluatoren verarbeitet werden. Es kann daher sein, dass Evaluatoren als vorbereitende Tätigkeit umfangreiche Dokumente lesen und die brauchbaren Informationen daraus exzerpieren müssen. Auf diese Art und Weise lassen sich alle grundlegenden Informationen zusammenstellen, die für die weitere Konzeption der Evaluationsstudie verwendet werden können.

Ein spezieller Fall in der Vorbereitungsphase tritt ein, wenn der Vergabe einer Evaluationsstudie ein **wettbewerbliches Verfahren** oder öffentliche Ausschreibung vorgelagert ist. Evaluatoren oder Evaluatorenteams können sich mit einem entsprechenden Angebot um die Evaluationsstudie bewerben. In diesem Fall schließt sich an die Informationssammlung die Ausarbeitung der Antrags- bzw. Wettbewerbsunterlagen durch das Evaluatorenteam an. Die Antragsunterlagen setzen sich üblicherweise aus dem Evaluationskonzept, einem Arbeits- und Zeitplan sowie einer Darstellung der benötigten Ressourcen in finanzieller und personeller Hinsicht zusammen. Das Evaluationskonzept kann wiederum eine strukturierte Beschreibung der Ziele der Evaluation, der Evaluationsmethoden, der Vorgehensweise sowie der anvisierten Produkte der Evaluationsstudie beinhalten. Handelt es sich um eine Auftragsevaluation, muss sich der Bewerber schon zum Zeitpunkt der Angebotsausarbeitung dezidiert mit evaluationsmethodischen Fragestellungen, wie z.B. Zugangswege zur Teilnehmergruppe, die Auswahl geeigneter Instrumente sowie die Auswertung und Präsentation der Ergebnisse auseinandersetzen.

Besonders bei öffentlichen Ausschreibungen werden die Ziele des Programms und der Evaluation veröffentlicht. Es erscheint empfehlenswert, Informationen zu den Programm- und Evaluationsziele selbst im Rahmen von wettbewerblichen Auswahlverfahren genau zu bewerten. In Dokumenten der Programminitiatoren werden häufig strategische Ziele genannt, die für die Planung eines Evaluationsdesigns zu unspezifisch sein können. Bewerber sollten daher Möglichkeiten nutzen, mit den Initiatoren bzw. Programmverantwortlichen in direkten Kontakt zu treten, um offene Fragen zu den Programm- und Evaluationszielen, der Intentionen der Auftraggeber und den Rahmenbedingungen zu erörtern. Die frühe Kontaktaufnahme mit den Auftraggebern der Evaluation signalisiert zudem Interesse für das Evaluationsprojekt und Engagement seitens des Bewerbers. In manchen Fällen gibt es für Bewerber keine Möglichkeit, mit den Auftraggebern in Kontakt zu treten bzw. detailliertere Informationen zu erhalten. In diesen Fällen muss aus den vorliegenden Informationen ein Evaluationskonzept entwickelt werden.

Die Planung des Evaluationsverfahrens kann jedoch fortgeführt werden, wenn kein Auswahlverfahren mit der Vergabe einer Evaluationsstudie verbunden ist. In diesem Fall kann direkt mit dem nächsten Schritt – der Erarbeitung von Programm- und Evaluationszielen begonnen werden.

4.2. Detaillierte Erarbeitung von Programm- und Evaluationszielen

Ein entscheidender Aspekt der Vorstrukturierung ist die Aufnahme der **Intentionen und Ziele**, die Auftraggeber und Programmmanager mit dem Programm verfolgen. Welche Ziele verfolgen der Programmverantwortliche mit den Evaluationsergebnissen? Welche Vorüberlegungen existieren hinsichtlich der Ausweitung des Programms? Wie wird die Nachhaltigkeit des Programms sichergestellt? In der Phase der Informationssammlung werden die Ziele des Programms erfasst, die aus den Dokumenten und Materialen zum Programm hervorgehen.

Angelehnt an Projektmanagementmethoden ist eine auf Ziele fokussierte Vorgehensweise sehr hilfreich. Zieldefinitionen erfüllen, bildlich gesprochen, die Funktion von Bahnschienen: Sie helfen, dass man nicht unkontrolliert von einem eingeschlagenen Weg abweicht. Mithilfe von Zielen können auch Weichenstellungen vorgenommen werden. Ändern sich Programminhalte und Strategien, lassen sich die Ziele von sozialen Programmen entsprechend anpassen. Das Gleiche gilt entsprechend auch für die Evaluations- und Programmziele.

Der amerikanische Evaluationstheoretiker Scriven forderte als ein zentrales Element seines Evaluationsansatzes eine bis dato neue Herangehensweise an die Auseinandersetzung mit einem sozialen Programm und seinen Zielen. Als Vertreter eines Evaluationsverständnisses, das den Hauptzweck von Evaluation in der qualitativen Beurteilung von Programmen hinsichtlich eines Mehrwerts für die Teilnehmer betrachtet, liegt eines seiner Interessen in der analytischen Auseinandersetzung mit den Zielen eines Programms. Im Vorfeld der Evaluation definierte Ziele sollten von den Stakeholdern und Evaluatoren nicht als gegeben hingenommen sondern sollten im Rahmen der Vorbereitung einer Evaluationsstudie in Anbetracht ihres Nutzens analysiert werden. Ein Schlagwort der Evaluationsforschung, das im Zusammenhang der Zielklärung von Scriven geprägt wurde, ist der Begriff **goal free evaluation**. Unter dem Begriff soll **nicht** ein Verfahren verstanden werden, das ziellos verläuft, sondern **die differenzierte Herausarbeitung von relevanten und nicht relevanten Zielen des Programms** während der ersten Phase der Evaluation. Die Planung von Evaluationsstudien solle nach Scriven *goal free* begonnen werden, d.h. der Evaluator sollte sich nicht durch existierende Ziele und Programmatiken beeinflussen lassen. Scriven fordert, dass als Resultat einer mit Programmverantwortlichen begangenen Analyse

nicht brauchbare Ziele für die Evaluation des Programms verworfen und an deren Stelle erreichbare, realistische Evaluationsziele gesetzt werden sollten. Die Ziele sollten „Es ist jedoch recht unwichtig, wie gut man Ziele erreicht, wenn sie überhaupt nicht wert sind, erreicht zu werden“ (Scriven 1972, S. 72).

Aus der Betriebswirtschaftlehre und dem Controlling hat eine Art und Weise der Definition von Zielen Eingang gefunden, die heutzutage im Projektmanagement breite Anwendung findet (Doran G.T., 1981) und die für den Ansatz der goal free evaluation nach Scriven gut genutzt werden kann. Mit dem Begriff **SMART** wird eine bestimmte Vorgehensweise bei der Definition verstanden, die zu einer eindeutigen Festlegung eines Ziels führt. SMART ist die Abkürzung für „Specific, Measurable, Accepted, Realistic, Timely“; die Wörter repräsentieren zugleich die Hauptkriterien für die Formulierung von smarten Zielen. Bei Zieldefinitionen muss stets darauf geachtet werden, dass auf alle Kriterien semantisch ein Bezug hergestellt wird. Die Formulierung von smarten Zielen kann für das Evaluationsteam zum Start einer Evaluationsstudie zur schwierigen Aufgabe werden. Zu einer Herausforderung wird es insbesondere dann, wenn aus der Kommunikation mit den Programminitiatoren hervorgeht, dass unterschiedliche Vorstellungen hinsichtlich der Kriterien von smarten Zielen existieren. In diesem Fall besteht aber die Chance, dass aufgrund der Transparenz und Eindeutigkeit, die smarte Ziele mit sich bringen, schnell der Konsens gefunden wird. Der unschlagbare Gewinn der Methode der smarten Zieldefinition für die Entwicklung von Evaluationskonzepten liegt in dem Umstand begründet, dass gleich mehrere Erkenntnisse der Evaluationsforschung in den Kriterien zur Definitionsbildung integriert werden. Mit „specific und measurable“ sind die empirischen Standards von Campbell und Cook adressiert. Das Kriterium „accepted“ setzt einen Konsensbildungsprozess voraus, ähnlich den responsiven Evaluationsansätzen von Patton und Stake. Mit „realistic“ und „timely“ sind pragmatische Evaluationsansätze thematisiert, wie beispielsweise von Rossi und Fremann (1983, 1999), Weiss (1972) und Pawson und Tilley (1997).

Die Anwendung von smarten Zieldefinitionen kann am Beispiel des Projekts „Spielend lernen“ illustriert werden. In der Planungsphase für die Evaluation von „Spielend lernen“ setzte sich der Evaluator mit den von der Stadt Nürnberg zur Verfügung gestellten Informationen auseinander. Aus den Unterlagen war zu entnehmen, dass im Programm „Spielend lernen“ mehrere Fördermaßnahmen zusammengeführt werden, die „in ihrem Zusammenspiel eine neue Kultur des Aufwachsens im sozialen Nahraum zum Gegenstand haben“ (Wolf 2006, S. 2). Ein strategisches Ziel war es u.a., die Chancengleichheit von Kindern im Bildungssystem durch frühkindliche Förderung und Stärkung der elterlichen Erziehungskompetenz zu erreichen. Legt man die Kriterien von smarten Zieldefinitionen für die Analyse dieses Ziels zugrunde, ist die Zieldefinition für die weitere

Gestaltung eines Evaluationsvorhabens ungeeignet. So ist die gesamte Zielformulierung zu unkonkret, nicht eindeutig und zu vage formuliert (Kriterium „Specific"). Die Definition enthält keine Anhaltspunkte dazu, wie das Erreichen der „Chancengleichheit von Kindern im Bildungssystem" gemessen werden kann (Kriterium „Measurable"). Des Weiteren werden keine Aussagen zu zeitlichen Aspekten gemacht, z.B. in welchem Zeitraum die Förderung stattfindet (Kriterium „Timely"). Auch wenn das Programmziel von den Projektbeteiligten gleichermaßen als relevant gesehen wird (Kriterium „Accepted"), kann aufgrund der unkonkreten Formulierung nicht abgeschätzt werden, ob das Ziel machbar ist (Kriterium „Realistic").

Um brauchbare Zielformulierungen zu erarbeiten, mussten mehr Details über das Programm „Spielend lernen" in Erfahrung gebracht werden. Neben der Auswertung von Maßnahmenbeschreibungen wurde mit der Gesamtkoordination von „Spielend lernen" ein halbstandardisiertes Leitfadeninterview durchgeführt. Aus den Ergebnissen des Interviews sowie weiteren Dokumenten zum Programm wurden folgende **Maßnahmenschwerpunkte** identifiziert, die für die weitere Planung der Evaluationsstudie relevant waren:

- Messung von Sprachkenntnissen bei Kindern im Vorschulalter,
- Inhaltliche und organisatorische Verknüpfung von Fördermaßnahmen im Frühförderbereich,
- Vernetzung der Akteure und mitwirkenden Institutionen im Projekt (z.B. Stadtteilmoderatoren) zum Zweck der verbesserten Zielgruppenerreichung,
- Elternarbeit, Empowerment und Qualifizierung von Personal im Sozialbereich,
- Initiierung weiterer Maßnahmen zur gesellschaftlichen Integration benachteiligter Familien.

Die Maßnahmenschwerpunkte beinhaltenen wiederum eine oder mehrere Maßnahmen. Gemeinsam mit der Gesamtkoordination wurde danach entschieden, welche Maßnahmenschwerpunkte und Einzelmaßnahmen Gegenstand der Evaluationsstudien sein sollten. Im Maßnahmenschwerpunkt „Messung von Sprachkenntnissen bei Kindern im Vorschulalter" fiel die Wahl auf das Projekt „Phono-logisch" – ein Phonologie-Training für Kinder im Vorschulalter zwischen 5 und 6 Jahren. Für dieses Projekt wurde daraufhin die folgende smarte Definition des Evaluationsziels erarbeitet:

Das Ziel der Evaluation ist es, die Entwicklung der Sprachkenntnisse bei allen Kindern, die in beiden „Spielend lernen"-Stadtteilen St. Leonhard/Schweinau und Langwasser in den Kindertagesstätten am Förderprojekt „Phono-logisch"

teilnehmen, im Verlauf des letzten Kindergartenjahres vor der Einschulung zu erfassen. Dabei soll anhand eines standardisierten, spezifisch für das Förderprojekt geeigneten Messinstruments die individuellen Fähigkeiten im Bereich der phonologischen Bewusstheit dokumentiert sowie die Anzahl der „Risikokinder" zum Zeitpunkt vor der Einschulung, d.h. Kinder mit fehlender oder mangelhafter phonologischer Bewusstheit, identifiziert werden.

Die Definition des Evaluationsziels muss nicht zwingend in einem Satz abgeschlossen werden. Wie dieses Beispiel zeigt, dient der zweite Satz einer Konkretisierung des Evaluationsziels. Die Zielfestlegung entspricht allen Kriterien von smarten Zielen: Es werden Angaben zur Messbarkeit des Outcomes des Förderprojekts (Dokumentation individueller Fähigkeiten, Anzahl Risikokinder) sowie Angaben zum Zeitraum der Evaluationstätigkeit gemacht (letztes Kindergartenjahr). Die Zieldefinition hat einen ausreichenden Konkretisierungsgrad, wurde gemeinsam mit den Projektverantwortlichen erarbeitet und ist insgesamt als realistisch einzuschätzen.

Wie das Beispiel der Zieldefinition aus dem „Spielend lernen"-Projekt zeigt, erscheint es zielführend für die Gestaltung von Evaluationsverfahren, dass sich Evaluatoren eingehend mit den Zielen und Arbeitsweisen von Programmen auseinandersetzen. Dabei muss ggf. ein erhöhter Aufwand seitens des Evaluators erfolgen, um Insbesondere bei komplexeren Programmen, die mehrere Einzelmaßnahmen kombinieren und/oder an mehreren Standorten umgesetzt werden, ist es notwendig, die **Struktur und die Organisation des Programms im Detail anhand der vorgeschlagenen Instrumente zu untersuchen**.

Der Prozess der Erarbeitung von Evaluationszielen kann auch zu einer Neubewertung und Anpassung der Programmziele und -inhalte führen. Dies kann beispielsweise bei Programmen der Fall sein, die sich in einer frühen Entwicklungsstufe befinden und bei denen sowohl das Konzept als auch die Organisation noch erprobt werden. Bei der Pilotierung der Integrationskursreihe „In Deutschland zu Hause" hat die Auseinandersetzung mit den Evaluationszielen Lücken im Konzept der Kursreihe offenbart. So war ein Ziel des Kursprogramms, Einbürgerungskandidaten Kenntnisse über gesellschaftliche Mitwirkungsmöglichkeiten zu vermitteln und dies sollte auch als Evaluationsziel aufgegriffen werden. Tatsächlich war im Kursprogramm der Aspekt „bürgerschaftliches Engagement" unterrepräsentiert und das Curriculum wurde dahingehend überarbeitet.

Wie im Kapitel zuvor gezeigt, ist die Anwendung des UTOS-Konzepts von Cronbach in jedem Fall für die Strukturierung der einzelnen Programmkomponenten zielführend. Parallel oder daran anschließend folgt die Analyse der Programmziele hinsichtlich ihrer Verwendung zur Entwicklung des Evaluationskonzepts. Beide Vorgänge erfolgen idealerweise in enger Abstimmung mit den Programmverantwortlichen und -beteiligten.

4.3. Einbindung von Programmbeteiligten und Programmverantwortlichen

Stake (1975) betont in seinem Evaluationsansatz der *responsiven Evaluation* die Wichtigkeit der **Einbindung von Programmverantwortlichen und Programmbeteiligten** während der gesamten Dauer einer Evaluationsstudie. Die **Art und Intensität der Zusammenarbeit** zwischen Evaluatoren und Programmbeteiligten bzw. -initiatoren kann jedoch variieren.

Bei dem hier untersuchten Programm „Spielend lernen" handelt es sich um ein komplexes soziales Programm mit vielen Einzelmaßnahmen und mehreren städtischen Dienststellen und freien Trägern, die am Umsetzungsprozess der Programminhalte beteiligt waren. Der Evaluator nahm regelmäßig an Sitzungen diverser Gremien der Stadtverwaltung teil, in denen die Fortführung des Programms behandelt wurde. Es erfolgte zudem von Beginn an eine enge Absprache mit den Programmverantwortlichen zu den einzelnen Umsetzungsschritten. Der Autor wurde im „Jour fixe" regelmäßig in die strategische und organisatorische Programmplanung und -steuerung involviert. Evaluationsvorhaben, Durchführung sowie die Präsentation der Ergebnisse erfolgte im Rahmen dieser Treffen sowie auch bei diversen stadtteilbezogenen Sitzungen. Neben der Durchführung der Evaluation nahm der Evaluator **die Aufgabe eines externen Beraters** ein.

Ein weiteres Beispiel für die Zusammenarbeit mit den Auftraggebern einer Evaluation stellt das Modellprojekt *Integrationskurs*e dar. Das europäische forum für migrationsstudien (efms) übernahm hier eine **Doppelfunktion**: In enger Kooperation mit dem Bildungszentrum Nürnberg wurde das curriculare Konzept für die Kursreihe entwickelt. Zugleich war das efms auch für die Evaluation der Kurse verantwortlich. Die Planung der Evaluationsstudie startete zum gleichen Zeitpunkt wie die Konzeption der Kursinhalte, so dass die Evaluationsmethoden genau auf die Kursziele abgestimmt werden konnten. Die Doppelfunktion des efms – Evaluation und Mitwirkung bei der Programmentwicklung – stellte sich im Fall der Integrationskurse als gewinnbringend heraus. Die Kursreihe konnte effizient ausgearbeitet und auf Basis der fortlaufend generierten Evaluationsergebnisse weiterentwickelt werden.

Diese Doppelfunktion erscheint dagegen bei Wirkungsevaluationen von sozialen Programmen wenig sinnvoll. Hier wird die neutrale Haltung einer Institution seitens der Programmverantwortlichen gewünscht, die mit wissenschaftlichen Methoden die Wirkungen einer Maßnahme untersucht. Das Projekt *frühstart* zeichnete sich durch eine klare Rollentrennung zwischen Evaluationsaufgaben und inhaltlicher Programmentwicklung aus. Der Evaluationsauftrag an das efms lautete u.a., mithilfe einer Wirkungsevaluation eine Reihe von Fragen der Pro-

gramminitiatoren zu den Effekten der Sprachförderung der Kinder zu beantworten. Davon abweichende oder ergänzende Leistungen seitens des Evaluators (z.B. Erfüllung von Beratungsaufgaben) wurden nicht festgelegt.

Neben der grundsätzlichen Art der Zusammenarbeit mit den Stakeholdern ist die Einbindung der Stakeholder zu bestimmten Phasen im Verlauf einer Evaluationsstudie relevant. Wie in den folgenden Kapiteln zu zeigen sein wird, sind Formen des Austauschs und der Zusammenarbeit insbesondere in der Eingangsphase einer Evaluationsstudie bedeutsam.

4.4. Entwicklung einer Programmtheorie

Theoriegeleitete Evaluationsansätze stellen eine Möglichkeit dar, Evaluationsstudien zu großen, komplexen sozialen Programmen durchzuführen. Sie eignen sich insbesondere auch – wie in Kapitel 2.3.3. beschrieben wurde – als Instrument zur **Konzeption** von Evaluationsstudien (vgl. Rogers 2000, 2008). In der Vorbereitungsphase einer theoriegeleiteten Evaluationsstudie wird dazu das zu untersuchende Programm in seine Einzelteile zerlegt. Für jedes Programm kann auf diese Weise eine individuelle Programmtheorie entwickelt werden. Chen unterscheidet zwischen zwei zeitlichen Phasen einer theoriegeleiteten Evaluationsstudie: die normative Phase und die summative Phase (vgl. Chen 1990, 2004). In der normativen Phase werden die Evaluationsziele und die grundlegenden Wirkungsannahmen über die Funktionsweise des Programms in Form einer Programmtheorie entwickelt. In der summativen Phase wird die Richtigkeit der darin verorteten Wirkungsannahmen durch die Anwendung von empirischen Methoden überprüft. Im Folgenden soll es zunächst um die normative Phase, in der die Grundsätze von theoriegeleiteten Evaluationsansätzen für die Planung von Evaluationsstudien verwendet werden.

Das Konzept der theoriegeleiteten Evaluation wurde bereits Anfang der 70er Jahre von Weiss (1972) thematisiert. Weiter ausgearbeitet wurde das Konzept dann in den 80er Jahren von Rossi und Chen (vgl. u.a. 1983, 1985, 1990) und später von Pawson und Tilley (1997). Theoriegeleitete Evaluationsansätze arbeiten immer mit einer so genannten **Programmtheorie**. Dabei handelt es sich um einen konzeptionell-strukturierenden Rahmen für Evaluationsstudien, der die Organisationsstruktur, die Wirkungstheorie und die Kontextbedingungen von sozialen Programmen erfasst. Die Hauptkomponenten von Programmtheorien wurden bereits in Kapitel 2.3.3. vorgestellt. Chen (2005) unterscheidet zwischen dem **Action Model** und dem **Change Model** – Begrifflichkeiten, die im Folgenden weiter verwendet werden. Beide Bereiche der Programmtheorie sind in einem **Programmkontext** eingebettet. Alle drei Elemente bilden zusammen die Programmtheorie.

Das **Action Model** enthält eine schematische Darstellung der Beziehung und Interaktion der Akteure und Organisationen untereinander sowie zu ihrer Umwelt. Aus dem Action Model sollte deutlich hervorgehen, welche **Rollen und Funktionen** die Akteure und Organisationen für das Programm einnehmen. Dies können direkt an der Programmdurchführung beteiligte Akteure sein, aber auch Programmträger bzw. -initiatoren, die lediglich eine indirekte Funktion in der Programmdurchführung wahrnehmen (z.B. Beratungsfunktion eines Beirats zu einem Programm). Schließlich ist im Action Model dargestellt, wie die Teilnehmer in das Programm eingebunden werden. **Organigramme** stellen eine Möglichkeit dar, Strukturen, organisatorische Beziehungen sowie institutionelle und personelle Zuständigkeiten im Programm strukturiert zu erfassen und zu beschreiben.

Neben den Organisationsstrukturen ist für die Konzeption einer Programmtheorie eine Beschreibung der geplanten Maßnahmenumsetzung relevant. Hier sollte die **Form und Intensität der Maßnahmen** (z.B. eine Maßnahmen zur Sprachförderung im Kindergarten) konkret beschrieben werden, mit denen die Programmteilnehmer konfrontiert werden. Aus der Beschreibung sollte entnehmbar sein, wie das Programm die Teilnehmer erreicht (z.B. Veranstaltungsformen), wie der Kontakt zu den Teilnehmern aufgenommen und fortgesetzt wird und wie und wann dieser Kontakt endet. Auch diese Beschreibung ist Bestandteil des Action Models nach Chen. Im Projekt *frühstart* wurde auf diese Weise die Form der Förderung erfasst. In *frühstart* wurden Gruppen aus Kindern in den beteiligten Kitas gebildet, die anschließend mit einem speziellen Konzept in ihrer Sprachentwicklung gefördert wurden. Mit den beteiligten Kitas wurde vereinbart, an welchen Tagen in der Woche und für wie lange die *frühstart*-Förderung stattfinden soll.

Der zweite Bereich der Programmtheorie wird nach Chen durch das **Change Model** dargestellt. Nach Rossi et al. (1999) setzt sich das Change Model – also die **Wirkungstheorie eines Programms** – aus einer Reihe von Aussagen zusammen, die beschreiben sollen, wie die Programmangebote die angestrebten Veränderungen bewirken sollen. Dabei kann es sich entweder um eine durch die Projektbeteiligten strukturierte Festlegung von Ziel- und Wirkungsannahmen zum Programm handeln, oder um die Anwendung einer wissenschaftlich fundierten Theorie auf die schematische, schrittweise Beschreibung der beabsichtigten Wirkungsweise eines Programms. Bei der Konstruktion eines Change Models sollten sich Evaluatoren nach Chen (2005, auch früher 1987) durch folgende Prinzipien leiten lassen:

- Das Change Model sollte nach Möglichkeit auf der Grundlage von wissenschaftlichen Prinzipien, wissenschaftlichen Theorien oder Modellen erarbeitet werden.

- Die Formulierung von Annahmen zur Programmtheorie sollte zukunftsgerichtet erfolgen.
- Die Programmtheorie sollte sich durch einen hohen Konkretisierungsgrad aus-zeichnen.

Der Beitrag von Pawson und Tilley (1997) zur Weiterentwicklung des theoriegeleiteten Evaluationsansatzes besteht nach Weiss (2007, S. 69) u.a. in der Betonung der **Notwendigkeit der Verwendung von wissenschaftlichen Theorien** anstelle von „plausiblen" Annahmen über die Wirkungsweise eines Programms. Im Idealfall ist die zu evaluierende Intervention auf Basis von wissenschaftlichen Erkenntnissen entwickelt worden, so dass die Konstruktion des Change Models besonders leicht fällt. Bei einem vorschulischen Förderprogramm können für die Konstruktion des Change Models beispielsweise Erkenntnisse und Theorien aus der Entwicklungspsychologie oder empirischen Pädagogikforschung herangezogen werden.

In der Evaluationspraxis kann jedoch nicht im Standardfall davon ausgegangen werden, dass Programme auf Basis von wissenschaftlichen Theorien entwickelt wurden. Wenn dies nicht der Fall ist, konstituiert sich das Change Model aus einem **Set von subjektiven Annahmen der Stakeholder über die beabsichtigte Funktionsweise des Programms.** Diese subjektiven Annahmen beinhalten Vorstellungen und Wissensbestände über den Gegenstandsbereich und sind mehr nach Plausibilität als nach innerer Widerspruchsfreiheit formuliert. Pawson und Tilley (1997) betonen für den gesamten Erarbeitungsprozess der Programmtheorie aus subjektiven Annahmen die sich daraus ergebende Aufgabe für den Evaluator, die Sichtweise der Stakeholder aufzunehmen, zu strukturieren und den weiteren Entwicklungsprozess zu moderieren. Das Change Model zu einem Programm entsteht somit in einem iterativen Prozess zwischen Evaluator und Auftraggeber bzw. Programmverantwortlichen, an dessen Ende die subjektiven Vorstellungen über das Programm auf einen gemeinsamen Nenner gebracht werden.

Im Gegensatz zum Action Model erscheint es bei der Entwicklung des Change Models wichtig, durch die Arbeit des Evaluators von Beginn an einen hohen Konkretisierungs- und Strukturierungsgrad anzuwenden. Je höher die Konkretisierung des Change Models erfolgt, desto leichter kann sich anschließend die Auswahl der Evaluationsmethoden herausstellen. Es kann jedoch nicht in allen Fällen davon ausgegangen werden, dass kurzfristig ein ausreichender Konkretisierungsgrad der Programmtheorie erarbeitet werden kann. Im Folgenden werden zwei Beispiele für Change Models vorgestellt, die Bestandteil der im Rahmen dieser Arbeit untersuchten Evaluationsstudien sind.

Im ersten Beispiel wurde im Zusammenhang mit der Evaluation des Programms „Spielend lernen" ein sehr allgemeines Change Model für alle Maßnahmen im

gesamten Programm entwickelt, das in Abbildung 4 dargestellt ist. Das Change Model wurde durch den Evaluator ausgearbeitet und in mehreren Arbeitsschritten mithilfe des Feedbacks der Programmbeteiligten weiterentwickelt. Es hatte zum Zeitpunkt der Evaluationsstudie die Funktion, dass sich Programmbeteiligte, Programmverantwortliche sowie der Evaluator über die erhofften Ziele und Funktionsweisen der Einzelmaßnahmen austauschen konnten und diente schließlich als **Orientierungsraster für die Entscheidungsfindung**, welche Einzelmaßnahmen überhaupt mit dem Evaluationsvorhaben überprüft werden sollten.

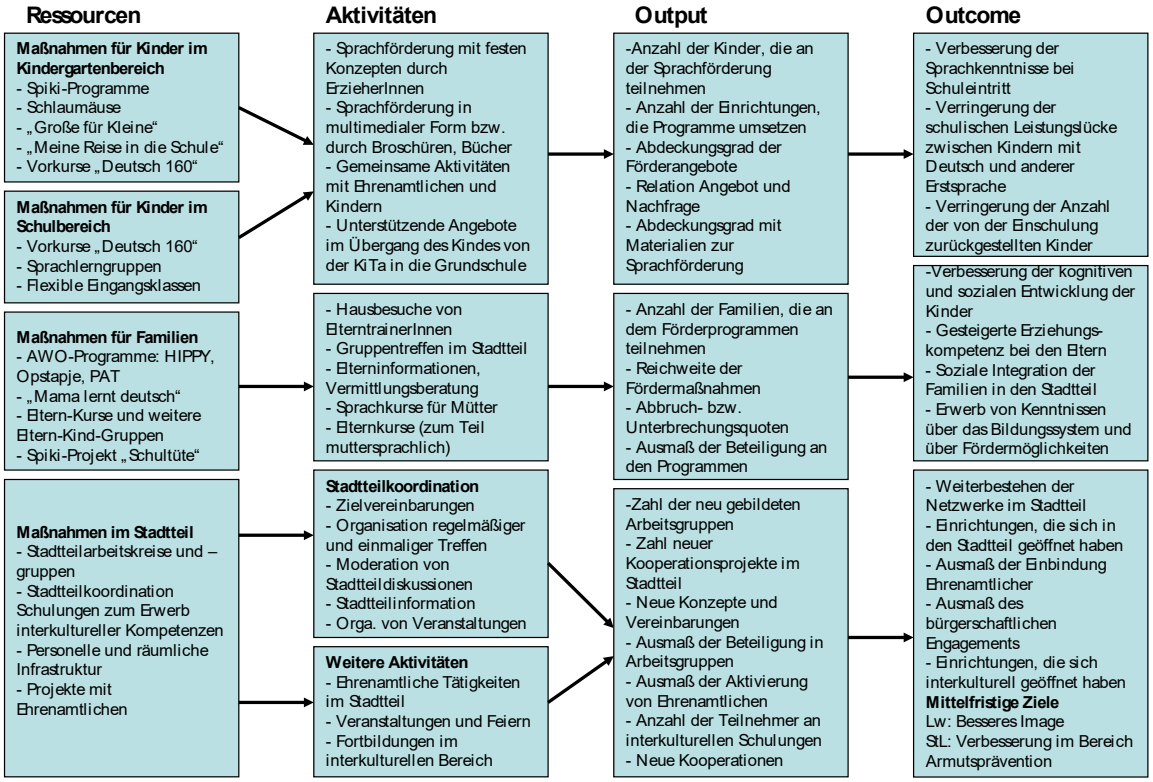


Abbildung 4: Beispiel: Allgemeines Change Model als ein Bestandteil einer Programmtheorie zum Projekt „Spielend lernen“

Das Change Model wurde als Ablaufmodell dargestellt. Die Verknüpfung mit Pfeilen kann sich als hilfreiches Instrument erweisen, um eventuell auftretende logische Brüche im Bereich der Wirkungsannahmen zu identifizieren. Von links nach rechts lesend wurden schematisch die einzelnen Schritte beginnend mit dem Ressourceneinsatz bis zu den gewünschten Wirkungen des Programms beschrieben. In der ersten Spalte sind die Ressourcen in „Spielend lernen“ dargestellt. Unter Ressourcen sind in diesem Fall Konzepte, Einzelprojekte sowie die personelle und räumliche Infrastruktur zu verstehen, die für die Gestaltung des Programms notwendig waren. Die mit den Ressourcen verbundenen Aktivitäten

sind in der zweiten Spalte abgebildet. Die abgebildete Programmtheorie repräsentiert die Vorstellung der Programmbeteiligten, wie die verschiedenen Einzelmaßnahmen zu einem Fördererfolg bei den Teilnehmern führen sollten. Die erwünschten Ergebnisse des Programms sollten nach der Konstruktion der Programmtheorie leicht ablesbar sein; z.B. „Durch die Teilnahme im Programm „SpiKi – Sprachförderung im Kindergarten" soll die sprachliche Leistungslücke zwischen Kindern mit Migrationshintergrund und deutschen Kindern im Kindergartenalter bis zum Schuleintritt geschlossen werden".

Mit den Programmverantwortlichen von „Spielend lernen" wurde auf Basis des Change Models diskutiert, welche Maßnahmen für die Bearbeitung in einer Evaluationsstudie überhaupt in Frage kommen. Die zum damaligen Zeitpunkt verfügbaren zeitlichen und finanziellen Ressourcen stellten schließlich die ausschlaggebenden Kriterien für die Auswahl der Maßnahmen dar. Durch das allgemeine Change Model hat man sich z.B. darauf geeinigt, die Wirkungen der Sprachförderung mit der Maßnahme „Phono-Logisch – Hand in Hand" (als Teil der Spiki-Programme in der ersten Spalte des Change Models) in der Evaluationsstudie zu untersuchen. Die Evaluationsmethode und die Ergebnisse werden in Kapitel 6.3. beschrieben. Das Change Model diente im Fall von „Spielend Lernen" als notwendiger Zwischenschritt in der Planung der Evaluationsstudie.

Um das **Change Model für die Ausarbeitung der Evaluationsmethode** zu verwenden, ist jedoch – anders als im Fall des ersten Beispiels – ein weit höherer Konkretisierungsgrad der einzelnen Bestandteile notwendig. Owen (2007) weist darauf hin, dass nach Weiss (1998) es das primäre Ziel für den Evaluationsforscher sein sollte, die Programmtheorie in allen Einzelschritten vollständig zu beschreiben. Dabei sind besonders die Erfassung der einzelnen Programmmechanismen sowie die Details hinsichtlich der Funktionsweise in der Praxis wichtig. Weiss nennt diese Form der Evaluation, die die Untersuchung einer Programmtheorie zum Anlass hat, theories of change (TOC) evaluations. (vgl. Owen 2007, S. 197). Um Programmtheorien möglichst realitätsnah zu entwickeln, hat Funnell (2000) eine so genannte *Program Theory Matrix* entwickelt, die ein geeignetes Instrument zur detaillierten Ausarbeitung einer Programmtheorie von den Evaluatoren verwendet werden kann. Die Anwendung der Matrix ist insbesondere für die **Planung von Wirkungsevaluationen** zu empfehlen. Im zweiten Beispiel wird daher die Ausarbeitung eines Change Models mithilfe der *Program Theory Matrix* am Beispiel des Projekts *frühstart* beschrieben. Für die Anwendung der Matrix als Systematisierungsrater spricht außerdem, dass bereits gesammelte Erkenntnisse und Informationen zum Programm aus den vorhergängigen Teilschritten bei der Planung von Evaluationsstudien direkt einfließen können. Zu diesen zählen das UTOS-Konzept nach Cronbach und die Festlegung von Evaluations- und Programmzielen nach der SMART-Systematik.

Die Ausarbeitung der Matrix nach Funnell sollte maßnahmenspezifisch mit der Benennung der angestrebten Ergebnisse beginnen und dann mit der Festlegung der Erfolgsfaktoren fortfahren, d.h. Evaluationskriterien, anhand derer festgestellt werden kann, ob mit der Maßnahme die angestrebten Ergebnisse erreicht wurden. Für die Maßnahme Sprachförderung im *frühstart*-Programm wurde als Ergebnis festgelegt, dass mit *frühstart* geförderte Kinder zum Zeitpunkt der Einschulung nachweislich bessere Deutsch-Sprachkenntnisse besitzen. Im zweiten Punkt *Erfolgskriterien* werden konkrete Förderziele festgelegt, deren Erreichungsgrad dann in der Evaluationsstudie mit geeigneten Methoden gemessen werden soll (z.B. ob die geförderten Kinder zu den Sprachkompetenzen von Kindern mit Deutsch als Erstsprache aufgeschlossen haben). Um die Mechanismen der Maßnahme zu analysieren, sind als nächstes die Programm- bzw. Maßnahmenfaktoren zu nennen (Schritte drei), die direkten Einfluss auf das Ergebnis der Maßnahme haben können. Für die Benennung der Programmfaktoren lässt sich das Förderkonzept von *frühstart* als Grundlage heranziehen. Wurden zuvor in der Phase der Informationssammlung mit dem UTOS-Konzept bereits Informationen zum Programm gesammelt, können diese ggf. für die Bearbeitung der folgenden Arbeitsschritte bei der Konstruktion der *Program Theory Matrix* angewendet werden.

Die Vorstellung, wie ein Programm theoretisch zu funktionieren hat, kann stark von dem abweichen, wie es dann tatsächlich in der Praxis umgesetzt wird und wie es sich langfristig entwickelt. Die Veränderung des Programms durch externe Faktoren lässt sich nicht ausschließen, jedoch kann bei der Planung der Evaluationsstudie eine detaillierte Aufstellung relevanter externer Programmfaktoren als wertvoll für die anschließende Ausarbeitung der Evaluationsmethode erweisen. In Schritt vier wurden am Beispiel *frühstart* alle benennbaren Faktoren außerhalb des Programms aufgezählt, die Einfluss auf das Ergebnis der Sprachfördermaßnahme haben könnten. Die fünfte Spalte enthält eine Beschreibung der Ressourcen und Aktivitäten, die für die Maßnahmenumsetzung notwendig sind (z.B. die Art und Weise der Förderung in den frühstart-Gruppen).

Die Schritte eins bis fünf beinhalten somit alle Informationen, die für die Formulierung eines ersten Change Models für eine Maßnahme im Programm relevant erscheinen. Sollen weitere Maßnahmen evaluiert werden, so kann nach dem gleichen Muster verfahren werden. Evaluatoren und Programmverantwortliche können außerdem bei Bedarf (z.B. komplexe Wirkungsannahmen) die in der Matrix in Spalten gruppierten Aussagen in einem Pfadmodell schematisch miteinander verknüpfen und abbilden, wenn das Ziel der Evaluation die detaillierte Überprüfung des Mechanismus beginnend von den Aktivitäten bis zum Ergebnis sein soll. Im Fall des Programms *frühstart* erschien dieser Schritt nicht sinnvoll, da es sich um die Erstevaluation des Programms handelte und sich das Ziel der

Evaluation auf die generelle Erfassung der Wirkungsweise der Sprachförderung konzentrierte.

Die Bearbeitung der *Program Theory Matrix* wird mit Angaben in den Spalten fünf und sechs abgeschlossen, in denen auf die zu erfassenden Informationen während der Maßnahmendurchführung sowie auf potenzielle Datenquellen eingegangen wird. Neben den Grundinformationen zu den *frühstart*-Gruppen (u.a. Anzahl geförderter Kinder, Alter der Kinder) wurden hier für die Bewertung der Sprachförderung schon Ergebnisse aus Sprachscreenings oder vergleichbare Instrumente als mögliche Datenquellen genannt. Die Arbeit mit der *Program Theory Matrix* im Rahmen des hier vorgestellten Prozessmodels kann für den Evaluator den arbeitserleichternden Vorteil mit sich bringen, dass inhaltlich ein gleitender Übergang von der Eingangsphase von Evaluationsstudien in die nächstfolgende Phase der Entwicklung und Auswahl der Evaluationsmethoden erreicht werden kann.

Abbildung 5: Beispiel eines Change Models für die Maßnahme Sprachförderung im Projekt frühstart (nach Funnell 2000, S. 94f.; [Anm. des Autors: eigene Übersetzungen aus dem Englischen]).

Nach Weiss (1972) hat die Anwendung einer Programmtheorie in einem Evaluationsverfahren die primäre Funktion, die Blackbox von Programmen zu öffnen. Die soeben beschriebene systematische Vorgehensweise bietet die Möglichkeit, hierfür die Grundvoraussetzungen in einer Evaluationsstudie zu schaffen. Liegt eine erste Version der Programmtheorie vor, sind die ersten Planungsschritte zu einer Evaluationsstudie abgeschlossen. Die Entwicklungsphase von Evaluationsstudien wird mit einer so genannten Evaluierbarkeitsprüfung abgeschlossen.

4.5. Durchführung einer Evaluierbarkeitsprüfung

Das Prinzip der Evaluierbarkeitsprüfung (Evaluability Assessment) basiert auf Überlegungen der amerikanischen Evaluationsforscher Wholey (Wholey et al. 1970, 1977, Wholey et al. 2004, Wholey & Strossberg 1983) und Suchman (1967), die erstmals in den 60er Jahren in die Evaluationsforschung eingeführt wurden[34]. Suchman sieht die Rolle des Evaluators als Vermittler zwischen der Forschersicht – verbunden mit der Notwendigkeit eines strengen, fundierten Methodeneinsatzes – und den Zielen und Interessen der Auftraggeber, die nicht zwingend mit wissenschaftlichen Standards übereinstimmen müssen. Auftraggeber von Evaluationsstudien sind im Untersuchungskontext von Suchman die öffentlichen Verwaltungen, deren Interesse es sei, Programme umzusetzen, die nützliche Ergebnisse zur Lösung von sozialen Problemen beitragen.

Die Evaluierbarkeitsprüfung hat nach Wholey das Ziel, auf Basis der zur Verfügung stehenden Informationen eine Grundlage für die Entscheidung zu bereiten, ob und wie ein Programm sinnvoll evaluiert werden kann (Wholey et al. 2004, S. 35). Daher seien im Zuge der Entwicklung eines Evaluationskonzepts unter Mitwirkung von Programmverantwortlichen bestimmte Bedingungen zu prüfen, bevor mit der Evaluation fortgefahren wird. „Evaluability assessment not only shows whether a program can be meaningfully evaluated (any program can be evaluated) but also whether evaluation is likely to contribute to improved program performance" (Wholey et al. 2004, S. 35). Nach Wholey geht es bei der Evaluierbarkeitsprüfung nicht darum, ob ein Programm überhaupt evaluiert werden kann (was der Name *Evaluability Assessment* u.a. vermittelt), sondern um die Prüfung, wie der Nutzen der geplanten Evaluationsstudie für die Programmweiterentwicklung maximiert werden kann. Die Evaluierbarkeitsprüfung stellt daher eine **wichtige Prüfweiche** zu einem Zeitpunkt dar, bevor mit der Ausarbeitung

[34] Bei der Auseinandersetzung mit deutschsprachiger Evaluationsliteratur wurde deutlich, dass der Aspekt Evaluierbarkeitsprüfung wenig thematisiert wird. Eine Fundstelle in der Literatur zum Thema Evaluierbarkeitsprüfung sind Publikationen von Margrit Stamm (2003, 2008), Erziehungswissenschaftlerin, die auf die Bedeutung der Prüfung für die Qualität der Evaluationsstudie hinweist.

der Evaluationsmethoden begonnen wird. Die zentrale Frage, zu deren Beantwortung eine Evaluierbarkeitsprüfung durchgeführt wird, lautet:

Kann die Evaluation mit den geplanten Ressourcen unter den gegebenen Charakteristika des Programms durchgeführt werden, so dass fundierte Aussagen zu den zuvor definierten Erkenntnisinteressen (z.B. Zielerreichung, Überprüfung der Programmtheorie) möglich sind und auf deren Basis das Programm weiterentwickelt wird?

Da sich der Evaluator bei der Durchführung der Evaluierbarkeitsprüfung mit den potenziell in Frage kommenden Evaluationsmethoden auseinandersetzt, findet dieser Arbeitsschritt im Rahmenmodell streng genommen im Übergang zwischen der Planung und der Entwicklung der Evaluationsmethoden statt. Davon abweichend wird aber die Evaluierbarkeitsprüfung im Folgenden – aus Gründen der besseren Veranschaulichung des Prozessmodells – als letzter Arbeitsschritt in der Planungsphase behandelt.

Die **Einbindung der Programmverantwortlichen** ist bei den Vorarbeiten zur Evaluierbarkeitsprüfung sowie bei der Klärung des Prüfergebnisses entscheidend – z.B. bei der Frage, wie das Ergebnis der Evaluierbarkeitsprüfung für die Gestaltung des Programms und der Evaluationsstudie weiter verwendet werden soll. Für die Durchführung der **Evaluierbarkeitsprüfung bei komplexen Programmen** eignen sich **Workshops mit Stakeholdern**, in denen die Informationen zum Programm sowie die Evaluationsabsichten systematisch behandelt werden. Eine schrittweise Vorgehensweise für die Aufbereitung der Informationen durch den Evaluator für die sich anschließende Evaluierbarkeitsprüfung schlägt Smith (1989, 2005) vor, die sich als Arbeitsschritt gut in das hier zu behandelte Prozessmodell in der Planungsphase von Evaluationsstudien integrieren lässt. Die einzelnen Punkte nach Smith wurden jedoch für die Verwendung im Phasenmodell in ihrer Reihenfolge geändert und auf die Evaluierbarkeitsprüfung als Schlusspunkt der Checkliste ausgerichtet.

1. „Genaue Analyse der Programmdokumente: Sichtung aller verfügbaren Dokumente über das zu evaluierende Programm. Einordnung der Informationen nach dem UTOS-Prinzip nach Cronbach.
2. Festlegung Evaluationszweck: Identifikation der Programmziele; Analyse der programmatischen, strategischen Programmziele hinsichtlich der Zielerreichung. Eventuell sich ergebende Gegensätzlichkeiten oder Widersprüche sollten hier behandelt werden. Formulierung SMARTer Ziele.
3. Beschreibung der Programmtheorie: Erörterung der Programmtheorie mit den Programmbeteiligten. Vorgehensweise nach Chen sowie Pawson und Tilley. An-wendung der Program Theory Matrix nach Funnell.

4. Workshops mit Programmbeteiligten zur Klärung offener Fragen: Durchführen von Interviews mit Programmbeteiligten, um ihre Sichtweise zur Programmtheorie zu erfassen.
5. Evaluierbarkeitsprüfung:
   - Festlegung der zu evaluierenden Maßnahmen: Mit den Programmverantwortlichen werden anhand der entwickelten Programmtheorie diejenigen Maßnahmen festgelegt, die evaluiert werden sollen.
   - Identifikation von geeigneten Evaluationsmethoden: Die zu evaluierenden Maßnahmen werden in einem mit den passenden Evaluationsmethoden verknüpft. Der Vorgang wird wiederum mit den Programmverantwortlichen abgestimmt.
   - Identifikation der Aufwand-Nutzen-Relation: Auch hier ist die Sichtweise der Stakeholder relevant. Der Aufwand für die Durchführung der Evaluationsstudie sollte zu verwertbaren Ergebnissen führen. Umgekehrt ist zu prüfen, ob die Evaluationsstudie mit den zur Verfügung stehenden Ressourcen durchführbar ist.“ (vgl. die Aufzählung bei Smith 2005, S. 138f.; [Anm.: Übersetzung aus dem Englischen durch den Autor]).

Die Vorbereitung der Evaluierbarkeitsprüfung sollte sich an den Einzelschritten in der Planungsphase von Evaluationsstudien orientieren. In der Praxis kann sich das Vorgehen auch als schleifenartiger Prozess herausstellen. So erfolgt die Entwicklung der Programmtheorie, wie im Kapitel zuvor beschrieben, zumeist in einem iterativen Verfahren der Ausarbeitung zwischen Evaluator und Stakeholdern. Die Evaluierbarkeitsprüfung selbst umfasst drei Aspekte: auf Basis der generierten Informationen (Grunddaten, Ziele und Programmtheorie) werden die zu evaluierenden Maßnahmen eines Programms festgelegt, geeignete Evaluationsmethoden identifiziert sowie der Aufwand und der erwartbare Nutzen für die Programmweiterentwicklung geschätzt. Eine effektiv und transparent durchgeführte Evaluierbarkeitsprüfung erzeugt im besten Fall das Ergebnis, dass die bisher erfolgten Planungsschritte für die Umsetzung der Evaluationsstudie sich als praktikabel und effektiv im Sinne der Evaluationsziele erwiesen haben. Mit der Entwicklung der Evaluationsmethoden kann in diesem Fall direkt fortgefahren werden.

Die Evaluierbarkeitsprüfung kann aber auch erst nach mehreren Arbeitsschleifen abgeschlossen werden, wenn die gewählten Maßnahmen und Evaluationsmethoden sich nicht mit Aufwand-Nutzen-Bewertungen in Übereinstimmung bringen lassen. Führt die Evaluierungsprüfung zu dem Ergebnis, dass keine konnte für keine Maßnahmen im Programm passende Evaluationsmethoden gefunden werden, die im Rahmen gemeinsamer Aufwand-Nutzen-Überlegungen liegen, sollte gemeinsam mit den Auftraggebern der Evaluationsstudie nach Lösungsansätzen

gesucht werden. Wholey et al. sehen für diesen Fall weitere Schritte zum Abschluss der Evaluierbarkeitsprüfung vor:

- "Reach agreement on any needed changes in program activities or goals.
- Explore alternative evaluation designs.
- Agree on evaluation priorities and intended uses of information on program performance" (Wholey et al. 2004, S. 36).

Die Evaluierbarkeitsprüfung kann im Einzelfall dazu führen, dass Änderungen am Programm und seinen Zielen und/oder an den gewählten Evaluationsmethoden erfolgen, bevor mit der Evaluationsstudie fortgefahren wird. Eine weitere Konsequenz aus den Ergebnissen der Evaluierbarkeitsprüfung wäre, die Durchführung der Evaluationsstudie auf einen Zeitpunkt zu verschieben, ab dem sich das Programm durch einen höheren Grad der Weiterentwicklung auszeichnet.

Ein Beispiel für eine Evaluierbarkeitsprüfung, die zu einer **Konkretisierung der ursprünglich vorgesehenen Evaluationskonzepts** führte, ist für das Programm „Spielen Lernen" zu nennen. Die Konzeption der Evaluationsstudie zum Programm stellte sich zunächst als Herausforderung für die Planung heraus, da das Programm eine Vielzahl von Maßnahmen vereinigte und damit die Entscheidung getroffen werden musste, welche Maßnahmen überhaupt Gegenstand einer Evaluationsstudie sein sollten. Bei der Evaluierbarkeitsprüfung mussten zwei Aspekte berücksichtigt werden:

- die Heterogenität der zahlreichen Maßnahmen im Programm sowie
- der finanzielle Spielraum für die Evaluationsstudie.

Die Maßnahmen des Programms „Spielend Lernen" reichten von spezifischen Förderprogrammen mit Pilotcharakter nach festen Curricula bzw. Konzepten über stadtteilbezogene, soziale Beratungsangebote bis hin zu einfachen Versorgungsangeboten (z.B. Mittagsessen für Schulkinder). Das Ergebnis der Evaluierbarkeitsprüfung auf Basis des im Kapitel zuvor beschriebenen allgemeinen Change Models machte deutlich, dass eine Wirkungsevaluation aller Maßnahmen innerhalb des anvisierten Zeitraums für die Evaluation und unter den gegebenen Ressourcen nicht realisierbar gewesen wäre. Nach Prüfung aller vorliegender Informationen und intensiven Diskussionen mit den Programmverantwortlichen einigte man sich auf die Evaluation von einigen ausgesuchten Maßnahmen innerhalb des Programms von „Spielend Lernen". So wurde dann „Phono-Logisch Hand in Hand" als einzige Fördermaßnahme in Form einer Wirkungsevaluation konzipiert und durchgeführt. Des Weiteren wurde bei einer Auswahl weiterer Maßnahmen formative Evaluationen durchgeführt (z.B. in Form von Feedbackbefragungen), die dem Zweck der Maßnahmenverbesserung dienten.

Geht es bei der Evaluierbarkeitsprüfung außerdem um die Entscheidung, ob mithilfe einer Wirkungsevaluation die **Effekte eines Programms** erfasst werden sollen, empfiehlt es sich, weitere Kriterien in die Prüfung aufzunehmen. Geprüft wird in diesem Fall die Machbarkeit einer Wirkungsevaluation. Die Leitfrage hierzu lautet:

Kann eine Wirkungsevaluation effizient durchgeführt werden, so dass valide und objektive Ergebnisse zu erwarten sind?

Der Autor schlägt vor, folgende Fragen im Rahmen der zweiten Evaluierbarkeitsprüfung zu bearbeiten.

- Welche Maßnahme(n) des Programms soll(en) hinsichtlich Ihrer Wirkungen evaluiert werden?
- Sind die Charakteristika der Maßnahmen (Programmtheorie und Ziele) ausreichen konkret beschreiben? (siehe *Program Theory Matrix*)
- Ist die Umsetzung der Maßnahmen im Programm in Form einer Programmtheorie unter Einbezug von Wirkungsannahmen im Detail dokumentiert?
- Über welchen Zeitraum und wie intensiv (Turnus, Dauer) haben die Teilnehmer Kontakt mit der Maßnahme (Exposure)?
- Lassen sich die Teilnehmer an der Maßnahme, deren Wirkungen gemessen wer-den soll, eindeutig charakterisieren und ggf. von anderen Teilnehmergruppen im Programm abgrenzen?
- Ist der Zeitrahmen für die Durchführung der Maßnahmen ausreichend, um eine Wirkungsevaluation auch mit mehreren Messzeitpunkten durchzuführen?

Bei experimentellen und quasi-experimentellen Wirkungsevaluationen:

- Wie groß ist die Stichprobe / Fallzahl für die Untersuchungs- bzw. Kontroll-gruppe zu wählen, wenn mit einem bestimmten Anteil an Ausfällen gerechnet wird?
- Wie soll das Verfahren zur Bildung der Untersuchungs- und Kontrollgruppen genau verlaufen?
- Welches Design der Wirkungsanalyse erscheint zum Zeitpunkt der Evaluierbarkeitsprüfung zielführend zu sein?
- Wie viele Messungen sind für die Bildung einer soliden Datengrundlage not-wendig?
- Wurden programminterne und programmexterne Störfaktoren benannt und können diese im Verlauf der Evaluation kontrolliert werden?

- Ist die Programmdurchführung im Evaluationszeitraum stabil gegenüber nicht-intendierten Veränderungen von außen?

Ausgangspunkt für die Evaluierbarkeitsprüfung von Wirkungsmessungen ist die in der Phase zuvor ausgearbeitete Programmtheorie inklusive der smarten Ziele des Programms. Die Programmtheorie beschreibt die Wirkungsannahmen – idealerweise in Form von aufeinander aufbauenden Aussagen. Lassen sich die obigen Fragen schlüssig beantworten, kann ebenso mit der Konzeption der Evaluationsmethode weiterverfahren werden.

Als ein Beispiel für die Anwendung der Evaluierbarkeitsprüfung für die Behandlung der Frage, **ob eine Wirkungsevaluation realisierbar ist**, kann die Evaluation zur Integrationskursreihe „In Deutschland zu Hause“ genannt werden. Im ursprünglichen Evaluationskonzept, das vor dem Start der Studie entwickelt wurde, wurde ein besonderes Augenmerk auf die Identifikation von Wirkungen als ein wichtiges Evaluationsziel gelegt. Da es sich um ein Modellprojekt handelte, bestand die Absicht darin, bereits in der Erprobungsphase herausfinden, ob die gewählte Kursform zu signifikanten Lernzuwächsen bei den Teilnehmern geführt hat. Um diese Wirkungen bei den Teilnehmern zu erfassen, war ein freiwilliger Wissenstest zum Ende der Kursreihe geplant. Die Konzeption einer geeigneten Wirkungsevaluation musste jedoch noch vor dem Beginn der Kurse im Herbst 2001 verworfen werden, da nicht der richtige Veranstaltungsrahmen für die Messung von Wirkungen gefunden wurde. Das erworbene Wissen hätte beispielsweise mit einem standardisierten schriftlichen Test am letzten Kursabend erfolgen können. Ein Wissenstest am Ende der Kursreihe hätte den Kursentwicklern Hinweise darüber gegeben, inwieweit die angestrebten Lernergebnisse mit den tatsächlich realisierten Lernergebnissen übereinstimmen.

Da der Wissenstest zur damaligen Zeit aus organisatorischen Gründen nicht umsetzbar war, wurde der Fokus der Evaluation auf andere Bereiche des Programms gelegt und nach Alternativen gesucht. So war ein weiteres didaktisches Ziel der Integrationskurse – neben der Wissensvermittlung – Zugehörigkeitsgefühle und Einstellungsänderungen zur deutschen Gesellschaft zu fördern. In den Integrationskursen wurde ein Ansatz in diese Richtung verfolgt, indem ein so genanntes semantisches Differenzial eingesetzt wurde, um Einstellungsmuster der Teilnehmergruppe zu verschiedenen Kursthemen zu beschreiben.

5. Entwicklung des Evaluationsdesigns

Nach Abschluss der Evaluierbarkeitsprüfung wird mit der Konstruktion des Evaluationsdesigns fortgefahren. Ein Evaluationsdesign wird hier folgendermaßen definiert: Die vollständige, in sich abgeschlossene Beschreibung der gewählten Methoden, Instrumente und Zeitpunkte des Einsatzes sowie der Bildung und Zusammensetzung der Teilnehmer an einer Evaluationsstudie (u.a. Auswahlverfahren, Stichproben Grundgesamtheit) für einen festgelegten Zeitraum auf Basis einer Programmtheorie oder Fragestellungen zu einem sozialen Programm.

Wie bereits im vorherigen Kapitel dargestellt, sind für die Evaluierbarkeitsprüfung eine Reihe von Vorarbeiten notwendig; von der Informationssammlung zu dem Programm, der Zieldefinition bis hin zur Formulierung der Programmtheorie. Als Ergebnis dieser Arbeitsschritte sollte für den Evaluator erkennbar werden – vorbehaltlich der Evaluierbarkeitsprüfung – welche empirischen Methoden für die Ziele der Evaluationsstudie in Frage kommen. Für die Konfiguration des Evaluationsdesigns steht prinzipiell das gesamte Repertoire der empirischen Sozialforschung zur Verfügung. In diesem Kapitel wird die Entwicklung des Evaluationsdesigns von allgemeinen Kriterien bis hin zur Auswahl spezifischer Instrumente auf Basis der Praxiserfahrung aus den untersuchten Studien beschrieben. Der Einstieg in die Entwicklung des Evaluationsdesigns kann durch die Auseinandersetzung des Evaluators mit der Entwicklungsstufe eines zu evaluierenden Programms erleichtert werden.

5.1. Evaluationsformen und Entwicklungsstufen des Programms

Nicht alle Erkenntnisinteressen können zu allen Zeitpunkten in einem Evaluationsdesign sinnvoll adressiert werden. Für die Ergebnisqualität von Evaluationsstudien ist – neben der strukturierten Planung des Vorhabens – die Analyse des **Entwicklungsstands des Programms zum Zeitpunkt der Planung einer Evaluationsstudie** entscheidend. Es erscheint nachvollziehbar, dass für ein Programm, das schon seit mehreren Jahren umgesetzt wird, ggf. andere Methoden ausgewählt werden, als dies bei Programmen mit Pilotcharakter der Fall wäre. Unter Umständen wurde das betreffende Programm in einem früheren Entwicklungsstadium bereits evaluiert. Die Erkenntnisse aus der Evaluation ließen sich in diesem Fall als eine Informationsgrundlage für die Entwicklung des Evaluationsdesigns und die Auswahl der Methoden heranziehen.

Beim Pilotieren einer neuen Maßnahme (erste Umsetzung und somit Erprobung) zeichnet sich diese bezüglich des dahinterliegenden Konzepts sowie den Arbeitsweisen zumeist durch einen frühen Entwicklungsstand aus. Soll eine

Maßnahme oder ein Programm in diesem Entwicklungsstadium evaluiert werden, so können mit der Evaluation theoretisch unterschiedliche Ziele verfolgt werden. Ist das primäre Ziel die Weiterentwicklung des Programms, werden die ausgewählten Evaluationsmethoden dem Charakteristikum einer **formativen Evaluation** folgen (vgl. Scriven 1991). Dazu werden zunächst Daten und Informationen zu dem Programm gesammelt und eingeordnet. Dies kann mithilfe der in den Unterkapiteln zuvor beschriebenen Systematiken und Instrumente (z.B. Anwendung der UTOS-Kategorisierung) erfolgen. Eine ausgearbeitete Programmtheorie kann die Grundlage für die Formulierung der Fragestellungen für die formative Evaluation darstellen. Anschließend werden zu den Evaluationszielen passende Evaluationsmethoden ausgewählt und die Erhebungsinstrumente entwickelt. Die in der Evaluationsstudie erhobenen Daten und Informationen werden ausgewertet und für die **Weiterentwicklung und Verbesserung des Programms** verwendet. Ist das Ziel der Evaluationsstudie, die Wirkungen eines Programms wissenschaftlich zu erfassen und zu untersuchen, kommen andere Evaluationsmethoden in Frage als bei formativen Evaluationen und müssen auf eine andere Art und Weise eingesetzt werden. Wirkungsevaluationen werden in Form von **summativen, zusammenfassenden Evaluationen** erfasst, d.h. die Ergebnisse beziehen sich auf eine bestimmte zeitliche Periode, in der das Programm durchgeführt wurde.

Bei der Evaluation des Projekts *frühstart* setzte die Evaluationsstudie erst nach einem halben Jahr ein, nachdem die Förderung der ersten *frühstart*-Gruppen bereits begonnen hatte. Eine Herausforderung für die Evaluation stellte die Tatsache dar, dass das Evaluationsdesign nach dem Programmstart von *frühstart* entwickelt und mit dem laufenden Programm sinnvoll verbunden werden musste. Zudem betonte der Auftraggeber der Evaluationsstudie sein besonderes Interesse an fundierten Ergebnissen im Bereich der Wirkungsmessung. Eine Evaluation der Wirkungen kann jedoch nicht ohne bestimmte vorbereitende, organisatorische Tätigkeiten erfolgen, vor allem mussten zahlreiche Detailinformationen über die Programmstruktur in den *frühstart*-Städten und Kindertagesstätten eingeholt werden. Das efms als Auftragnehmer der Evaluationsstudie musste daher gemeinsam mit den Programmverantwortlichen einen schnellen Einstieg in die Evaluation finden. Das später entwickelte Design beinhaltete einen Vergleich von standardisiert erhobenen Sprachständen bei geförderten und nicht geförderten Kindern in einem Kontrollgruppenvergleich. So wurden als einer der ersten Schritte im Evaluationsprojekt die Instrumente der Wirkungsmessung ausgesucht, Kontrollgruppenkindergärten gesucht und die organisatorische Umsetzung der Datenerhebung geplant. Für eine höhere Qualität der Ergebnisse der Wirkungsevaluation wäre ein Pretest der für die Förderung vorgesehenen Kinder noch vor der ersten Fördereinheit vorteilhaft gewesen, um einen Vergleich des Sprachstandes der Kinder vor und nach ihrer Förderung durch das Programm zu

erhalten. Dieser Vorbereitungsschritt war jedoch aufgrund des späten Einstiegs des Evaluators in das Programm nicht mehr möglich.

Die Entwicklung des Evaluationsdesigns zu der Integrationskursreihe „In Deutschland zu Hause“ ist als Gegenbeispiel zur Vorgehensweise im Programm *frühstart* zu nennen. Hier konnte das Design mit ausreichend zeitlichem Vorlauf vor der Pilotierung der Kurse entwickelt werden. Für die erste Kursreihe war es wichtig, anhand geeigneter Evaluationsmethoden Informationen zur Durchführung der einzelnen Kurseinheiten zu erhalten. Evaluationsergebnisse sollten nach jeder Kurseinheit vorliegen, damit kurzfristig Anpassungen am Curriculum sowie dem didaktischen Aufbau vorgenommen werden konnten. Das primäre Interesse des Evaluators sowie der Kursverantwortlichen lag während der Pilotphase in der Durchführung einer formativen Evaluation, um eine Weiterentwicklung und Verbesserung der Kursinhalte zu realisieren. Für die Datenerhebung wurde hauptsächlich das Instrument der teilnehmenden Beobachtung angewendet. Zum Abschluss der Pilotphase hatten die Ergebnisse den Charakter einer summativen Evaluation, bei der die Anzahl der Teilnehmer pro Kursreihe, Bewertungen der Teilnehmer zum gesamten Kurs sowie Angaben zum subjektiv geschätzten Lernerfolg in einem Ergebnisbericht dargestellt wurden. Für die summative Evaluation wurde u.a. die Methode der standardisierten Befragung der Kursteilnehmer gewählt.

Die beiden Praxisbeispiele sollen verdeutlichen, dass die **inhaltliche Gestaltung des Evaluationsdesigns auf dem Entwicklungsstand eines Programms abgestimmt werden sollte.** Basierend auf den Erkenntnissen aus den Evaluationsstudien lässt sich für den Einstieg in die Entwicklung des Evaluationsdesigns eine Einordnung der Funktionen und Hauptformen von Evaluation vornehmen. In Abbildung 6 ist die Entwicklung eines Programms ab dem Zeitpunkt der Implementierung den idealtypischen Hauptformen und Funktionen von Evaluation gegenübergestellt.

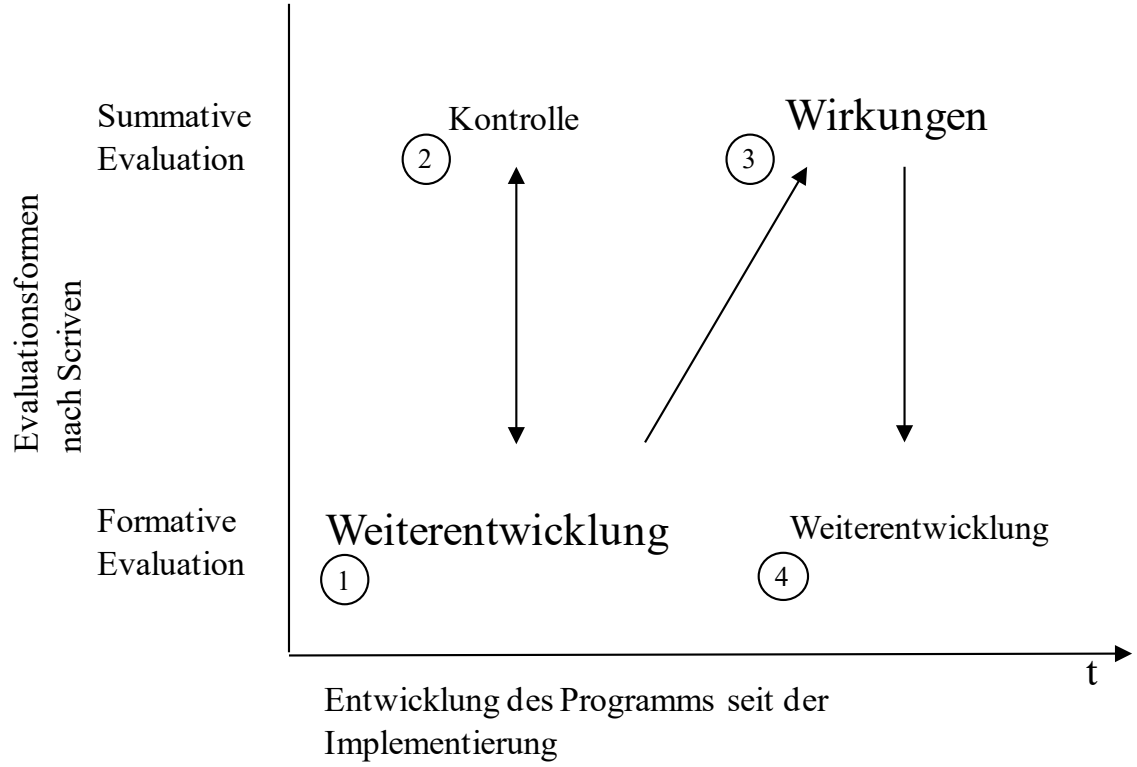


Abbildung 6: Typische Funktionen und Formen von Evaluation unter Berücksichtigung der Entwicklungsstufe des Programms

Typischerweise ist die primäre Funktion von Evaluationsstudien bei Programmen, die z.B. pilotiert werden, deren Weiterentwicklung (Ziffer 1). So dominierten im Evaluationsdesign zu der Kursreihe „In Deutschland zu Hause" Evaluationsmethoden mit dem Formen der formativen Evaluation (z.B. Beobachtungsprotokolle, um das didaktische Konzept flexibel anzupassen). Nach dem Abschluss der Kursreihe (10 Kursabende) wurden u.a. standardisierte Befragungen der Kursteilnehmer in Form einer summativen Evaluation durchgeführt, um ein zusammenfassendes Feedback zu erhalten (Ziffer 2). Die summative Evaluation erfüllte in der Pilotierungsphase der Kursreihe für die Programmverantwortlichen und den Evaluator die Funktion der Kontrolle – z.B. ob die Kursinhalte von den Teilnehmern als nützlich bewertet wurden.

Durchläuft ein Programm die Pilotierungsphase und sind Anpassungen am Programmkonzept erfolgt, ändert sich meist die Schwerpunktsetzung der Erkenntnissinteressen bei Evaluationsstudien. Bei Förderprogrammen ist es beispielsweise die Erfassung der Wirkungen der Förderung auf die Teilnehmer, die an Relevanz gewinnt (Ziffer 3). Im Programm *frühstart* wurde in Form einer summativen Evaluation die Wirkungen der Sprachförderung in den *frühstart*-Gruppen mithilfe von Sprachstandsmessungen unternommen. Auch hier wurden die Ergebnisse zur Weiterentwicklung des Programms verwendet. Würden die Er-

gebnisse einen Fördereffekt des *frühstart*-Konzepts belegen, gäbe dies für die Programmverantwortlichen Anlass zur Erörterung von Weiterentwicklungspotentialen des Programms (Ziffer 4).

Wie lässt sich die Entwicklung des Evaluationsdesign planerisch vorbereiten? In einem idealen Fall kann die Programmtheorie als Vorlage für die Ausarbeitung des Evaluationsdesigns herangezogen werden. Aus der Programmtheorie gehen die Ziele, die angestrebte Funktionsweise und der Entwicklungsstand des Programms hervor, so dass gemeinsam mit den Programmverantwortlichen der Zweck bzw. die Funktion der Evaluationsstudie festgelegt werden kann: Evaluation zu Weiterentwicklungszwecken, zu Kontrollzwecken und/oder zur Messung von Wirkungen. Diese Unterscheidung von Evaluation nach verschiedenen Funktionen ist deckt sich mit der Unterscheindung nach Chelimsky (1997, S. 10ff.), wie sie in der Einleitung beschrieben wurde.

Wie sehr jedoch die Anforderungen an ein Evaluationsdesign in der Realität von den idealen Zuständen – wie in der Abbildung 6 – dargestellt, abweichen können, zeigt die Evaluationsstudie zum Programm *frühstart*. Hier galt das Hauptinteresse schon zu Beginn der Pilotierung an der Wirkungsmessung der Sprachförderung. Formative Evaluationsformen nahmen eine nachrangige Rolle ein, z.B. in Form von Feedbackbefragungen von Erzieherinnen während ihrer Teilnahme an Schulungseinheiten. Auch sah sich der Evaluator mit der Herausforderung konfrontiert, die Programmtheorie und das Evaluationsdesign zeitgleich zu entwickeln, da der Evaluationsauftrag erst nach dem Start des Programms vergeben wurde. In anderen Fällen (z.B. wenn das Programm als Pilotprojekt initiiert wurde) kann das spätere Einsetzen einer Evaluationsstudie dazu führen, dass potentiell wichtige Informationen aus der Anfangsphase des Programms für die Evaluatoren nicht mehr zugänglich sind. Diese nicht intendierten und vom Auftraggeber nicht in diesem Maße überschaubaren Folgen sollten Evaluatoren im Vorfeld versuchen zu antizipieren. In solchen Fällen müssen Evaluatoren bereit dazu sein, auf Basis der Bedingungen und vorhandenen Ressourcen flexibel Entscheidungen zum Evaluationsdesign zu treffen.

Im folgenden Kapitel werden Evaluationsmethoden vorgestellt, die typischerweise in Evaluationsstudien zur Erfolgskontrolle und Weiterentwicklung von Programmen Anwendung finden (in Abbildung 6 Ziffern 1, 2 und 4). Daran schließt sich ein Kapitel mit der Darstellung von Evaluationsmethoden an, die insbesondere für Wirkungsevaluationen geeignet sind (Ziffer 3).

5.2. Methoden zum Zweck der Erfolgskontrolle und Weiterentwicklung von Programmen

Unter Evaluationsmethoden zum Zweck der Erfolgskontrolle und Weiterentwicklung von sozialen Programmen sind hauptsächlich Verfahren zu subsummieren, mit Hilfe derer die Konzeption, Umsetzung, Zielerreichung sowie die die Funktionsweise von sozialen Programmen analysiert werden können. Im Fall von Programmevaluationen beruhen Evaluationsmethoden auf der systematischen Anwendung von Instrumenten der empirischen Sozialforschung. Die Instrumente zur Informations- und Datenerfassung können einzeln oder als Kombination mehrere Instrumente (so genannter Methoden-Mix-Ansatz) in einem Evaluationsdesign zusammengefasst werden. Nach Patton (2002) werden in Evaluationsstudien die folgenden fünf Instrumente häufig in Evaluationsstudien eingesetzt: [35]

Schriftliche Befragungen: Die gesammelten Informationen lassen sich inhaltsanalytisch nach Kategorien auswerten. Das Ziel der Analyse ist die Identifikation von Mustern gleichen Antwortverhaltens. Eine strukturierte Ergebnisauswertung ist mit dieser Methode möglich, jedoch erlaubt dieses Verfahren keine Nachfragen. Bei der Einführung von neuen Programmen sind Feedbackbefragungen der Teilnehmer anhand von standardisierten Fragebögen sehr beliebt. Die Fragebögen können auch offene Fragen beinhalten, so dass sich Teilnehmer ausführlich zum Programm äußern können.

Interviews (z.B. standardisiert oder teilstandardisiert, Leitfadeninterviews, narrative Interviews): Abhängig von der Interviewart werden zumeist offene Fragen gestellt, die anschließend – wie bei schriftlichen Fragebögen – inhaltsanalytisch ausgewertet werden. Interviews in Face-to-Face-Situationen bieten außerdem den Vorteil gegenüber schriftlichen Befragungen, dass der Gegenstandsbereich der Befragung detailliert erhoben werden kann. Durch Nach- und Verständnisfragen lassen sich Unsicherheiten abklären und das Gesagte kann in eine inhaltliche Beziehung zueinander gesetzt werden. Bei Evaluationen stellen Interviews in der Implementierungsphase von Programmen eine beliebte Erhebungsmethode dar. Dem konstruktivistischen Wissenschaftsverständnis folgend können Teilnehmer, Mitarbeiter sowie weitere Stakeholder zum Programm befragt werden. Das Befragungskonzept und die Auswertung können dann triangulativ, d.h. aus der Perspektive der am Programm beteiligten Stakeholder, erfolgen. Eine Spezialform von Interviews sind Experteninterviews, bei denen Wissensträger zu

[35] An dieser Stelle wird einer Auswahl von relevanten Evaluationsinstrumenten vorgestellt, die häufig im Rahmen von Evaluationsstudien verwendet werden. Für eine detaillierte Darstellung der verschiedenen Methoden der qualitativen Sozialforschung sei auf die entsprechende Fachliteratur zur qualitativen Sozialforschung verwiesen (z.B. Lamnek 1996, Mayiring 2000).

bestimmten fachlichen Themen befragt werden. Ein Schwachpunkt von Interviews ergibt sich daraus, dass der Interviewer teilweise unbewusst durch die Art der Fragestellung und die Interviewatmosphäre die Antworten des Befragten beeinflusst.

Beobachtung (z.B. standardisierte und teilstandardisierte Beobachtung): Die Beobachtung liefert als klassisches Instrument der empirischen Sozialforschung Beschreibungen zu Verhaltensweisen, Interaktionen und Situationen innerhalb von sozialen Gruppen. Die Beobachtung wird von einem oder mehreren Forschern mithilfe eines Beobachtungsrasters protokolliert und anschließend inhaltsanalytisch ausgewertet. Das Instrument der Beobachtung unterscheidet sich danach, ob es teilnehmend bzw. nicht-teilnehmend ist, sowie nach dem Grad des strukturierten Erfassens der Ergebnisse. Beobachtungsverfahren sind im Vergleich zu den anderen Evaluationsinstrumenten relativ zeitintensiv, sowohl hinsichtlich des Verfahrens als auch bei der Auswertung. Wie bei Interviews können durch teilnehmende Beobachtungen die Ergebnisse durch den Evaluator verzerrt werden. Das Instrument der Beobachtung spielt bei Evaluationsstudien eher eine Nebenrolle, da aufgrund von eng gesetzten Zeitplänen die Entscheidung für den Einsatz alternativer Methoden gefällt wird.

Dokumentenanalyse (z.B. Auswertung von Literatur, Fachzeitschriften, Statistiken etc.): Das Verfahren der Dokumentenanalyse ist bei allen Evaluationsverfahren nahezu unentbehrlich geworden. Hauptsächlich in der Vorbereitung einer Evaluationsstudie ist die Aufbereitung umfangreicher Informationen notwendig. Diese Informationen bieten die Grundlage für die Entwicklung einer geeigneten Evaluationsstrategie.

Gruppendiskussionen (z.B. Fokusgruppen, Workshops): In moderierten Gruppenveranstaltungen können durch die Beteiligung einer Mehrzahl von Stakeholdern relevante Informationen für die Evaluation von Programmen gewonnen werden. Von zentralen Fragestellungen ausgehend, können die Sichtweisen der Diskussionsteilnehmer gesammelt und kategorisiert werden. Diese qualitative Methode eignet sich besonders als Kontrollinstrument, wenn ein Feedback der Programmbeteiligten notwendig ist. Wie bei Interviews birgt diese Methode die Gefahr, dass durch die Moderation der Ausgang der Gruppendiskussion und somit die Ergebnisse beeinflusst und gesteuert werden.

Für die Auswahl der Evaluationsinstrumente für das Evaluationsdesign sollten idealerweise die Angaben aus dem zuvor entwickelten Change Models der Programmtheorie (siehe Spalten 1-3 der *Program-Theory-Matrix* auf Seite 93) herangezogen werden. In der folgenden Tabelle ist beispielgebend das Evaluationsdesign zum Modellprojekt „In Deutschland zu Hause“ abgebildet. Darin sind die u.a. die Instrumente der Datenerhebung, die Zielgruppen sowie die Erhebungszeitpunkte festgelegt. Aus der Darstellung des Evaluationsdesigns wird deutlich,

dass mit der Anwendung der Evaluationsinstrumente ursprünglich drei Funktionen erfüllt werden sollten: die Weiterentwicklung der Kursreihe durch geeignete formative Evaluationsverfahren, die Kontrolle der Kursqualität sowie eine Wirkungsmessung bei den Teilnehmern. Zu dem Modellprojekt Integrationskurse wurde keine Programmtheorie entwickelt. Die Programmziele wurden aber vor der Entwicklung der Kurse schriftlich festgehalten und sind in der Tabelle in Spalte zwei beschreiben.

Instrumente und Datenquellen	Gemäß Programmziele oder Programmtheorie	Gegenstand der Erhebung	Zielgruppen	Erhebungszeitpunkte	Funktion der Evaluation
Kurzfeedbackfragebögen	Kontinuierliche Verbesserung der Kursinhalte sowie der Didaktik	Subjektives Feedback zu den Kursinhalten und zum Kursleiter	Teilnehmer	Nach jeder Kurseinheit	Weiterentwicklungsfunktion (formative Evaluation)
Beobachtungsraster	Kontinuierliche Verbesserung der Kursinhalte sowie der Didaktik	Strukturierte Erfassung des Kursverlaufs durch den Evaluator	Teilnehmer und Kursleiter	Während jeder Kurseinheit	Weiterentwicklungsfunktion (formative Evaluation)
Standardisierter Fragebogen	Einbürgerungswillige als Hauptzielgruppe	Fragen zur Teilnahme-Motivation und Einbürgerungsintention	Teilnehmer	Vor dem Start der ersten Kurseinheit	Kontrollfunktion (summative Evaluation)
Standardisierter Fragebogen	Hohe Qualität des Kurses. Kriterien: Kursleitungsverhalten, Aufbau und Didaktik, Inhalte	Feedback der Teilnehmer zu den Kurskriterien	Teilnehmer	Nach Abschluss der letzten Kurseinheit	Kontrollfunktion (summative Evaluation)
Daten zu den Kursteilnehmern	Ausgewogene Aufwand-Nutzen-Relation	Anzahl Anmeldungen, Teilnehmer jeder Kurseinheit, Fluktuation, Kursabbruch, Anzahl Absolventen	-	Nach Abschluss der letzten Kurseinheit	Kontrollfunktion (summative Evaluation)
Standardisierter Fragebogen zum Abschluss der Kursreihe	Vollziehen von Einstellungsänderungen, Erwerb von sozialkundlichen Kenntnissen, Aneignung von Orientierungswissen	Einstellungserhebung mittels Semantischem Differential, Angabe von subjektiv geschätzte Lernergebnisse,	Teilnehmer	Nach Abschluss der letzten Kurseinheit	Wirkungsmessung

		Wissenstest (konnte nicht umgesetzt werden): 10 Multiple-Choice-Fragen zu den Kursinhalten			

Tabelle 2: Evaluationsdesign zum Programm Modellprojekt Integrationskurse „In Deutschland zu Hause“

Wenn durch den Evaluator die Strukturierung des Evaluationsdesigns in Tabellen- bzw. Matrixform konsequent angewendet wird, bietet dies den Vorteil, dass die Erhebungsinstrumente direkt auf die Programmziele und die Zielgruppen ausgerichtet werden können. Die Darstellung des Evaluationsdesign in Matrixform kann generell für die Vorbereitung einer Evaluationsstudie bei allen Programmen angewendet werden.

Die Vorgehensweise bei formativen Evaluationen soll im Folgenden am Beispiel der kontinuierlichen Verbesserung der einzelnen Kurseinheiten näher beschrieben werden. Wie in Kapitel 2 erläutert wurde, geht aus der Entwicklung der Evaluationsforschung in den USA hervor, dass formative Evaluationsformen seit den 70er Jahren deutlich an Relevanz und Häufigkeit zugenommen haben. Der Bedeutungszuwachs findet seinen Ausdruck in diversen Evaluationsansätzen und führte zu einer „Umorientierung von der Wissenschaftlichkeit (als Hauptkriterium für die Konzeption und Bewertung von Evaluationen) zur Anwendungsorientierung“ (Flick 2006, S. 12). Das von Stake in den 70er Jahren entwickelte Konzept der responsiven Evaluation bezieht Stakeholder beispielsweise durch die beständige Rückmeldung der Evaluationsergebnisse in den Programmentwicklungsprozess ein.

Die Evaluation des Kursprogramms „In Deutschland zu Hause“ hatte einen **primär formativen Charakter** und folgte den Prinzipien einer responsiven Evaluation nach Stake. In der Pilotphase der Kurse wurde ein so genannter Evaluationszyklus angewendet. Durch die Anwendung des Zyklus wurde sichergestellt, dass jede abgehaltene Kurseinheit von dem Kursleiter und dem Evaluator gemeinsam reflektiert und Verbesserungsmaßnahmen kurzfristig initiiert wurden. Das Funktionsschema eines formativen Evaluationszyklus ist in der folgenden Grafik nach Balk dargestellt (vgl. Balk 2000, S. 36)[36]:

[36] Der von Balk vorgestellte Zyklus orientiert sich maßgeblich an den PDCA-Zyklus (Plan-Do-Check-Act) nach Deming (1982) und Shewhart (1931). Der PDCA-Zyklus ist ein Instrument des statistikgesteuerten Prozessmanagements, das in Qualitätsmanagementsystemen in Unternehmen Anwendung findet.

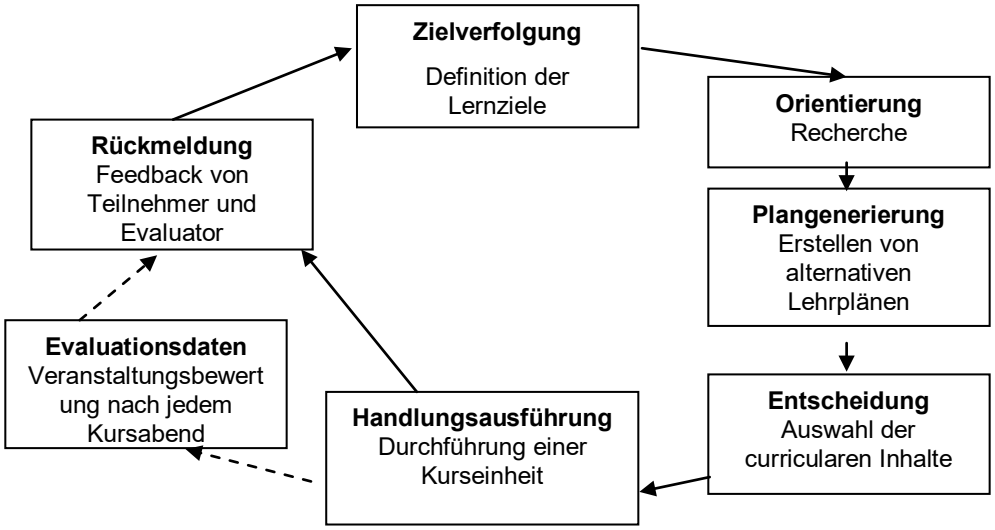


Abbildung 7: Schema der formativen Evaluation des Integrationskurses „In Deutschland zu Hause“

Die Funktion dieses abstrakten Schemas lässt sich wie folgt erläutern. In den Evaluationsverlauf ist während des gesamten Kurszeitraums der Kursleiter stark eingebunden. Unter dem Aspekt Zielverfolgung ist in diesem konkreten Beispiel die Festlegung von Lernzielen für die Teilnehmer gemeint, die als Strukturierungshilfen bei der Ausarbeitung der curricularen Inhalte dienen. Die Aspekte Orientierung, Plangenerierung und Entscheidung umreißen den Entstehungsprozess des Curriculums und die Auswahl der Lehrmethoden. Das Curriculum wurde vor den einzelnen Kurseinheiten ausgearbeitet. Zu jeder Kurseinheit wurde ein standardisiertes Feedback von den Teilnehmern eingeholt. Aus der parallel verlaufenden Entwicklung der Kurs- und Evaluationsstrategie sind die Instrumente für die programmbegleitende Evaluation entstanden, wie z.B. Kurzfeedbackbögen an die Teilnehmer, die nach jeder Kurseinheit eingesetzt wurden. Sowohl die Feedbackbögen als auch die teilstandardisierte, teilnehmende Beobachtung in den einzelnen Kurseinheiten durch den Evaluator lieferten nützliche Informationen für eine kontinuierliche Optimierung der Kursinhalte und deren Vermittlung. Die Auswahl der didaktischen Methoden sowie der Medienformen beruhte in den ersten Kursreihen zu einem guten Teil auf dem Teilnehmerfeedback. Die Anwendung des Evaluationszyklus erlaubte eine effektive qualitative Weiterentwicklung des gesamten Kurskonzeptes, so dass zum Ende des Evaluationszeitraumes ein aus der Praxiserprobung weiterentwickeltes Kursmaterial zur Verfügung stand. Das Kursmaterial sowie das Kurskonzept wurden in mehreren Evaluationsschleifen weiterentwickelt (z.B. wurde das Kursbuch einmal komplett überarbeitet).

Der Evaluationszyklus kann zum Zweck der formativen Evaluation auf prinzipiell alle Programme angewendet werden, wobei die Zyklusdauer unterschiedlich sein

kann. Bei den Integrationskursen bezog sich der Zyklus auf jede einzelne Kurseinheit. An anderen Fällen können sich die Entwicklungszyklen bis hin auf das Management eines gesamten Programms erstrecken.

5.3. Evaluationsmethoden für Wirkungsmessungen

Sollen die Effekte eines Programms oder einer Maßnahme erfasst werden, eignet sich die Anwendung von Wirkungs- und Nutzenevaluationen. Beachtet man das zuvor Gesagte, macht dieser Evaluationsverfahren erst nach Abschluss der frühen Implementationsphase Sinn (vgl. Wottawa & Thierau 1998; Chen 1990), wenn das Programm eine gewisse Entwicklungszeit durchlaufen hat und sich als recht stabil erweist. Im Gegensatz zur formativen Evaluation, die eine entscheidende Rolle bei der programmbegleitenden Evaluation einnimmt, wird bei Wirkungsevaluationen von summativer Evaluation gesprochen. Bei summativen Evaluationen erfolgt die Auswertung der erhobenen Daten zusammenfassend und bewertend (Stockmann & Meyer 2014).

Bei Wirkungsevaluationen im engeren Sinn handelt es sich um verschiedene Varianten von ex post Evaluationen mit mindestens zwei Erhebungszeitpunkten unter zusätzlichem Einbezug einer Kontrollgruppe. In der Evaluationsforschung besteht weitgehende Einigkeit darüber, dass experimentelle und quasi-experimentelle Versuchsanordnungen die besten Designs für die Wirkungsevaluation von Programmen sind (Campbell & Stanley 1966, Cook & Campbell 1979).

Die Fragen „Zeigt das Programm Wirkung?“ und „Ist es das Programm, das eine Veränderung in der Untersuchungsgruppe verursacht oder ist es etwas anderes?“ sind die zentralen Fragestellungen, mit denen sich Wirkungsevaluation beschäftigt. Da Evaluationsforscher die Auswirkungen von sozialen Programmen nicht simulieren können, müssen Methoden der empirischen Sozialforschung angewandt werden, um die Wirkungen zu untersuchen. Aussagen zu Wirkungen können in der Sozialforschung nicht mit absoluter Sicherheit, sondern nur innerhalb definierter Fehlergrenzen und mit unterschiedlicher Plausibilität getroffen werden.

Es lassen sich viele Ursachen für die Schwierigkeit benennen, die Wirkungen von sozialen Programmen zu bestimmen. Soziale Phänomene haben meist viele Ursachen, d.h. bei der Bestimmung von Ursache-Wirkungs-Zusammenhängen müssen eine Vielzahl von Einflüssen in dem zu untersuchenden Modell berücksichtigt werden. Die Schwierigkeit bei der Durchführung von Wirkungsevaluationen besteht gerade darin, die Wirkung der programminternen Einflussfaktoren (endogene Variablen) getrennt von den externen Einflussfaktoren (exogene Variablen) zu analysieren. So ist beispielsweise denkbar, dass das soziale Problem

durch eine Reihe von (exogenen) Variablen beeinflusst wird, jedoch nicht durch das Programm selbst.

Die für die Wirkungsevaluation entscheidende Frage ist, ob die Intervention eines Programms häufiger oder vermehrt zu bestimmten Effekten führt als es ohne oder mit einer anderen Intervention zu erwarten wäre. Nach Rossi und Freeman (1988) kommen Wirkungsmessungen bei sozialen Programmen immer dann in Frage, wenn die Wirkungen nicht eindeutig feststellbar, messbar oder zu benennen sind. Das besondere Interesse der Programmträger ist es, dabei zu erfahren, ob und wenn ja, wie stark die Maßnahme oder Intervention Wirkungen bei der Zielgruppe erzeugt hat. Wirkungsevaluationen erscheinen nach Rossi und Freeman außerdem immer dann eingesetzt werden, wenn die Wirkungsweise einer Intervention oder Maßnahme unter verschiedenen Bedingungen überprüft werden soll (z.B. an mehreren Standorten und/oder mit unterschiedlichen Zielgruppen). Die Daten für die Wirkungsanalyse werden in systematischen, objektiven Verfahren gesammelt. Im Folgenden wird eine Auswahl dieser Verfahren beschrieben.

5.3.1. Experimente

Das wissenschaftliche Experiment kann definiert werden als „wiederholbare Beobachtungen unter kontrollierten Bedingungen, wobei eine (oder mehrere) unabhängige Variable(n) derartig manipuliert wird (werden), dass eine Überprüfungsmöglichkeit der zugrunde liegenden Hypothese (Behauptung eines Kausalzusammenhangs) in unterschiedlichen Situationen gegeben ist“ (Zimmermann 1972, S. 37; vgl. für ähnliche Definition auch Friedrichs 1980, S. 333). Das Experiment ist die exakteste Form wissenschaftlicher Forschung, da es im Gegensatz zu anderen quantitativen Methoden verschiedene Vorteile bietet. Die Ergebnisse von Experimenten erlauben Kausalaussagen über die Beziehung zwischen abhängigen und unabhängigen Variablen: „Im Mittelpunkt des klassischen Experiments steht das Bemühen, für die Datenerhebung Bedingungen zu schaffen, in denen nur das Ursache-Wirkungs-Prinzip zwischen Maßnahme und Effekt zur Geltung kommen kann“ (Kromrey 2002, S. 93). Durch die Möglichkeit der Manipulation der Versuchsbedingungen kann der Einfluss einer oder mehrerer unabhängiger Variablen auf eine abhängige Variable errechnet werden. Die wichtigste Anforderung an den Evaluator lautet, die hypothetisch angenommene Ursache für die Veränderung („Maßnahme“, „Treatment“, „Stimulus“ oder in diesem Fall ein soziales Programm) kontrolliert in die Experimentalsituation einzuführen und alle bestimmbaren Einflüsse abzuschirmen oder durch spezielle Methoden (Randomisierung) unwirksam zu machen. Bei der Prüfung von Hypothesen werden im Idealfall alle bedeutsamen Variablen kontrolliert. Dies kann jedoch nur durch so genannte Laborexperimente erreicht werden, bei denen ein einziges

Merkmal in kontrollierter Weise eingeführt wird und alle anderen Einflussfaktoren ausgeblendet werden. Situationen, die eine Durchführung von Laborexperimenten erlauben würden, sind so in der sozialen Realität nicht anzutreffen. In der Sozialforschung und Psychologie wird deshalb das Konzept des Feldexperiments verfolgt. Dabei versuchen die Forscher, die Bedingungen aus den Laborexperimenten in die Feldforschung zu übertragen. Die Kontrolle möglicher Einflussfaktoren ist jedoch in realen sozialen Situationen fast niemals vollständig möglich. Einsatzgebiete von Wirkungsevaluationen unter der Verwendung von Experimenten sind vor allem Reformprogramme mit geringem Umfang oder Modellprojekte mit einem zeitlich beschränkten Erprobungscharakter.

Wenn die Wirkung einer Politik, eines Programms oder einer Maßnahme gemessen werden soll, ermöglichen Experimente genaue Antworten über Wirkungszusammenhänge zu erzeugen (vgl. Cook 2006). In randomisierten Experimenten unterscheidet sich im Idealfall die Untersuchungsgruppe von der Kontrollgruppe nur durch die Tatsache, dass bei Ersteren ein Treatment (ein Projekt, eine Maßnahme, ein Curriculum, etc.) angewendet wird. Die Ergebnisse randomisierter Experimente erlauben daher Aussagen zu kausalen Zusammenhängen zwischen Treatment und Untersuchungsgruppe bzw. zwischen Nicht-Treatment- und Kontrollgruppe. Bei echten randomisierten Experimenten im sozialen Bereich müssen zwei formale Bedingungen erfüllt werden: (1) die Auswahl der Einheiten (z.B. Personen oder Familien) für die Untersuchungsgruppe und für die Kontrollgruppe muss zufällig erfolgen; und (2) die Anzahl der Einheiten je Gruppe sollte ausreichend groß sein (idealerweise mehrere hundert). Durch die zufällige Auswahl der Einheiten und einer angemessenen Gruppengröße ist es möglich, den intervenierenden Einfluss objektiver Merkmale wie Alter, Einkommen, Geschlecht, Bildungsgrad, ethnische Zugehörigkeit und den Einfluss von latenten Merkmalen wie Intelligenz und Motivation für die Wirkungsmessung zu minimieren.

In der folgenden Tabelle sind Beispiele für verschiedene Formen von experimentellen Designs abgebildet. Jede dieser Untersuchungsdesigns basiert auf der zufälligen Zuordnung von Personen (Zielgruppe der Evaluation) zur Gruppe der Teilnehmer an der Intervention (Experimentgruppe) und einer weiteren Gruppe, die die Intervention nicht erhält (Kontrollgruppe). Sozialwissenschaftliche Experimente machen sich den Vorteil des „kontrollierten Zufalls“ zunutze (Kromrey 2002), indem die Teilnehmer zufällig der Experiment- bzw. Kontrollgruppe zugeordnet werden. Diese Methode für die Vorbereitung eines Experiments wird auch „Randomisierung“ genannt. Kromrey (2002) weist aber darauf hin, dass dieses Vorgehen der zufälligen Zuordnung nicht mit einer Zufallsauswahl der Teilnehmer verwechselt werden darf. In der Realität findet sich meist eine pragmatische Herangehensweise an die Durchführung von sozialen Experimenten. Für das Experiment werden zunächst potentielle Teilnehmer rekrutiert. Die für das

Experiment bedeutsamen Merkmale (z.B. Alter, Geschlecht, Erwerbstätigkeit) der Teilnehmer werden erhoben und anschließend gruppiert[37]. In einem Zufallsverfahren (z.B. durch eine Rechenprozedur in der Datenbank) werden die Teilnehmer aus den verschiedenen Gruppen (Schichten) der Experiment- bzw. Kontrollgruppe zugewiesen. Nach der Randomisierung dürfen sich die Verteilungen der ausgewählten Merkmale zwischen Experiment- und Kontrollgruppe nicht statistisch signifikant voneinander unterscheiden. Dadurch wird gewährleistet, dass gruppenbezogene Unterschiede nicht das Ergebnis des Experiments verzerren.

Experimentelles Design	Symbolische Darstellung	Typische Fragestellungen
1. Nur Posttest-Design mit einer Kontrollgruppe	X O O	Hat das Programm bei den Teilnehmern zu einer Veränderung geführt?
2. Design mit zwei Programmteilen und einer Kontrollgruppe	O X O Y O X Y O	Haben beide, ein oder kein(e) Bereich(e) des Programms zu Veränderungen bei den Teilnehmern geführt? Ist der Einfluss auf die Programmwirkung in beiden Bereichen gleich groß?

Legende: o = Messzeitpunkt; x,y = Intervention oder Maßnahme

Tabelle 3: Beispiele für experimentelle Evaluationsdesigns[38]

Das einfachste Design eines richtigen Experiments sieht die Messung nach der Intervention bei Experiment- und Kontrollgruppe vor, jedoch ohne eine Messung zu einem Zeitpunkt vor der Intervention durchgeführt zu haben (Boruch 1997). Da die Gruppenzusammensetzung durch eine Zufallsauswahl von Personen erfolgte, ist ein Pretest nicht notwendig, um die Wirkungen des Programms eindeutig zu identifizieren. Selbst bei einem einfachen Experiment ist die interne Validität der Ergebnisse höher als beispielsweise bei quasi-experimentellen Designs, die mit nicht-äquivalenten Kontrollgruppen arbeiten. Die hohe interne Validität von randomisierten Experimenten lässt sich durch den Ausschluss von

[37] Denkbar ist auch ein weiteres Vorgehen: die Teilnehmer werden von Beginn an hinsichtlich bestimmter Merkmale und Quoten rekrutiert. Diese Methode der vorgeschalteten Quotenauswahl bietet dem Evaluationsforscher eine größere Möglichkeit, die Zusammensetzung der Gruppe zu kontrollieren. In der sozialen Realität ist diese Herangehensweise wohl kaum zu realisieren: für Evaluationsforscher ist es meist nicht möglich, aus einem Pool von Interessenten die Teilnehmer für das Experiment auszusuchen. Man nimmt oft das, was vorhanden ist.

[38] Die in der Tabelle gewählte Darstellungsweise folgt der von Posavac und Carey (1992, S. 216ff.).

programminduzierten Einflussfaktoren erklären, die die Ergebnisse des Experiments verzerren können. Die Experiment- und Kontrollgruppe unterscheiden sich nicht hinsichtlich der Alterszusammensetzung. Reifungsprozesse finden in beiden Gruppen während des Untersuchungszeitraumes im selben Ausmaß statt und können die Ergebnisse nicht verändern. Die Verzerrung der Experimentergebnisse aufgrund von Selektionsprozessen (Selbstselektion und Fremdselektion) kann durch die zufällige Zuteilung der Programmteilnehmer in Experiment- und Kontrollgruppe ausgeschlossen werden. Dasselbe gilt auch für statistische Effekte (z.B. Regression zum Mittelwert), die sich aus der Häufigkeitsverteilung bestimmter Merkmale in den Gruppen ergeben können (Decken- oder Bodeneffekte, z.B. eine Gruppe von Schülern mit sehr schlechten schulischen Leistungen).

Die Messung der Nettowirkungen eines Programms kann nach Rossi und Freeman (Rossi et a. 1988, S. 95 ff.) anhand einer einfachen Formel berechnet werden:

Nettowirkung = Ergebniswerte für die randomisierte Versuchsgruppe nach der Intervention – Ergebniswerte für die randomisierte Kontrollgruppe nach der Intervention ± stochastische Effekte

Die Größe der stochastischen Effekte ist von der Zahl der Beobachtungen bzw. Programmteilnehmer abhängig. Der *Zufallsfehler* wird kleiner, wenn mehr Personen am Experiment teilnehmen. Je mehr Personen sich in der Kontroll- und Experimentgruppe befinden, desto höher ist die Wahrscheinlichkeit, dass die Wirkungen einer Intervention als statistisch signifikant interpretiert werden. Neben dem Zufallsfehler spielt die *Effektstärke* eine entscheidende Rolle für die Interpretation der Nettowirkung. Wird in der Vorbereitungsphase des Experiments erwartet, dass die Effektstärke klein sein wird, müssen Experiment- und Kontrollgruppe einen großen Umfang haben, damit die Nettowirkungen als praktisch relevant interpretiert werden können.

Weiterhin ist denkbar, dass ein Programm inhaltlich in zwei Bereiche unterteilt wird. Für diesen Fall bietet es sich an, ein Evaluationsdesign anzuwenden, das beide Bereiche des Programms individuell berücksichtigt. Hierbei handelt es sich in den meisten Fällen um komplexe Experimente, mit denen der Versuch unternommen wird, gleichzeitig mehrere Programmvarianten zu analysieren. Realisiert wurden solch komplexe Experimente in der Realität selten[39].

[39] In den USA wurden in den 70er Jahren einige Experimente mit Reformprogrammen im sozialpolitischen Bereich durchgeführt. Im New-Jersey-Pennsylvania Income Maintenance Experiment wurden acht Sozialhilfeprogramme hinsichtlich ihrer Wirkungen untersucht (Kershaw & Fair 1976). In ähnlicher Weise wurden im Rahmen des Housing Allowance Demand Experiment (Kennedy 1980) 23 Versuchsgruppen gebildet.

Bei einigen Wirkungsanalysen kann es sinnvoll sein, vor dem eigentlichen Experiment einen Pretest durchzuführen. Experimente, die nur aus einer Posttest-Untersuchung bestehen, produzieren Aussagen, ob ein Programm wirkt. Durch den Pretest lassen sich zusätzlich Informationen über das quantitative Ausmaß der Veränderung sammeln. Ein weiterer Vorteil von Pretest-Posttest-Experimenten liegt im statistischen Bereich. Wenn Daten vor und nach der Einführung des Programms erhoben werden, können individuelle Unterschiede innerhalb jeder Gruppe bei der Analyse der Programmwirkung kontrolliert werden. Ein geeignetes Instrument für die Analyse von Experimenten mit zwei Erhebungswellen und einer Kontrollgruppe bietet die so genannte Kovarianzanalyse.

Ein Beispiel für ein Experiment mit einem Pretest ist das Solomon Vier-Gruppen-Design. Das Vier-Gruppen-Design eines randomisierten Experiments bietet Evaluatoren den folgenden Vorteil: ist beispielsweise nicht sichergestellt, dass die zufällige Aufteilung der Programmteilnehmer in Experiment- bzw. Kontrollgruppe strukturiert nach methodischen Kriterien erfolgt, wodurch nicht ausgeschlossen wird, dass sich die Gruppen voneinander unterscheiden, kann mit Hilfe der Daten aus dem Pretest eine Wirkungsanalyse ähnlich der quasi-experimenteller Untersuchungen vorgenommen werden.

Pretests in experimentellen Wirkungsanalysen können sich aber auch nachteilig auf die Ergebnisse auswirken: Programmteilnehmer könnten durch die Vorerhebung sensitiver auf das Programm reagieren. Dies bedeutet, dass Pretests zu einer höheren Reaktivität bei den Programmteilnehmern führen können und dass im Extremfall der Pretest zu einem Teil der Intervention bzw. des Programms werden kann. Campbell und Stanley (1966) erwähnen, dass ein Pretest die externe Validität der Evaluationsergebnisse schwächen kann, da nicht auszuschließen ist, dass der Pretest mit den Programmbestandteilen interagiert und auf diese Weise zu den Veränderungen bei den Teilnehmern führt. Obwohl in einigen Situationen – wie oben beschrieben – Posttests für die Analyse von Wirkungen ausreichend sind, erscheint es zielführender, mehr Messungen der Zielvariablen vor und nach der Intervention vorzunehmen. Neben der Zuverlässigkeit der Messung werden Aussagen zur Nachhaltigkeit eines Programms gewonnen. Auch regelmäßige Nachuntersuchungen sind mit dem so genannten experimentellen Zeitreihendesign möglich.

Es kann festgehalten werden, dass Experimente hinsichtlich ihrer Vorbereitung und Durchführung sehr aufwändig sind. Damit ein Experiment zu einem sozialen Programm in Realität durchführbar ist, müssen eine Reihe begünstigender Faktoren vorliegen (z.B. die Möglichkeit der randomisierten Zuordnung von Teilnehmern zur Untersuchungs- bzw. Kontrollgruppe) sowie eine umfassende Kontrolle von externen Störfaktoren erfolgen – also das Herstellen von Laborbedingungen.

5.3.2. Quasi-experimentelles Design

In der empirischen Sozialforschung ist die Planung und Durchführung von Experimenten schwer zu realisieren. Am Beispiel der Sprachförderung kann dies deutlich gemacht werden: Bei der Untersuchung des Spracherwerbs von Vorschulkindern gilt es, eine Vielzahl von Bedingungen zu berücksichtigen, z.B. ob es sich um den Erwerb der Erst- oder Zweitsprache handelt, wie die Sprachförderung im Kindergarten abläuft oder wie die Kindergruppen hinsichtlich bestimmter Merkmale – wie ethnische Zugehörigkeit – zusammengesetzt ist. Individuelle Faktoren (z.B. Intelligenz), unterschiedliche Treatments sowie Intensität und Zeitraum der Förderung können die Sprachentwicklung in unterschiedlicher Art beeinflussen.

Die Wirkungsmessung kann im Rahmen von Evaluationsuntersuchungen alternativ mit einem quasi-experimentellen Design erfolgen. Diese unterscheiden sich von Experimenten durch die Art und Weise der Bildung von Experiment- und Kontrollgruppe. Programmteilnehmer werden bei **quasi-experimentellen Verfahren ohne Randomisierung der Experiment- bzw. Kontrollgruppe zugeordnet**, d.h. sie haben nicht die gleiche Wahrscheinlichkeit, in die Experimentgruppe platziert zu werden (Campbell & Stanley 1966; Campbell et al. 2001). Aufgrund der fehlenden Randomisierung kann davon ausgegangen werden, dass sich die Gruppen – abhängig von dem gewählten Zuteilungsverfahren – hinsichtlich ihrer Merkmalskonfiguration voneinander unterscheiden. Daher wird während der Planungsphase eines Quasi-Experiments der Versuch unternommen, Unterschiede in der Merkmalsverteilung zwischen Experiment- und Kontrollgruppe zu minimieren. Der Terminus Experiment wird im Namen dieses Designs beibehalten, da die Versuchsanordnung den klassischen Experimenten entspricht, wie sie im Kapitel zuvor erläutert wurden. Ein anderes Unterscheidungsmerkmal ist, dass in der Literatur zu quasi-experimentellen Untersuchungen die Kontrollgruppe häufig als Vergleichsgruppe bezeichnet wird. Die Autoren Rossi und Freeman äußern sich positiv über die Möglichkeiten eines quasi-experimentellen Ansatzes: „The term ‚quasi-experiment' does not imply that the procedures described are necessarily inferior to the randomized controlled experiment in terms of reaching plausible estimates of net effects [...]. However, quasi-experiments, properly constructed, can provide information on impact that is free of most, if not all, of the confounding processes (threats to validity)" (Rossi & Freeman 1988, S. 267).

Die häufigsten Arten von quasi-experimentellen Untersuchungen sind das Pretest-Posttest Design mit einer Kontrollgruppe, das Nur-Posttest Design mit einer Kontrollgruppe und das Zeitreihendesign mit einer Kontrollgruppe. Das Pretest-Posttest Design (häufig auch als nichtäquivalentes Gruppendesign bezeichnet) ist vergleichbar mit dem klassischen Experiment. Die Validität der Ergebnisse

hängt davon ab, wie genau die Kontrollgruppe die Zusammensetzung der Experimentgruppe replizieren kann. In der folgenden Tabelle sind die zwei wichtigsten quasi-experimentellen Designs formal dargestellt (vgl. Posavac und Carey (1992, S. 192f.)).

Quasi-experimentelles Design	Symbolische Darstellung	Typische Fragestellung
1. Nur Posttest-Design mit einer Kontrollgruppe	X O O	Gibt es einen Unterschied zwischen den Teilnehmern in der Untersuchungs- bzw. Kontrollgruppe?
2. Pretest-Posttest-Design mit einer Kontrollgruppe	O X O O O	Gibt es einen Unterschied zwischen den Teilnehmern in der Untersuchungs- bzw. Kontrollgruppe? Wie groß ist dieser Unterschied?

Legende: o = Messzeitpunkt; x = Intervention oder Maßnahme

Tabelle 4: Beispiele für quasi-experimentelle Evaluationsdesigns

Der Evaluator erwartet bei einem Quasi-Experiment, dass die Unterschiede zwischen Erst- und Zweiterhebung deutlich höher ausfallen als bei einem randomisierten Experiment. Dies hängt vielfach damit zusammen, dass mögliche Einflussfaktoren auf das Untersuchungsdesign größer sind, die zugleich nicht Bestandteil des Treatments sind. Durch den Einbezug einer Kontrollgruppe in die Analyse lassen sich die Einflüsse eines sozialen Programms hinsichtlich endogener und exogener Störfaktoren kontrollieren. Wenn Wirkungen bei den Teilnehmern der Kontrollgruppe zu den gleichen Zeitpunkten gemessen werden, kann davon ausgegangen werden, dass Reifungsprozesse die Vergleichbarkeit zwischen den Gruppen nicht stören. Potenzielle Ereignisse während der Evaluationsphase bzw. Programmdurchführungsphase wirken sich dann auf beide Gruppen im gleichen Maße aus. Die Ausfälle von Programmteilnehmern sind bei der Pretest- und Posttest-Untersuchung bekannt und können hinsichtlich systematischer Unterschiede untersucht werden. Da die Teilnehmer nicht durch Selbstselektion der Experimentgruppe zugeordnet werden, stellen Quasi-Experimente eine gute Vergleichsgrundlage für die Untersuchung von kausalen Zusammenhängen dar. Diese Situation ist jedoch in der Realität oft nicht gegeben. Wie bereits zuvor beschrieben, ist die Möglichkeit nicht auszuschließen, dass die Ausprägungen der abhängigen Variable durch programmfremde Faktoren beeinflusst wird.

Zur Bildung konstruierter Kontrollgruppen benötigt man ein gewisses Maß an Erfahrung mit Methoden der empirischen Sozialforschung. Dies betrifft vor allem die Auswahl der spezifischen Merkmale, hinsichtlich derer sich Experiment- und Kontrollgruppe ähnlich sein müssen. Welche Merkmale besonders relevant erscheinen, kann von Evaluation zu Evaluation unterschiedlich sein. Stellen Personen die Teilnehmer beider Gruppen, sind demographische Variablen von besonderer Bedeutung für die Generierung der Gruppen. Daneben erscheint es sinnvoll, weitere Merkmale für die Kontrolle von Unterschieden heranzuziehen. Diese lassen sich durch das Studium entsprechender Literatur bzw. langjährige Erfahrung im thematischen Bereich identifizieren. Rossi und Freeman (1988) bemerken, dass die Auswahl möglichst vieler Variablen weder effizient noch sinnvoll ist. Viele Merkmale – beispielsweise im Bildungsbereich – korrelieren stark untereinander. Soll beispielsweise das kulturelle Kapital der Eltern von Migrantenkindern erfasst werden, genügt der höchste Bildungsabschluss der Eltern. Vielfach ist es hilfreich, eine Liste von relevanten Merkmalen in der Vorbereitungsphase der Evaluation aufzustellen. Anschließend ist zu prüfen, inwieweit die relevanten Merkmale mit überschaubarem Kosten- und Zeitaufwand erhoben werden können. Solche Listen sind jedoch kein Ersatz für hinreichende Kenntnisse des Untersuchungsgegenstandes. Durch die Auseinandersetzung mit einschlägiger Literatur kann deutlich werden, dass bestimmte Merkmale für die Evaluation zwar brauchbar, jedoch in der gegebenen Situation nicht optimal sind.

Die Bildung einer Kontrollgruppe für quasi-experimentelle Evaluationen kann auf verschiede Weise erfolgen. Eine weit verbreitete Methode ist das Parallelisieren, womit die Art und Weise einer systematischen Auswahl von Untersuchungseinheiten bezeichnet wird mit dem Ziel einer größtmöglichen Angleichung der Merkmale in Experiment- und Kontrollgruppe. Das Parallelisieren der Gruppen kann in zwei Schritten erreicht werden. Im ersten Schritt wird der Untersuchungsbereich sozial-räumlich abgesteckt. Soll die Wirkung von Sprachfördermaßnahmen auf die Deutschkenntnisse bei Kindern mit Migrationshintergrund erhoben werden, eignen sich für die Kontrollgruppe Kindertageseinrichtungen aus denselben Stadtteilen bzw. aus Stadtteilen mit ähnlichen Strukturmerkmalen (z.B. demographische Zusammensetzung, Arbeitslosenquote, Segregationsindex). Auch sollten im Idealfall die ausgewählten Kindertageseinrichtungen für die Kontrollgruppe vergleichbar mit den Experiment-Einrichtungen sein (Anzahl der Kinder mit Migrationshintergrund, Anzahl der ErzieherInnen mit Migrationshintergrund). Der zweite Schritt der Parallelisierung wird auf individueller Ebene durchgeführt. Hier bieten sich wiederum zwei mögliche Vorgehensweisen an: das individuelle „Matching“ und die Quotierung.

Beim **Matching** wird hinsichtlich bestimmter, als besonders relevant erscheinender individueller (bei Personen Alter, Geschlecht) und untersuchungsrelevanter Merkmale, nach Zwillingen für die Wirkungsmessung gesucht. Soll die Wirkung

eines Förderprogramms in Schulen auf leistungsschwache Schüler untersucht werden, kann das individuelle Matching durch die Auswahl von Schulkameraden der am Programm teilnehmenden Schüler erreicht werden. Ein Nachteil ist die zeitintensive und nicht effizient erscheinende Vorbereitungsphase der Wirkungsmessung (vgl. „Housing"-Studien von Wilner et al. 1962). Das Matching ist jedoch nach aktueller Meinung das beste Verfahren zur Bildung von konstruierten Kontrollgruppen[40].

Das zweite Verfahren mit dem Ziel der Parallelisierung ist die **Quotierung**. Im Gegensatz zum Matching wird das Verfahren nicht auf individueller Ebene durchgeführt. Mit der Quotierung wird der Versuch unternommen, bei spezifischen Merkmalen eine annähernd gleiche Verteilung der Merkmalsausprägungen vorzunehmen. Da die Untersuchungspraxis häufig ein individuelles Matching nicht erlaubt, werden Kontrollgruppen in den meisten Fällen durch Berücksichtigung festgelegter Quotierungsmerkmale konstruiert.

Der Nachteil beider Verfahren ist der unter Umständen **hohe Verlust an Untersuchungseinheiten in der Kontrollgruppe**. Sowohl durch das individuelle Matching als auch durch die Quotierung kann aufgrund mangelnder Vergleichbarkeit eine Vielzahl von Untersuchungseinheiten wegfallen. Bei der Quotierung hingegen kann nicht ausgeschlossen werden, dass – trotz aller Vorsicht – Experiment- und Kontrollgruppe unterschiedlich zusammengesetzt sind.

Das quasi-experimentelle Design wird auch deshalb so häufig angewendet, weil sich keine Möglichkeiten ergeben, randomisierte Experimente durchzuführen. Dies muss nicht unbedingt mit ethischen Gründen zusammenhängen. Häufig handelt es sich bei den Teilnehmern an sozialen Programmen um Freiwillige, die neben dem Programmservice keine weiteren Leistungen erhalten. Die Bildung einer Kontrollgruppe gestaltet sich vor allem dann schwierig, wenn die Teilnehmer keinen direkten Nutzen aus ihrer Beteiligung ziehen können. Je schwieriger es in der Vorbereitung eines Quasi-Experiments war, eine Kontrollgruppe zusammenzustellen, desto wahrscheinlicher sind die Ergebnisse nach der Wirkungsmessung zu interpretieren.

Der wissenschaftlich gesicherte Umgang mit Pretest-Posttest-Untersuchungen mit konstruierten Kontrollgruppen gehört zu den Designs, die am schwierigsten herzustellen und beizubehalten sind. Das Design wird als so schwierig beschrieben, dass einige Evaluationsforscher sogar davon abraten: „Although, this is probably a minority opinion, I recommend against the use of nonequivalent control groups, especially for beginning researchers" (Barker Bausell 1986,

[40] Beim individuellen Matching sollen für jedes Individuum der Untersuchungsgruppe ein oder mehrere, den Matchingkriterien entsprechende, Individuen in die Kontrollgruppe aufgenommen werden.

S. 138). Die Schlussfolgerungen aus Quasi-Experimenten sind gemäß der Argumentation von Bausell selten schlüssig, da sich zu viele alternative Erklärungen für die vorliegenden Ergebnisse ergeben können. Dies hängt überwiegend mit den Störfaktoren bei Wirkungsanalysen zusammen, die in Kapitel 5.3.4. erläutert werden.

5.3.3. Nur Posttest-Designs und Pretest-Posttest-Untersuchungen ohne Kontrollgruppe

Das primäre Interesse bei Wirkungsanalysen ist in den meisten Fällen, zunächst Anhaltspunkte zu gewinnen, inwieweit sich Programmteilnehmer in einer Weise verändert haben, die mit den intendierten Zielen des Programms zusammenhängt. Haben die Teilnehmer an einem Integrationskurs nach der letzten Einheit ihren Wissensbestand der unterrichteten Themen erhöhen können? Wie bewerten Eltern, die an einem Familien-Empowermentprojekt teilgenommen haben, ihre Erziehungskompetenz?

Art des Evaluationsdesigns	Symbolische Darstellung	Typische Fragestellungen
1. Nur Posttest-Design	X O	Wie ist das Niveau der Teilnehmer nach der Intervention? Wurden die minimalen Zielvorgaben des Programms erreicht?
2. Pretest-Posttest-Design	O X O	Die beiden obigen Fragen. Wie groß ist die Veränderung seit ihrer Teilnahme am Programm?

Legende: o = Messzeitpunkt; x = Intervention oder Maßnahme

Tabelle 5: Einfache Evaluationsdesigns ohne Kontrollgruppe

Wenn Evaluatoren in Erfahrung bringen wollen, ob sich Programmteilnehmer seit ihrem Eintritt in das Programm verändert haben, wird häufig das Pretest-Posttest-Design gewählt (siehe Tabelle 5 (vgl. Posavac und Carey (1992, S. 171))). Die Ergebnisse der Pretest-Posttest-Messung sind jedoch schwer zu interpretieren. Das Programm kann eine Wirkung bei den Teilnehmern gehabt haben, trotzdem ist das Pretest-Posttest-Design nicht ausreichend, um eine solche Schlussfolgerung zu ziehen. Es handelt sich also streng genommen nicht um ein Design für Wirkungsevaluationen. Das Design kann aber im Rahmen einer Voruntersuchung zu einer quasi-experimentellen Untersuchung umgesetzt werden, um das Ausmaß der Veränderung bei den Teilnehmern festzuhalten.

Die Begründung für die Unzulässigkeit von kausalen Schlussfolgerungen bei einfachen Designs ohne Kontrollgruppe erfolgte in der Evaluationsforschung durch Campbell und Stanley (1966) unter dem Titel „threats to internal validity". Der Begriff interne Validität bezeichnet die Möglichkeit, von einer unabhängigen Variable (z.B. die Teilnahme an einem Programm) kausal auf die Veränderung einer oder mehrerer abhängiger Variablen zu schließen (z.B. Merkmale, deren Veränderung als Verbesserung bzw. Verschlechterung im Programm interpretiert werden). Von „interner Validität" unterscheiden Campbell und Cook den Begriff „externe Validität": „Internal validity refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause. External validity refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times" (Cook & Campbell 1979, S. 37).

Die interne Validität der Ergebnisse von Programmevaluationen kann im Bereich der Feldforschung nie vollständig hergestellt werden. Ziel des Evaluators ist es, durch Auswahl eines geeigneten Designs und durch die stringente Kontrolle einer Vielzahl von Störfaktoren der internen Validität der Ergebnisse möglichst nahe zu kommen. Nach Campbell (1975) kommt der internen Validität von Programmen eine höhere Bedeutung zu als der externen Validität. Da die Funktionsweise und Wirkung sozialer Programme nicht unter Laborbedingungen – wie beispielsweise in der klinischen Psychologie – erprobt werden kann, sind diese besonders anfällig für Störfaktoren. Wirkungsanalysen sollten sich aus diesem Grund Methoden bedienen, die Störfaktoren und alternative Ergebnisinterpretationen Schritt für Schritt eliminieren und schließlich nur noch das soziale Programm als einzig mögliche Variable identifizieren, die zur Veränderung bei den Programmteilnehmern geführt hat.

5.3.4. Störfaktoren bei Wirkungsanalysen

Zur Überprüfung der internen Validität haben Campbell und Stanley eine Checkliste für experimentelle und quasi-experimentelle Studien entwickelt. Bei quasi-experimentellen Studien kann die interne Validität vor allem durch Selektionseffekte gefährdet sein: Personen oder Gruppen, die nicht durch ein zufälliges Auswahlverfahren der Untersuchungsgruppe zugewiesen wurden, können einen nicht quantifizierbaren Einfluss auf die Evaluationsergebnisse haben. Neben Selektionseffekten aufgrund nicht-zufälliger Auswahlverfahren kann zudem auch die interne Validität bei experimentellen Anordnungen durch (a) Reifungspro-

zesse; (b) unvorhergesehene Ereignisse zwischen Erhebungswellen und (c) Regressionseffekte bei extremen Verteilungen gefährdet sein (Campbell & Stanley 1966)[41].

Zur Kontrolle der internen Validität gehören auch die Berücksichtigung stochastischer Effekte und die Reliabilität der Messungen. Inwieweit bei der Evaluation eines sozialen Programms die unabhängige mit der abhängigen Variable kovariiert, und welche Schlussfolgerungen sich daraus über den kausalen Zusammenhang ergeben, wird mit dem Konzept der „statistical conclusion validity" untersucht. Zu diesem Konzept gehört die Untersuchung der Gefährdung der internen Validität durch Störfaktoren. Wenn die gemessenen Veränderungen statistisch signifikant und zudem groß genug sind, um als praktisch bedeutsam eingestuft zu werden, müssen zusätzlich nicht-programmspezifische Ursachen berücksichtigt werden, die möglicherweise eine Erklärung für die Veränderungen darstellen.

a) Reifungsprozesse

Bis zum Erreichen des Erwachsenenalters sind Kinder und Jugendliche einem konstanten Lernprozess ausgesetzt. Kinder lernen mit zunehmendem Alter komplexere Fähigkeiten; ältere Menschen brauchen umgekehrt länger, um bestimmte Handlungen durchzuführen. Im Leben dieser Personen können sich binnen kurzer Zeit große Veränderungen ergeben. Wirkungsevaluationen von Sprachförderprogrammen bei Kindern im Vorschulalter müssen den Prozess des Sprachenlernens berücksichtigen, da dieser auch ohne das Programm stattfindet.

Das Vorhandensein von Reifungsprozessen in einer wirkungsanalytischen Evaluation bedeutet jedoch nicht, dass dies der einzige Störfaktor ist. Auch ist daraus nicht die Schlussfolgerung zu ziehen, dass das Programm keine Wirkungen hat. Die Identifikation und Kontrolle von Reifungsprozessen kann am besten durch Evaluationsdesigns mit Kontrollgruppen erfolgen. Durch eine einfache Pretest-Posttest-Untersuchung ohne Personen, die nicht am Programm teilnahmen, kann dieser Störfaktor nicht kontrolliert werden.

b) Historische Ereignisse

Dieser Störfaktor tritt in Erscheinung, wenn zwischen dem Pretest und dem Posttest ein Ereignis eintritt, welches mit der Programmmaßnahme nicht zusammenhängt aber dennoch Einfluss ausübt. Die Ergebnisse der zweiten Messung können dadurch verstärkt oder abgeschwächt werden. In der Literatur finden sich verschiedene Erläuterungen, was unter historischen Ereignissen zu verstehen ist. Campbell und Stanley (1966) definieren diese als „the specific events occurring

[41] Näheres zum Einfluss von Regressionseffekten siehe Nachtigall und Suhl (2002): „Der Regressionseffekt. Mythos und Wirklichkeit".

between the first and second measurement in addition to the experimental variable" (1966, S. 5). Die meisten Evaluationsforscher beziehen sich auf diese Definition, hingegen bieten Rossi und Freeman (1988) eine genauere Unterscheidung von Ereignissen hinsichtlich des zeitlichen Kontextes an. Sie trennen dabei zwischen langfristigen Ereignissen als *„secular drift"* und kurzfristigen Ereignissen als *„interfering events"* (1988, S. 192f.). Die zeitliche Komponente spielt eine entscheidende Rolle bei der Evaluation von Wirkungen. Es ist leicht nachvollziehbar, dass – je länger ein Programm angeboten wird – das Auftreten von Ereignissen umso wahrscheinlicher ist.

Die statistische Kontrolle von historischen Ereignissen kann nur dann durchgeführt werden, wenn dementsprechend Daten gesammelt wurden. Reifungsprozesse sind hingegen vorhersehbar und treten nicht plötzlich ein. Posavac und Carey (1992) geben Evaluatoren den Ratschlag, die sozialen und gesellschaftlichen Veränderungen im Umfeld des Programms genau zu dokumentieren. Neben historischen Ereignissen, die als endogene Faktoren auf das Programm wirken, nennen Cook und Campbell (1979) auch endogene (programminterne) Faktoren (local history). Ein Dozentenwechsel in einem Fortbildungskurs für Erwachsene kann beispielsweise großen Einfluss auf den Lernerfolg der Teilnehmer haben.

c) Testung

Wird für die Wirkungsevaluation dasselbe Testinstrument an mehreren Messzeitpunkten verwendet, können sich Programmteilnehmer bei der zweiten Messung an die Vorgehensweise bei der ersten Messung erinnern. Eine Gefahr für die interne Validität besteht vor allem dann, wenn Testinstrumente im Posttest zur Überprüfung von Kenntnissen oder dem Abfragen eines bestimmten Wissensbestands eingesetzt werden. Bei Erwachsenen spielt dieser Störfaktor eine größere Rolle als bei Kindern, da erstere im Allgemeinen Prüfsituationen stärker reflektieren. Die Reaktivität ist ein zweiter Aspekt, der in diesem Zusammenhang nicht zu vernachlässigen ist. Personen verhalten sich anders, wenn sie sich der Situation bewusst sind, dass sie beobachtet werden. Wird eine Überprüfung von Sprachkenntnissen bei Vorschulkindern im Kindergarten nicht von den ErzieherInnen, sondern von externem Personal durchgeführt, ist mit hoher Wahrscheinlichkeit damit zu rechnen, dass selbst mit dem zuverlässigsten Sprachtest die Sprachkenntnisse der Kinder unterschätzt werden. Einflüsse des Testinstrumentariums lassen sich am besten durch das Solomon Vier-Gruppen-Design kontrollieren. Dabei werden die Ergebnisse einer quasi-experimentellen Untersuchung mit einem Vergleich zwischen Untersuchungs- und Kontrollgruppe mit zwei reinen Posttestuntersuchungen der Untersuchungs- und Kontrollgruppe verglichen. Dieser Vergleich verschiedener Untersuchungsvarianten bietet die Mög-

lichkeit, Pretest-Einflüsse (z.B. Anwendung des Testinstrumentariums) zu kontrollieren. Das Pretest-Posttest-Design ist dagegen zu schwach, um diese Art von Störfaktoren in befriedigender Weise berücksichtigen zu können.

d) Fehlwerte

Es ist selten der Fall, dass alle Teilnehmer, die mit dem Programm begonnen haben, dieses auch in der regulären Zeit beenden. Die Ausfallquoten variieren von Programm zu Programm, jedoch ist meistens die Anzahl der Teilnehmer, die das Programm vorzeitig beenden, signifikant hoch. Teilnehmer, die ein Programm verlassen, können sich in mehrfacher Hinsicht von den Teilnehmern unterscheiden, die im Programm bleiben. Evaluationsdesigns mit nur einer Posttest-Messung tendieren dazu, die Wirkungen von sozialen Programmen zu überschätzen, wenn nur die besonders erfolgreichen und motivierten Teilnehmer am Ende überprüft werden.

e) Selektionseffekte

Die Teilnahme an sozialen Programmen ist freiwillig. Dadurch ergibt sich die Gefahr, dass sich Experiment- und Kontrollgruppen unkontrolliert bilden. Unkontrollierte Gruppenbildung (Selektion) kann z.B. bedeutet, dass einige Teilnehmer (oder entsprechende Untersuchungsobjekte) eine größere Chance haben, an dem Programm teilzunehmen und somit in die Untersuchungsgruppe zu gelangen als andere.

Fasst man das bisher berichtete zusammen, so kann gesagt werden, dass zunächst für die Konstruktion des Evaluationsdesigns bei Wirkungsevaluationen nur ganz spezifische Verfahren in Frage kommen. Die aus der Psychologie oder Medizin übertragenen Verfahren für die Evaluation von sozialen Programmen haben strenge methodische Anforderungen. Auf den Evaluator kommt die Aufgabe zu, in der Phase der Evaluationsdesignentwicklung zu prüfen, inwieweit die methodischen Anforderungen bei der Bildung des Verfahrens sowie bei der Durchführung der Evaluation eingehalten werden können. Campbells Störfaktoren bei sozialen Programmenzeigen, mit welchen potenziellen Schwierigkeiten experimentelle und quasi-experimentelle Designs für Wirkungsevaluation ausgesetzt sind. Mit der zweiten Evaluierbarkeitsprüfung lassen sich zumindest die Faktoren benennen und darauf aufbauend können Strategien entwickelt werden, um den Einfluss dieser Faktoren zu kontrollieren. So können validitätsbeschränkende Einflüsse durch die Auswahl eines erprobten und im testtheoretischen Sinne als gut identifizierten Testinstrumentariums minimiert werden. Reifungsprozesse der Teilnehmer im Evaluationszeitraum können durch die Erfassung der Charakteristika wie z.B. Alter und die Bildung von ausreichend großen Untersuchungs- und Kontrollgruppen kontrolliert werden.

Die bisherige Ausarbeitung des Prozesses lässt sich folgendermaßen zusammenfassen. Die Ergebnisse der Evaluationseingangsphase bilden die Ausgangslage für die Entwicklung des Evaluationsdesigns. Zu der Eingangsphase zählen die Erarbeitung von Evaluationszielen, die Zusammenarbeit mit den Programmverantwortlichen zur Festlegung eines Projektplans, die Durchführung einer Evaluierbarkeitsprüfung sowie die Ausarbeitung einer Programmtheorie. **Die Programmtheorie dient als Grundlage für die Entscheidung für oder gegen ein Evaluationsdesign**. Eine Wirkungsanalyse sollte dann angestrebt werden, wenn zuvor die Machbarkeit in der Evaluierbarkeitsprüfung festgestellt wurde. Kriterien der Evaluierbarkeitsprüfung sind: angemessener Zeitrahmen, Möglichkeit der Bildung von Kontrollgruppen, Voraussetzungen für die Bildung der Kontrollgruppe sind gegeben (z.B. Teilnehmer mit vergleichbaren Merkmalen wie in der Untersuchungsgruppe), Anwendung von zuverlässigen Messinstrumenten möglich. In Abhängigkeit von den operativen Evaluationszielen, den Charakteristika und Zugänglichkeit der Zielgruppen, den Kontextbedingungen (z.B. räumliche Strukturierung des Programms) sowie den **Zeit- und Budgetrestriktionen** sollte die Festlegung von Evaluationsmethoden erfolgen. Schließlich ist der **Reifegrad des zu evaluierenden Programms** ein wichtiges Kriterium bei der Entscheidung für ein Evaluationsdesign. In Abhängigkeit von den Evaluations- und Programmzielen sowie der Programmtheorie werden geeignete Evaluationsmethoden ausgewählt und die Evaluationskriterien in die Erhebungsinstrumente eingearbeitet[42]. Das Evaluationsdesign sollte anschließend hinsichtlich der Konformität mit den Erkenntnisinteressen der Auftraggeber diskutiert und angepasst werden. Sind diese Schritte absolviert, kann mit der Planung der Durchführungsphase begonnen werden.

[42] Auf die Operationalisierung der Erhebungsmerkmale sowie die Gestaltung des Erhebungsinstrumentariums wird an dieser Stelle nicht eingegangen, da dies in einschlägigen Fachpublikationen der empirischen Sozialforschung im Detail behandelt wird. Verwiesen sei an dieser Stelle beispielsweise auf das Lehrbuch „Forschungsmethoden und Evaluation“ von Bortz und Döring (2003).

6. Durchführung der Evaluation und Auswertung der Evaluationsdaten

Zu Beginn der dritten Phase steht zunächst die Organisation der Erhebungsdurchführung. Dazu gehört z. B. die Erstellung von Zeit- und Projektplänen, die Information der Stakeholder sowie ggf. die Schulung von Mitarbeitern an der Evaluationsstudie (z.B. Interviewer). Die Erhebungsinstrumente werden in dieser Phase getestet und optimiert. Mit dem Beginn der Datenerhebung entsteht zugleich ein erster Rücklauf von Datenmaterial. Für die Auswertung des ersten groben Datenmaterials empfiehlt sich die Erstellung eines Auswertungsplans. Aufgenommen werden sollten die Auswertungsmerkmale, die Darstellungsform der Ergebnisse sowie die Auswertungstiefe. Auftraggeber sind sehr daran interessiert, erkenntnisbringende Evaluationsergebnisse so schnell wie möglich zur Verfügung gestellt zu bekommen. Eine grobe Erstauswertung der Daten nach einem mit dem Auftraggeber abgestimmten Auswertungsplan kommt schließlich diesem Wunsch entgegen.

6.1. Organisation und Vorbereitung der Erhebungsphase

Als erster Arbeitsschritt sollte zu Beginn der Erhebungshase die Erstellung eines Erhebungsplans erfolgen. Der Erhebungsplan muss bestimmten Anforderungen genügen, die sich aus dem Evaluationsdesign und dem Ziel der Evaluation ableiten lassen. Generell kann die Gestaltung des Erhebungsplans nach folgender Frage geleitet ausgearbeitet werden: Welche Methoden werden zu welchen Zweck wann und wie oft eingesetzt? Zwar können die Evaluationsmethoden, wie im Kapitel zuvor beschrieben, schon ausgewählt sein, jedoch können z.B. rein praktische Gründe oder Situationsbedingungen vorliegen, die Anpassungen der Vorgehensweise bei der Erhebung der Daten notwendig machen.

6.1.1. Vorbereitung der Erhebungsphase

Die Dauer der Vorbereitung der Erhebungsphase kann zeitlich leicht unterschätzt werden. Der Schwerpunkt der Tätigkeit des Evaluators konzentriert sich mit dem Start der Evaluationsstudie bewusst auf die Entwicklung des Evaluationsdesigns. Dies erscheint notwendig und richtig, da die Auswahl und Konzeption der passenden Methoden letztendlich die Qualität der Evaluationsstudie zum größten Teil beeinflussen. Allerdings sollte zugleich nicht unterschätzt werden, wie der Ablauf von Evaluationsstudien durch unvorhergesehene bzw. nicht im Vorfeld antizipierbare Ereignisse (z.B. Fluktuation von Personal) oder durch Änderungen in den Kontextbedingungen sich zeitlich in die Länge ziehen kann.

In den Projekten „Spielend Lernen“ und *frühstart* waren die Instrumente für die Sprachstandserfassung schon ausgewählt, da startete eine in zeitlicher Hinsicht

nicht unerhebliche Phase von mehreren Monaten der Organisation und Vorbereitung der Datenerhebungen. In *frühstart* nahm der Evaluator an mehreren Weiterbildungseinheiten für die ErzieherInnen der *frühstart*-Kitas teil, um sich zum einen ein Bild von den Inhalten des Sprachförderkonzepts zu machen und zum anderen die ErzieherInnen selbst mit Informationen zum geplanten Ablauf der Datenerhebungen zu versorgen sowie um deren Mitwirkung bei der Evaluation zu werben.

Auf Basis der gesammelten Erfahrungen in den untersuchten Studien sind die folgenden relevanten Schritte beim Einstieg in die Datenerhebungsphase zu beachten:

- *Vorbereitung eines Pretests* oder von vorbereitenden Untersuchungen (z.B. Fallstudien) zur Optimierung des Erhebungsinstrumentariums.
- *Abgleich des Erhebungsplans mit der Programmtheorie*: In einem Erhebungsplan wird festgelegt, zu welchen Zeitpunkt und Evaluationszweck Datenerhebungsinstrumente eingesetzt werden. Bei Wirkungsanalysen besteht der Erhebungsplan mindestens aus der Festlegung der Messzeitpunkte, den zusätzlichen parallel verlaufenden Datenerhebungen sowie ggf. genaue Angaben zu den Zeitpunkten für die Durchführung des Pretests.
- *Vor-Ort-Besuche:* Evaluatoren sollten direkte Einsicht in die Durchführung der zu evaluierenden Maßnahme erhalten; idealerweise durch Formen der teilnehmenden, nicht intervenierenden Beobachtung. Der Autor vertritt die Ansicht, dass reine Projektbesuche, bei denen Evaluatoren die Art und Weise der Durchführung der Maßnahmen nur beschrieben wird, nicht ausreicht, um ein differenziertes Bild von der intendierten Funktionsweise der zu evaluierenden Maßnahme zu erhalten.
- *Information an die Befragungsteilnehmer und an die sonstigen Beteiligten:* Mit ausreichend zeitlichem Vorlauf sollten alle über den Startzeitpunkt, die Dauer und die genaue Vorgehensweise bei der Erhebung informiert werden. Hier bieten sich für größere Personengruppen Informationsveranstaltungen an.
- *Abklärung der rechtlichen Voraussetzungen*: Sollen personenbezogene Daten erhoben und gespeichert werden, muss geprüft werden, inwieweit ein Daten-schutzfreigabeverfahren initiiert werden muss. Sind Minderjährige in die Erhebungen involviert, muss zusätzlich eine formale Zustimmung der Erziehungsberechtigten eingeholt sowie ggf. weitere datenschutzrechtliche Vorgehen zur Wahrung der Anonymität der Teilnehmer beachtet werden.
- *Planung von Verfahren mit mehreren Erhebungswellen* bzw. bei einer begleitenden Evaluation, bei der die Datenerhebung in regelmäßigen Abständen er-folgt. Dies ist zugleich ein weiterer Bestandteil des Erhebungsplans.

- *Zusätzliche personelle Unterstützung für die Datenerhebung:* Evaluationsstudien haben einen um eine mehrfach höheren zeitlichen und personellen Auf-wand, wenn die zu evaluierende Maßnahme an mehreren Standorten durchgeführt wird und alle Standorte in die Evaluation einbezogen werden sollen. Die Herausforderung besteht für die Evaluation darin, dass die Datenerhebungen an allen Standorten nach dem gleichen zeitlichen Verfahren ablaufen müssen.
- *Abklärung der Unterstützung durch weitere Stellen:* z.B. Behörden, Datenschutzbeauftragte, Entscheidungsgremien.

6.1.2. Reaktion auf kurzfristige Änderungen bei der Organisation der Erhebungsphase

Alle potentiellen Veränderungen in organisatorischer Hinsicht, die eine Gefahr für die weitere Durchführung der Evaluation bedeuten, lassen sich durch die Durchführung der Evaluierbarkeitsprüfungen im Vorfeld weitgehend kontrollieren bzw. vermeiden.

Sicherlich können auch bei einer perfekt geplanten Erhebungsphase Probleme auftauchen, die gelöst werden sollten, bevor die Datenerhebung startet. Mit der Durchführung der Evaluierbarkeitsprüfungen ist zumindest sichergestellt, dass keine Qualitätseinbußen zum Zeitpunkt der Fertigstellung der Evaluationsstudie zu erwarten sind. Ein Beispiel: Für eine Wirkungsanalyse hat man sich Evaluator und Auftraggeber auf die Durchführung einer quasi-experimentellen Untersuchung geeinigt. Im Laufe der Vorbereitung der Untersuchung konnten jedoch nicht genügend Teilnehmer für die Kontrollgruppe rekrutiert werden. Dadurch steht die Umsetzung der gesamten Wirkungsanalyse und damit der gesamten Evaluationsstudie in Frage. Diese für das Projekt fatale Entwicklung kann nur durch die konsequente Durchführung der zweiten Evaluierbarkeitsprüfung in der Phase der Entwicklung des Evaluationsdesigns vermieden werden.

Eine Schwierigkeit, mit der Evaluatoren im Verlauf einer Evaluationsstudie rechnen müssen, sind auftretende Änderungen am Programm sowie in der Umwelt des Programms. Dies können beispielsweise Reifungsprozesse sein, indem sich ein Programm in der Praxis etabliert hat. Daher sind insbesondere Modellvorhaben, d.h. Programme die zum ersten Mal getestet werden, im besonderen Maße von Veränderungen betroffen.

Andere Veränderungen können sein: „Mehr Geld wird verfügbar – oder weniger. Mitarbeiter scheiden aus und neue mit anderen Auffassungen oder Qualifikationen werden angestellt. Die politischen Winde wechseln und alte Beziehungen werden abgebrochen“ (Weiss 1972, S. 125). Auch können sich die im Kapitel zuvor genannten Störfaktoren nach Campbell manifestieren. Spätestens wenn dies eintritt, müssen Programmverantwortliche und Evaluatoren steuernd eingreifen.

Das für die Evaluation eigentlich *Problematische* ist dabei, dass Veränderungen auf die Teilnehmerzusammensetzung Auswirkungen auf die Maßnahme und somit auf die Aktivitäten im Programm haben können (vgl. auch Weiss 1972, S. 125). Letztendlich stellen Veränderungen die Richtigkeit einer zuvor ausgearbeiteten Programmtheorie in Frage. Evaluatoren können nur einen Überblick über Programmveränderungen behalten, wenn ein kontinuierliches Daten-Monitoring der Programmcharakteristika und Programmaktivitäten erfolgt. Treten die oben beschriebenen Veränderungen am Programm sowie im Kontext des Programms auf, muss der Evaluator prüfen, inwieweit dadurch das geplante Vorgehen und das Evaluationsdesign gefährdet sind. Das Resultat dieser Prüfung sollte anschließend mit den Programmverantwortlichen gemeinsam mit dem Ziel erörtert werden, ob und wenn ja welche Anpassungen am Evaluationsdesign vorgenommen werden sollten.

Wie soll der Evaluator mit dem Problem der potenziellen Programmänderungen zum Zeitpunkt der Evaluation umgehen? Ein Beispiel aus den *Evaluationsstudien* des Autors für eine einschneidende Veränderung der externen Bedingungen ist im Fall der Evaluationsstudie zum Integrationskurs „In Deutschland zu Hause" zu nennen. Die ursprüngliche Planung sah vor, die Kurse verpflichtend für Einbürgerungskandidaten mit fehlenden sozialkundlichen Kenntnissen vorzuschreiben. Eine verpflichtende Teilnahme von Einbürgerungswilligen in der Modellerprobungsphase der Kurse konnte nicht realisiert werden. Für die Teilnahme an den Kursen herrschten nun neue Bedingungen als ursprünglich vorgesehen. Die Kurse wurden somit letztendlich auch von politisch interessierten Migranten besucht. Dies war jedoch nicht die im Vorfeld anvisierte Zielgruppe. Die Kurse wurden trotzdem weiter geplant und schließlich der veränderten Zielgruppe angeboten.

6.1.3. Probleme der Akzeptanz der Evaluatoren bei Programmbeteiligten

Die Durchführung von Evaluationsstudien und das damit verbundene Auftreten von Evaluatoren können bei Programmbeteiligten nicht nur auf Zustimmung und Wohlwollen stoßen. Häufig werden Eingriffe in die Berufspraxis befürchtet, Angst vor Konsequenzen aufgrund von Evaluationsergebnissen sowie Unsicherheitsgefühle und Aversionen gegenüber Veränderungen („Warum etwas verändern, wenn es bisher auch gut lief?").

Ein prominentes Beispiel aus dem Hochschulkontext ist die Einführung von hochschulweiten Evaluationssystemen zur Bewertung der Lehre. Bis Ende der 90er Jahre war Lehrevaluation nur an wenigen Hochschulen verbreitet. Im Zuge der Bologna-Reform und Novellierungen der länderspezifischen Hochschulgesetzgebungen wurde die regelmäßige, standardisierte Form der Lehrevaluation durch studentische Bewertungen zum Regelinstrument. Die Einführung von Evaluationsformen stieß auf ein geteiltes Echo. Während manche Dozenten an

der Hochschule die neuen Instrumente der Bewertung eigeninitiativ aufgriffen, formierten sich anderenorts lobbyähnliche Vereinigungen von Professoren und Fachbereichen, um sich gegen die studentische Bewertung zur Wehr zu setzen. Ungeachtet der methodischen Diskussionen zum Sinn und Unsinn des Einsatzes von standardisierten Kurzfragebögen mit geschlossenen Antwortvorgaben als herkömmliches Instrument der Lehrevaluation stellte sich in den Hochschulverwaltungen die Frage, wie mit dem Thema Evaluation in Studium und Lehre umgegangen werden soll. In externen Vorgaben (z.B. Akkreditierungen, Hochschulgesetze) wird von den Hochschullehrern gefordert, sich aktiv mit den Evaluationen auseinanderzusetzen, um die eigenen Lehre weiterzuentwickeln. Wie soll jedoch eine Evaluationskultur an einer Hochschule entstehen, wenn kein proaktiver Partizipationswille seitens der Dozenten vorhanden ist?

Vor diesem Hintergrund erscheinen **vertrauensbildende Verhaltensweisen und Maßnahmen seitens der Evaluatoren** ein erster notwendiger Schritt. Solche Maßnahmen können auf die frühe Einbindung der Programmbeteiligten in den Evaluationsprozess abzielen, in dem sie schon zum Zeitpunkt der Vorbereitung möglichst viel über das Ziel, die Inhalte, den zeitlichen Umfang sowie viel über ihre Rolle im Evaluationsprozess erfahren. So sollten die einzelnen Schritte der Evaluatoren transparent gehalten werden, entweder, indem die Teilnehmer direkt mit den Informationen versorgt werden oder durch die Veröffentlichung des Evaluationsdesigns im Internet (z.B. gemeinsame Projekthomepage), der allen Teilnehmern an der Evaluation zugänglich gemacht wird. Ein weiterer und ggf. der wichtigste Punkt ist, das Evaluationsdesign auf potentielle Erfahrungen aus vergangenen Evaluationsstudien aufzubauen. Dadurch werden Synergien schaffen, unnötige Datenerhebungen und -analysen vermieden und den Programmbeteiligten signalisiert: Man nimmt die Vorarbeiten ernst und bezieht diese in die Planung der Evaluation ein. Ein weiterer Aspekt ist die Planung des Umfangs der Evaluationsstudie, der im Falle von geringer Kooperationswilligkeit zu berücksichtigen ist. Das Evaluationsdesign sollte so schlank wie möglich gehalten werden, der Datenerhebungsplan übersichtlich ausfallen, indem Datenerhebungen in zeitlicher Hinsicht auf wenige Zeitpunkte reduziert und Instrumente gebündelt angewendet werden.

In den folgenden drei Unterkapiteln wird die Durchführung von Evaluationsstudien anhand von Praxisbeispielen erläutert. Es wird zunächst auf das Evaluationsdesign eingegangen, bevor anschließend die zentralen Ergebnisse der Studien vorgestellt und diskutiert werden. Aus der metaanalytischen Betrachtung der Wirkungsanalyse zum Programm *frühstart* soll hervorgehen, welche methodischen Aspekte sich bei der Umsetzung der Evaluationsdesigns bewährt haben und welche Herausforderungen sich im Evaluationsverlauf ergeben haben. Aus

den gewonnenen Erkenntnissen lassen sich wiederum Ansatzpunkte für die Ausarbeitung des Leitfadens für die Durchführung von Evaluationsprojekten ableiten, der im Schlusskapitel vorgestellt wird.

Bei den vorgestellten Ergebnissen handelt es sich um Beispiele für einzelne der oben genannten fünf Zwecke der Datenerhebung darstellen. Um die Bandbreite der Evaluationsmöglichkeiten bei Programmevaluationen zu illustrieren, handelt es sich zuerst um die Darstellung einer programmbegleitenden Evaluation, an die sich eine standardisierte Sprachstandsmessung anschließt. Die Beispiele für die Durchführung enden mit einer detaillierteren Beschreibung einer Wirkungsmessung sowie einer kritischen Betrachtung des angewendeten Evaluationsdesigns.

6.2. Durchführungsbeispiel 1: Weiterentwicklung des Integrationskurses „In Deutschland zu Hause“

Die Evaluation der Integrationskurse mit dem Titel „In Deutschland zu Hause“ verdeutlicht, zu welchen Ergebnissen die programmbegleitende Evaluation kam und wie das Kurskonzept **kontinuierlich weiterentwickelt** wurde. Das Durchführungsbeispiel 1 steht stellvertretend für ein Evaluationsprojekt, bei dem die **Weiterentwicklung des Evaluierungsobjektes** (hier die Kurse) im Vordergrund stand. Das Evaluationsziel war es, das junge Konzept der Kurse auf Basis der Evaluationsergebnisse weiterzuentwickeln. Daher kamen hauptsächlich Teilnehmerfeedbackorientierte Methoden zum Einsatz, wie z.B. kurze Fragebögen und Formen der teilnehmenden Beobachtung in den Kursen. Im Sinne des responsiven Evaluationsansatzes von Stake (1975) wurden das Design und Vorgehen mit den Programmverantwortlichen der Stadt Nürnberg abgesprochen. Der Evaluationsansatz zeigt zudem wie auf kreative Art und Weise Evaluationsmethoden eingesetzt werden und wie Dozenten und Teilnehmer in den Evaluationsprozess eingebunden werden können.

Im Folgenden werden die eingesetzten Methoden und die zentralen Ergebnisse des formativen Evaluationsprozesses vorgestellt, der sich über eine Dauer von 1,5 Jahren (Oktober 2001 bis Mai 2003) erstreckte. Seit Beginn der ersten Kursreihe im Oktober 2001 war das fertige Curriculum thematisch in zwei Teile getrennt. Der erste Teil behandelte schwerpunktmäßig die Themen Geschichte, Landeskunde, Migration und Demographie. Im zweiten Teil wird der Fokus verstärkt auf die Vermittlung „demokratischer Werte“ gelegt. Die inhaltliche Gliederung der Kursreihe stellt sich folgendermaßen dar:

Kapitel 1: Deutschland und die Migration	Kapitel 6: Der Rechtsstaat
Kapitel 2: Landeskundliche Einführung	Kapitel 7: Der Sozialstaat
Kapitel 3: Deutsche Geschichte: von 1914 bis 1945	Kapitel 8: Staatsbürgerschaft als Mitgliedschaft/Möglichkeiten politischer Mitwirkung
Kapitel 4: Deutsche Geschichte: Deutschland seit dem zweiten Weltkrieg	Kapitel 9: Das deutsche Wahlsystem
Kapitel 5: Grund und Menschenrechte/Rechte und Pflichten	

Tabelle 6: Kapitel des Teilnehmerskripts

6.2.1. Angewandte Evaluationsmethoden

Das Evaluationsdesign und der Erhebungsplan umfassen eine Reihe von Evaluationsinstrumenten. Es sollten sowohl die einzelnen Kurseinheiten als auch die Kursmaterialien nach mehreren Durchläufen verbessert werden. Im Verlauf der Kurse wurden fortlaufend Informationen gesammelt und ausgewertet. Das Evaluationsdesign umfasste hauptsächlich den Einsatz von Feedbackbefragungen und Beobachtungsprotokolle.

a) Feedbackbefragungen zum Verlauf der Kurse

Die Teilnehmer der ersten beiden Abendkurse im Wintersemester 2001/02 wurden zu bestimmten Merkmalen der Unterrichtsgestaltung befragt. Für diese Art der Befragung wurden so genannte Kurzfeedbackfragebögen erstellt, die jeweils am Ende einer Kurseinheit zum Einsatz kamen. Die Fragebögen bestanden aus vier Fragen, die sich auf die generelle Evaluation der Unterrichtsgestaltung und Dozentenbewertung konzentrierten. Alle Teilnehmer konnten den Grad ihrer Zustimmung angeben, inwieweit

- die behandelten Themen interessant waren,
- die Erklärungen des Dozenten verständlich waren,
- der Unterricht im Allgemeinen und die Sprache der gelesenen Texte einfach oder schwer waren.

Das Instrument der Kurzfeedbackfragebögen erlaubt direktes Eingreifen und Variation der Unterrichtsgestaltung ohne eine weitere Befragung der Teilnehmer und war dadurch ein fester Bestandteil der prozessorientierten intervenierenden Evaluation. Auf Basis der Ergebnisse konnten nach jeder Kurseinheit inhaltliche Anpassungen am Konzept vorgenommen werden[43].

b) Beobachtungsprotokolle

Durch Kurzfeedbackfragebögen sind unmittelbare Bewertungen einzelner Kurseinheiten möglich, wodurch eine schnelle Intervention bei der Unterrichtsgestaltung, der Auswahl eines Gastdozenten oder bei bestimmten Inhalten möglich ist. Feedbackfragebögen liefern aber keinen Einblick in den Unterrichtsverlauf, da sie keine Informationen über die Kommunikationssituation zwischen den Teilnehmern und dem Kursleiter widerspiegeln. Gerade aus der Art und Weise, wie der Dozent anhand didaktischer Mittel den Stoff Kursteilnehmern präsentiert und mit ihnen diskutiert, lassen sich Schlussfolgerungen für eine angemessene

[43] So wurde im Kapitel eins mehr Zeit für die Behandlung des Themenblocks „Migration“ eingeräumt. Kapitel sechs und sieben wurden mit deutlich mehr Praxisbeispielen als zunächst geplant ergänzt.

didaktische Gestaltung der Kurse treffen. Durch Beobachtungsprotokolle gewinnt man ein differenzierteres Bild von der gesamten Kurssituation.

Die Beobachtungsprotokolle beschreiben einen Zeitraum von einem Jahr (Sommersemester 2002 bis Sommersemester 2003), in dem der Verlauf von drei Abendkursen und einem Wochenendkurs in insgesamt 28 Protokollen erfasst wurde. Dabei erwies es sich als sehr nützlich, unmittelbar nach der Beobachtung die gewonnenen Erkenntnisse und Eindrücke aus der sozialen Situation niederzuschreiben. Die Protokolle wurden anschließend nach den Kriterien des Beobachtungsrasters ausgewertet. Bei der Erstellung der Beobachtungsprotokolle wurde das folgende Raster für die zu beobachtenden Kriterien nach Lamnek festgelegt (Lamnek 1995, S. 299f.):

Die Teilnehmer:

- Anzahl der Teilnehmer
- aktiv bzw. passiv am Kurs Teilnehmende
- Grad der Teilnahme bzw. Nichtteilnahme

Der Dozent:

- Stil des Dozenten (Vortragstil vs. Teilnehmer aktivierend)
- angewandte didaktische Methoden
- Abwechslung der Didaktik

Lernerfolge:

- Angemessenheit der didaktischen Methoden
- spontane Lernprozesse

Das Ziel der Beobachtung war die Erfassung der angewandten didaktischen Methoden und der Interaktion zwischen Teilnehmer und Kursleiter. Anhand der Beobachtungsprotokolle lassen sich Variation und Angemessenheit der Lernformen im Kursverlauf darstellen. Die Ereignisse im Verlauf einer Kurseinheit wurden im nächsten Schritt den im Beobachtungsraster definierten Kriterien zugeordnet. Die Auswertung aller Beobachtungsprotokolle bildete die Ausgangslage für die Entwicklung methodisch-didaktischer Gestaltungskriterien.

Aus den Erfahrungen und Beobachtungen der Kurse ergaben sich Anregungen für den methodisch-didaktischen Aufbau der Integrationskurse. Die Präsentation der Inhalte sowie die Reaktion der Teilnehmer wurden in den Beobachtungsprotokollen strukturiert festgehalten.

c) Allgemeine Kursbewertung seitens der Teilnehmer mittels offener Fragen

Generell kann bei der Analyse von offenen Antworten zwischen quantitativen und qualitativen Aspekten unterschieden werden. Die Auswertung zu den quantitativen Aspekten „Anzahl der Anmerkungen" und „Umfang der Anmerkungen" wird als erstes vorgestellt. Im Folgenden werden (zur besseren Veranschaulichung) die drei gestellten Fragen auf Kurzbezeichnungen reduziert:

- „Was war gut in der Veranstaltung?" wird als „Positive Kommentare" bezeichnet.
- „Was hat Ihnen an der Veranstaltung nicht gefallen?" wird als „Kritik" bezeichnet.
- „Was sollte Ihrer Meinung nach in der Veranstaltung geändert werden?" wird kurz als „Verbesserungsvorschläge" bezeichnet.

Die deutliche Betonung der „Positiven Kritik" (78 Nennungen) bestätigt die Ergebnisse der zuvor beschriebenen geschlossenen Fragen zur allgemeinen Kurskritik (siehe Abbildung 8). Bei der qualitativen Analyse der drei offenen Fragen zur Kurskritik ließen sich die Antworten zu Grob- und Unterkategorien zusammenfassen. Dabei wurden Statements, die dem Inhalt nach identisch waren, nur einmal wiedergegeben und durch eine entsprechende Häufigkeitsangabe in Klammern gekennzeichnet. Sehr lange Statements wurden zusammengefasst.

Die Kategorie „Positives" zeichnet, aufgrund der meisten abgegebenen Kommentare, das differenzierteste Bild der Antwortstruktur. An dieser Stelle lassen sich die Antworten den Unterkategorien „Allgemein positive Kursbewertungen", „Dozentenleistung", „Teilnehmerselbstbewertung", „Unterricht" und „Inhalte/Themen" zuordnen. Bei der Kategorien „Kritik" und „Verbesserungsvorschläge" ließen sich zwei Grobkategorien unterscheiden: „Ablauf und Organisation der Kurse" und „Unterricht". Auf den folgenden zwei Seiten sind die Ergebnisse der offenen Fragen in Abbildungen dargestellt. In den Abbildungen werden die Bewertungen der Kurse durch die Zuordnung der Antworten in Über- bzw. Unterkategorien ersichtlich.

Die Auswertung der Statements zu den offenen Fragen ergibt eine ausdrücklich positive Beurteilung der Kurse seitens der Teilnehmer. Besonders zufrieden zeigten sich die Teilnehmer mit der Themenauswahl und der Leistung des Dozenten, die Inhalte verständlich zu erklären („Unser Dozent war sehr gut", „Offenheit des Dozenten").

Bei der negativen Kurskritik fällt zunächst auf, dass die Teilnehmer diese Frage mit einer verneinenden Ansicht beantworteten, d.h. an dieser Stelle ausdrücklich bestätigten, dass ihnen der Kurs insgesamt sehr gut gefallen hat („Alles war in Ordnung"). Dagegen wurde die Kategorie Unterricht und Themen am häufigsten

mit negativen Statements belastet. An dieser Stelle standen konkrete Themen, die Didaktik oder der allgemeine Zeitmangel im Mittelpunkt der Kritik. Als Verbesserungsvorschläge wurden häufigere Diskussionen und mehr Anschauungsmaterial verlangt. Die Weiterentwicklung des Kurskonzepts wurde außerdem in der Programmplanung und hinsichtlich des Prinzips des Anschlusslernens vorangetrieben.

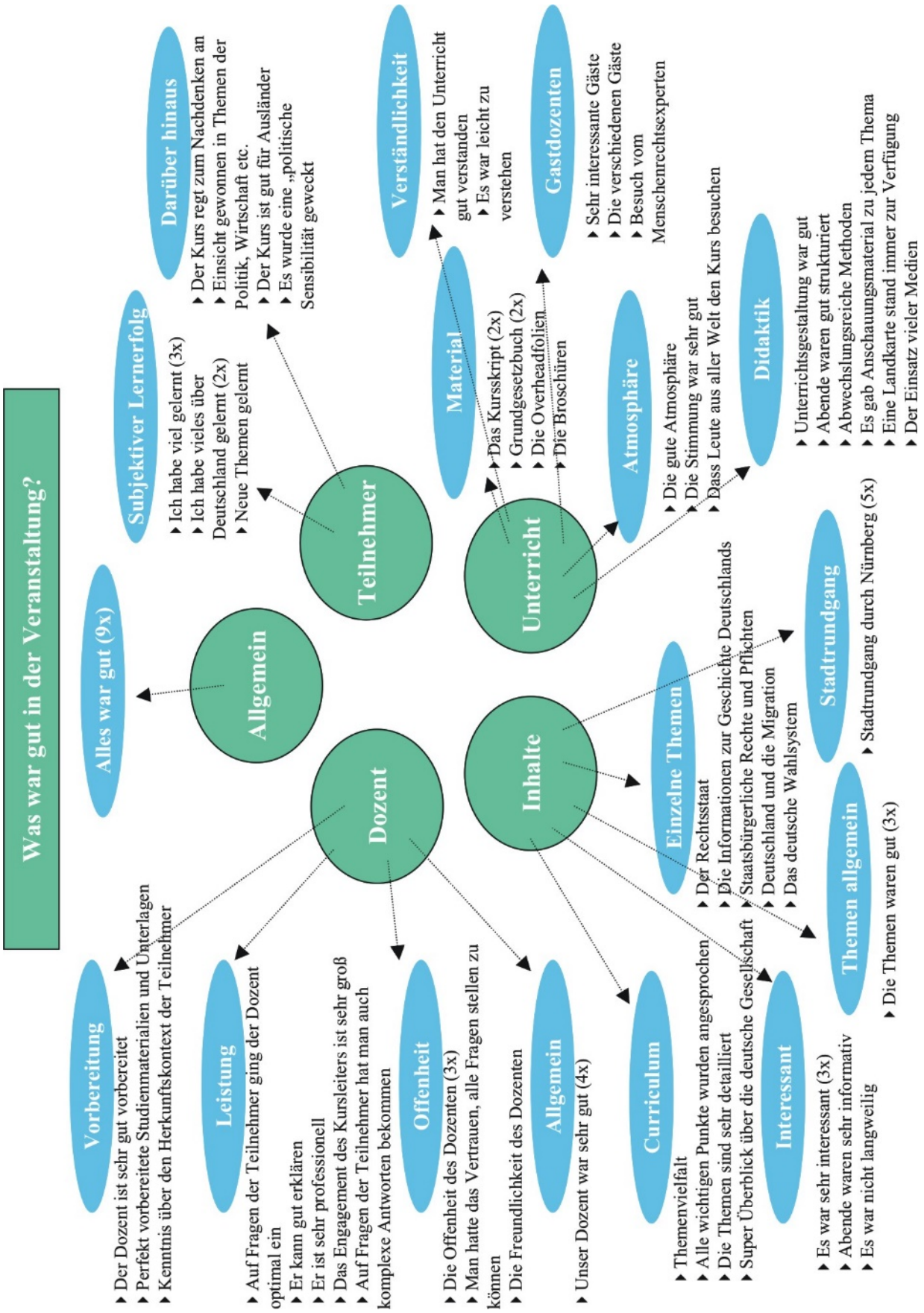


Abbildung 8: Positive Bewertungen durch die Kursteilnehmer

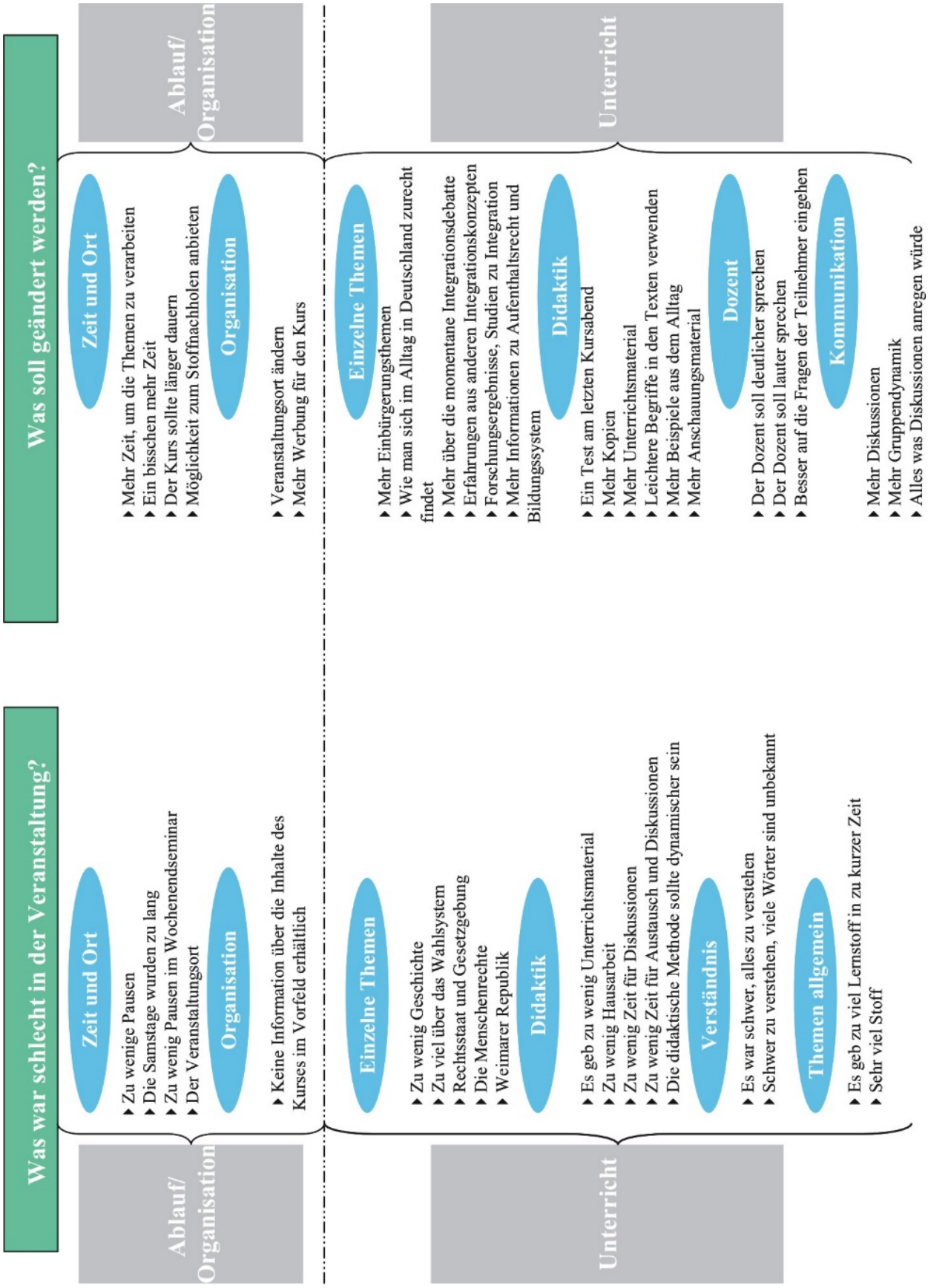


Abbildung 9: Negative Bewertungen durch die Teilnehmer und Verbesserungsvorschläge

6.2.2. Didaktische Grundprinzipen

Das Evaluationsprojekt hatte – wie bereits berichtet – den Zweck, die Kurse inhaltlich weiterzuentwickeln. Bei der Weiterentwicklung wurde seitens der Kursleitung Wert auf zwei didaktische Aspekte Wert gelegt: die vorausschauende Programmplanung sowie die Anwendung des Prinzips des Anschlusslernens.

Vorausschauende Programmplanung bedeutet in dem hier vorgestellten Zusammenhang, dass den individuellen Voraussetzungen, Lerninteressen und Erwartungen der zukünftigen Teilnehmer bei der Entwicklung des Kurskonzepts Rechnung getragen wird. In der Praxis der Erwachsenenbildung wird durch die Anwendung dieses Prinzips versucht, „sozialbiographische Daten, kognitive Strukturen und Erwartungshaltungen" der Teilnehmer bei der Kursgestaltung zu berücksichtigen (Tietgens 1984, S. 446). Mit der Bedeutung des Begriffs vorausschauender Programmplanung ist daher das Prinzip der Teilnehmerorientierung eng verbunden. Durch die Erhebung von Teilnehmerstrukturdaten im Vorfeld der Kurse kann der Unterrichtsverlauf vom Kursleiter und den Organisatoren auf die speziellen Anforderungen der Teilnehmer maßgeschneidert werden. Die vorausschauende Programmplanung bietet im Idealfall den Ausgangspunkt für die weitere didaktische Konzeption der Kurse. Bei den Nürnberger Kursen wurde im Rahmen der vorausschauenden Programmplanung das Konzept in der Vorbereitungsphase auf die Gruppe der Einbürgerungswilligen bzw. Einbürgerungskandidaten angepasst.

Die Erfahrung im Projekt hat gezeigt, dass eine Bedarfsanalyse bei ähnlichen Kursen unabdingbar ist. Für die erfolgreiche Positionierung eines ähnlich freiwilligen Kursangebots in der Zukunft wird die Durchführung einer lokalen Bedarfsanalyse vorgeschlagen, um eine Mindestgröße der Zielgruppe für die Durchführung der Kurse sicherstellen zu können. Die zentralen Fragen, die im Zusammenhang mit der Bedarfsanalyse beantwortet werden müssen, sind: Ist das konzipierte Angebot am Bedarf orientiert? In welcher Weise müssen die Inhalte der Kurse an die Bedürfnisse der Teilnehmer angepasst werden? Zur Klärung dieser Fragen muss eine Analyse der Heterogenität der potentiellen Teilnehmer nach den folgenden Aspekten erfolgen:

- Alter
- Geschlecht
- Sprachkenntnisse
- Bildungsstand

Hier bietet es sich an, statistische Informationen über die Zielgruppe der Kurse vom Statistischen Amt der Stadt (oder einer Abteilung der öffentlichen Verwal-

tung mit ähnlicher Funktion) zu beziehen. Generell ist eine Sammlung von Informationen über bestehende Maßnahmen im Bereich Erwachsenenbildung und ihrer Teilnehmerstruktur zur Planung des eigenen Angebots nützlich. Aus Gesprächen mit Vertretern von Migrantenorganisationen sowie Verbänden können zusätzlich Informationen über den zu erwartenden Teilnehmerkreis bezogen werden.

Unter dem **Prinzip des Anschlusslernens** ist die Rückanbindung neuer Informationen an frühere Erfahrungen und vorhandene kognitive Muster und Einstellungen zu verstehen. Bei den Integrationskursen versuchte der Kursleiter, situationsbezogen an kognitive Strukturen der Teilnehmer anzuknüpfen, indem Vergleiche zu der Situation im Herkunftsland hergestellt wurden. Durch Vergleiche der deutschen Gesellschaftsstrukturen mit denen im Herkunftsland werden Unterschiede aber auch Gemeinsamkeiten deutlich: Die Teilnehmer können ihre Erfahrungen und Kenntnisse aktiv in das Kursgeschehen einbringen. Bei den Integrationskursen wurde dies am deutlichsten im Kapitel Menschenrechte. Teilnehmer aus afrikanischen Staaten berichteten meist eindringlich von ihren Erlebnissen.

6.2.3. Fazit zur angewendeten Evaluationsmethode

Die beschriebene Evaluation der Integrationskurse einen nach der Begriffsdefinition von Scriven stark **formativ ausgerichteten Evaluationsverlauf**. Dies bedeutet, dass in der Pilotphase nach jeder Kurseinheit die Feedbackergebnisse direkt zur Planung und Verbesserung der nächstfolgenden Kurseinheit herangezogen wurden. Diese „Luxusevaluation“ hatte zur Folge, dass bereits nach der Durchführung einer Kursreihe ausreichend Hinweise vorlagen, um die Kursmaterialien und Lehrformen zu überarbeiten. Die aus den Feedbackbefragungen gewonnenen Erkenntnisse konnten für die kontinuierliche Weiterentwicklung des Kurskonzepts genutzt werden. Die Erhebungsinstrumente wurden gezielt auf die spezifischen Erkenntnisinteressen ausgesucht und entwickelt. Auf diese Weise konnten simultan mehrere Methoden (z.B. teilnehmende Beobachtung, standardisierte Befragung der Kursteilnehmer) eingesetzt werden. An dessen Ende stand ein ausgereiftes Kurskonzept inklusive Kursmaterialien zur Verfügung, das zwei Jahre später in die Überlegungen zur Gestaltung von Orientierungskursen im Rahmen des Deutschkursprogramms der Stadt Nürnberg herangezogen wurde.

Die in diesem Beispiel gelebte Flexibilität im Einsatz von Evaluationsmethoden kann im Prinzip nur bei Evaluationsstudien mit dem **Ziel einer Weiterentwicklung der Inhalte des Programms** angewendet werden. Vorausgesetzt, die Erhebungsinstrumente werden unter Wahrung der Methoden und Standards der empirischen Sozialforschung konstruiert und angewendet, bleibt es dem Evaluator und den Programmverantwortlichen schließlich überlassen, wie die Ergebnisse

interpretiert und weiterverwendet werden. Dagegen ist bei den folgenden Beispielen der Handlungsspielraum, was die Auswahl und Gestaltung des Untersuchungsdesigns betrifft, stärker eingeschränkt. In beiden Beispielen sollen die Sprachkompetenzen von Kindern in spezifischen und abgrenzbaren Sachverhalten erfasst werden. Um dies zu erreichen, kommen nur wenige Untersuchungsanordnungen in Frage.

6.3. Durchführungsbeispiel 2: Ergebnisse einer Pretest-Posttest-Untersuchung zur phonologischen Bewusstheit von Kindern im Kindergartenalter im Programm „Spielend lernen“

Bei der nachfolgend vorgestellten Methode handelt es sich um eine Pretest-Posttest-Untersuchung von Sprachkenntnissen bei Kindern im Vorschulalter in einem bestimmten Bereich der Sprachentwicklung, der Phonologie. Es handelt sich um eine von mehreren kleinen Evaluationsuntersuchungen, die im Rahmen der Evaluationsstudie zu dem Programm „Spielend lernen“ in Nürnberg durchgeführt wurde. Die hier angewendete Methode der Pretest-Posttest-Untersuchung ist vergleichbar mit dem Design von quasi-experimentellen oder experimentellen Studien, unterscheidet sich aber hauptsächlich dadurch, dass keine Kontrollgruppe gebildet wird (vgl. Campbell 1988, S. 152). Anhand dieser Methode kann auf strukturierende Art und Weise die Entwicklung der phonologischen Fähigkeiten von Kindern im Vorschulalter aufgezeigt werden. Das hier angewandte Verfahren der Pretest-Posttest-Untersuchung eignet sich nicht – im Gegensatz zum noch folgenden Durchführungsbeispiel 3 –, um Wirkungen einer sozialen Intervention, von externen Störfaktoren bereinigt, eindeutig zu quantifizieren. Das Verfahren kann aber Anwendung finden, wenn die Wirkungen einer sozialen Maßnahme bereits in einer zu einem früheren Zeitpunkt durchgeführten Evaluation eindeutig festgestellt wurden und das Programm unter ansonsten unveränderten Bedingungen weiterhin angeboten wird. In diesem Fall kann die Pretest-Posttest-Untersuchung eine **Kontrollfunktion** übernehmen.

Ein Ziel von „Spielend lernen“ war es, bereits erprobte Projekte in den Stadtteilen in die Breite zu bringen. Zu diesen Projekten gehört auch „Phono-Logisch – Hand in Hand“, eines von vier Projekten im Programm „Sprachförderung in Kindertagesstätten“ (SpiKi) des Jugendamts der Stadt Nürnberg. Durch „Phono-Logisch – Hand in Hand“ wird in Kindertagestätten und Kindergärten spielerisch in der täglichen pädagogischen Arbeit und in speziellen Fördergruppen die phonologische Bewusstheit trainiert[44]. Unter phonologischer Bewusstheit versteht man in sprachwissenschaftlicher Hinsicht die Fähigkeit, bei der Aufnahme, der Verarbeitung, dem Abruf und der Speicherung von sprachlichen Informationen

[44] Das Programm „Phono-Logisch – Hand in Hand“ ist eine Weiterentwicklung des Würzburger Trainingprogramms zur phonologischen Bewusstheit „Hören, lauschen, lernen“. Im Herbst 2002 wurde eine Erprobungsphase mit 16 Kindergärten des Jugendamtes in Nürnberg durchgeführt, um erste Erfahrungen mit der Wirkungsweise des Würzburger Trainingsprogramms zu sammeln.

Wissen über die lautliche Struktur der Sprache heranzuziehen (vgl. Wagner und Torgesen 1987)[45].

Ergebnisse der Grundschulforschung im deutschsprachigen Raum deuten in den letzten Jahren darauf hin, dass phonologische Bewusstheit bei Kindern eine Kernvoraussetzung für erfolgreichen Schriftspracherwerb darstellt (Einsiedler & Kirschhock 2003). Umgekehrt bedeutet dies, dass Kinder mit geringer phonologischer Fähigkeit mit hoher Wahrscheinlichkeit gefährdet sind, in der Grundschule Lese- und Schreibprobleme zu entwickeln (Einsiedler, Helbig & Treinies 2002). Um die Fortschritte in der Herausbildung der phonologischen Bewusstheit zu überprüfen, wurde das Erhebungsverfahren zur phonologischen Bewusstheit (ARS) entwickelt. ARS ist die Abkürzung für die Aufgaben, die Kinder während des Testverfahrens ausführen: Anlaute hören; Reime finden; Silben klatschen[46].

Das ARS-Verfahren sieht vor, Kinder zu Beginn des letzten Kindergartenjahres mit dem Erhebungsverfahren zu überprüfen. Aufgrund der Ergebnisse von ARS und weitergehenden pädagogischen Überlegungen werden Fördermaßnahmen für Kinder mit mangelnder phonologischer Bewusstheit in den Einrichtungen erarbeitet. Das ARS-Verfahren wird dann ein zweites Mal bei den überprüften Kindern zu Schulanfang eingesetzt. Durch diese Vorgehensweise und einem Vergleich zwischen beiden Erhebungen wird die Entwicklung der phonologischen Bewusstheit im letzten Kindergartenjahr individuell bei jedem überprüften Kind deutlich[47]. 2005 führte das Jugendamt in Koordination mit den Stadtteilkoordinatorinnen weitere Schulungen im Umgang mit dem Verfahren besonders für Einrichtungen mit freier Trägerstruktur durch.

[45] Es wird zwischen der phonologischen Bewusstheit im weiteren und engeren Sinne differenziert. Phonologische Bewusstheit im „weiteren Sinne" bezeichnet die Fähigkeit, zwischen Wörtern und Silben zu unterscheiden und Reime wahrzunehmen. Mit im „engeren Sinne" wird die Fähigkeit verstanden, konkrete Laute in Wörtern und Silben zu erkennen und zu unterscheiden. Die letztgenannte Unterscheidungsfähigkeit wird zumeist durch Erfahrungen im Rahmen des Schriftspracherwerbs erworben.

[46] Das ARS-Verfahren hat sich als ein zeitlich effektives Diagnoseinstrument herausgestellt, das von den ErzieherInnen in den Einrichtungen problemlos umgesetzt werden kann. Das Verfahren wurde im Jahr 2002 vom Jugendamt und dem Staatlichen Schulamt der Stadt Nürnberg in Kooperation mit der Erziehungswissenschaftlichen Fakultät der Universität Erlangen/Nürnberg (EWF), der Universität Koblenz-Landau sowie der Universität Passau entwickelt. Martschinke, S.; Kammermeyer, G.; Pickelein, M.; Forster, M.: Anlaute abhören, Reime finden, Silben klatschen (ARS) - Erhebungsverfahren zur phonologischer Bewusstheit. Donauwörth: Auer 2004.

[47] Die phonologische Bewusstheit kann auch bei Kindern mit anderer Erstsprache als Deutsch mit dem ARS-Verfahren überprüft werden. Hierzu wurde eine Hör-CD entwickelt, auf der

6.3.1. Ergebnisse der Pretest-Untersuchung von Kindern in „Spielend lernen"-Stadtteilen

Es sollten mit einer Pretest-Posttest-Untersuchung Erkenntnisse zu der Frage zusammengetragen werden, ob und wenn ja, wie stark sich die phonologischen Fähigkeiten der Kinder verändert haben. Zunächst startete die Vorbereitung der Pretest-Untersuchung. Im Rahmen der Evaluation von „Spielend lernen" wurden im Sommer 2005 Kindergärten und Kindertagesstätten in den Stadtteilen St. Leonhard/Schweinau und Langwasser in Informationsveranstaltungen motiviert, das ARS-Verfahren in ihren Einrichtungen zur Überprüfung der phonologischen Bewusstheit bei den Kindern anzuwenden.

Das efms hat die Ergebnisse der Überprüfung mit dem ARS-Verfahren von 10 Kindergärten bzw. Kindertagesstätten im Herbst 2005 zugeschickt bekommen. Die für die Auswertung des ARS-Verfahrens genutzten Einrichtungen können daher auch als Stichprobe aus allen relevanten Einrichtungen in den Stadtteilen gesehen werden[48]. Insgesamt konnten 129 Kinder in die Auswertung der Ersterhebung einbezogen werden. Es handelte sich dabei bei allen Kindern um die Ersterhebung ein Jahr vor Schulbeginn. 4 Einrichtungen lagen geographisch im Stadtteil St. Leonhard/ Schweinau und 5 Einrichtungen im Stadtteil Langwasser. Die Zusammensetzung der Einrichtungen hinsichtlich der Trägerstruktur war ausgeglichen: bei 6 von 10 Kindergärten bzw. Kindertagesstätten handelte es sich um städtische Einrichtungen[49].

Alle untersuchten Einrichtungen wiesen sich durch einen relativ hohen Anteil von Kindern mit Migrationshintergrund aus – die Spannweite reichte von 48% bis zu 86% Anteil in einem Kindergarten. Dementsprechend hoch war auch der Anteil von Kindern mit Migrationshintergrund, die im Herbst 2005 mit dem ARS-Verfahren getestet wurden.

Das Erhebungsverfahren war ein Einzeltestverfahren, für das pro Kind etwa zwischen 10 und 20 Minuten eingeplant werden sollte. Die Überprüfung der phono-

alle Anweisungen von Muttersprachlern übersetzt und gesprochen werden. Das ARS-Verfahren kann dadurch in 10 verschiedenen Sprachen durchgeführt werden.

[48] Bei der Auswahl handelt es sich jedoch nicht um eine Zufallsstichprobe, sie ist daher keinesfalls statistisch repräsentativ für die Gesamtsituation in den Stadtteilen. Vielmehr bieten die Auswertungsergebnisse einen Eindruck über die Ausprägung der phonologischen Bewusstheit bei Kindern im Vorschulalter.

[49] Die untersuchten Einrichtungen variierten sehr stark in Bezug auf die Anzahl der Kinder, dem Anteil an Kindern mit Migrationshintergrund und der Anzahl mit dem ARS-Verfahren überprüften Kinder. Die größte Einrichtung betreute 105 Kindern in 5 Kindergruppen, die kleinste dagegen 23 Kinder in nur einer Kindergruppe.

logischen Bewusstheit wurde durch eine Erzieherin in der Einrichtung durchgeführt[50]. Wurden bei der Überprüfung 10 oder weniger Punkte erreicht, war das Kind als so genanntes „Risikokind“ einzustufen. Ein Risikokind zeichnet sich durch eine geringere phonologische Bewusstheit aus.

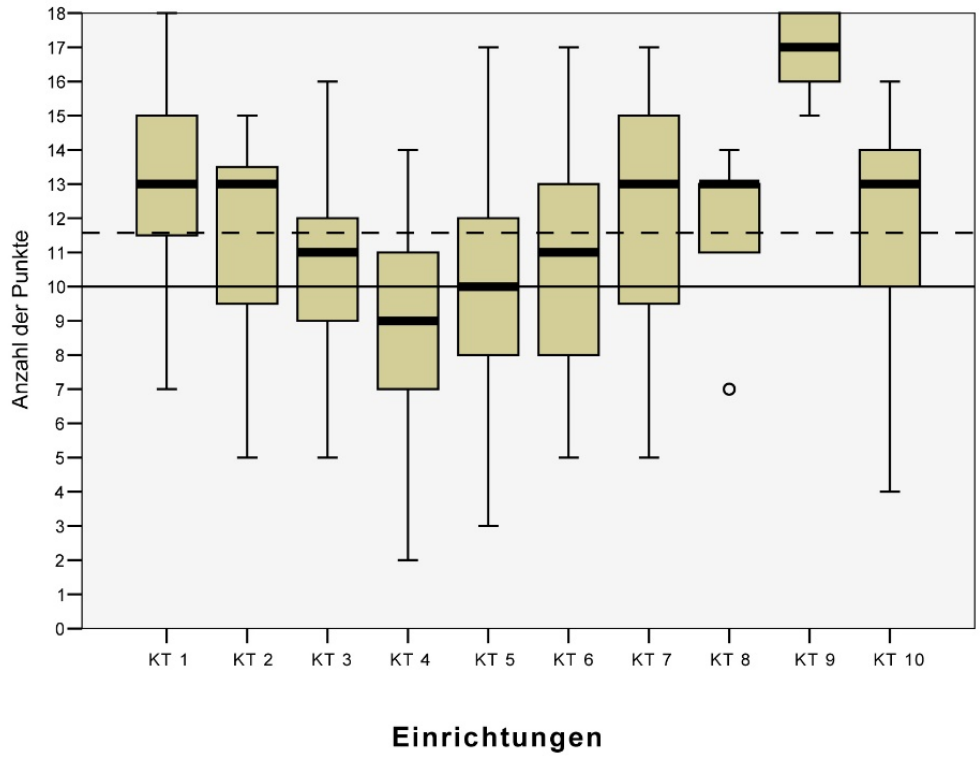


Abbildung 10: Ergebnisse des ARS-Verfahrens nach einzelnen Einrichtungen (KT = Kindertagesstätte)

In der obigen Abbildung sind die Ergebnisse des ARS-Verfahrens in Form von so genannten Boxplots dargestellt[51]. Anhand von Boxplotdarstellungen wird die Verteilung der Werte in der Gesamtgruppe deutlich. Jeder Boxplot beschreibt die Ergebnisse einer Kindergarten- bzw. Kindertageseinrichtung[52]. Zudem ist in der Abbildung der Mittelwert aller ARS-Tests (11,5) als gestrichelte Linie und der Schwellenwert 10 als durchgezogene Linie dargestellt. Kinder im Vorschulalter

[50] Das Prüfverfahren umfasste drei Bereiche: Silben klatschen, Anlaute hören und Reime finden. In jedem Prüfbereich waren sechs Aufgaben vorgegeben. Jede richtig beantwortete Frage wurde mit einem Punkt bewertet, so dass insgesamt 18 Punkte erreicht werden konnten.

[51] Ein dicker horizontaler Strich in jeder Box repräsentiert den Median. Die Verteilung aller Werte wird durch den Median in der Mitte halbiert. Die Box umschließt die mittleren 50% der erfassten Kinder einer Einrichtung. Der Winker nach oben zeigt die 25% der Kinder an, die am besten abgeschnitten haben. Der Winker nach unten geht bis zum Minimalwert. Kreise repräsentieren einen Extrem- bzw. Minimalwert.

[52] Es handelt sich um die Addition der gesammelten Punkte.

müssen mindestens den Wert 11 erreichen, um nicht als Risikokind eingestuft zu werden.

Bei der Betrachtung der Abbildung wird deutlich, dass es sichtbare Unterschiede in der Verteilung der Werte zwischen den verschiedenen Einrichtungen gibt. Es lassen sich zwei Einrichtungen nennen, in denen bis auf eine Ausnahme keine Risikokinder durch das ARS-Verfahren identifiziert werden konnten. Die Verteilung der Werte zeigt weiterhin, dass die Unterschiede zwischen Einrichtungen recht groß ausfallen können. In einem Kindergarten wurden mit dem ARS-Verfahren über 50% Risikokinder identifiziert.

6.3.2. Ergebnisse der Posttest-Untersuchung von Kindern in den „Spielend lernen"-Stadtteilen

Genau ein Jahr später wurde die Untersuchung mit dem gleichen Instrument in denselben Kindergärten unter der Teilnahme der nun ein Jahr älteren Kinder wiederholt. Im Rahmen der Zweiterhebung der Evaluation konnten 106 Kinder mit dem ARS-Verfahren überprüft werden, zu denen bereits Daten aus der Ersterhebung vorlagen. Insgesamt neun Einrichtungen haben die Ergebnisse der ARS-Untersuchung der Evaluation auch bei der Zweiterhebung zur Verfügung gestellt. Aufgrund der ausgeglichenen Beteiligung von Einrichtungen nach Stadtteilen ergab sich für Langwasser eine Gesamtzahl von 46 und für St. Leonhard/Schweinau von 50 überprüften Kindern. Von den insgesamt 106 Kindern waren 58 Jungen und 48 Mädchen. Bei der Zweituntersuchung war der Großteil der Kinder (etwa 80%) in einem Alter zwischen sechs und sieben Jahren. Dies entspricht auch der konzeptionell in „Phono-Logisch – Hand in Hand" vorgesehenen Zielgruppe der Kinder im letzten Kindergartenjahr. Die erlernte Erstsprache ist ein wichtiges Unterscheidungsmerkmal zur differenzierten Untersuchung der sprachlichen Entwicklung bei Kindern. Es wurden insgesamt 16 verschiedene Erstsprachen identifiziert, die von den Kindern der Zweituntersuchung gesprochen werden. Deutsch war die Erstsprache bei etwa der Hälfte aller überprüften Kinder, gefolgt von Türkisch und Russisch.

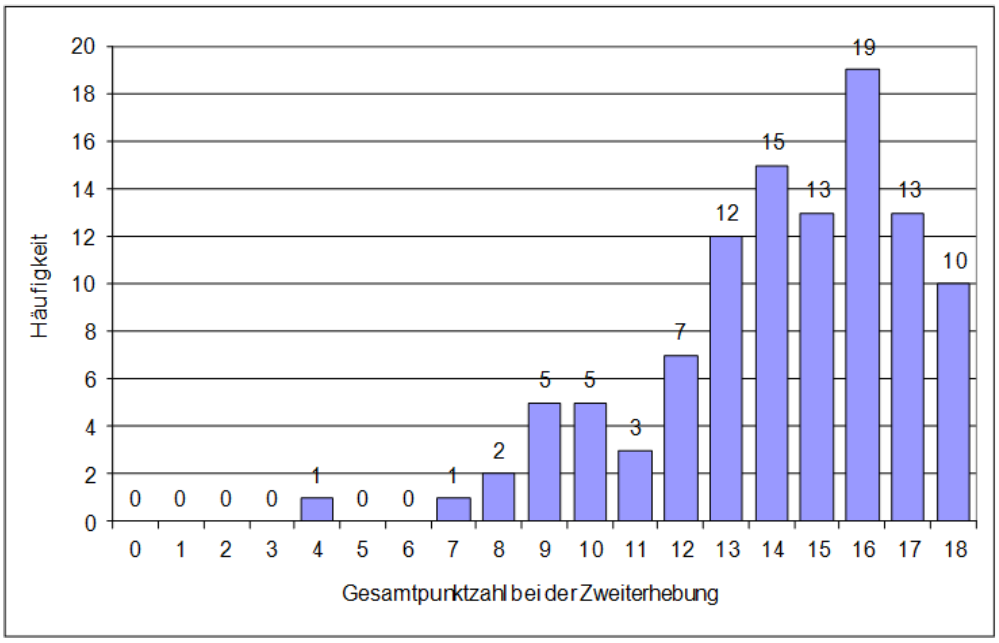


Abbildung 11: Häufigkeitsverteilung der erzielten Gesamtpunktezahlen in der Zweiterhebung (N=106)

Die Häufigkeitsverteilung macht deutlich, dass ein sehr großer Teil der überprüften Kinder zum Zeitpunkt der Zweiterhebung über eine ausreichende bis sehr gute phonologische Unterscheidungsfähigkeit verfügte. Statistisch mit dem Lagemaß Median ausgedrückt bedeutet dies, dass 50% der Kinder mindestens eine Gesamtpunktzahl von 15 erreicht haben. In die Kategorie „Risikokinder“ lassen sich den Ergebnissen zufolge noch 14 Kinder (etwa 13%) einordnen. Der Mittelwert aller erzielten Punkte liegt bei 14,2. Die Ergebnisse der Zweiterhebung lassen sich des Weiteren nach drei Subtests auswerten. Hier wird deutlich, dass es erhebliche Leistungsunterschiede hinsichtlich der drei zentralen Bereiche der phonologischen Bewusstheit gibt. Während das Klatschen von Silben kaum noch einem Kind Probleme bereitete (Mittelwert = 5,6), war es eher der Bereich Anlaute hören (Mittelwert = 3,8), der über die Höhe des Testergebnisses bei vielen überprüften Kindern entschied. In der folgenden Tabelle sind die Mittelwerte der Testergebnisse nach Erstsprachen dargestellt.

	Erstsprache		
	Deutsch	Russisch	Türkisch
Gesamtpunktzahl	14,9	13,6	12,6
Silben segmentieren	5,7	5,7	5,3
Anlaute hören	4,1	3,7	3,0

Reime bilden	5,2	4,2	4,3

Tabelle 7: Mittelwerte der Punktzahlen nach Erstsprache und Testbereich

Das Ergebnis zeigt, dass die Kinder mit Deutsch als Erstsprache am besten in den drei Testbereichen abgeschnitten hatten. Betrachtet man die Ergebnisse genauer, so werden große Unterschiede zwischen den untersuchten Gruppen erkennbar. Während Kinder mit deutscher Muttersprache in allen drei Bereichen hohe Durchschnittswerte erzielt hatten, waren es die Kinder mit anderen Erstsprachen, die mit niedrigeren Werten in den Bereichen Anlaute hören und Reime bilden die Gesamtdurchschnittswerte drücken. Hier fällt zudem der Bereich Anlaute hören auf, bei dem hauptsächlich Kinder mit Türkisch als Erstsprache Schwierigkeiten im Test hatten.

Eine weitere Möglichkeit der Analyse bestand für die Einrichtungsebene. In der folgenden Abbildung sind die Ergebnisse in Form von Boxplots dargestellt.

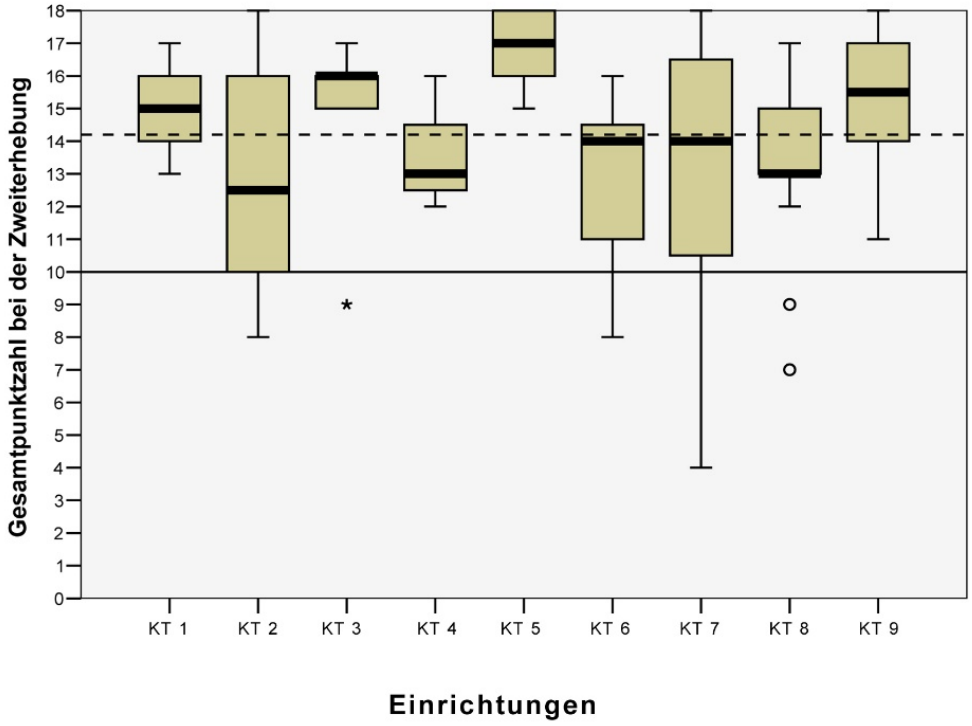


Abbildung 12: Ergebnisse des ARS-Verfahrens nach einzelnen Einrichtungen nach der Zweiterhebung (KT = Kindertageseinrichtung)

Wie schon bei den zuvor beschriebenen Ergebnissen verdeutlicht die Abbildung noch einmal, dass die Mehrheit der Kinder nicht in den Risikobereich unter einem Wert von 10 fällt. Im Gegensatz zur Ersterhebung ist außerdem festzuhalten, dass die Testergebnisse der Kinder eine geringe Streuung aufweisen.

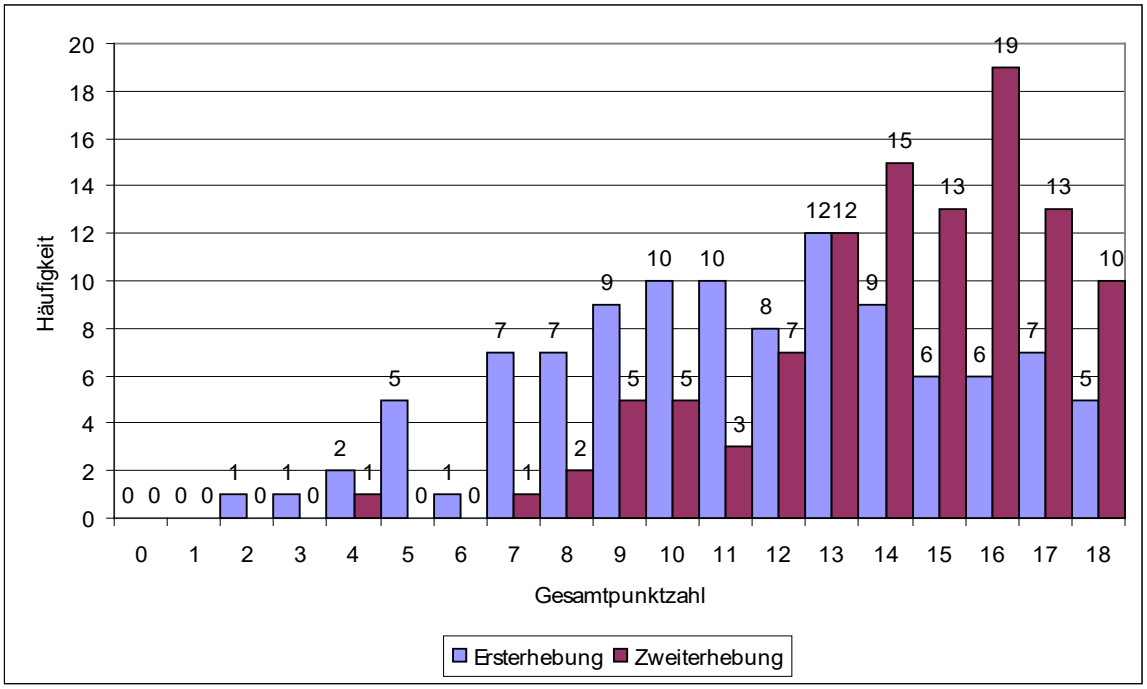


Abbildung 13: Die Verteilung der Testergebnisse bei Erst- und Zweiterhebung im Vergleich (N=106)

Entsprechend den Kriterien des ARS-Verfahrens verfügt die Mehrheit der Kinder über gute bis sehr gute phonologische Unterscheidungsfähigkeiten. Ein Vergleich beider Erhebungswellen zeigt in der Abbildung 13 deutliche Unterschiede hinsichtlich der erreichten Gesamtpunktzahl. Um zu überprüfen, ob sich die Ergebnisse der Erst- und Zweiterhebung statistisch signifikant voneinander unterscheiden, wurde der nicht-parametrische Wilcoxon-Test für verbundene Stichproben gerechnet[53]. Der Wilcoxon-Test wurde angewendet, um die Untersuchungsgruppe auf Unterschiede in der zentralen Tendenz zwischen der Erst- und Zweiterhebung zu überprüfen. Der Test setzt nicht zwingend eine Normalverteilung sowie ein metrisches Skalenniveau der erreichten Punktewerte voraus[54].

	Median Ersterhebung	Median Zweiterhebung	Z-Wert (Prüfwert)	Signifikanz (2-seitig)
Gesamtpunktzahl (t2-t1)	11,5	15	-6.825	p < .001
Anlaute hören (t2-t1)	3,0	4,0	-5.157	p < .001

[53] Näheres zu nicht-parametrischen Testverfahren wie dem Wilcoxon-Test und dem Mann Whitney U-Test ist im Kapitel 6.4. zur Evaluation der Sprachförderung im *frühstart*-Projekt erläutert.

[54] Bis auf die erreichte Gesamtpunktzahl bei der Ersterhebung (Z = .85, p = .460) zeigt der Kolmogorov-Smirnov-Anpassungstest für alle weiteren Punkteverteilungen wie z.B. Gesamtpunktzahl der Zweiterhebung (Z = 1.37, p = .048), Anlaute hören Ersterhebung (Z = 1.78, p = .004) keine annäherungsweise Normalverteilung.

Reime bilden (t2-t1)	4,0	5,0	-5.349	p < .001
Silben segmentieren (t2-t1)	5,0	6,0	-5.195	p < .001

Tabelle 8: Ergebnisse des Wilcoxon-Rangsummen-Tests für verbundene Stichproben (t2-t1, N=106)

Das Ergebnis des Wilcoxon-Tests ist vor dem Hintergrund der soeben präsentierten Daten wenig überraschend und auch eindeutig interpretierbar. Die Test-Ergebnisse untermauern die schon in Abbildung 13 erkennbaren deutlichen Unterschiede zwischen beiden Erhebungswellen. Hochsignifikant zeigt das Ergebnis in Tabelle 8, dass die überprüften Kinder zum Zeitpunkt der Zweiterhebung eine deutlich besser ausgeprägte phonologische Bewusstheit entwickelt haben. Diese Schlussfolgerung trifft nicht nur nach Analyse der Gesamtpunktzahl zu sondern auch für die drei einzelnen Testbereiche des ARS-Verfahrens.

Zwar fehlte zu Vergleichszwecken eine Kontrollgruppe, jedoch war es außerdem interessant zu erfahren, durch welche Merkmalskonfiguration sich Kinder auszeichnen, die bei der Zweiterhebung besonders gut abschnitten. Zu diesem Zweck wurde mit den Stichprobendaten eine mehrfaktorielle, univariate Varianzanalyse (ANOVA) gerechnet. Bei dem Verfahren der Varianzanalyse wird der Einfluss von mindestens zwei Faktoren (unabhängige Variablen) und bei Bedarf deren Wechselwirkung auf eine abhängige, zu erklärende Variable quantifiziert. Die erreichten Gesamtpunktzahlen bei der Zweiterhebung stellen die abhängige Variable dar. Die Faktoren (unabhängige Variablen) stellten in diesem Fall das Alter der Kinder, das Geschlecht, die Erstsprache sowie das Stadtteil der besuchten Kita dar[55]. Das Ergebnis zeigt, dass bis auf das Geschlecht der Kinder, keine der anderen Faktoren isoliert betrachtet einen signifikanten Einfluss auf das Ergebnis der Zweiterhebung hatte. Mädchen schneiden in der Stichprobe hinsichtlich der erreichten Gesamtpunktzahl signifikant besser ab als Jungen ($N = 106$, $F = 4.42$, $p = .038$, $df = 1$). Der Zusammenhang verstärkt sich zudem, wenn zusätzlich die Merkmale „Erstsprache“ und „Stadtteil“ als wechselwirkende Faktoren in die Rechnung einbezogen werden. So erreichen Jungen in der Stichprobe mit einer anderen Erstsprache aus Deutsch, die in Langwasser eine Kita besuchten, eine signifikant schlechtere Gesamtpunktzahl bei der Zweiterhebung ($N = 106$, $F = 7.36$, $p = .008$, $df = 1$). Insbesondere das letztgenannte Ergebnis führte innerhalb der Projektgruppe zu Diskussionen über geschlechtsspezifische Förderangebote.

[55] Zwischen den unabhängigen Faktoren bestand hinsichtlich der Verteilung der Merkmale kein statistischer Zusammenhang, so dass eine wichtige Voraussetzung für das Verfahren der Varianzanalyse gegeben war.

6.3.3. Methodische Schlussfolgerungen aus der Untersuchung

Das Design der Evaluation im Bereich der Überprüfung der phonologischen Bewusstheit sah ursprünglich vor, im Rahmen der Wirkungsevaluation mit einer Kontrollgruppe zu arbeiten. Für die Bildung der Kontrollgruppe wären Einrichtungen in Frage gekommen, die nicht in „Spielend lernen“ beteiligt gewesen wären und zudem keine Förderung mit „Phono-Logisch Hand in Hand“ durchgeführt hätten. Für die Evaluation der phonologischen Bewusstheit konnte jedoch im Projektzeitraum keine Einrichtung gefunden werden, die mit „Spielend lernen“-Kitas vergleichbar war und bereit gewesen wäre, an der Evaluation mitzuwirken. Infolgedessen entspricht die dann tatsächlich umgesetzte Evaluationsstrategie einer Pretest-Posttest-Untersuchung ohne Kontrollgruppe. Ohne Kontrollgruppe kann eine Veränderung der phonologischen Bewusstheit bei den überprüften Kindern aufgezeigt werden, jedoch lassen sich diese Veränderungen aufgrund der fehlenden Kontrollgruppe nicht ausschließlich bestimmten Förderfaktoren (z.B. tägliche Förderung mit dem Konzept „Phono-Logisch – Hand in Hand“) zuschreiben. Das Ziel der Evaluation war daher, die Veränderungen bei den phonologischen Fähigkeiten innerhalb eines Jahres zu erfassen und darüber hinaus Aussagen zu formulieren, wie stark sich diese Veränderung bei unterschiedlichen Teilnehmergruppen manifestierte.

Es ist anzunehmen, dass alterstypische Entwicklungs- und Reifungsprozesse (vgl. Campbell 1988, S. 153) einen relativ großen Anteil an der Erklärung der Ergebnisse der Zweiterhebung einnehmen. Im Umkehrschluss ist aber auch die Schlussfolgerung falsch, dass das Programm „Phono-Logisch – Hand in Hand“ keine positive Wirkung im Förderzeitraum gezeigt hat. Die Wirkung konnte mit dem gewählten Evaluationsdesign aufgrund der fehlenden Kontrollgruppe nicht quantifiziert werden. Die hier vorgestellten Ergebnisse können zudem durch die Art und Weise der Durchführung beider Messungen beeinflusst worden sein. Der Umgang der ErzieherInnen mit dem ARS-Verfahren kann – trotz vorheriger Schulung – variieren. Fluktuationen in der Untersuchungsgruppe können hingegen ausgeschlossen werden, da es sich um dieselbe Gruppe von Kindern handelt, die hinsichtlich ihrer phonologischen Fähigkeiten bei beiden Messungen überprüft wurde.

Im Kreis der Projektbeteiligten und Kommunalinitiativen wurden die Ergebnisse intensiv diskutiert. So konnte gezeigt werden, dass der überwiegende Teil der Kinder in der Altersgruppe vor der Einschulung nicht zu den so genannten Risikokindern zu zählen ist. Außerdem wurde deutlich, dass Jungen in der Stichprobe eine signifikant schlechtere phonologische Bewusstheit als Mädchen hatten. Dies warf im Kreis der Projektbeteiligten die Frage auf, ob eine auf den geschlechtsspezifischen Bedarf ausgerichtete Förderung angemessen wäre.

Als Kontrollevaluation, ob innerhalb eines Jahres Veränderungen bei den phonologischen Fähigkeiten bei den Teilnehmern eingetreten sind, ist das verwendete Evaluationsdesign ausreichend. Anhand der Pretest-Posttest-Untersuchung der phonologischen Bewusstheit in „Spielend lernen“ kann illustriert werden, dass die alleinige Durchführung von zwei standardisierten Messungen bei ein und derselben Teilnehmergruppe nicht mit einer Wirkungsevaluation gleichzusetzen ist. Für eine Wirkungsanalyse in Form eines experimentellen bzw. quasi-experimentellen Designs fehlen insbesondere noch folgende methodischen Elemente:

- Bildung einer in quantitativer Hinsicht ausreichend großen Kontrollgruppe, die sich aus Teilnehmern zusammensetzt mit vergleichbaren Charakteristika (z. B. Altersstruktur, Vorbildung, Stadtteilbezug) zu der der Untersuchungsgruppe.
- Entwicklung einer Programmtheorie (Ziele, Ressourcen, Maßnahmen, Outcome-Annahmen) für die Darstellung der theoretischen Wirkungsweise des Pro-gramms „Phono logisch – Hand in Hand“.
- Erfassung von bereits diagnostizierten Sprachentwicklungsverzögerungen bzw. -störungen sowie von demographischen Daten zur sozialen und familiären Situation der überprüften Kinder.
- Eingehende Schulung der ErzieherInnen in der Anwendung des Überprüfungsinstruments sowie die Festlegung eines standardisierten Verfahrens der Überprüfungsdurchführung.

Die hier beschriebenen Erkenntnisse aus der vorgestellten Phonologie-Überprüfung flossen in die Entwicklung des Leitfadens für die Evaluationspraxis ein, der im Schlusskapitel dieser Arbeit vorgestellt wird. Die Ergebnisse geben Anlass dazu, die Relevanz von Evaluierbarkeitsprüfungen im Vorfeld der Entwicklung des Evaluationsdesigns nochmals hervorzuheben. Insbesondere die methodischen Elemente der Phonologie-Überprüfung, die fehlen, um die Evaluation als Wirkungsmessung zu bezeichnen, betonen die Notwendigkeit einer eingehenden Auseinandersetzung mit den Programmmerkmalen in der Vorbereitungsphase von Evaluationsstudien. Auch die Entwicklung einer Programmtheorie und die Anwendungskompetenz im Umgang mit dem Messinstrument sind notwendige Einzelschritte im Vorfeld einer Wirkungsevaluation. Andererseits muss auch festgehalten werden, dass bei aller methodischen Sorgfalt, Maßnahmen von externen Einflüssen abhängig sein können, die sich einer Kontrolle oder Einflussnahme in einer wissenschaftlichen Evaluationsuntersuchung entziehen können. Als ein Beispiel ist das nicht Zustandekommen einer adäquaten Kontrollgruppe zu nennen. Im nächstfolgenden Beispiel wurde ein Schritt weiter gegangen und eine quasi-experimentelle Wirkungsanalyse eines Förderprogramms mit Kontrollgruppen durchgeführt.

6.4. Durchführungsbeispiel 3: Beispiel einer Wirkungsanalyse unter Anwendung eines quasi-experimentellen Designs im Programm *frühstart*

In einer aktuellen Metaanalyse zur Sprachkompetenz von Kindern mit Migrationshintergrund wird – wie schon berichtet – festgestellt, dass Defizite beim Erwerb und Gebrauch der deutschen Sprache einen negativen Einfluss auf die Bildungsentwicklung im späteren Leben und der Arbeitsmarktintegration haben (Kiziak, Kreuter, Klingholz 2012). Das Projekt *frühstart* möchte dieser Entwicklung entgegenwirken und setzt dabei mit gezielter Förderung zu einem frühen Zeitpunkt in Kindertagesstätten an. *frühstart* fußt auf einem pädagogischen Ansatz, der Weiterbildungen von Erzieher(innen), ein Konzept für Elternbegleitung sowie einem Sprachförderansatz zu einem Programm kombiniert. Die folgenden Ausführungen zur Evaluation beziehen sich nur auf die Komponente der Sprachförderung in *frühstart*. Gleichwohl wurden in Realität alle drei Komponenten zur gleichen Zeit umgesetzt und sind inhaltlich miteinander verknüpft.

Das in methodischer Hinsicht Besondere an der *frühstart*-Studie ist die komplette Planung und Durchführung einer quasi-experimentellen Wirkungsanalyse. Das für die Wirkungsanalyse entwickelte Design orientierte sich stark an den Evaluationsansätzen der amerikanischen Evaluationsforschung in der Phase der methodenzentrierten Evaluation und hier besonders an den Ansätzen von Campbell und Cook (2001).

Um die Wirkung der Sprachförderung zu prüfen, wurden in zwei Erhebungswellen alle *frühstart*-Kinder und alle Kinder in der Kontrollgruppe mit einem standardisierten Sprachstandsmessverfahren überprüft. Die Sprachuntersuchung erfolgte mit dem Marburger Sprach-Screening (MSS), ein in Hessen zwei Jahre zuvor entwickeltes Verfahren zur Messung des Sprachstandes von Kindern im Alter von 4 bis 6 Jahren[56]. Im Vorfeld wurden verschiedene Sprachtests für das Quasi-Experiment in Betracht gezogen. Die Entscheidung fiel letztendlich auf das MSS aus den folgenden Gründen:

- Das Verfahren deckt alle relevanten Bereiche der Sprachentwicklung ab.
- Das Verfahren ist für Kinder im Alter zwischen 4 und 6 Jahren geeignet und entsprach somit der Altersgruppe der frühstart-Kinder.
- Das Verfahren kann von ErzieherInnen nach vorheriger Schulung in Eigenregie in der Einrichtung durchgeführt werden.

[56] Berger, Holler-Zittlau, Dux: Marburger Sprach-Screening für 4- bis 6-jährige Kinder (MSS): Ein Sprachprüfverfahren für Kindergarten und Schule. Persen Verlag, 2006. Das Sprachscreening wird seit dem Jahr 2007 in Hessen zur Erhebung des Sprachstandes von Kindern vor der Einschulung empfohlen.

- Das Konzept des Verfahrens ähnelt inhaltlich dem Sprachförderkonzept „Wir verstehen uns gut“, da in beiden Fällen mit Zeichnungen bzw. Wimmelbildern gearbeitet wird, die den Kindern mehr oder weniger komplexe Alltagssituationen zeigen (z.B. Spielplatzsituation).
- Das MSS sowie das Förderkonzept sind eher dem pädagogischen als dem sprachwissenschaftlichen Bereich zuzuordnen.

Ungeachtet der zahlreichen Argumente, die für die Verwendung des MSS sprachen, wurden zur damaligen Zeit auch einige kritische Punkte diskutiert. So wurde das Verfahren nicht speziell für die Überprüfung von Migrantenkindern mit Deutsch als Zweitsprache entwickelt und es blieb unklar, inwieweit sprachwissenschaftliche Aspekte bei der Entwicklung berücksichtigt wurden. Auch lagen keine Metaanalysen des Verfahrens vor, anhand derer Aussagen zur Reliabilität, Validität und Objektivität des Verfahrens abgeleitet werden konnten.

Die Sprachüberprüfung nahmen die ErzieherInnen in den Einrichtungen vor. Das Verfahren erfasst detailliert Schlüsselkompetenzen in der Sprachentwicklung der deutschen Sprache, wie z.B. Kommunikationskompetenz, Artikulation und Wortschatz sowie die Entwicklung diverser grammatischer Fähigkeiten[57]. Diese einzelnen sprachlichen Bereiche werden in so genannten Subtests erhoben. Das Marburger Sprach-Screening basiert bei der Auswertung auf dem Prinzip, dass generell im Ergebnis nach sprachauffälligen und sprachunauffälligen Kindern unterschieden wird. Die Gesamtpunktzahl in den Subtests wird mit einer Mindestpunktzahl in Beziehung gesetzt, die jedes Kind erreichen muss, um nach den vorgegebenen Testkriterien in der Sprachentwicklung als unauffällig eingestuft zu werden.

Die Beteiligten am Projekt *frühstart* hatten sich aus mehreren Gründen für die Anwendung des Marburger Sprach-Screenings entschieden. In der Auswahlphase wurden verschiedene Tests in Betracht gezogen, die zur damaligen Zeit deutschlandweit zur Verfügung standen. Ein wichtiger Grund für die endgültige Entscheidung für das Marburger Screening war der Detailreichtum der Subtests. Neben Wortschatz und Sprachproduktion können auch genauere Aussagen zur grammatischen Entwicklung getroffen werden. Dies war besonders wichtig für die Wirkungsmessung der Sprachförderung im Projekt *frühstart*. Ein weiterer Grund für die Auswahl war das Prüfmaterial. Das Prüf-Wimmelbild mit einer

[57] Berger, Holler-Zittlau, Dux. Untersuchungen zum Sprachstand vierjähriger Vorschulkinder. Aktuelle phoniatrisch-pädaudiologische Aspekte 2004/2005. Bd.12. Berger R, Holler-Zittlau I. Ergebnisse einer Folgeuntersuchung der im Jahre 2003 ermittelten sprachauffälligen Vorschulkinder. 21. wissenschaftliche Jahrestagung der DGPP 2004. http://www.egms.de/static/en/meetings/dgpp2004/04dgpp68.shtml letzter Zugriff: 06.12.16). Nach der Entwicklung des Verfahrens wurde in einer Auswahl von Kindertageseinrichtungen in Hessen eine Kindergartenkohorte zu zwei Zeitpunkten mit dem MSS überprüft, um Rückschlüsse zur sprachlichen Entwicklung zu erhalten. Eine Kontrollgruppe wurde nicht untersucht.

Spielplatzsituation hat Ähnlichkeit mit Bildern im Begleitordner zum Förderprogramm „Wir verstehen uns gut – Spielerisch Deutsch lernen“ von Elke Schlösser. Hier wird die Sprachförderung durch Illustrationen (z.B. Personen und Gegenstände in einer Küche) hilfreich unterstützt.

Eine besondere Bedeutung erhalten die im Folgenden beschriebenen Ergebnisse der Sprachstandsprüfung durch die Tatsache, dass es zu Beginn der Evaluationsstudie gelungen ist, die Motivation und Bereitschaft von drei Kindertageseinrichtungen zu gewinnen, sich an der quasi-experimentellen Studie als Kontrollgruppe zu beteiligen. Es handelte sich dabei um Kindertageseinrichtungen, die sich für eine Förderung mit *frühstart* in ihrer Einrichtung bei den Stiftungen beworben hatten, letztendlich jedoch keine Förderzusage erhalten hatten. Die Kontrolleinrichtungen (jeweils aus Frankfurt, Gießen und Wetzlar) sind Nachbareinrichtungen in den gleichen Stadtteilen wie die *frühstart*-Einrichtungen. Die Kitas erklärten sich bereit, Kinder der gleichen Altersgruppe mit dem Marburger Sprachscreening in beiden Untersuchungswellen zu überprüfen. In den Kontrollgruppen-Kitas sollten diejenigen Kinder getestet werden, die für eine Förderung in *frühstart* in Frage gekommen wären[58].

Die Kontrollgruppeneinrichtungen wurden vor der ersten Erhebungswelle gebeten, jeweils 20 Kinder für die Untersuchung auszuwählen, die nach Einschätzung der ErzieherInnen Sprachförderung benötigen und potenziell für die Förderung in *frühstart* in Frage kämen. Die gemessenen Sprachkenntnisse der *frühstart*-Kinder ließen nach der zweiten Erhebungswelle und einem Vergleich mit der Kontrollgruppe Rückschlüsse über die Wirkungen des Projekts zu. Die Wirkungsevaluation konnte durch den Vergleich der beiden Erhebungszeitpunkte zu differenzierten Aussagen über Wirkungen im Projekt *frühstart* gelangen. Dieses quasi-experimentelle Evaluationsdesign war notwendig, um Wirkungen des Programms im zeitlichen Verlauf zu identifizieren. Die Datenerhebung während der Erhebungswellen betraf alle 12 *frühstart*-Kitas und zusätzlich 3 Kontroll-Kitas.

Nachdem die Zusage der Kontrollgruppen-Kitas vorlag, wurde das konkrete Evaluationsdesign ausgearbeitet. Hierzu fand eine Begehung in den einzelnen Kitas statt, damit sich der Evaluator ein Bild von der Situation, der Ausstattung und der Arbeitsweise vor Ort machen konnte.

6.4.1. Aufbau und Durchführung des Marburger Sprach-Screenings (MSS)

Die Materialien zur Durchführung des Verfahrens setzen sich wie folgt zusammen: a) ein Handbuch, b) eine Bildvorlage mit einer Spielplatzsituation und c)

[58] In der Vorbereitungsphase der Erhebungswellen wurde eine Eintagesschulung der ErzieherInnen – sowohl der *frühstart*- als auch der Kontrollgruppen-Kitas – im Umgang mit dem Marburger Sprachscreening durchgeführt.

Überprüfungsbögen. Das Marburger Sprach-Screening (MSS)[59] ist ein Einzelprüfverfahren, bei dem die Überprüfung pro Kind mindestens 20 Minuten dauert. Die sprachlichen Bereiche werden anhand eines Wimmelbildes geprüft, auf dem eine Spielplatzsituation abgebildet ist. Während des Screenings werden dem Kind Fragen zu Details, Vorgängen und Situationen im Bild gestellt. Die Äußerungen des Kindes werden in einer Reihe von Subtests protokolliert. Die Subtests im Marburger Sprach-Screening sind wie folgt untergliedert:

- Spontansprache
- Sprachverständnis
- Sprachproduktion
- Wortschatz / Artikulation
- Adjektive (Farben, Eigenschaften und Formen)
- Verben (Tätigkeiten)
- Pluralbildung
- Satzbildung
- Präposition im Akkusativ- und Dativkontext
- Nebensatzbildung mit Konjunktion
- Partizipbildung
- Phonologische Diskriminierung (nur bei 5- bis 6-Jährigen)

Die richtigen Antwortreaktionen der Kinder sind in Form von erwarteten Antworten im Überprüfungsbogen standardisiert. Dabei ist die Durchführung des Tests so angelegt, dass bestimmte Antworten der Kinder angeregt bzw. provoziert werden. Die Anweisungen zur Durchführung der Subtests sollen daher von den ErzieherInnen wörtlich übernommen werden. Das Screening erlaubt zudem auch spontane Äußerungen und Ideen der Kinder während des Tests zu berücksichtigen. Der Test wird abgebrochen, wenn ein Kind nach mehrmaliger Wiederholung einzelner Subtests keine Antwortreaktionen zeigt[60].

[59] Genaue Informationen zur Entwicklung sowie zur Durchführung des Marburger Screenings sind einem Handbuch zu entnehmen. Quelle: Holler-Zittlau, Dux, Berger: Marburger Sprach-Screening für 4- bis 6-jährige Kinder (MSS). Ein Sprachprüfverfahren für Kindergarten und Schule, Persen Verlag, Hornburg, 2004.

[60] Das Protokollieren der Antworten des Kindes erfolgt durch einfaches Notieren der Antworten sowie durch die Zeichen (+) für richtig und (-) für falsch im Sinne der Aufgabenstellung.

Die Erzieherin zeigt dem Kind das Bild mit der Spielplatzsituation und fordert dieses auf, einzelne Personen/Gegenstände und Situationen zu zeigen. Die Prüferin sagt: *„Zeige mir einen Baum... ein Buch...einen Jungen mit einer blauen Hose...etc."*. Die Prüferin protokolliert dabei, ob das Kind die richtigen oder falschen Personen oder Gegenstände benennt und notiert bei einer falschen Antwort die abweichende Reaktion. Für jede richtige oder erwartete Antwort des Kindes wird ein Punkt notiert. Diese Punkte addieren sich schließlich für jeden Subtest zu einer Gesamtpunktzahl[61]. Bei den 5- bis 6-Jährigen ist die Messlatte deutlich höher gelegt, wobei hier drei Subtests zur phonologischen Bewusstheit noch dazukommen und überprüft werden. Grundsätzlich gelten die gleichen Beurteilungskategorien für Kinder mit Deutsche Muttersprache und für Kinder mit Deutsch als Zweitsprache.

6.4.2. Ergebnis einer Voruntersuchung zur sozialen Interaktion in den *frühstart*-Gruppen

Bevor die erste Sprachstandserhebung mit den *frühstart*-Kinder startete, wurden im Rahmen einer vorbereitenden Fallstudie die Interaktionen der Kinder in den einzelnen Fördergruppen untersucht. Die Ergebnisse stellten eine wichtige **Informationsgrundlage für die Planung und Umsetzung des quasi-experimentellen Untersuchungsdesigns** dar. Da interkulturelle Erziehung das Eintreten in interkulturelle Beziehungen bedeutet, wurde ein Instrument für die Evaluation dieses Bereichs ausgesucht, anhand dessen in den drei Intensiv-Kitas die Interaktionen der Kinder standardisiert erhoben und in einem so genannten Soziogramm dargestellt werden konnten. Segregationen innerhalb der Fördergruppen können sich nachteilig auf den Sprachfördererfolg auswirken. Eine Hypothese in diesem Zusammenhang würde lauten, dass sich Fördereffekte durch frühstart nicht festigen können, wenn Kinder der Fördergruppe außerhalb der Förderzeiten in anderen Sprachen als Deutsch untereinander kommunizieren. Bei der Soziometrieanalyse handelt es sich um ein Instrument für formative Evaluationen, d.h. die Ergebnisse konnten für Anpassungen der Förderpraxis verwendet werden.

Die ErzieherInnen beobachteten während eines festgelegten Zeitraumes die Kind-Kind-Interaktionen in einer Kindergruppe basierend auf der folgenden Frage: „Mit wem spielt das Kind sehr gerne?". Die ErzieherInnen notierten ihre

[61] Die Gesamtpunktzahl in den Subtests wird mit einer Mindestpunktzahl in Beziehung gesetzt, die jedes Kind erreichen muss, um nach den gegebenen Testkriterien in der Sprachentwicklung unauffällig eingestuft zu werden. Je nach Alter des Kindes können hierzu zwei Auswertungsbögen bearbeitet werden: jeweils ein Bogen für 4- bis 5-Jährige und für 5- bis 6-Jährige. Die Mindestpunktzahl variiert dabei nach Altersstufe.

Beobachtungen in einer Soziometrietabelle[62]. Die erhobenen Daten wurden für die Auswertung mithilfe einer Software für Soziometrieanalysen eingelesen[63]. Bei der Untersuchung von Kind-Kind-Interaktionen war das Ziel der Evaluation im *frühstart*-Projekt, Segregationen in den Kindergruppen hinsichtlich ausgesuchter Merkmale zu betrachten. Den Knotenpunkten in den Soziogrammen wurden nach bestimmten Merkmalen unterschiedliche Farben und Formen gegeben. Das Ergebnis der Auswertung ist beispielgebend in der folgenden Soziometriedarstellung abgebildet.

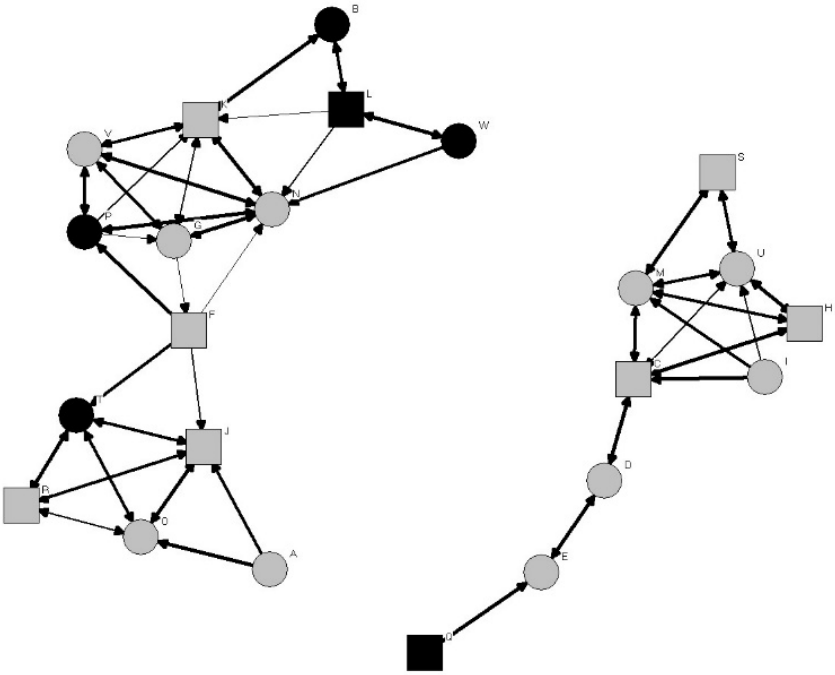


Abbildung 14: Beispiel: Soziogramm der „Tigergruppe“ einer Kindertagesstätte in Gießen[64]

Im Soziogramm sind die Kontaktbeziehungen der Kinder innerhalb einer Kindergruppe erkennbar. Im obigen Fall wurde deutlich, dass sich die Kindergruppe aus zwei Untergruppen zusammensetzte. Die Verbindungspfeile zwischen den Netzwerkknoten repräsentieren die Wahlen der Kinder, wobei die Wahlen auch gegenseitig erfolgen können. Je dicker der Pfeil, desto enger war die Beziehung zwischen den Kindern.

[62] In der Soziometrietabelle waren die Namen der Kinder in Zeilen angegeben. Die Erzieherin konnte dann aufgrund ihrer Einschätzung die Spielpartner der Kinder in die Tabellenspalten eintragen. Die Präferenzen der Kinder wurden in eine Rangfolge von der ersten bis zur vierten Wahl gebracht. Zusätzlich wurden noch weitere Daten zum Kind erhoben, wie Teilnahme an *frühstart*, Migrationshintergrund und Geschlecht.

[63] Für die Analyse der Soziogrammtabellen wurde das Computerprogramm UCINET 6 verwendet: https://sites.google.com/site/ucinetsoftware/home (letzter Zugriff am; 24.07.15).

[64] Kreis = Junge; Quadrat = Mädchen; Grau = Migrationshintergrund; Schwarz = kein Migrationshintergrund.

In der obigen Abbildung sind alle Kinder mit Migrationshintergrund grau, alle Jungen mit einem Kreis und alle Mädchen mit einem Quadrat dargestellt. Die in den Soziogrammen dargestellten Kinder mit Migrationshintergrund nahmen größtenteils auch an der *frühstart*-Förderung teil. Das oben abgebildete Soziogramm zeigt keine Segregationen hinsichtlich der ausgesuchten Merkmale. Da die meisten Kinder in der Gruppe Migrationshintergrund hatten, waren Urteile bezüglich Segregationen dieser Gruppe nicht sinnvoll.

Die genauere Betrachtung von allen Kindergruppen in den Intensiv-Kitas hat ergeben, dass Segregationen hinsichtlich der ausgewählten Merkmale teilweise in einzelnen Kindergruppen vorhanden waren. Diese Segregationen betrafen allerdings vor allem Gruppenbildungen hinsichtlich des Geschlechts. Relevant große Segregationen hinsichtlich des Migrationshintergrundes der Kinder konnten in keiner Gruppe festgestellt werden. Diese Einschätzung entsprach auch dem Ergebnis von Vorgesprächen der Evaluation mit den ErzieherInnen im *frühstart*-Projekt.

6.4.3. Ergebnisse des Pretests

Das Sprachscreening selbst wurde bei allen 12 *frühstart*-Einrichtungen und 3 Kontrolleinrichtungen in einem Zeitraum zwischen Mai und Juli 2005 durchgeführt. In jeder Einrichtung sollten alle *frühstart*-Kinder in den zwei *frühstart*-Fördergruppen getestet werden. Bis Mitte August standen die Testergebnisse von insgesamt 208 *frühstart*-Kindern und 53 Kontrollgruppen-Kindern für die Auswertung zur Verfügung. Dies entsprach einer sehr guten Ausschöpfungsquote von 96% aller mit *frühstart* geförderten Kinder in 11 Einrichtungen.

Hinsichtlich der Gesamtzahl der Kinder in den Einrichtungen waren große Unterschiede feststellbar: relativ kleine Einrichtungen (wie z.B. eine Kita mit 43 Kindern in Gießen) standen großen Kindertagesstätten mit mehr als 100 Kindern und bis zu 8 Kindergruppen gegenüber. Auch der Anteil der Kinder mit Migrationshintergrund variierte stark nach Einrichtung. Bei den evaluierten *frühstart*-Einrichtungen lag der Anteil von Kindern mit Migrationshintergrund zwischen 12% und 88%. Im Durchschnitt konnte davon ausgegangen werden, dass rund zwei Drittel aller Kinder in den *frühstart*-Einrichtungen einen Migrationshintergrund hatten[65]. Zu Qualitätsmerkmalen der einzelnen Einrichtungen wurden keine Informationen gesammelt. In der Wirkungsanalyse wurden Qualitätsmerkmale der Einrichtungen wie Ausstattung, Betreuungsrelation oder Raumangebot nicht berücksichtigt. Es hätte aufgrund der geringen Fallzahlen kein sinnvoller Vergleich der Förderergebnisse auf Einrichtungsebene vorgenommen

[65] Der Migrationshintergrund wurde für die Studie folgendermaßen definiert: Wenn mindestens ein Elternteil nach Deutschland gezogen ist und/oder das Kind selbst wurde nicht in Deutschland geboren.

werden können. Daher konnte die Hypothese nicht überprüft werden, inwieweit die Förderbedingungen in Kindertageseinrichtungen einen Einfluss auf den Fördererfolg haben.

Die insgesamt 208 überprüften *frühstart*-Kinder ließen sich etwa zur Hälfte in Jungen und Mädchen aufteilen. Die folgende Abbildung bietet einen Überblick zu den Verteilungen nach Geschlecht für Kinder mit deutscher Muttersprache und für Kinder mit Migrationshintergrund in der Untersuchungs- und Kontrollgruppe.

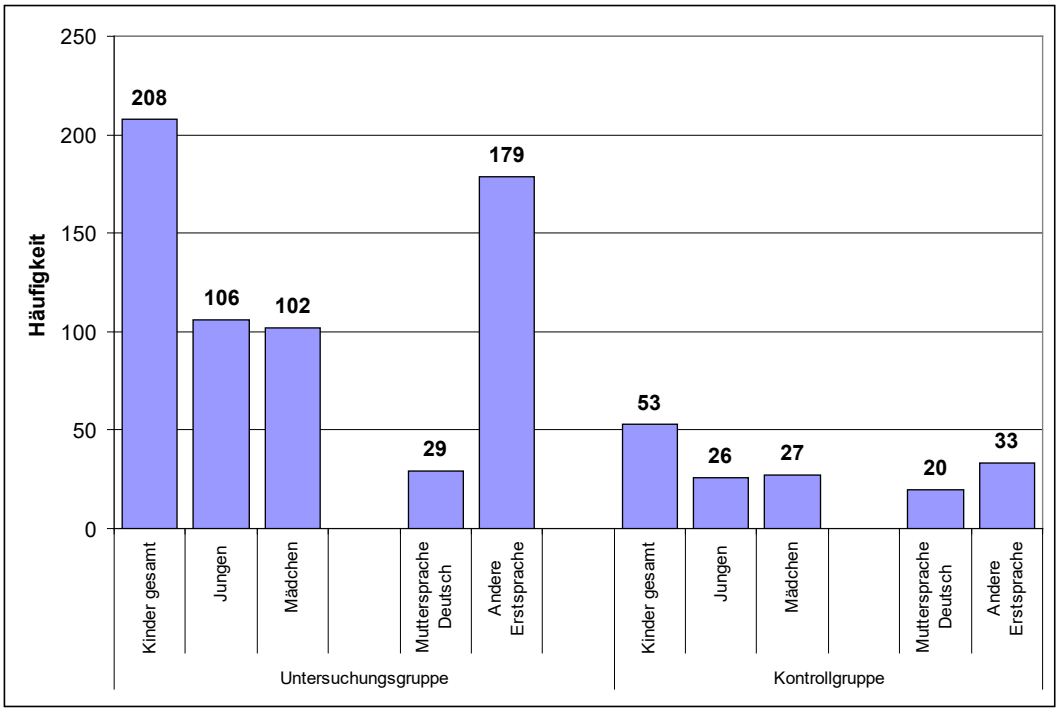


Abbildung 15: Ersterhebung: Überprüfte Kinder nach Geschlecht und Erstsprache in der Untersuchungs- und Kontrollgruppe

Wie aus der Grafik ersichtlich wird, ergab sich sowohl bei der Gruppe der deutschen Muttersprachler sowie bei der Gruppe mit anderer Erstsprache eine Gleichverteilung nach Geschlecht (51% Jungen, 49% Mädchen). Dieses Zahlenverhältnis in der Untersuchungsgruppe entsprach ungefähr dem der Kontrollgruppe, die sich aus 26 Jungen und 27 Mädchen zusammensetzte.

Unabhängig von Merkmalen wie Alter oder Erstsprache konnten zunächst allgemeine Aussagen zum Sprachstand aller überprüften *frühstart*-Kinder gemacht werden. Es ließen sich sowohl Gruppen von *frühstart*-Kindern benennen, die über geringe Sprachkenntnisse verfügten als auch Gruppen von *frühstart*-Kindern mit sprachentwicklungskonformen und **teilweise sehr guten sprachlichen Fähigkeiten** (z.B. haben etwa 10 *frühstart*-Kinder das Sprach-Screening mit voller Punktzahl abgeschlossen). Besonders viele sprachliche „Auffälligkeiten“ zeigten *frühstart*-Kinder allerdings in den Bereichen „Eigenschaften/Formen erkennen und

benennen“ und im gesamten grammatischen Prüfbereich. Der Anteil der sprachlich „unauffälligen“ Kinder lag in diesen Subtests teilweise unter 25%.

Um einen Gesamteindruck der sprachlichen Kompetenzen aller *frühstart*-Kinder zu erhalten, war es notwendig, eine weitere Unterscheidung der Ergebnisse nach erlernter Erstsprache zu treffen. Da sich das Projekt *frühstart* primär an Kinder mit Migrationshintergrund richtete, wurden vergleichsweise wenige Kinder mit Deutsch als Muttersprache gefördert – nur 26 Kinder mit Deutsch als Muttersprache konnten in die Auswertungen einbezogen werden.

Die sprachlichen Fähigkeiten der Kinder mit Deutsch als Muttersprache waren weiter entwickelt als die der Kinder mit einer anderen Muttersprache. *frühstart*-Kinder mit Muttersprache Deutsch zeigten in allen Subtests prozentual weniger sprachliche „Auffälligkeiten“ als Kinder mit anderer Erstsprache. Bis auf den Bereich Lautbildung zeigten in allen anderen Subtests Kinder mit Migrationshintergrund 15% bis 20% mehr Auffälligkeiten in der Sprachentwicklung. Bei der Anwendung von Adjektiven und im grammatischen Prüfbereich waren die Unterschiede in der Entwicklung am stärksten. Kinder mit Migrationshintergrund waren in diesen Bereichen teilweise doppelt so häufig sprachlich „auffällig“.

6.4.4. Ergebnisse des Posttests

Nach der zweiten Erhebungswelle (ein Jahr nach der Ersterhebung) standen insgesamt 260 Sprachscreening-Bögen für die Auswertung der Ergebnisse zur Verfügung. Aufgrund fehlender Angaben zu Alter und Geschlecht einzelner Kindern der Untersuchungsgruppe beziehen sich die folgenden Auswertungen auf 245 Kinder.

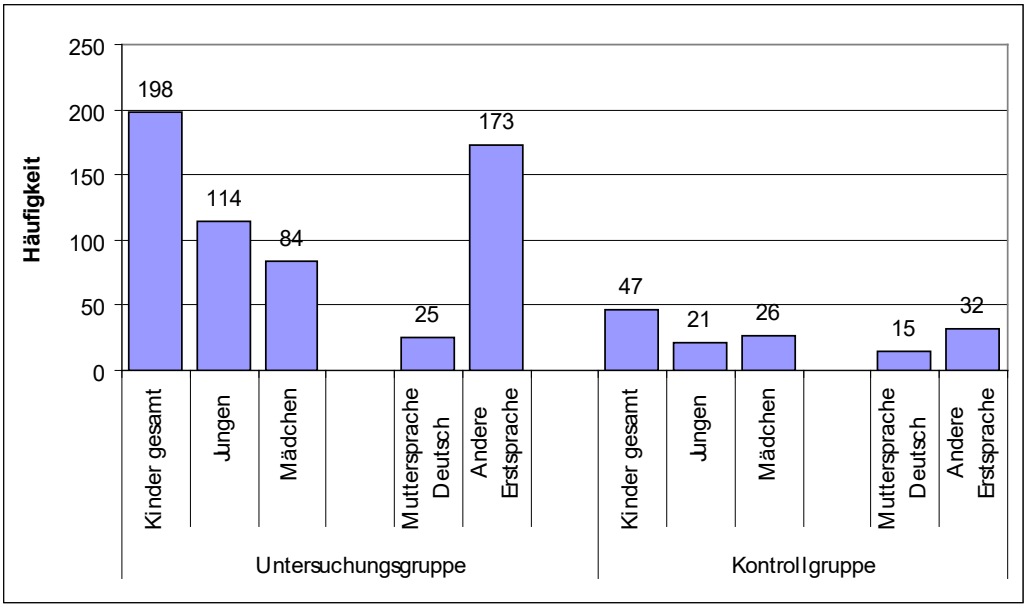


Abbildung 16: Zweiterhebung: Überprüfte Kinder nach Geschlecht und Erstsprache in der Untersuchungs- und Kontrollgruppe

Die Untersuchungsgruppe setzt sich aus 114 Jungen und 84 Mädchen zusammen. Mit 47 Kindern war die Kontrollgruppe schwächer besetzt als bei der Ersterhebung (N=53). Es handelt sich dabei um dieselben Kinder wie bei der Ersterhebung. Die Zusammensetzung der Gruppen hinsichtlich der erlernten Erstsprache war dagegen – wie schon bei der Ersterhebung – höchst unterschiedlich. Während in der Untersuchungsgruppe bei 13% der Kinder die Muttersprache Deutsch war, lag dieser Anteil in der Kontrollgruppe bei etwa 33%. Dieser hohe Anteil der Deutsch-Muttersprachler machte eine differenzierte Betrachtung der Ergebnisse der Sprachstandsuntersuchung notwendig. Aufgrund der ungleichen Gruppenverteilung wurde bei der Analyse der Ergebnisse Wert darauf gelegt, Kinder mit nicht-deutscher Erstsprache und Deutsch-Muttersprachler getrennt voneinander zu betrachten.

Neben der allgemeinen Unterscheidung zwischen Deutsch und Nicht-Deutsch als Erstsprache war für die Auswertung der Ergebnisse von Bedeutung, die Erstsprachen der *frühstart*- und Kontrollgruppen-Kinder mit Migrationshintergrund zu identifizieren. Die Kategorie „Andere Erstsprache" setzte sich aus 26 unterschiedlichen Sprachen zusammen. Dominierend waren Türkisch, gefolgt von Deutsch, Russisch, Sprachen aus dem Serbo-Kroatischen-Sprachraum sowie Albanisch. Kinder mit türkischer Muttersprache stellten mit 35% (N=70) die Mehrheit in der Untersuchungsgruppe dar. Russisch und Serbo-Kroatisch sprachen in der Untersuchungsgruppe jeweils 9% der Kinder (N=18)[66]. In der Kontrollgruppe waren hingegen hauptsächlich türkisch- und deutschsprachige Kinder (jeweils 35%) vertreten.

23 ErzieherInnen, die *frühstart*-Fördergruppen betreuten, wurden zur Durchführung der Förderung in den Einrichtungen befragt. Jede Erzieherin förderte in ihrer Gruppe regelmäßig 7 bis 12 *frühstart*-Kinder. Die Förderung erfolgte – gemäß den Vorgaben im Konzept von Elke Schlösser *„Wir verstehen uns gut, spielerisch Deutsch lernen"* – zweimal die Woche im Durchschnitt jeweils 37 Minuten lang.

Die Netto-Förderdauer der *frühstart*-Kinder betrug in dem Zeitraum zwischen Beginn der Förderung (April 2004) und dem Zweitpunkt der Zweiterhebung (Juli 2006) im Durchschnitt etwa 21 Monate (~ 635 Tage)[67]. Der durchschnittliche Zeitaufwand, den die ErzieherInnen benötigten, um eine Fördereinheit vor- und nachzubereiten, betrug 40 Minuten. In den meisten *frühstart*-Gruppen hatte sich die Förderdauer seit Beginn des Projekts nicht verändert.

[66] Die Gruppe der sonstigen Sprachen setzte sich aus pakistanischen Dialekten, Vietnamesisch, Tamilisch, Afghanisch und Spanisch zusammen (jeweils eine Nennung).

[67] Die Netto-Förderdauer errechnet sich aus der Differenz zwischen dem individuellen Zeitraum der Förderung (Brutto-Förderdauer) und der individuellen Absenz.

Um die Ergebnisse einer Vergleichsanalyse mit der Kontrollgruppe sinnvoll interpretieren zu können, war es notwendig, nähere Informationen über die Situation der Sprachförderung bei den Kontrollgruppen-Kindern zu erhalten. Die ErzieherInnen in den Kontroll-Einrichtungen gaben an, dass 31 Kinder (66%) bereits in der Vergangenheit in ihrer sprachlichen Entwicklung mit verschiedenen Konzepten gefördert wurden. Dabei handelte es sich um eine Mischung aus verbreiteten festen (pädagogisch/didaktisch ausgearbeiteten) Konzepten sowie Konzepten, die in Eigeninitiative von den ErzieherInnen entwickelt wurden. 8 Kinder aus einer Kindertageseinrichtung wurden bereits mit dem Konzept „Wir verstehen uns gut – spielerisch Deutsch lernen" gefördert, welches die Grundlage der Sprachförderung in frühstart bildet. **Die Tatsache, dass bereits 8 Kinder in der Kontrollgruppe mit „Wir verstehen uns gut – spielerisch Deutsch lernen" gefördert wurden, wurde erst nach der ersten Erhebungswelle bekannt**[68]. Eine solche Situation war praktisch nicht zu verhindern, da viele Bundesländer – nicht zuletzt ausgelöst durch die damals aktuellen ersten PISA-Ergebnisse – während des Projektzeitraums Sprachförderprogramme in breiter Form implementierten. Aufgrund der Zusammensetzung der Gruppen nach Erstsprache und Alter galten folgende Einschränkungen bei der Auswertung der Ergebnisse:

- Die Auswertung des Marburger Sprach-Screenings wurde getrennt hinsichtlich der Kategorien „Kinder mit Deutsch als Muttersprache" und „Kinder mit einer anderen Erstsprache" durchgeführt. Zwar ist der sukzessive Zweitspracherwerb im Gegensatz zur monolingualen Sprachentwicklung noch nicht umfassend erforscht, jedoch ist der Umstand zu berücksichtigen, dass Kinder mit Deutsch als Zweitsprache verspätet mit dem Spracherwerb beginnen (vgl. Kracht & Rothweiler 2003).
- Um den Alterseffekt zu berücksichtigen, wurden quantitative Vergleiche zwischen Untersuchungs- und Kontrollgruppe nur mit den zwei Altersgruppen „5-6 Jahre" und „6-7 Jahre" vorgenommen. Diese Altersgruppen waren zudem in der Untersuchungs- und Kontrollgruppe ausreichend besetzt.

6.4.5. Veränderung der Sprachkompetenz bei *frühstart*-Kindern nach der Zweiterhebung

Die Zweiterhebung wurde genau ein Jahr später nach der Ersterhebung bei denselben Kindern Bei der Zweiterhebung wurden die Ergebnisse des Marburger Sprach-Screenings in drei Altersgruppen miteinander verglichen: „bis 5 Jahre", „5 bis 6 Jahre" und „6 bis 7 Jahre". Dabei zeigte sich nicht überraschend, dass der

[68] Die Auswertung der Angaben lieferte folgendes Ergebnis: in der Untersuchungsgruppe befanden sich 65 Kinder (33%), die zum Zeitpunkt der Erhebung oder in der Vergangenheit Formen von Sprachauffälligkeiten mit therapeutischem Förderbedarf zeigten. In der Kontrollgruppe war dieser Anteil mit 11 Kindern (23%) etwas kleiner.

sprachliche Entwicklungsstand abhängig von dem Alter der Kinder ist. Je älter die überprüften Kinder waren, desto wahrscheinlicher wurden sie in der sprachlichen Entwicklung als „unauffällig" eingestuft. Die nachfolgende Abbildung zeigt die Verteilung der bestandenen Subtests bei *frühstart*-Kindern in einem Vergleich zwischen Erst- und Zweiterhebung[69]. Die Auswertung umfasst alle Kinder der Untersuchungsgruppe (N = 179).

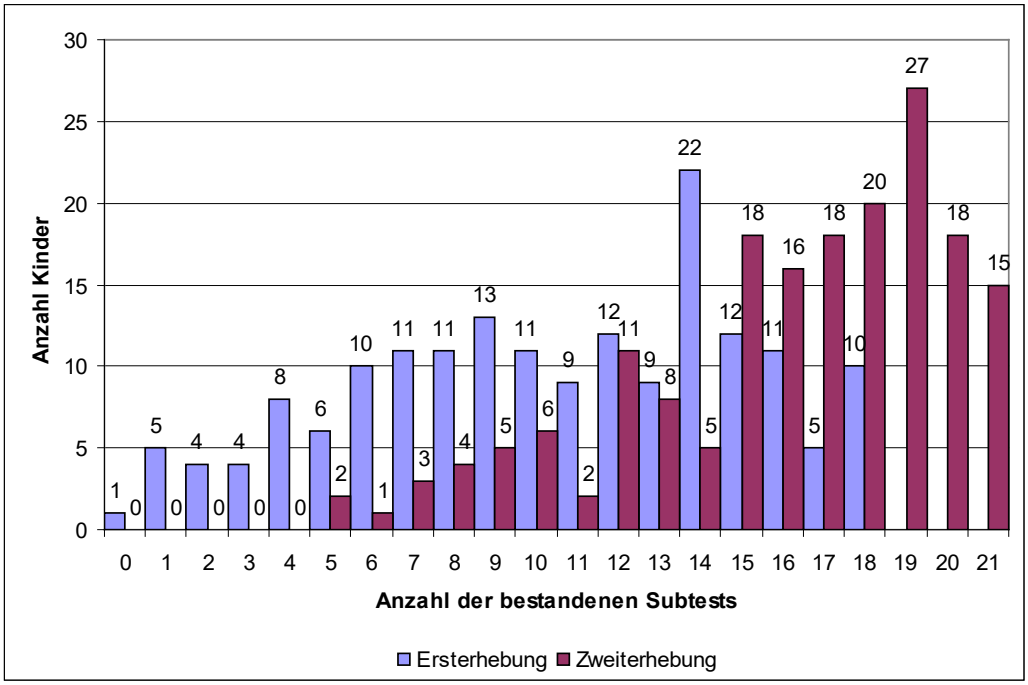


Abbildung 17: Bestandene Subtests bei allen Kindern in der Untersuchungsgruppe im Vergleich zwischen Erst- und Zweiterhebung

Die Grafik zeigt, wie deutlich sich die sprachliche Kompetenz der *frühstart*-Kinder zwischen Erst- und Zweiterhebung verbessert hat: 50% der *frühstart*-Kinder waren nach der Zweiterhebung in mindestens 17 Subtests des MSS in der sprachlichen Entwicklung unauffällig. *frühstart*-Kinder mit einer anderen Erstsprache als Deutsch hatten im Zeitraum zwischen den Erhebungswellen (1 Jahr) zum Sprachkompetenzniveau der Kinder mit Deutsch als Muttersprache aufgeschlossen. Nach den Ergebnissen des Sprach-Screenings in den Bereichen Sprachverständnis, Sprachproduktion, Lautbildung und Wortschatz war die sprachliche Entwicklung **bei der überwiegenden Mehrheit der Kinder sowohl mit und ohne Erstsprache Deutsch als unauffällig einzustufen**. Auch im Prüfbereich Grammatik war diese Entwicklung feststellbar: verglichen mit der Ersterhebung beherrschten jetzt annähernd doppelt so viele Kinder Pluralbildung, Satzbildung, die dritte Person Singular, Nebensatzbildung und Partizipbildung. Nur in den

[69] Im Rahmen der Ersterhebung wurde der phonologische Bereich (Subtests 19, 20 und 21) nicht überprüft, da die *frühstart*-Kinder zu diesem Zeitpunkt noch zu jung waren. Phonologische Fähigkeiten werden meist ein Jahr vor der Einschulung getestet.

Prüfbereichen Wortschatz, Eigenschaften und Formen benennen und in Teilen der Grammatik (Präposition im Dativ- und Akkusativkontext) waren noch relevant große Unterschiede zwischen der Gruppen erkennbar (10-15%).

6.4.6. Vergleich der Screening-Ergebnisse mit der Kontrollgruppe

Um nähere Informationen zu Effekten im Bereich der Sprachförderung durch *„Wir verstehen uns gut – spielerisch Deutsch lernen"* in *frühstart* zu erhalten, wurden die Screening-Ergebnisse in beiden Erhebungswellen mit denen der Kontrollgruppe verglichen. Da die Auswahl der Kinder für die Teilnahme an *frühstart* systematisch (d.h. nicht zufällig) durch die Einrichtungen vorgenommen wurde und die Bildung der Kontrollgruppen davon unabhängig zu einem späteren Zeitpunkt im Projekt stattfand, war davon auszugehen, dass die Merkmale der *frühstart*-Kinder mit denen aus der Kontrollgruppe nicht vollständig vergleichbar sein würden.

Vor der Anwendung eines Testverfahrens muss daher eine Vergleichbarkeit zwischen Untersuchungs- und Kontrollgruppe durch Parallelisieren nach bestimmten Merkmalen hergestellt werden. Der Vergleich mit der Kontrollgruppe wurde daher nur bei Kindern mit einer anderen Erstsprache als Deutsch durchgeführt. Es wurden nur Screening-Ergebnisse von Kindern in Betracht gezogen, die sowohl bei der Erst- als auch bei der Zweiterhebung beteiligt waren.

Es wurde ein Zusammenhangsmaß zwischen Altersstruktur und Gruppenzugehörigkeit berechnet, um auszuschließen, dass die Ergebnisinterpretation aufgrund von Alterseffekten fehlerbehaftet ist. Die Altersgruppen 5-6 und 6-7 Jahren unterschieden sich nicht hinsichtlich ihrer Zugehörigkeit zur Untersuchungs- oder Kontrollgruppe χ^2 (df = 1; N = 145) = .237, p = .626. Darüber hinaus gab es keinen Unterschied zwischen den Variablen Geschlecht und der Zugehörigkeit zur Untersuchungs- oder Kontrollgruppe χ^2 (df = 1; N = 145) = .563, p = .453.

Das Parallelisieren stellt zwar eine Vergleichbarkeit zu einem gewissen Grad her, jedoch verringern sich die Gruppengrößen deutlich. Durch die Fokussierung auf einen Teil der überprüften Kinder wurden für die folgende Analyse die Screening-Ergebnisse von 116 Kindern aus der Untersuchungsgruppe mit 29 Kindern aus der Kontrollgruppe verglichen. Konkret wurden die erreichten Punktezahlen in den einzelnen Subtests des Sprach-Screenings mit der Zugehörigkeit des Kindes zur *frühstart*- bzw. Kontrollgruppe in Beziehung gesetzt und interpretiert. Zunächst wurden die durchschnittlich erreichten Punktezahlen in jeweils der Erst- und Zweiterhebung für beide Gruppen grafisch dargestellt, um mögliche Veränderungen zu visualisieren. In den folgenden Abbildungen sind beispielhaft die Ergebnisse von sechs Subtests dargestellt.

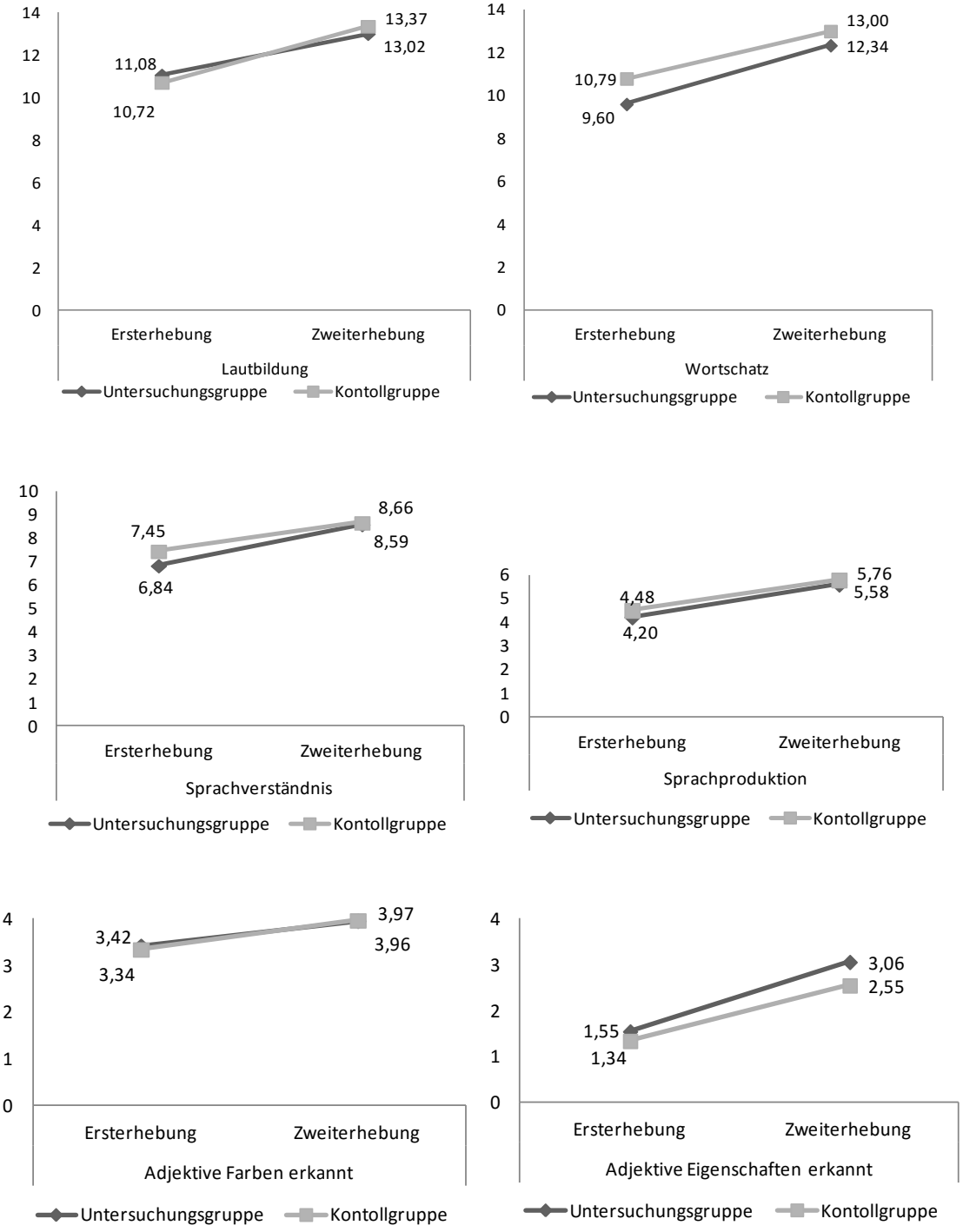


Abbildung 18: Vergleich der durchschnittlich erreichten Punktwerte bei sechs Subtests, Kinder mit einer anderen Erstsprache als Deutsch (Untersuchungsgruppe N = 116; Kontrollgruppe N = 29)

Aus der Betrachtung der Abbildungen wird deutlich, dass sowohl die Kinder in der Untersuchungs- als auch in der Kontrollgruppe im Durchschnitt eine ähnliche sprachliche Entwicklung zwischen den Messzeitpunkten zeigen. Beim Sub-

test „Wortschatz“ ist nach schlechteren Screening-Ergebnissen in der Ersterhebung ein deutliches „Aufholen“ in der Grafik erkennbar. Mittelwertvergleiche bei den übrigen Subtests zeigten ein vergleichbares Bild.

Für die Untersuchung von Daten aus quasi-experimentellen Designs mit hinsichtlich der Merkmale unterschiedlich zusammengesetzten Untersuchungs- und Kontrollgruppen (das so genannte non-equivalent control group design) (vgl. Campbell & Stanley 1966, Dugard & Todman 1995) eignet sich eine Kovarianzanalyse (ANCOVA). Die Kovarianzanalyse verbindet Elemente der multiplen Regressionsanalyse mit Elementen der Varianzanalyse. Sie bietet sich für die Analyse von nicht-experimentellen Evaluationsdesigns mit Kontrollgruppen geradezu an, bei denen der Einfluss mehrerer kategorialer unabhängiger Variablen (UVn) auf eine metrische abhängige Variable (AV) untersucht und um den Einfluss einer metrischen Kontrollvariable (Kovariate) bereinigt werden soll. Im vorliegenden Fall wäre es die Bearbeitung der Untersuchungsfrage, inwieweit das Ergebnis der Zweiterhebung (AV) durch Faktoren (UVn) wie die Zugehörigkeit zur Untersuchungs- bzw. Kontrollgruppe und/oder weiterer Faktoren (z.B. Geschlecht, Alter) unter gleichzeitiger Berücksichtigung der Ergebnisse der Ersterhebung (Kovariate) beeinflusst ist. Vor der Anwendung des ANCOVA-Modells muss das Datenmaterial bestimmte Voraussetzungen erfüllen (vgl. Bonate, Peter L. (2000)):

- Die AV muss ebenso wie die Kovariate eine metrische Skalierung aufweisen.
- Die Stichproben müssen unabhängig voneinander sein.
- Normalverteilung und Varianzhomogenität der AV muss gegeben sein.
- Homogene Regressionskoeffizienten (d.h. keine Korrelation der Kovariate mit den UVn).
- Korrelation der Kovariate mit der AV.
- Vergleichbare Gruppenstärke der Untersuchungs- und Kontrollgruppe.

Bei den vorliegenden Daten im Projekt *frühstart* werden nicht alle Voraussetzungen für die Anwendung des ANCOVA-Modells erfüllt. So liegt im konkreten Fall keine Varianzhomogenität der Überprüfungsergebnisse in der Untersuchungs- und Kontrollgruppe vor. Zudem folgen die Überprüfungsergebnisse auch in den einzelnen Subtests nicht einer Normalverteilung[70]. Auch aufgrund der unterschiedlich starken Besetzung der Untersuchungs- und Kontrollgruppe wurde die

[70] Beispiel Subtest "Wortschatz": Der Levene-Test auf Varianzhomogenität ist hoch signifikant $F(1,142) = 9.14$, $p = .003$, d. h. die Varianzen beider Gruppen sind heterogen. Ein *t*-Test könnte zwar gerechnet werden, jedoch zeigt der Test auf Normalverteilung kein positives

Anwendung der Kovarianzanalyse verworfen und anstatt dessen nicht-parametrische Tests zur Identifikation von Gruppenunterschieden gerechnet. Für die Überprüfung, ob sich die Ergebnisse der Untersuchungsgruppe von der Kontrollgruppe nach sowohl Erst- als auch Zweiterhebung signifikant voneinander unterscheiden, wurde der Man-Whitney-U-Test für unabhängige Stichproben herangezogen. Zur Überprüfung von signifikanten Veränderungen innerhalb einer Gruppe (Untersuchungs- oder Kontrollgruppe) zwischen Erst- und Zweiterhebung wurde der Wilcoxon-Test für verbundene Stichproben verwendet. Bei beiden Verfahren handelt es sich um Test-Verfahren zur Überprüfung, ob die zentrale Tendenz von zwei Stichproben unterschiedlich ist und können bei ordinal- oder intervallskalierten Merkmalen angewendet werden, deren Verteilungen nicht normalverteilt sind. Durch nicht-parametrische Tests können nur Gruppenunterschiede festgestellt werden. In die Modelle können nicht – wie im Fall der Kovarianzanalyse – zusätzliche Faktoren aufgenommen werden, um ihre Beziehungen zur einer abhängigen Variable zu berechnen. In der folgenden Tabelle sind alle Ergebnisse zum Mann-Whitney-U-Test abgebildet.

Ergebnis als Voraussetzung dafür. Der Kolmogorov-Smirnov-Test zeigt mit n = 144, Z =2.96, p < .001, dass die Ergebnisse des Subtests "Wortschatz" nicht einer Normalverteilung folgen. Der Test auf Normalverteilung brachte bei allen anderen Subtests ein nahezu identisches Ergebnis.

	Ersterhebung				Zweiterhebung			
	N	Mann-Whitney-U	Z	Zweiseitiges exaktes p	N	Mann-Whitney-U	Z	Zweiseitiges exaktes p
Sprachverständnis	145	1375	-1.535	.125	145	1678	-.020	.984
Sprachproduktion	145	1610	-.235	.814	145	1568	-.980	.327
Lautbildung	145	1520	-.813	.416	133	1001	-.592	.554
Wortschatz	145	1357	-1.616	.106	144	1552	.607	.544
Adjektive Farbe erkennen	145	1665	-.101	.919	144	1666	-.016	.987
Adjektive Farbe benennen	145	1638	-.271	.786	144	1637	-.401	.688
Adjektive Eigenschaften erkennen	145	1553	-.664	.507	143	1233	-2.251	.024*
Adjektive Eigenschaften benennen	145	1648	-.173	.862	143	1234	-2.222	.026*
Adjektive Formen erkennen	145	1660	-.110	.913	141	1615	-.058	.954
Adjektive Formen benennen	145	1386	-1.541	.123	141	1235	-2.187	.029*
Verben Tätigkeiten erkennen	145	1493	-1.106	.269	143	1541	-1.170	.242
Verben Tätigkeiten benennen	145	1334	-1.902	.057*	143	1613	-.349	.727
Pluralbildung	145	1297	-1.934	.053*	141	1479	-.770	.441
Satzbildung 3. Person Singular	145	1374	-1.568	.117	142	1420	-1.252	.210
Präposition im Akkusativkontext	145	1465	-1.223	.221	141	1348	-1.493	.135
Präposition im Dativkontext	145	1626	-.338	.735	139	1582	-.071	.944
Nebensatzbildung	145	1198	-2.577	.010*	138	1562	-.112	.910
Partizipbildung	145	1342	-1.801	.072	138	1346	-1.278	.201
Auditive Wahrnehmung					108	841	-.036	.972

Reimwörter erkennen					110	752	-.925	.355
Wortlänge erkennen					107	576	-2.247	.025*

** Statistisch signifikantes Test-Ergebnis, d.h. p <.05 (bei einer angenommenen Irrtumswahrscheinlichkeit von 5%).*

Tabelle 9: Mann-Whitney-U-Test für alle Subtests des Sprach-Screenings auf Rangunterschiede zwischen Untersuchungs- und Kontrollgruppe, Kinder mit einer anderen Erstsprache als Deutsch (Ngesamt=145)[71]

Zum Zeitpunkt der Ersterhebung unterscheiden sich Untersuchungsgruppe und Kontrollgruppe in den meisten Subtests nicht signifikant voneinander. In drei Subtests zeichneten sich Kontrollgruppen-Kinder signifikant durch eine geringfügig bessere Sprachkompetenz aus. Dies sind bei Betrachtung der Unterschiede zwischen den mittleren Rangwerte die Subtests „Verben Tätigkeiten benennen" (m. Rang UG = 70,00; m. Rang KG = 85,00), „Pluralbildung" (m. Rang UG = 69,69; m. Rang KG = 86,26) und „Nebensatzbildung" (m. Rang UG = 68,83; m. Rang KG = 89,67) (Prüfgrößen siehe Tabelle 9).

Nach der zweiten Messung – ein Jahr später – konnte zunächst festgestellt werden, dass die Entwicklung der Sprachkompetenz in beiden Gruppen deutlich vorangeschritten ist. Der nicht-parametrische Wilcoxon-Rangsummen-Test für verbundene Stichproben, der angewendet wurde, um die einzelnen Gruppen auf Unterschiede in der zentralen Tendenz zwischen der Erst- und Zweiterhebung zu überprüfen, liefert eindeutig interpretierbare Ergebnisse. Im Fall der Untersuchungsgruppe weisen ***frühstart*-Kinder in allen Subtests nach der Zweiterhebung eine im Vergleich zur Ersterhebung signifikant weiter vorangeschrittene Sprachentwicklung auf.** Wie schon in Abschnitt f) erläutert, zeigt der Großteil der Kinder keine Auffälligkeiten in der sprachlichen Entwicklung. Beispielhaft für diese deutlich ausgeprägte, positive Entwicklung können hier die Ergebnisse der Subtests „Sprachverständnis" und „Wortschatz" genannt werden. In beiden Subtests erreichten die Kinder nach den Werten in der Ersterhebung (Sprachverständnis: *Md* = 7,0; Wortschatz: *Md* = 10,0) in der Zweiterhebung (Sprachverständnis: *Md* = 9,0; Wortschatz: *Md* = 13,0) deutlich bessere Summenwerte. Der Unterschied ist hoch signifikant (Sprachverständnis: Z = -7.61, p < .001; Wortschatz: Z = -7.27, p < .001). Alle anderen Subtests im MSS zeigen ebenfalls hoch

[71] Der Mann-Whitney-U-Test beinhaltet einen Vergleich von Rangreihen der Überprüfungsergebnisse in der Untersuchungsgruppe mit denen der Kontrollgruppe. Der empirische Wert U ist die Summe aller Rangplatzüberschreitungen bei einem Vergleich aller Rangplätze zwischen beiden Gruppen. Z ist der in SPSS ausgegebene negative Wert für die Prüfgröße, der leichter interpretierbar ist. Je größer der Wert, desto größer ist der Unterschied der Rangreihen. Weitere Angaben zum genauen Prüfverfahren des Mann-Whitney-U-Tests siehe Rasch, Friese, Hofmann & Naumann (2006).

signifikante Unterschiede bei einem Vergleich der Ergebniswerte beider Erhebungen. Bei der Kontrollgruppe konnte ebenfalls eine weiter vorangeschrittene Sprachentwicklung zwischen beiden Erhebungszeitpunkten festgestellt werden, die jedoch bei 7 von 18 Haupt-Subtests nicht statistisch signifikant ausfiel. Dies bedeutet, dass die Differenz der im Test gebildeten Rangreihen klein ist, so dass Zufallsfehler bis zu einem vertretbaren Niveau (unter 5%) nicht ausgeschlossen werden können (z. B.: Subtest „Verben Tätigkeiten benennen“ Z = -1.14, p = .253).

Für die Interpretation der Wirkungen der Sprachförderung in *frühstart* ist der Vergleich zwischen Untersuchungs- und Kontrollgruppe nach der Zweiterhebung besonders interessant. Die Prüfgröße Z des Mann-Whitney-U-Tests zeigt nach der Zweiterhebung in zahlreichen Subtests einen Unterschied zwischen den gebildeten Rangreihen. In vier Subtests waren die festgestellten Sprachfortschritte der *frühstart*-Kinder signifikant weiter ausgeprägt als in der Kontrollgruppe: „Adjektive Eigenschaften erkennen“ (m. Rang UG = 75,68; m. Rang KG = 57,53 r = -.18), „Adjektive Eigenschaften benennen“ (m. Rang UG = 75,67; m. Rang KG = 57,57; r = -.18), „Adjektive Formen benennen“ (m. Rang UG = 74,47; m. Rang KG = 57,59; r = -.18) sowie in einem Subtest zur phonologischen Bewusstheit „Wortlänge erkennen“ (m. Rang UG = 56,95; m. Rang KG = 40,34; r = -.22) (Prüfgrößen siehe Tabelle 9). Die berechneten Werte für die relative Effektstärke r sind nach Cohen (1988) jedoch als schwach einzustufen[72]. Neben drei Subtests, bei denen das signifikant bessere Abschneiden der Kontrollgruppe in der Ersterhebung aufgehoben wurde, d.h. die Untersuchungsgruppe hat in der Zwischenzeit zur Kontrollgruppe aufgeschlossen, so dass keine Unterschiede zwischen beiden Gruppen nach der Zweiterhebung feststellbar waren.

Aus der Ergebnisanalyse kann geschlussfolgert werden, dass beide Gruppen zwischen den Messzeitpunkten in der sprachlichen Entwicklung deutlich vorangeschritten sind. Der Vergleich zwischen Untersuchungs- und Kontrollgruppe zeigt, dass *frühstart*-Kinder in vier Subtests signifikant besser als die Kontrollgruppenkinder abgeschnitten haben. In drei weiteren Subtests konnten *frühstart*-Kinder im Vergleich zur Kontrollgruppe schlechtere Ergebnisse der Ersterhebung wieder ausgleichen.

[72] Die relative Effektstärke ist ein Maß für die Beurteilung, wie stark der signifikante Unterschied zwischen den Gruppenwerten ist, so dass eine Aussage zur praktischen Relevanz des identifizierten Effekts getroffen werden kann. Die relative Effektstärke r errechnet sich bei nicht-parametrischen Testverfahren aus der Teilung der Prüfgröße Z durch die Quadratwurzel der Stichprobengröße: $r = Z/\sqrt{N}$. r kann einen Wert zwischen -1 und +1 annehmen. Je näher der Wert bei 1 (-1) liegt, desto stärker ist der Effekt zu beurteilen.

6.4.7. Angaben zur internen Konsistenz des Messinstruments

Mit Hilfe einer Reliabilitätsanalyse können Aussagen zur Zuverlässigkeit eines Tests gemacht werden. Die Ergebnisse sind gerade für den vorliegenden Fall besonders interessant, da die Kinder nicht immer von denselben ErzieherInnen bei der zweiten Erhebung überprüft wurden. Die interne Konsistenz der Skalen im Messinstrument wurde hier mit Cronbachs Alpha gemessen. Alpha errechnet sich durch die mittlere Korrelation zwischen den einzelnen Items und den Items insgesamt (Bortz & Döring 2003, S 195ff.). Die Erkenntnisabsicht bei *frühstart* war, durch die Reliabilitätsanalyse Anhaltspunkte über die Güte der einzelnen Subtests in Bezug auf die Erfassung des intendierten inhaltlichen Konstrukts zu erhalten. Der Alphawert sollte entsprechend methodischer Standards über einem Wert von 0,700 liegen, damit von einer ausreichenden Qualität der Skalen ausgegangen werden kann. Die nachfolgende Tabelle gibt Auskunft über die Messgenauigkeit der verschiedenen Subtests.

Testbereich	Anzahl der Items	Cronbachs α
Wortschatz	14	0,831
Lautbildung	14	0,706
Sprachverständnis: „*Zeige mir…*“	10	0,549
Eigenschaften erkennen, Farbe	4	0,731
Adjektive benennen, Farbe	4	0,667
Eigenschaften erkennen, Fühlen	4	0,762
Adjektive benennen, Fühlen	4	0,798
Eigenschaften erkennen, Form	4	0,833
Adjektive benennen, Form	4	0,832
Tätigkeiten erkennen	4	0,545
Verben benennen	4	0,714
Pluralbildung	5	0,784
Verben 3. Person Singular	4	0,768
Akkusativ	3	0,596
Dativ	3	0,701
Nebensatzbildung	3	0,814

Partizipbildung	3	0,623
Auditive Wahrnehmung	10	0,894
Reimwörter	3	0,74
Wortlänge	3	0,78

Tabelle 10: Ergebnis der internen Konsistenz-Analyse der Subtests nach der Zweiterhebung

Das Sprachverständnis, die Adjektive von Farben benennen, die Tätigkeiten erkennen, die Akkusativbildung und die Partizipbildung wiesen ein α aus, das kleiner als 0,700 ist, womit die Tests als ungenau eingestuft werden mussten. Zusammengefasst bescheinigen die Ergebnisse der Reliabilitätsanalyse dem Screening-Instrument in methodischer Hinsicht eine recht hohe Zuverlässigkeit der einzelnen Subtests. Die Ergebnisse der Reliabilitätsanalyse konnten in der Evaluationsstudie als ein zusätzliches Kriterium für die Einschätzung genutzt werden, inwiefern mit dem Screening-Verfahren valide Daten erhoben wurden.

6.4.8. Ergebnisinterpretation der quasi-experimentellen Untersuchung in *frühstart*

Ein Vergleich zwischen beiden Erhebungswellen zeigt zusammengefasst eine deutlich vorangeschrittene Sprachentwicklung und einen abnehmenden Anteil von Kindern mit „Sprachauffälligkeiten“. Dieses Ergebnis lässt sich auf Basis der durchgeführten Analysen einwandfrei feststellen.

Sowohl in der Ersterhebung als auch in der Zweiterhebung konnten auch statistisch **signifikante Unterschiede zwischen der Untersuchungs- und Kontrollgruppe** identifiziert werden. Diese Schlussfolgerung kann nur für die analysierte Gruppe von Kindern mit einer anderen Erstsprache als Deutsch gezogen werden, da nur für diesen Fall ein Vergleich zwischen Untersuchungs- und Kontrollgruppe hergestellt werden konnte. Während bei der Ersterhebung Kontrollgruppenkinder bei drei Subtests besser abgeschnitten haben als *frühstart*-Kinder, konnte dieses Ergebnis nach der Zweiterhebung umgekehrt werden. Bei vier Subtests erzielten Kinder aus der Untersuchungsgruppe signifikant bessere Ergebnisse als Kinder aus der Kontrollgruppe. Zudem haben *frühstart*-Kinder zu Kontrollgruppenkindern zum Zeitpunkt der Zweiterhebung in drei Subtests wieder aufgeschlossen. Die berechneten Effektstärken sind jedoch in allen Fällen als gering einzustufen, so dass man nicht von Unterschieden einer praktisch relevanten Größenordnung zwischen Untersuchungs- und Kontrollgruppe ausgehen kann. Die Ergebnisanalyse deutet somit auf einen Effekt durch die Förderung in frühstart hin, jedoch konnte die schwach ausgeprägten Effekte nur bei einer kleinen Zahl der Subtests im MSS festgestellt werden.

Um die Ergebnisinterpretation abzurunden, ist es des Weiteren notwendig, mögliche Störfaktoren zu diskutieren, um die Güte der Untersuchung (interne Validität) nach Campbell und Stanley (1966) abzuschätzen:

- *Förderdauer und -intensität:* Die Förderdauer war mit durchschnittlich 37 Minuten pro Woche sehr kurz. Der Zeitraum zwischen den Erhebungen war mit einem Jahr sehr kurz angesetzt, möchte man die Auswirkungen eines Förderkonzepts erfassen. Die Evaluationsstudie startete jedoch zu einem Zeitpunkt, als die Förderung der Kinder bereits in vollem Gange war. Meta-Analysen von Förderprogrammen verdeutlichen, dass Förderung im Kindergarten durchschnittlich ab einer Dauer von 119 Minuten Effekte zeigt (vgl. Leseman 2002, S. 24f.). Es ist wahrscheinlich davon auszugehen, dass bei der festgestellten Förderdauer pro Woche nach einem Jahr keine substantiell großen Fördereffekte zu erwarten sind.
- *Interne Konsistenz der Subtests:* Das Marburger Sprachscreening wurde einer Reliabilitätsanalyse unterzogen mit überwiegend positivem Ergebnis.
- *Kontext-Merkmale der Förderung:* Die Förderung kann auch mit Faktoren zusammenhängen, die sich nicht primär auf die Teilnehmer oder das Förderinstrument beziehen. Zu diesen können Kita-bezogene Faktoren zählen wie z.B. der Betreuungsschlüssel, die Kindergruppengrößen, die Ressourcenausstattung oder das Engagement der Kita-Leitung sowie der MitarbeiterInnen. Zudem können familiäre Kontextfaktoren (z. B. Bildungsniveau der Eltern, Förderung der Kinder im häuslichen Kontext) eine große Rolle bei der Erklärung des Fördererfolgs einnehmen (vgl. hierzu Tietze et al. 1998).
- *Weitere Merkmale bei den überprüften Kindern:* Des Weiteren können bestimmte Merkmale – zu denen nicht für alle mit dem MSS überprüften Kindern Informationen gesammelt werden konnten – wie z. B. in der Vergangenheit ärztlich diagnostizierte Sprachentwicklungsverzögerungen und -störungen sowie das Ausmaß der bereits erfolgten Förderung mit anderen Sprachförderinstrumenten einen im Nachhinein nicht mehr bestimmbaren Einfluss auf das Ergebnis gehabt haben. Dagegen konnte – wie eingangs beschrieben – mithilfe von Soziometrieanalysen die Interaktionsstruktur der Kinder innerhalb der Fördergruppen erfasst werden, um Rückschlüsse über das Kommunikationsverhalten der Kinder in ihrer Erst- und Zweitsprache zu erhalten.
- *Reaktivität bei der Anwendung des MSS:* Zwar haben Schulungen im Umgang mit dem MSS für alle ErzieherInnen in den Untersuchungs- und Kontrollgruppeneinrichtungen stattgefunden, das Auftreten von Reaktivität ist jedoch nicht vollkommen ausschließbar. Da die ErzieherInnen sich

der Evaluationssituation bewusst waren, die Förderung zwischen den Erhebungen sowie die Überprüfung mit dem MSS selbst durchführen, können die Ergebnisse entsprechende Verzerrungen aufweisen.

- *Auswahl und Zusammensetzung der frühstart- und Kontrollgruppe:* In *frühstart* wurden die ErzieherInnen aus den Kontrollgruppen-Kitas gebeten, etwa 20 Kin-der für die Überprüfung auszuwählen, die nach Einschätzung der Einrichtung eine intensivere Sprachförderung benötigen würden. Idealerweise sollte die Zuteilung der Kinder in die Untersuchungs- bzw. Kontrollgruppe nach dem Zufallsprinzip erfolgen. Ist dieses Verfahren in der Praxis nicht möglich, sollte in der Vorbereitungsphase darauf geachtet werden, dass für die Kontrollgruppe ausreichend Kinder mit vergleichbaren Merkmalen zu denen in der Untersuchungsgruppe gewonnen werden (Matching-Verfahren).

Ein zentrales methodisches Resümee aus dem *frühstart*-Projekt ist, dass die Durchführung von quasi-experimentellen Wirkungsanalysen mit **ausreichendem zeitlichem Vorlauf** geplant werden sollte. Im idealen Fall ist ausreichend Zeit vorhanden, die **Vorbereitung sowie die Entwicklung der Evaluationsstudie auf Basis einer ausgearbeiteten Programmtheorie** schrittweise zu vollziehen. Eine im Detail ausgearbeitete Programmtheorie kann optimal für die Konstruktion des Evaluationsdesigns verwendet werden. Korrekturen am Förderkonzept oder an der Durchführung der Förderung sind dadurch vor der ersten Wirkungsmessung noch möglich. Somit kann als ein weiteres methodisches Resümee die **besondere Relevanz der Evaluierbarkeitsprüfung** genannt werden.

Bei *frühstart* kam der positive Umstand dazu, dass Kontrollgruppen-Kitas aus denselben Stadtteilen mit Kindern zur Teilnahme gewonnen werden konnten, die vergleichbare Merkmale wie die *frühstart*-Kinder aufwiesen. Ohne solche begünstigende Faktoren aus dem Programmkontext, auf die das Evaluationsteam die Wirkungsanalyse aufbauen kann, sind experimentelle Untersuchungen schwer zu realisieren. Zudem zeigt sich, dass der Aufwand – sowohl für die Vorbereitung als auch für die Durchführung – einer quasi-experimentellen Studie immens werden kann. Viel wichtiger erscheint es daher, mit programmbegleitenden (Umsetzung-) Evaluationen, wie im Durchführungsbeispiel 1 beschrieben, eine spätere Wirkungsevaluation vorzubereiten. Noch bevor auf die Herausforderungen in der Ergebnisphase eingegangen wird, folgt zunächst ein kurzes Resümee der soeben diskutierten Praxisbeispiele.

6.5. Erkenntnisse aus den Ergebnissen von „Phono logisch – Hand in Hand“ und *frühstart* für die Planung von Wirkungsevaluationen

Für das Projekt „Spielend lernen“ wurde der Stand der phonologischen Bewusstheit von Kindergartenkindern anhand eines eigens dafür entwickelten Instruments zu zwei Zweitpunkten mit einem Abstand von einem Jahr erhoben. Aus der Sprachentwicklungsforschung ist bekannt, dass die phonologischen Kompetenzen von Kleinkindern ein guter Prädikator für spätere schulische Lese- und Schreibkompetenzen sind (vgl. Landerl & Wimmer 1994; Marx et al. 1993). Die Messung der phonologischen Bewusstheit im Rahmen der Evaluationsstudie zeigte im Ergebnis eine deutliche Verringerung der Anzahl an Risikokindern (Kinder, die zum Messzeitpunkt eine schwach ausgeprägte phonologische Bewusstheit hatten). Da die Voraussetzungen für ein experimentelles Untersuchungsdesign nicht gegeben waren, lässt sich der Förderanteil durch das Konzept „Phono logisch – Hand in Hand“ nicht quantifizieren. Dies bedeutet jedoch um Umkehrschluss nicht, dass die Förderung mit dem Konzept „Phono logisch – Hand in Hand“ zu keinen Wirkungen geführt hat, sie lassen sich aufgrund des umgesetzten Evaluationsdesigns in empirischer Hinsicht nicht ausreichend differenziert beschreiben.

Im zweiten Beispiel wurde beim Projekt *frühstart* ein quasi-experimentelles Design für die Untersuchung von Sprachkompetenzen bei Kindergartenkindern entwickelt. Der organisatorische Aufwand sowie der benötigte zeitliche Vorlauf waren im Vergleich zur phonologischen Untersuchung deutlich höher. Vergleicht man die Ergebnisse beider Untersuchungen miteinander, fällt eine große Gemeinsamkeit auf. In beiden Untersuchungen hat die Sprachkompetenz aller getesteten Kinder innerhalb eines Zeitraums von zwei Jahren zugenommen. Diese Entwicklung gilt auch für die Kinder mit Migrationshintergrund in beiden Studien. Die Kernfrage in Bezug auf die Sprachentwicklung lautete daher: Welchen Anteil an der Steigerung der Sprachkompetenz hat das spezifische Förderprogramm, das vor allem bei der Arbeit mit Kindern mit Migrationshintergrund eingesetzt wurde?

Im Projekt *frühstart* erlaubte das Evaluationsdesign mit einer Kontrollgruppe, Hinweise zur Beantwortung dieser Frage zu liefern. Durch alterstypische Reifungsprozesse der überprüften Kinder lässt sich der deutliche Sprachentwicklungsfortschritt zwischen beiden Erhebungszeitpunkten überwiegend erklären. Für die Interpretation der Ergebnisse kommt erschwerend hinzu, dass in den Kontrollgruppen-Kitas im Untersuchungszeitraum verschiedene Formen von

Sprachförderung stattgefunden haben[73]. Das Ergebnis zeigt jedoch auch signifikant weiter vorangeschrittene der geförderten Kinder in einzelnen Subtests gegenüber Kindern aus der Kontrollgruppe. Dies lässt auf einen gewissen Fördererfolg durch das Konzept „Wir verstehen uns gut – spielerisch Deutsch lernen" schließen. Die Sprachförderung in den *frühstart*-Kitas repräsentiert neben den Weiterbildungen der ErzieherInnen zur Förderung der pädagogischen Kompetenz sowie der Elternbegleitung von *frühstart*-Kindern einen Baustein des gesamten Programms.

Wie ist das Ergebnis beider Studien im Kontext von Initiativen zur Sprachförderung zu bewerten? Die Sichtung von Ergebnissen aus wissenschaftlichen Evaluationen zu den Wirkungen von kompensatorischen Sprachförderprogrammen zeigt nur wenige signifikante Fördererfolge in einer praktisch relevanten Größenordnung (Polotzek et al. 2008; Roos & Schöler 2007; Roos et al. 2010a, 2010b; Wolf, Stanat & Wendt 2010). In den Studien wird zudem darauf hingewiesen, dass selbst das Aufschließen einer geförderten Gruppe von Migrantenkindern zum Sprachstand einer deutschen Vergleichsgruppe von Reifungsprozessen sowie vielen weiteren familiären und umweltbezogenen Einflüssen abhängig sein kann.

Der Aufwand für die Durchführung von Wirkungsanalysen wirft vor dem Hintergrund der betrachteten Untersuchungsergebnisse Fragen zur Aufwand/Nutzen-Aspekten auf. Um Wirkungsanalysen nach allen Standards der empirischen Sozialforschung durchzuführen, ist – neben den weiteren projektstrukturbedingten und organisatorischen Voraussetzungen – ein **hoher Ressourcen- und Zeitaufwand** notwendig. Im Rahmen der zweiten Evaluierbarkeitsprüfung sollte daher gemeinsam mit den Auftraggebern geprüft werden, ob sich der Aufwand für die Messung von Programmwirkungen anhand einer experimentellen Versuchsanordnung mit den zu erwartenden Ergebnissen zu rechtfertigen ist und welche Alternative es zu einem solchen Verfahren gibt. Auch sollte Klarheit über die möglichen Evaluationsergebnisse und die damit verbundenen Folgeentscheidungen für das Programm hergestellt werden. Selbst sorgfältig geplante experimentelle Studien sind mit einer Irrtumswahrscheinlichkeit behaftet. Aus den genannten Aspekten geht deutlich hervor, welche Komplexität die Auswertung und Interpretation der Ergebnisse einer quasi-experimentellen Studie annehmen kann.

Die Erfahrung aus den vorgestellten Studien zur Evaluation von Wirkungen legt nahe, dass besonders intensiv folgende Vorbereitungsschritte in der Eingangsphase von experimentellen Studien begangen werden sollten:

[73] Die Form und Dauer der Sprachförderung in den Kontrollgruppen-Kitas wurde in der Evaluationsstudie nicht erfasst. Dass die deutsche Sprache aktiv gefördert wurde – situativ und nach festen Konzepten – ergab sich aus Gesprächen des Evaluators mit den ErzieherInnen.

- *Auswahl des Testinstrumentariums:* Nur durch die Wahl des geeigneten Instruments lässt sich die Wahrscheinlichkeit des Auftretens von Messfehlern reduzieren. Idealerweise wurde das Instrument in der Vergangenheit in verschiedenen Untersuchungen erprobt, weiterentwickelt und es liegen Meta-Evaluationsergebnisse über das Instrument vor, die eine gute Konstruktvalidität der einzelnen Test-Items verdeutlichen, so dass eine hohe interne Validität der Ergebnisse zu erwarten ist.
- *Bildung von gut besetzten Kontroll- und Untersuchungsgruppen:* Die Bildung einer quasi-experimentellen Untersuchungsanordnung ist eine zeit- und ressourcenintensiver Prozess, der selbstverständlich streng nach den methodischen Standards der empirischen Sozialforschung durchgeführt werden sollte. Unabhängig davon, ob das Verfahren durch Zufallsauswahl oder nach bestimmten statistischen Merkmalen der Teilnehmer erfolgt, sollte die Bildung der Untersuchungs- und Kontrollgruppen durch den Evaluator der Studie durchgeführt wer-den. Dies ist wiederum ein Argument dafür, dass bei Wirkungsanalysen der Evaluator schon mit gutem zeitlichen Abstand zum Start des Programms mit den Vorbereitungen zum Evaluationsdesign starten sollte.
- *Mehr als zwei Messungen:* Für die Validität der Ergebnisse sowie die Betrachtung der Nachhaltigkeit der initiierten Maßnahmen sollten nach Möglichkeit mehr als zwei Messungen durchgeführt werden. Mehrere Messungen können sich unter bestimmten Umständen (z.B. wenn die Ausfallquote bei den Teil-nehmern erwartbar klein ausfällt) kompensierend auf Nachteile auswirken, die aufgrund von kleinen Untersuchungs- bzw. Kontrollgruppen erwachsen.
- *Standardisierte Erfassung möglicher Einflussfaktoren:* Um das vielfach bereits thematisierte Blackbox-Phänomen bei Wirkungsanalysen zu beleuchten, ist es anzuraten, eine genaue Programmtheorie aufzustellen. Die ausgearbeitete Programmtheorie kann Auskunft darüber geben, welche für die Wirkungsmessung relevant erscheinender Einflussfaktoren erhoben werden sollten. Bei Programmen zur Sprachförderung sind dies die individuelle Erfassung von diagnostizierten Sprachentwicklungsverzögerungen, Angaben zu bereits durchgeführten Sprachförderungen sowie von Kontextmerkmalen, innerhalb derer die Förderung durchgeführt wird.

7. Ergebnisphase der Evaluation

Nach dem Abschluss der gesamten Erhebung beginnt die vierte und letzte Phase im Evaluationsprozess. In der Ergebnisphase geht es um die detaillierte Interpretation und Präsentation der Ergebnisse. In enger Kooperation mit den Auftraggebern sollte auf Basis der Ergebnisse zunächst in einem ersten Schritt ein Abgleich mit dem Evaluationszielen erfolgen. Abhängig davon, ob mit der Evaluation das Programm weiterentwickelt oder ob Aussagen zu den Wirkungen formuliert werden sollen, sind die Ergebnisse entsprechend an die Auftraggeber zu kommunizieren. Die Ergebnisse sollten vor allem genutzt werden, um das Programm weiterzuentwickeln.

7.1. Verwendung der Ergebnisse für die Weiterentwicklung des Programms und der Programmtheorie

Nach der Auswertung und Aufbereitung der Ergebnisse sollte im Idealfall eine Überarbeitung der Programmtheorie durchgeführt werden. Die Programmtheorie, bestehend aus Action Model und Change Model, sollten in Bezug auf die Evaluationsergebnisse und den gesammelten Erfahrungen bei der Evaluationsdurchführung kritisch analysiert werden. Die Überarbeitung der Programmtheorie kann auf Basis der Beantwortung der folgenden Fragen erfolgen:

Erfassung der Programmergebnisse:

- Welche Ergebnisse hat das Programm nach einer bestimmten Laufzeit erzeugt?
- Inwieweit stimmen die Ergebnisse mit den in der Programmtheorie angestrebten Ergebnissen überein?
- Wird die intendierte Zielgruppe mit den Programmmaßnahmen erreicht? Wenn ja, in welchem Ausmaß?

Kontrolle der Programmdurchführung:

- Wurde das Programm konzeptgetreu und planungsgemäß durchgeführt?
- Haben sich Änderungen bei den Zielen, Maßnahmen oder in der Teilnehmerstruktur im Verlauf der Durchführung ergeben?
- Welche und wie viel Ressourcen sind in das Programm geflossen? Können Aus-sagen zum Kosten-Nutzen-Verhältnis gemacht werden?

In der folgenden Abbildung ist zum Zweck der Verdeutlichung des Überarbeitungsprozesses der Programmtheorie als schematischer Prozess dargestellt. In

dem Kapitel zur Eingangsphase der Evaluation wurde ausführlich der Erarbeitungsprozess der Programmtheorie dargestellt, beginnen mit der Erarbeitung eines ersten Entwurfs der Programmtheorie, der anschießend mit den Programmbeteiligten und -verantwortlichen diskutiert und ggf. angepasst werden kann. Diese Phase kann mehrere iterative Arbeitsschritte umfassen. Nachdem die Elemente der Programmtheorie zum Zeitpunkt t1 ausgehend von der Beschreibung eines Action Models entwickelt wurden und fest stehen, schließt sich die Überprüfung der Programmtheorie durch geeignete Evaluationsmethoden an. Dieser Schritt wurde anhand des Durchführungsbeispiels zum Programm *frühstart* eingehend erläutert.

Im Anschluss an die Auswertung in der Ergebnisphase folgt die Rückkopplung der Ergebnisse zum Zweck der Weiterentwicklung der Programmtheorie zum Zeitpunkt t2. So können die Ergebnisse der Evaluation beispielsweise die Folge haben, dass neben dem Change Model auch das Action Model hinsichtlich der Gültigkeit überarbeitet werden muss. Theoretisch ließe sich daran eine **Re-Evaluation des Programms** anschließen, um die weiterentwickelte Programmtheorie zu überprüfen. Bei quasi-experimentellen Verfahren können die in der Programmtheorie dargelegten Wirkungsannahmen durch x Re-Evaluationen validiert werden, d.h. mit der zunehmenden Zahl an Messungen lassen sich gesicherte Theorien über die Wirkungsweise des untersuchten sozialen Programms ableiten.

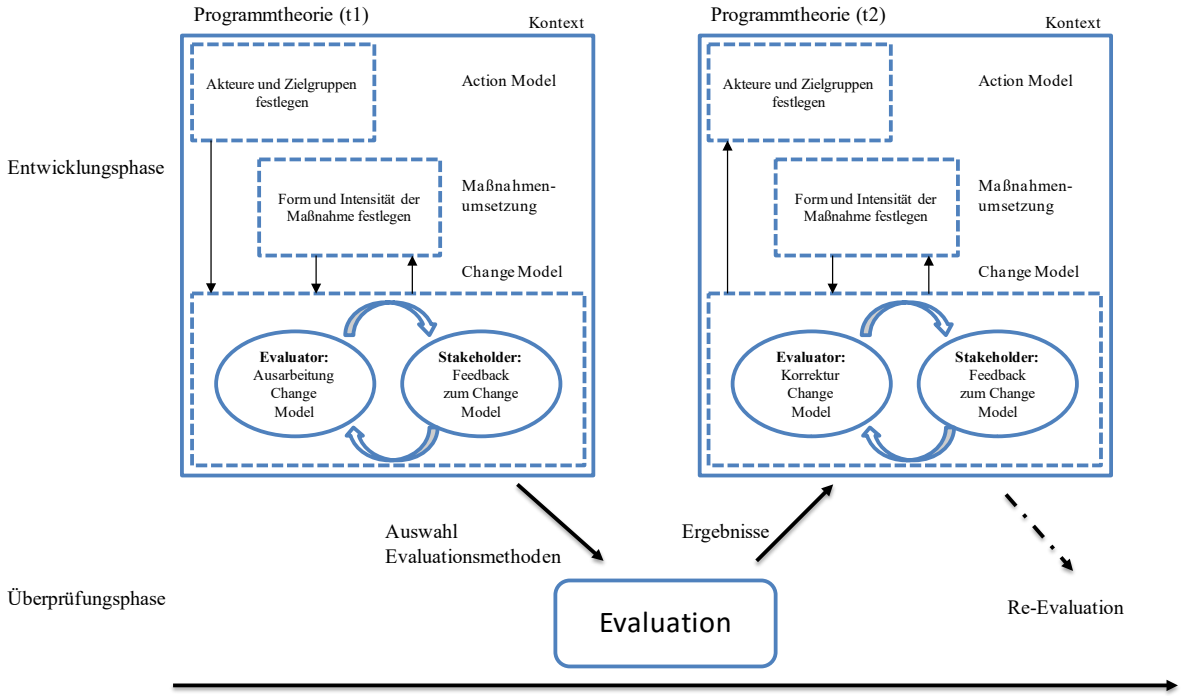


Abbildung 19: Entwicklung, Überprüfung und Weiterentwicklung einer Programmtheorie als schematischer Ablauf

Im Programm *frühstart* endete der Evaluationsauftrag nach der einmaligen Anwendung des quasi-experimentellen Designs, ohne dass die programmtheoretischen Annahmen überarbeitet wurden und eine Folgeevaluation geplant werden konnte.

Bei formativen Evaluationsformen, so wie bei der Evaluationsstudie zur Integrationskursreihe „In Deutschland zu Hause" hat der Überarbeitungsprozess kontinuierlich stattgefunden, d.h. während der Laufzeit der Integrationskurse hat sich die Entwicklungsphase mit der Überprüfungsphase ständig abgewechselt – teilweise nach jeder Kurseinheit. Summativen Evaluationen, wie dies bei der Evaluationsstudie zu *frühstart* oder „Spielend lernen" der Fall war, zeichnen sich dagegen durch zeitlich längere **Überarbeitungsschleifen der Programmtheorie** aus. Aufgrund der Rückkopplungsschleifen bei der Weiterentwicklung der Programmtheorie zeigt sich eine gewisse Ähnlichkeit mit gängigen Qualitätsmanagementsystemen (z.B. Total Quality Management (Malorny & Hummel 2011)) aus, die eine kontinuierliche, zyklische Weiterentwicklung des Gegenstandbereichs (PDCA-Zyklus: Plan – Do – Check – Act) als fundamentales Grundprinzip vorsehen.

7.2. Zusammenstellung der Ergebnisse, Kommunikation und Präsentation

Während der Endphase eines Datenerhebungsprozesses warten Auftraggeber meist ungeduldig auf erste Ergebnisse. Dabei interessiert zunächst einmal besonders, ob sich ihre Erwartungen und Annahmen hinsichtlich des Erfolgs des Programms in den Evaluationsergebnissen widerspiegeln. Es kann durchaus üblich sein, dass Auftraggeber schon im Vorfeld die Kommunikationsstrategie der Ergebnisse vorbereitet haben und Medien identifiziert haben, über die Erkenntnisse aus der Evaluationsstudie gestreut werden sollen.

Evaluatoren sollten daher nach Möglichkeit von den Kommunikationsplänen des Auftraggebers Kenntnis besitzen, um sich bestmöglich in der Ergebnisphase einer Evaluationsstudie darauf einstellen zu können. Die Vermittlung der Ergebnisse an die Programminitiatoren und Auftraggeber der Evaluationsstudie sollte beginnen, sobald die ersten Auswertungsergebnisse vorliegen. Dies können auch Kurzberichte zu geführten Interviews sein oder Häufigkeitsauszählungen bei standardisierten Befragungen. Mit einer frühzeitigen Fertigstellung der Grobauswertung bleibt mehr Zeit für den Evaluator, in Zusammenarbeit mit dem Auftraggeber Form und Schwerpunkt von Detailauswertungen zu entwickeln. Auf Basis der Erfahrungen aus den zuvor im Detail erläuterten Evaluationsstudien empfiehlt sich folgender mehrstufiger Prozess zur Präsentation und Kommunikation der Ergebnisse:

- *Erstellung eines Auswertungsplans:* Der Auswertungsplan sollte mit den Auftraggebern diskutiert werden; ggf. wird er mit den Auswertungswünschen komplettiert. Durch die Planerstellung wird der Auftraggeber transparent in die Ergebniszusammenfassung eingebunden.
- *Diskussion der Ergebnisse mit den Auftraggebern:* Hierzu bieten sich am besten gemeinsame Workshops an, in dem man bei vertrauter Atmosphäre genügend Zeit hat, gemeinsam die Ergebnisse zu analysieren und verschiedene Alternativen der Reaktion auf die Ergebnisse diskutieren kann.
- *Zusammenstellung der finalen Ergebnisse:* Hier sollte auch mit dem Auftraggeber vorab vereinbart werden, welche Form und welcher Umfang die Ergebnisberichte haben sollen. Werden Ergebnisse für Presseinformationen benötigt, ist ein kurzer Ergebnisbericht sinnvoller. Sollen die Ergebnisse einem breiten Publikum vorgestellt werden, können damit Publikationen, Veröffentlichungen im Internet und Referate auf entsprechenden Tagungen und Symposien verbunden sein.
- *Präsentation der Ergebnisse und Vereinbarung des weiteren Vorgehens mit den Auftraggebern:* Die enge Kommunikationsarbeit mit dem Auftraggeber kann schließlich dafür genutzt werden, um über das weitere Vorgehen nach dem Abschluss der Evaluationsstudie zu sprechen. Entsprechende Folgeevaluationen lassen sich auf der Grundlage von sachlichen Gesprächen in Workshops leichter vereinbaren.
- *Ende bzw. Weiterführung der Evaluation:* Das Ende einer Evaluationsstudie kann aber auch der gleichzeitige Beginn einer neuen Studie beim gleichen Auftraggeber bedeuten.

In jedem Fall sollte die Alternative zur Ergebnispräsentation vermieden werden, nämlich die Auswertung, Interpretation und Verschriftlichung der zentralen Ergebnisse der Evaluationsstudie durch den Evaluator im „stillen Kämmerlein“, ohne die Projektbeteiligten und den Auftraggeber in diesen Prozess zu involvieren. Dies birgt die Gefahr, dass auf diese Art und Weise generierte Ergebnisse im Ausschluss der Projektbeteiligten Gefahr laufen, prinzipiell in Frage gestellt zu werden. Die Evaluationsstudie wird nicht wahrgenommen und akzeptiert werden.

Die Wahrscheinlichkeit ist außerdem groß, dass Entscheidungsträger die vorgelegten Ergebnisse nicht akzeptieren, wodurch der Sinn der Evaluationsstudie insgesamt in Frage gestellt wird. Bei einer ansonsten korrekten methodischen Arbeit steht zudem die Reputation des Evaluators auf dem Spiel. Wie Evaluationsergebnisse in der Endphase einer Evaluationsstudie in der Praxis verwendet wurden, kann im positiven Sinne am Beispiel von „Spielend lernen“ beschrieben werden.

Die Ergebnisse der Evaluation wurden in zahlreichen städtischen Gremien bis hin zur Integrationskommission vorgestellt und diskutiert[74]. Die starke Beratungsfunktion durch der Evaluator und die enge Einbindung der Evaluation in Arbeitsprozesse haben sich in diesem Programm bewährt und stellten sich schließlich als Mehrwert für alle Projektbeteiligten heraus. Die Ergebnisse brachten einen detaillierten Einblick in das Engagement der Stadt Nürnberg und weiterer Akteure im Bereich der Integration von Migranten. Damit zeigt sich am Beispiel dieser Studie deutlich, dass sich Evaluationsstudien hinsichtlich der Zielorientierung im Laufe der Umsetzung des Programms verändern können. Im konkreten Fall „Spielend lernen" äußerten die Programminitiatoren einen anderen Bedarf an Unterstützung durch das Evaluationsteam, als dies in der Konzeptionsphase der Fall war. Nachdem Wirkungsevaluationen aus den beschriebenen Gründen nicht realisierbar waren, rückte die Beratungsfunktion in den Vordergrund. Für das Evaluationsteam bedeutet dies, das Evaluationsdesign flexibel auf Veränderungen bzw. sich verändernde Evaluationsbedarfe anzupassen.

Des Weiteren boten die Ergebnisse eine Grundlage für die Entscheidung, ob und wie das Programm „Spielend lernen" weitergeführt werden sollte. Die Stadt Nürnberg hat sich Anfang 2007 dazu entschlossen, „Spielend lernen" auf weitere Nürnberger Stadtteile auszudehnen und die Förderung fortzuführen. Das efms sollte die Folgeevaluation des Programms übernehmen und wurde wieder gebeten, einen Evaluationsvorschlag zu unterbreiten. Im Zentrum des Erkenntnisinteresses stand die Arbeit der Stadtteilkoordination.

7.3. Umgang mit nicht erwarteten Evaluationsergebnissen

Die Evaluationstheoretikerin Weiss weist Anfang der 70er Jahre auf potentielle Gegensätzlichkeiten bei Zielen und Erwartungen bezüglich der Evaluationsergebnisse hin (Weiss 1972). Vorausgesetzt, dass Evaluatoren den Standards der empirischen Forschung treu sind, kann das Interesse an und der Umgang mit Evaluationsergebnissen auf Seiten der Auftraggeber sehr unterschiedlich ausfallen. Die Palette der Handlungen in Folge von „schlechten" Evaluationsergebnissen reicht von „Augenwischerei" (Verharmlosung) über „Schönfärberei" (Hervorhebung der positiven Ergebnisse) bis hin zur absichtsvollen Fälschung von Daten (Weiss 1972, S. 33). Zur Vermeidung von unerwarteten bzw. nicht erwünschten Evaluationsergebnissen sollte die Kommunikation und der Austausch von Daten mit den Programmverantwortlichen und Auftraggebern der Evaluationsstudie während des gesamten Evaluationsprozesses aufrechterhalten

[74] Die Integrationskommission setzte sich zum Zeitpunkt der Evaluationsstudie in Nürnberg aus den Vertretern der verschiedenen Parteien im Stadtrat zusammen und wird unter dem Vorsitz des Oberbürgermeisters geleitet. In den Sitzungen werden Beschlussvorlagen u. a. im Themenbereich Migration und Integration behandelt.

werden. Missverständnisse und nicht erfüllte Erwartungen können nur vermieden werden, wenn Auftraggeber von Evaluationen und Programminitiatoren in Abstimmungsprozesse zur Entwicklung des Evaluationsdesigns eingebunden sowie fortlaufend mit Status-quo-Analysen informiert werden. Idealerweise entwickelt der Evaluator frühzeitig ein Gespür für die Erwartungen und Wünsche der Auftraggeber. Augenwischerei und Schönfärberei der Evaluationsergebnisse können somit vermieden werden, wenn der Evaluator den Auftraggeber schon im Planungsstadium über die Vor- und Nachteile bestimmter Evaluationsdesigns informiert und berät. Die Auswahl des Evaluationsvorgehens sollte dann gemeinsam erfolgen. Nach der Datenerhebung kann ein Auswertungsplan den Auftraggebern Transparenz vermitteln. Dadurch werden „Überraschungen" bei der Präsentation der Evaluationsergebnisse vermieden.

Wie sich Programme im Laufe einer Evaluationsstudie verändern können, kann am Beispiel des Integrationskursprojekts „In Deutschland zu Hause" verdeutlicht werden. Die Durchführung des Integrationskursprojekts führte letztendlich im Jahr 2003 nicht zu den von den Projektinitiatoren erhofften Ergebnissen. Das damalige – nach Abschluss der Kursreihen – zusammengefasste Ergebnis der Evaluation lautete folgendermaßen: Das Staatsangehörigkeitsgesetz aus dem Jahr 2000 verlangt von jedem Einbürgerungsbewerber eine Loyalitätserklärung gegenüber dem Grundgesetz. Eine solche Erklärung macht nur Sinn, wenn sie bewusst und in Kenntnis zumindest von Grundkenntnissen der Verfassung erfolgt. Diese Überlegungen führten die Initiatoren des Projekts „In Deutschland zu Hause" zur Entwicklung und Erprobung der hier beschriebenen und evaluierten Modellstudie beim Bildungszentrum der Stadt Nürnberg.

Im Rahmen gegebener Möglichkeiten und Restriktionen gelang es in der Modellstudie nur zu einem sehr geringen Teil, den Kreis der im Einbürgerungsprozess befindlichen Personen zu erreichen und für den Kurs zu rekrutieren. So wurde der Kurs zu einem Angebot im zu der damaligen Zeit wenig nachgefragtem Feld der politischen Bildung. Damit die Kurse angeboten werden konnten, mussten die Bewerbungsmaßnahmen auf alle interessierten Migranten im Raum Nürnberg ausgedehnt werden.

Die Evaluation hat gezeigt, dass der Kurs als freiwillig wahrzunehmendes Angebot im Rahmen sozialkundlicher/politischer Bildung nur kleinere Zahlen politisch Interessierter aus den gebildeten Schichten der Migranten erreichte. Soll die staatsbürgerliche Bildung und politische Integration zukünftiger Bürger aus dem Kreis der Migranten verbessert werden, bieten sich zwei Möglichkeiten an: (1) den Kursbesuch rechtlich verpflichtend zu machen; oder (2) bei weiter freiwilligem Besuch starke und motivierende Anreizstrukturen im Rahmen des Einbür-

gerungsvorgangs zu schaffen. Die Evaluationsergebnisse haben außerdem gezeigt, dass auch Migranten, die seit vielen Jahren in Deutschland leben, alltagspraktische Informationen und Beratungsangebote wünschen.

Am hier beschriebenen Beispiel der Evaluation der Integrationskurse wird deutlich, dass trotz eines inhaltlich ausgereiften Kurssystems im Sinne einer formativen Evaluation am Ende das Modellprojekt eingestellt werden musste, weil Anreize für eine Teilnahme an den Kursen fehlten. Es handelte sich um eine auf qualitativ hohem Niveau ausgearbeitete Kursreihe, das erworbene Wissen hatte jedoch für den individuellen Teilnehmer keinen Verwertungsnutzen.

7.4. Entscheidungen auf Basis von Evaluationsergebnissen

Die frühe Head Start-Phase hatte entscheidenden Einfluss auf die Weiterentwicklung der Evaluationsmethodologie in den USA (vgl. Ellsworth 1998, Ellsworth & Ames 1998). Maßgeblich in den Arbeiten von Carol Weiss wurde das Beziehungsgeflecht zwischen politischer Entscheidungsebene als Auftraggeber von Evaluationen und der wissenschaftlichen Ebene als Auftragnehmer thematisiert (Weiss 1972). Als eine zentrale Einsicht aus den Erfahrungen mit früheren Evaluationen stellte sich heraus, dass Evaluationsergebnisse zu den Wirkungen eines Programms nicht automatisch auf politischer Ebene zu Entscheidungen über die Fortführung des Programms führen. Diese „naive Vorstellung“ (Weiss 1972) der Arbeitsweise von politischen Entscheidungsprozessen kann nach den Erfahrungen mit Head Start nicht aufrechterhalten werden. Die Prozesse präsentieren sich in der Realität viel komplexer. Wissenschaftliche Untersuchungen haben einen Einfluss auf den politischen Entscheidungsprozess, jedoch sind die Effekte oft diffus und führen aus Forschersicht zu unabsehbaren Entwicklungen (Cohen & Garet 1975).

Ähnlich wie Carol Weiss ist Cronbach der Überzeugung, dass politische Entscheidungen nicht auf dichotomen „wird durchgeführt / wird nicht durchgeführt“-Entscheidungen beruhen. Cronbach stützt sich bei dieser Ansicht auf Belege aus der Forschung über Organisationen und das Zustandekommen von politischen Entscheidungsprozessen. Es zeigt sich dabei, dass Evaluationsstudien eng mit Entscheidungsprozessen verbunden sind. Folgende Schlüsse zieht Cronbach aus seiner Analyse der Verwertung von Evaluationsergebnissen. Seine Erfahrungen aus Evaluationsstudien zeigen Folgendes: „Selten werden Entscheidungen einfach getroffen, gibt es nur einen Entscheidungsträger, haben fundierte Empfehlungen als ein Produkt von Evaluationen einen größeren Einfluss auf Folgeentscheidungen als der politische Diskurs, gibt es „Stop-go“-Entscheidungen zu einem Programm…“ (Shadish, Cook, Leviton 1991, S. 335 [Anm.:

Übersetzung aus dem Englischen durch den Autor]). Auch die sorgfältigste Planung im Vorfeld und ein reibungsloses Projektmanagement können nicht verhindern, dass es im Verlauf von Evaluationsstudien zu zeitlichen Verschiebungen und Änderung der Evaluationsziele und Evaluationsstrategie kommt. Sehr wichtig ist es deshalb, dass das Evaluationsteam weiterhin auch in der Abschlussphase einer Evaluationsstudie den engen Kontakt zu den Auftraggebern der Studie sowie den Projektbeteiligten aufrecht hält. Bei der Aufbereitung der Ergebnisse in der Endphase der Evaluation sollte darauf geachtet werden, mit den Auftraggebern der Evaluation im engen kommunikativen Austausch zu stehen. Der Vorteil dieser Vorgehensweise ist, dass das Evaluationsteam den Auftraggebern bedarfsorientiert die Ergebnisse aufbereiten kann.

8. Ein Prozessmodell für die Planung und Durchführung von Evaluationsstudien

Evaluatoren können auf eine Fülle von Fachliteratur zu Methoden der Evaluation von sozialen Programmen zugreifen. In der Literatur werden Methoden aus der empirischen Sozialforschung im Detail dargelegt und ihre Tauglichkeit für die Evaluation von Programmen eingehend diskutiert. Eine Ausgangsthese dieser Arbeit war, dass in der Literatur – trotz der eingehenden Betrachtung von methodischen Konzepten und Evaluationsdesigns – ein Mangel an forschungspraktischen Hilfestellungen und Optimierungsmöglichkeiten für die Arbeit von Evaluatoren feststellbar ist. Aus der Fachliteratur sind im deutschsprachigen Raum eine Fülle von Anregungen zur Anwendung von geeigneten Methoden und Instrumenten für Evaluationsstudien identifizierbar. Zumeist in Evaluationslehrbüchern werden Anwendungsgebiete verschiedener Evaluationsmethoden im Detail diskutiert. Daneben werden Evaluationsansätze – zumeist aus dem angloamerikanischen Raum – beschrieben und Anwendungsbeispiele in diversen Themenfeldern genannt (z. B. Bildungsförderung, Arbeitsmarktpolitik, Entwicklungshilfe). In der Literatur lassen sich jedoch wenige Handlungsanleitungen identifizieren, die Evaluatoren ganz auf die Forschungspraxis ausgerichtet im Detail in ihrer Arbeit unterstützen.

Die intensive Beschäftigung des Autors mit der Evaluationspraxis offenbarte die Komplexität des Themenfelds. Zu den Hauptelementen einer Evaluationsstudie gehören neben der differenzierten Planung, die Intensität der Zusammenarbeit mit Projektbeteiligten und Auftraggebern, die Berücksichtigung von Effizienz- und Effektivitätskriterien sowie die zielorientierte Zusammenstellung von adäquaten Methoden zu einem Evaluationsdesign. Neben der Auswahl der passenden Methodik sieht sich der Evaluator bei der Durchführung von Evaluationsstudien mit einer Reihe von Herausforderungen kommunikativer und organisatorischer Art konfrontiert – wie dies bereits anhand der Praxisbeispiele in dieser Arbeit illustriert werden konnte. Das Ziel dieser Arbeit ist es, einen Beitrag zur Schließung dieser Lücken in der Praxis zu leisten und ein **theorie- und erfahrungsbasiertes Prozessmodell für die Begleitung von Evaluationsprojekten** zu entwickeln.

Korrespondierend zum Ziel der Arbeit wurden in der Einleitung zunächst mehrere Fragestellungen ausgearbeitet. Die Fragen hatten den Zweck, die Ausarbeitung auf die folgenden Produkte der Arbeit zu fokussieren:

- die Identifikation und Diskussion von relevanten Methoden, Verfahren und Handlungsanleitungen für die Durchführung von Programmevaluationen,

- die Einordnung dieser methodischen Aspekte in eine Abfolge von Handlungen, die in einem Prozessmodell zur Begleitung von Evaluationsprojekten abgebildet werden können und schließlich,
- die Nennung von konkreten Empfehlungen und Entscheidungshilfen, die Evaluatoren zu bestimmten Phasen im Evaluationsprozess nützlich sein können.

Die Arbeit wurde im zweiten Kapitel mit einer einleitenden Betrachtung der historischen Entwicklung der Evaluationstätigkeit in den USA und Deutschland begonnen. Der Schwerpunkt der Darstellung liegt auf der Beschreibung der Entwicklungen in den USA. Zum einen liegt dies im hohen Professionalisierungsgrad der Evaluationstätigkeit in den USA begründet. Zum anderen konnten dadurch Erfahrungen aus der Evaluation von Programmen zur Bildungsförderung für die Diskussion der Evaluationsansätze herangezogen werden.

Die erste Fragestellung der Arbeit wurde in einer chronologisch-systematischen Erörterung einer Reihe von Evaluationsansätzen behandelt. Die Analyse der Entwicklung der Evaluationstätigkeit, die in drei Zeitphasen erfolgte, schließt im darauffolgenden Kapitel mit einer Einteilung der beschriebenen Evaluationsansätze nach Zeitphasen und methodischen Unterscheidungsmerkmalen (u.a. wissenschaftstheoretische Einordnung, angewendete Instrumente, Zweck der Evaluation) ab.

Zunächst geht die Weiterentwicklung der Evaluationsmethode mit einer zunehmenden Komplexität der ausgearbeiteten Evaluationsansätze im Verlauf der letzten Jahrzehnte einher. Wie die Analyse zum Ende des zweiten Kapitels verdeutlicht, ist eine Entwicklung ausgehend von **methodischen und hin zu stakeholder- sowie nutzenorientierten Evaluationsansätzen identifizierbar**. Außerdem kann eine markante **Differenzierung der Evaluationsansätze nach Funktionen, Typen sowie Evaluationsinstrumenten** als ein Ergebnis der Entwicklung der Evaluationstätigkeit festgestellt werden. Die Hauptvertreter der Evaluationstätigkeit in den 60er Jahren wählten Evaluationsmethoden, die Wirkungen der Förderung bei Teilnehmern an sozialen Programmen messen sollten. So erhoffte man sich von quasi-experimentellen Designs mit Untersuchungs- und Kontrollgruppen im empirischen Sinn objektive Aussagen zu erhalten, ob die Durchführung von Förderprogrammen Wirkungen zeigt, die eine weitere Finanzierung der Programme rechtfertigt. Auch Vertreter der objektiven, auf quantitativen Verfahren ausgerichteten Evaluationstätigkeit setzten sich kritisch mit der bis dato angewendeten Evaluationsmethodik auseinander.

Erkenntnisse aus der ersten Phase der Evaluationstätigkeit – die sich hauptsächlich der Umsetzung von Wirkungsevaluationen gewidmet hat – wurden bei der Entwicklung des hier vorgestellten Prozessmodells berücksichtigt. Aus der ersten

Phase kommen insbesondere die methodischen Instrumente für die Durchführung von Wirkungsevaluationen und die damit verbundene Arbeit von Campbell und Scriven in Betracht. Hier spielen die experimentellen und quasi-experimentellen Untersuchungsdesigns nach Campbell eine hervorgehobene Rolle.

Auf die kritische Auseinandersetzung mit rein quantitativen Evaluationsmethoden folgte in den 70er Jahren eine grundlegende Umorientierung bezüglich der methodischen Konstruktion von Evaluationsdesigns. Neben der Integration von qualitativen Methoden in die Evaluationsansätze wuchs die Bedeutung von teilnehmerorientierten Evaluationsverfahren. Guba und Lincoln (1989) betonten die besondere Relevanz der Etablierung von Kommunikationsstrukturen mit Stakeholdern für die Qualität der Studien insgesamt. Evaluatoren wird demnach nahegelegt, gemeinsam mit Projektverantwortlichen Ziele für die Evaluation zu formulieren, Arbeitspläne mit Verantwortlichkeiten zu entwickeln und sich im Verlauf der Evaluation fortlaufend mit den Projektverantwortlichen abzustimmen. Stake schlägt beispielsweise die Entwicklung eines Evaluationsplans vor, der aus mehreren Komponenten besteht und neben methodischen Verfahren insbesondere Instrumente des Projektmanagements vorsieht (vgl. Stake 1975, S. 37). Hier kommt den qualitativen Methoden und Verfahren (z.B. Befragungen, Workshops, Beobachtungen) eine besondere Rolle zu: Aus der Auswertung der erhobenen Informationen soll die Funktionsweise des Programms besser verstanden, die Kontextbedingungen, in dem das Programm umgesetzt wird, erfasst und zudem die ggf. unterschiedlichen Intentionen der am Programm beteiligten Personen festgehalten werden. Die Produkte der zweiten Phase konnten daher für die Konzeption der Grundstruktur des Prozesses sowie für die Festlegung der Prozessteilschritte verwendet werden. Aus dem Evaluationsansatz der partizipativen Evaluation nach Stake ließen sich für das Prozessmodell Elemente herausarbeiten, die für die Zusammenarbeit zwischen Evaluatoren, Programmverantwortlichen und Programmbeteiligten wichtig sind.

Die dritte Phase der Entwicklung der Evaluationstätigkeit zeichnet sich durch die Einführung von theoriegeleiteten Evaluationsansätzen aus, die zielorientiert, auf Basis einer zuvor ausgearbeiteten Programmtheorie, quantitative und qualitative Verfahren kombinieren. Die theoriegeleiteten Ansätze von Rossi und Chen (1990) und der realistische Evaluationsansatz von Pawson und Tilley (1997) sind Beispiele für die differenzierte Erarbeitungen von Evaluationsverfahren, die sich an den Zielen, Zielgruppen und Kontextbedingungen des Programms orientieren. Die Evaluationsansätze der dritten Phase bieten schließlich Lösungen für die systematische Untersuchung von Programmmechanismen (theoriegeleitete Evaluation nach Rossi und Chen) sowie für Evaluationsstudien zu Programmen, die an mehreren Standorten durchgeführt werden. Die genannten Ansätze thematisieren außerdem Lösungen für die **Blackbox-Problematik** bei der Evaluation von sozialen Programmen an. Durch das Instrument der **Programmtheorie** ist, wie

dargestellt, eine detaillierte Behandlung der Programmabläufe in Evaluationsstudien möglich. Nicht umfassende Evaluationen sind in den Vordergrund gerückt, sondern flexible Ansätze, deren Methodenauswahl an die Charakteristika des Evaluationsgegenstands anknüpft und die einen hohen Grad an Teilnehmerorientierung aufweisen. In den Fokus der Betrachtung von Evaluationsprojekten ist der Nutzen von Programmevaluationen gerückt.

Des Weiteren wird bei der Betrachtung der drei Phasen deutlich, dass die Evaluationsforschung eine kontinuierliche Entwicklung durchlaufen hat. Die Schwerpunktsetzung hat sich im Laufe der Zeit verlagert: Insbesondere sind unter dem Stichwort theoriegeleitete Evaluation in der dritten Phase Evaluationsansätze für die Untersuchung von komplexen Programmen entstanden. Während in der ersten Phase Wirkungsmessungen dominierten, wurde das Instrumentarium der Evaluation im Laufe der Jahre um qualitative Methoden erweitert. Diese Entwicklung hatte zwei Konsequenzen: Evaluationsansätze wurden zwar komplexer, andererseits wurden Studien fokussierter auf die spezifischen Erkenntnisinteressen der Programminitiatoren konzipiert. Die Weiterentwicklung des Evaluationsinstrumentariums hatte zudem Konsequenzen auf die Gestaltung der Rolle, die Evaluatoren bei der Durchführung von Evaluationsprojekten einnehmen.

Neben der Integration von theoretischen Erkenntnissen in den Evaluationsprozess war ein weiteres Ziel der Arbeit, Anknüpfungspunkte zur Weiterentwicklung der bestehenden Evaluationspraxis zu liefern. Als zielführend erschien es daher, in Anlehnung an generische **Projektmanagementverfahren**, den typischen Verlauf von Evaluationsstudien in Form eines Prozessablaufes zu schematisieren. Anschließend wurde der zunächst grobe, generische Prozess in einzelne Hauptphasen untergliedert: Planung, Konzeption, Durchführung und Ergebnis. Die Vorteile der Anwendung von prozessorientierten Vorgehensweisen für die Praxis liegen auf der Hand: Evaluationsstudien werden häufig als Auftragswerke vergeben, wodurch bei eng kalkulierten Budgets und engen Zeitvorgaben ein striktes Projektmanagement für den Erfolg der Studien notwendig wird, um das Evaluationsprojekt möglichst effizient und effektiv durchzuführen.

Methodische Verfahren aus den analysierten Evaluationsansätzen wurden den einzelnen Phasen zugeordnet. Dadurch wurde der Prozess um weitere Einzelschritte im Detail erweitert. Zusätzlich wurden Erkenntnisse aus Evaluationsstudien des Autors im Rahmen seiner Tätigkeit am efms zur Vervollständigung des Prozesses herangezogen sowie die gewonnenen Praxiserfahrungen hinsichtlich der angewendeten Methoden erörtert. Das Ergebnis der vorgenommenen Analyse ist ein detaillierter Prozess der Durchführung von Evaluationsstudien. Bei den vorgestellten Praxisbeispielen handelt es sich um Auftragsevaluationen von

Programmen zur Bildungsförderung von Migranten. Alle beschriebenen Bildungsprogramme haben das gemeinsame Ziel, durch eine oder mehrere Maßnahmen auf Bildungsdefizite kompensatorisch zu wirken.

Des Weiteren wurde darauf geachtet, **Prinzipien von Qualitätsmanagementverfahren** in den Prozessverlauf zu integrieren. Dadurch haben die Evaluatoren Instrumente an der Hand, um ihre eigene Evaluationsarbeit zu überprüfen und ggf. zu verbessern, d.h. eine Qualitätsentwicklung ihrer eigenen Arbeit zu betreiben. Für Programmverantwortliche bieten Qualitätsmanagementprozesse die Möglichkeit, ihre Anforderungen an die Evaluation zu kommunizieren, den Ablauf des Verfahrens zu begleiten sowie mit den resultierenden Evaluationsergebnissen steuernd im Programm einzugreifen.

Zudem sind die Phasen im Prozessablauf so gestaltet, dass die Mehrzahl der Einzelschritte und Prüfmechanismen in der **Eingangsphase der Evaluationsstudie** verortet wurden. Dies ist mit den Analyseergebnissen der Evaluationsansätze in den Eingangskapiteln sowie den Praxiserfahrungen zu begründen. Fehler sowie ineffiziente Vorgehensweisen in späteren Phasen des Prozesses lassen sich durch geeignete Maßnahmen in der Eingangsphase des Prozesses minimieren. In diesem Sinne ist mit der Anwendung des Qualitätsmanagementprozesses eine **Optimierungsfunktion** der Evaluationstätigkeit verknüpft.

Im Rahmen dieser Arbeit ist auf diese Weise eine Prozessbeschreibung entstanden, die sich aus den vier Phasen **Eingangsphase der Evaluation, Entwicklung des Evaluationsdesigns, Umsetzungsphase und Ergebnisphase** zusammensetzt. Das spezifisch Neue an dem vorgeschlagenen Prozess ist die theorieorientierte und in der Praxis erprobte Vorgehensweise bei der Erarbeitung der Einzelschritte sowie die schematisch dargestellte Transparenz in der Kommunikation zwischen Evaluator und Projektverantwortlichen.

Phasen und Arbeitsschritte im Evaluationsprozess

Evaluator / Evaluationsteam

Auftraggeber / Programmmanager

1. Eingangsphase der Evaluation

1.1. Informationssammlung (S.77 ff.)

• Liegen Informationen zu Zielen und Inhalten in Form von Konzepten, Programmbeschreibungen etc. vor?
• Basiert das Konzept des Programms auf einer wiss. Theorie?
• Wurden bisherige Ergebnisse systematisch festgehalten?
• Gibt es Informationen zu den Evaluationszielen?
• Wurden zu dem Programm bereits Evaluationen durchgeführt?

1.2. Strukturierung der gesammelten Information:

• Eine Beschreibung des Programms nach den Kriterien von Cronbach (1980) anlegen: Units, Treatments, Observing operations, Settings
• Bei Ausschreibungen: Unterlagen prüfen (u.a. Vergabekriterien, Ziele der Evaluation, Fördervolumen und Förderbedingungen)

1.3. Detaillierte Erarbeitung von Programm- und Evaluationszielen

• Intention und Ziele der Evaluation erfassen, „goal free evaluation" nach Scriven (1991)
• Priorisierung der Ziele der Programmverantwortlichen
• Verständigung auf Evaluationszielen mit den Programmverantwortlichen
• Hilfsinstrument: SMARTe-Ziele definieren

1.4. Einbindung der Programmbeteiligten und Programmverantwortlichen

• Art und Grad der Einbindung von Programmbeteiligten festlegen; siehe Responsive Evaluation nach Stake
• Einholen der unterschiedlichen Zielvorstellungen

1.5. Entwicklung einer Programmtheorie (Chen 1987, Pawson & Tilley 1997)

• Erarbeitung des Action Model: Rollen und Funktionen der Akteure im Programm
• Erarbeitung des Change Model: Wirkungstheorie des Programms
• Hilfsinstrument: Program Theory Matrix nach Funnell (2000) auf Basis der zuvor entwickelten SMARTen Ziele. Nennung/Festlegung von:
  - Erfolgsfaktoren: Programmmerkmale zur Bestimmung des Erfolgs
  - Faktoren außerhalb des Programms, die Einfluss auf den Erfolg haben können
  - Form und der Intensität der Maßnahmen und Aktivitäten im Programm
  - Daten- und Informationsquellen zur Outcome-Erhebung und während der Programmdurchführung

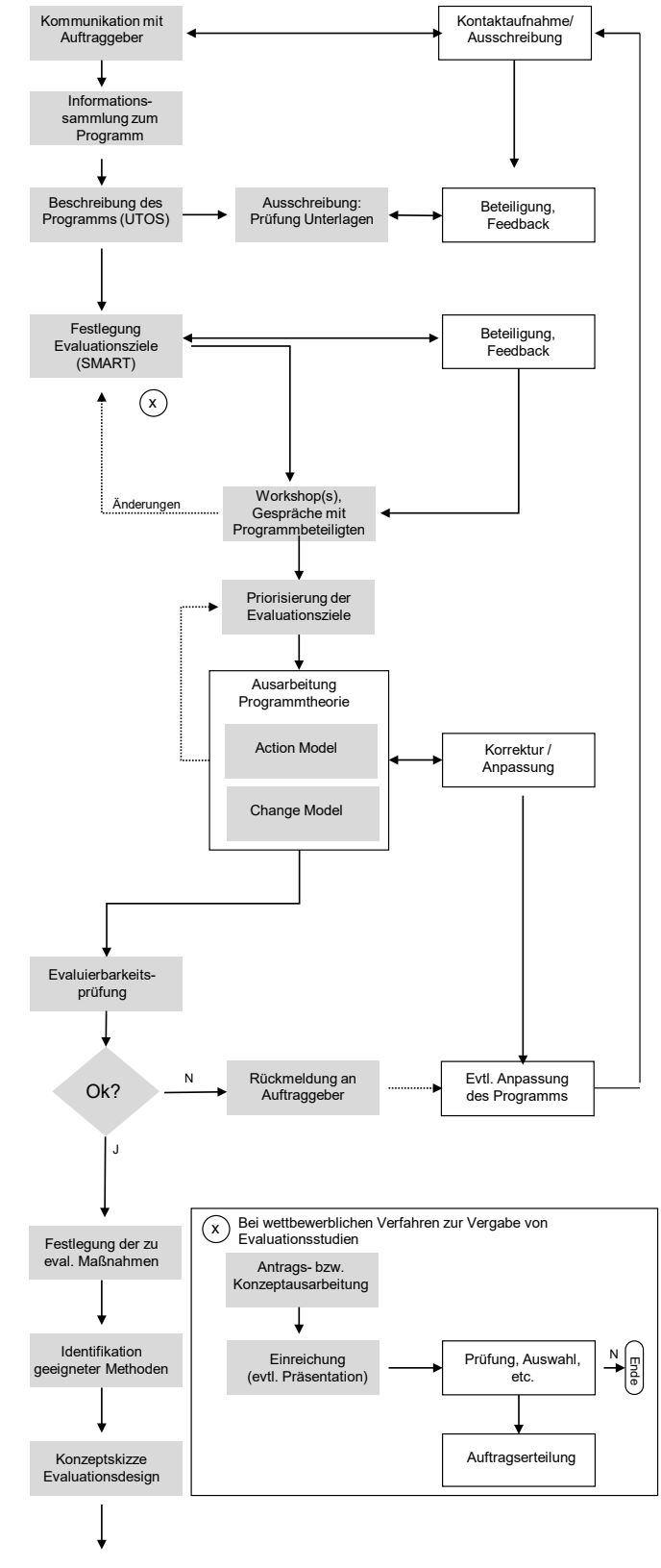
• Optional: Übertragung der Programmtheorie in einen schematischen Ablauf (Grafik) (siehe Chen 2004)

1.6 Durchführung einer Evaluierbarkeitsprüfung (Wholey 1970)

Leitfrage: *Kann die Evaluation mit den geplanten Ressourcen unter den gegebenen Charakteristika des Programms durchgeführt werden, so dass fundierte Aussagen zu den zuvor definierten Erkenntnisinteressen möglich sind?*
• Durchführung notwendiger Abpassungen im Programm
• Plausibilitätsprüfung der Programmtheorie
• Identifikation der Aufwand-Nutzen-Relation
• Prüfen des finanziellen Spielraums für die Evaluationsstudie
• Festlegung der zu evaluierenden Maßnahmen
• Erste Identifikation geeigneter Evaluationsmethoden, Identifikation von alternativen Verfahren

Speziell für die Wirkungsanalysen: *Kann eine Wirkungsanalyse effizient durchgeführt werden, so dass valide und objektive Ergebnisse zu erwarten sind?*

• Größe der Stichprobe/Fallzahlen für Untersuchungs- und Kontrollgruppe
• Ist eine detaillierte Charakterisierung der Teilnehmer möglich?
• Ist die Bildung von Untersuchungs- und ggf. Kontrollgruppen möglich?
• Ideales Design der Wirkungsmessung (u.a. Anzahl mehrerer Messungen)
• Ist eine Identifikation von programminternen und -externen Störfaktoren möglich?
• Stabilität der Programmdurchführung während der Evaluation erreichbar?
• Haben die Teilnehmer in der Vergangenheit bereits an ähnlichen Programmen teilgenommen?



Weiter mit der Entwicklungsphase

2. Entwicklung des Evaluationsdesigns

2.1. Evaluationsformen und Entwicklungsstufen des Programms

- Analyse des Entwicklungsstands des Programms
- Welche Funktion soll die Evaluationsstudie haben?
  - Informationen für die Weiterentwicklung von Programmen
  - Kontrolle der Programmdurchführung
  - Erfassung von Programmwirkungen bei den Teilnehmern

2.3. Auswahl von Methoden zum Zweck der Erfolgskontrolle und Weiterentwicklung von Programmen

- Entscheidung für bestimmte qualitative und quantitative Methoden auf Basis der entwickelten Programmtheorie und den Zielen der Evaluation
- Entwurf eines Erhebungsplans nach den folgenden Kriterien:
  - Funktion der Evaluation,
  - Zielgruppen,
  - Instrumente,
  - Gegenstand der Erhebung.
- Operationalisierung der Erhebungsmerkmale auf Basis der Programmtheorie
- Fertigstellung der Erhebungsinstrumente

- Auswahl der Evaluationsmethoden bei Wirkungsmessungen auf Basis der Programmtheorie (vgl. Campbell & Stanley 1963):
  - Experiment
  - Quasi-Experiment

- Verfahren zur Bildung der Untersuchungs- und Kontrollgruppe konzipieren (Campbell & Stanley 1963):
  - Randomisierung bei Experimenten
  - Matching-Verfahren bei Quasi-Experimenten
  - Ggf. Quotierung bei Quasi-Experimenten
  - Anzahl der Messungen festlegen
  - Umgang mit möglichen externen Störfaktoren
- Ggf. Auswahl von standardisierten, getesteten Erhebungs- bzw. Testverfahren
- Erstellung eines Zeitplans für die Erhebungen

3. Durchführungsphase

3.1 Vorbereitung der Erhebungsphase

- Planung des Vorgehens ggf. in mehreren Wellen
- Abgleich des Erhebungsplans mit der Programmtheorie
- Vor-Ort-Besuche durch den Evaluator/das Evaluationsteam
- Schulung von Personal, das die Evaluation unterstützt
- Abklärung rechtlicher Aspekte (z.B. Datenschutz)
- Information an Programmbeteiligte über die bevorstehende Datenerhebung
- Organisation zusätzlicher personeller Unterstützung für die Datenerhebung

3.2 Anwendung der Instrumente

- Ermittlung der Charakteristika der Teilnehmer; darunter auch spezifische Merkmale (z.B. demographische Daten zur sozialen und familiären Situation)
- Anwendung der Instrumente entsprechend des Evaluationsdesigns

Bei Wirkungsanalysen:
- Durchführung von Pretests
- Durchführung der begleitenden Datenerhebungen
- Auswertung der ersten vorliegenden Daten und Erstellung von Kurzberichten
- Analyse von möglichen externen Störfaktoren sowie deren potentiellen Einfluss auf das Programm
- Durchführung der Posttests
- Analyse von möglichen externen Störfaktoren sowie deren potentiellen Einfluss auf das Programm

4. Ergebnisphase

4.1 Vorbereitung der Ergebnisphase

- Erstellung eines Auswertungsplans
- Zusammenfassung der Grobergebnisse
- Auswertung der ersten vorliegenden Daten und Erstellung von Kurzberichten
- Detailauswertung der erhobenen Daten und Informationen hinsichtlich der Erkenntnisinteressen/SMARTen Zielen der Evaluationsstudie
- Analyse von möglichen externen Störfaktoren sowie deren potentiellen Einfluss auf das Programm:
  - Qualität des Test/Messinstruments
  - Kontext-Merkmale der Intervention
  - Resistenz der Untersuchungs- und Kontrollgruppe bezüglich externer Einflüsse
  - Diskussion der Ergebnisse mit dem Auftraggeber

4.2 Verwendung der Ergebnisse für die Weiterentwicklung der Programmtheorie

- Abgleich der Ergebnisse mit der Programmtheorie:
  - Überprüfung der SMARTen Ziele
  - Überprüfung der Förderdauer und -intensität
  - Analyse der Programmdurchführung
- Weiterentwicklung der Programmtheorie auf Basis der Erkenntnisse
- Ggf. Planung einer Follow-up Evaluationsstudie, um die Programmtheorie zu validieren
- Zusammenstellung der finalen Ergebnisse
- Evaluationsansatz von Weiss (1974) zum Umgang mit nicht erwarteten Evaluationsergebnissen beachten
- Erstellung von Berichten, Präsentationen, Publikationen

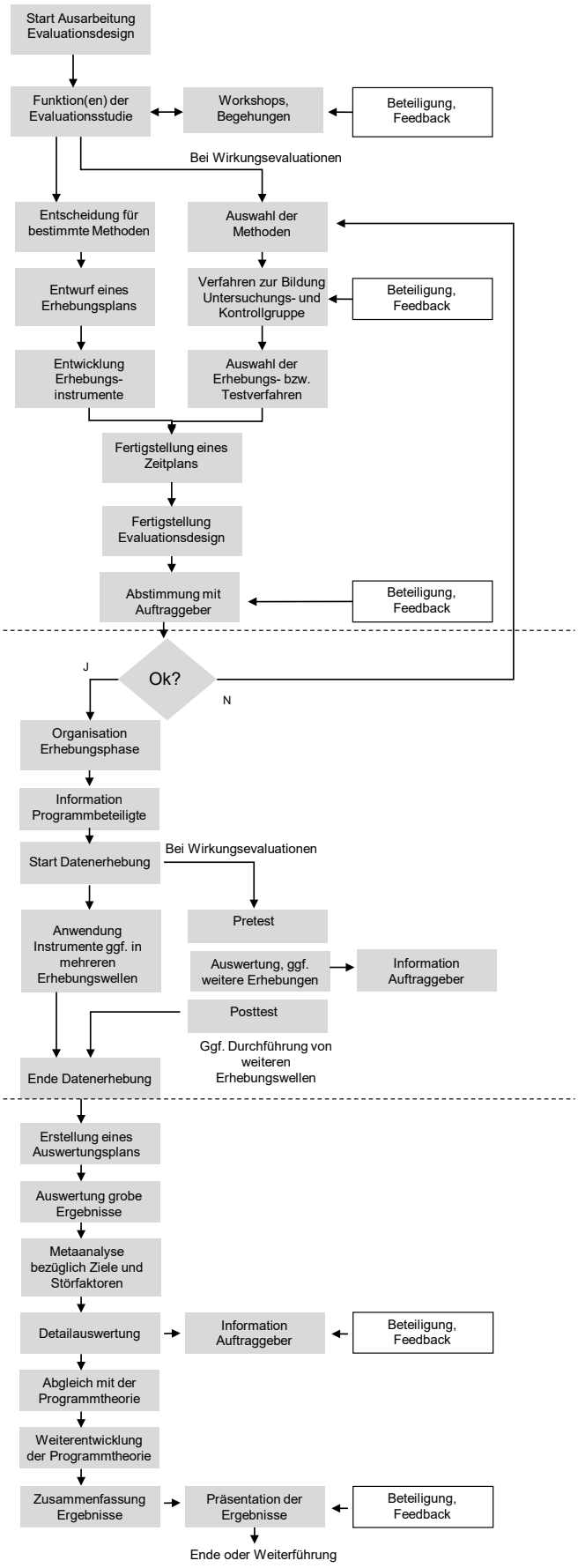


Abbildung 20: Leitfaden für die Konzeption, Durchführung und den Abschluss von Programmevaluationen

Die Abbildung gliedert sich in zwei Seiten. Auf der linken Seite sind die vier Hauptphasen und die jeweiligen Einzelschritte im Evaluationsprozess dargestellt. Relevante methodische Aspekte und Kriterien der Durchführung – die mit der ersten zentralen Fragestellung in dieser Arbeit adressiert wurden – sind stichpunktartig darunter aufgeführt. Auf der rechten Seite ist der Prozess als Flussdiagramm in einem idealtypischen Ablauf abgebildet. Die Kommunikation zwischen Auftraggeber der Evaluation und Evaluatoren wird durch die Pfeilrichtungen illustriert.

Ein zentrales Ergebnis der Metaanalyse der Evaluationsansätze ist, dass die **Eingangsphase eine zentrale Phase im Evaluationsprozess** darstellt und für das spätere Gesamtergebnis der Evaluationsstudie ausschlaggebend sein kann. Die Analyseergebnisse der untersuchten Evaluationsansätze sowie die Erfahrungen aus der Praxis lassen sich zu der Schlussfolgerung zusammenfassen, dass einige während dieser Phase getroffene Entscheidungen zur Planung und Entwicklung des Evaluationsdesigns später nicht mehr revidierbar sind. Werden bestimmte Aspekte und Arbeitsschritte in der Vorbereitungsphase nicht eingeplant, dann ist eine Korrektur oder Anpassung des Evaluationsdesigns zu einem späteren Zeitpunkt im Projekt zumeist nicht mehr möglich, ohne dass sich der Charakter und die Zielsetzung der Studie verändert.

Die beschriebenen Arbeitsschritte in der Anfangsphase einer Programmevaluation sowie die Wahl der geeigneten Evaluationsmethoden sind Stellschrauben für die Qualität einer Evaluationsstudie. Durch die intensive Auseinandersetzung mit dem prozesshaften Charakter einer Evaluationsstudie wurde deutlich, dass noch lange vor der Entwicklung von Erhebungsinstrumenten die eigentliche Arbeit des Evaluators sich zuerst auf die Kommunikation mit den Auftraggebern beschränken sollte sowie auf die Identifikation der Art und Weise, wie sich das zu untersuchende Programm in der Praxis manifestiert. In der Eingangsphase der Evaluation sind daher insbesondere drei Arbeitsschritte beachtet zu beachten:

(1) Die gemeinsame Definition der Evaluationsziele mit den Programmverantwortlichen.

(2) Das Zusammentragen aller zur Verfügung stehenden Informationen über das Evaluationsobjekt, um Informationslücken zu identifizieren.

(3) Die Entwicklung einer Programmtheorie gemeinsam mit dem Auftraggeber, um die Funktionsweise des Programms zu verstehen.

Wurde die Evaluation als Auftragswerk öffentlich ausgeschrieben, bildet die Vorbereitung auf den Wettbewerb um das Projekt einen weiteren Schwerpunkt in der ersten Phase. Die Evaluationsansätze von Scriven und Cronbach eignen sich besonders gut, um einen Einstieg in die methodische Auseinandersetzung mit

dem Programm und möglichen Evaluationsmethoden zu erhalten. Scriven plädiert mit dem Ansatz der „goal free evaluation", dass Evaluatoren aus einem sachlich abstrakten Blickwinkel ggf. existierende Zieldefinitionen zum Programm übergehen und das Evaluationsobjekt von Grund auf hinsichtlich des Konzepts und der Prozesse und Produkte analysieren.

Das Programm sollte in seiner Komplexität erfasst und beschrieben werden. Besonders viel Wert sollte bei der Beschreibung auf die Maßnahmen und die an dem Programm Beteiligten gelegt werden. Was genau ist die Intervention am Programm? Was soll gefördert werden? An wen richtet sich das Programm? Wie ist das Programm organisiert?

Die Intentionen der Programmverantwortlichen und der Programmbeteiligten erfassen. Entscheidend für die Arbeit von Evaluatoren ist es, ein Gefühl davon zu erhalten, wie das zu evaluierende Programm bei den Programmverantwortlichen wahrgenommen wird. Unterscheidet sich die Sichtweise auf das Programm bei Personen, die das Programm verantworten müssen von den Personen, die für die Umsetzung zuständig sind?

Es sollten messbare Ziele definiert werden. Es ist wahrscheinlich, dass in einem Programm gleich mehrere strategische und inhaltliche, mit einem Konzept verbundene Ziele verfolgt werden. Es ist des Weiteren anzunehmen, dass am Programm beteiligte Personen (z.B. Programmkoordinatoren) das Programm unter anderen Gesichtspunkten bewerten als Programmverantwortliche. Vorgeschlagen wird daher die Anwendung des Konzepts von SMARTen Zielen. Mit dieser Strukturierungstechnik können Ziele so festgelegt werden, dass sie spezifisch, messbar, eindimensional und zeitlich erreichbar sind.

Die Programmtheorie sollte zunächst in Zusammenarbeit mit den Stakeholdern erarbeitet werden. Aus den gesammelten Informationen zum Programm sowie den Zielen lässt sich ein erster Entwurf der Programmtheorie erarbeiten. Die Entwürfe der ausgearbeiteten Programmtheorie lassen sich im iterativen Verfahren mit Unterstützung der Programmbeteiligten weiterentwickeln. Die **Evaluierbarkeitsprüfung sollte frühzeitig durchführt werden.** Es empfiehlt sich, Wirkungsevaluationen nur dann durchzuführen, wenn alle Faktoren – wie in Kapitel 1.6. der Abbildung beschrieben – gegeben sind. Als Ausgangspunkt zur generellen Prüfung, ob das Programm evaluierbar ist, kann die ausgearbeitete Programmtheorie herangezogen werden.

Die ausgearbeitete Programmtheorie ist zugleich der Ausgangspunkt für eine Machbarkeitsuntersuchung in Form einer Evaluierbarkeitsprüfung. Grundlage für die Prüfung sind (SMARTe) Ziele, eine ausgearbeitete Programmtheorie sowie eine Kosten-Nutzen-Abschätzung der geplanten Evaluationsstudie. Das Ergebnis dieser Prüfung stellt die Erkenntnis dar, ob und mit welchen Methoden eine Evaluation des Programms grundsätzlich realisierbar wäre.

Das Prüfergebnis mündet in einer Vereinbarung des Evaluators mit dem Auftraggeber, welche Evaluationsziele und Aspekte eines Programms evaluiert werden sollen, z.B. kann die Entscheidung für eine programmbegleitende Evaluation zum Zweck der Qualitätsentwicklung gefällt werden, da das Programm sich in einer frühen Implementationsphase befindet. Dies war der Fall im Praxisbeispiel der evaluierten Kursreihe „In Deutschland zu Hause". Dort kamen mehrere Evaluationsmethoden simultan zum Einsatz. Korrespondierend zu den Zielen und zum Implementationsgrad der Kursreihe wurde zunächst der Schwerpunkt auf formative Evaluationsverfahren gelegt, um die kontinuierliche Weiterentwicklung der Kurscurricula sicherzustellen. Nach zwei Durchläufen der Kursreihe wurde der Fokus weg von teilnehmerorientierten Feedbackbefragungen und hin zur Methode der teilnehmenden Beobachtung verlegt. Ziel war es dann, motivationale Aspekte des Engagements der Kursteilnehmer sowie Lernprozesse im Detail zu erfassen.

In einem anderen Beispiel besteht der Wunsch des Auftraggebers nach einer **Wirkungsevaluation.** In diesem Fall empfiehlt sich bei der Evaluierbarkeitsprüfung außerdem zu beurteilen, ob genügend Ressourcen, Zeit sowie weitere relevante Rahmenbedingungen für die Durchführung von Wirkungsmessungen vorhanden sind. Im Unterschied zu Evaluationsstudien, die ein anderes Ziel als die Wirkungsmessung eines Programms bzw. einer Fördermaßnahme bei den Teilnehmern verfolgen, kommen für Wirkungsevaluationen nur wenige, spezielle Verfahren in Betracht. Die Evaluationsmethode beinhaltet in diesem Fall zumindest eine Messung der mit dem Programm intendierten Fördereffekte sowie einen Kontrollgruppenvergleich.

Evaluationsstudien zur Identifikation von Wirkungen werden insbesondere dann in Auftrag gegeben, wenn Programmwirkungen nicht offensichtlich sind. Das primäre Instrument zur Erhebung von validen und reliablen Daten sind daher standardisierte Test- oder Prüfverfahren unter direkter Beteiligung der Teilnehmer. Im Rahmen der zweiten Evaluierbarkeitsprüfung sollten bei experimentellen und quasi-experimentellen Designs folgende Aspekte geprüft werden:

- Möglichkeit der Erfassung von Teilnehmercharakteristika (sowohl Untersuchungs- als auch Kontrollgruppe)
- Realisierbare Größe der Untersuchungs- bzw. Kontrollgruppe
- Verfahren zur Bildung der Untersuchungs- und Kontrollgruppe
- Intensitätsgrad der Maßnahme (Dauer, Mitwirkung der Teilnehmer im Programm)
- Benötigte zeitliche und finanzielle Ressourcen
- Anzahl und Zeitpunkte realisierbarer Messungen

- Möglichkeit der Isolierung von Störfaktoren

Die Analyse der Praxisbeispiele hat ergeben, dass außerdem die Größe der Untersuchungs- bzw. Kontrollgruppe, die Anzahl und Zeitpunkte der Messungen sowie die Qualität der Wirkungsmessung einen großen Einfluss auf die Qualität der erfassten Daten hatten. Führt die Prüfung einer Wirkungsanalyse zu einem negativen Ergebnis (z.B. weil nicht genügend Teilnehmer für eine Kontrollgruppe rekrutiert werden können), sollte gemeinsam mit dem Auftraggeber geprüft werden, ob es eine alternative Zielsetzung für das Evaluationsvorhaben gibt.

Im Anschluss an die Eingangsphase der Evaluation folgt die Ausarbeitung des Evaluationsdesigns. Ausgehend von den Zielen, Zielgruppen und Erkenntnisinteressen werden die passenden Methoden zur Datenerhebung ausgewählt. Zu berücksichtigen ist dabei der Entwicklungsstand des Programms. Die Evaluationsstudie kann bis zu drei verschiedene Funktionen erfüllen: **Evaluation zur Zweck der Weiterentwicklung des Programms, Evaluation zur Kontrolle sowie Evaluation zur Wirkungsmessung**. Die Operationalisierung der Evaluationsziele und die Konstruktion der Instrumente sind weitere Einzelschritte, die mit den Stakeholdern abgestimmt werden. Bei Wirkungsanalysen fällt in dieser Phase die methodische Vorbereitung der Wirkungsmessungen an. Dazu zählen bei experimentellen und quasi-experimentellen Untersuchungen u.a. die Bildung der Untersuchungs- und Kontrollgruppe, die Auswahl bzw. Entwicklung des Messinstrumentariums sowie die Ausarbeitung eines Messzeitplans. Auch in dieser Phase ist eine enge Stakeholdereinbindung für die Qualität der Evaluationsstudie entscheidend. Dadurch können Erwartungen von Auftraggebern und Programmbeteiligten nochmals geprüft und ggf. bei der sich anschließenden Datenerhebung berücksichtigt werden.

Die eigentliche Umsetzung der Evaluation umfasst sowohl die Erhebungs- als auch die Auswertungsverfahren. Vergleichbar zur Erstellung eines Messzeitplans bei Wirkungsanalysen können Auswertungszeitpläne mehr Transparenz und Planungssicherheit für Auftraggeber vermitteln. Die dritte Evaluationsphase geht nahezu fließend in die vierte und abschließende Phase über, bei der es um die Kommunikation und Präsentation der Ergebnisse geht. Neben einer Grob- und Detailauswertung beinhaltet diese letzte Phase eine Metaevaluation der Programmtheorie. Hier sollte die ursprünglich, zu Beginn des Evaluationsverfahrens entwickelte Programmtheorie hinsichtlich ihrer weiteren Gültigkeit überprüft werden. Es bietet sich an, nach der Revision der Programmtheorie das Evaluationsverfahren mit dem Ziel der weiteren Validierung wieder aufzunehmen.

Als Resümee der inhaltlichen Betrachtung des gesamten Prozesses für die Evaluation von sozialen Programmen kann die Schlussfolgerung formuliert werden, dass für die Umsetzung in der ersten Evaluationsphase der **Verwertungsnutzen**

der späteren Evaluationsergebnisse im Vordergrund steht. Der Verwertungsnutzen der Ergebnisse spiegelt zudem das Hauptinteresse der Auftraggeber wider. Dementsprechend unterliegen die anschließende Planung des Evaluationsvorhabens, die Entwicklung der Evaluationsmethoden sowie die Zusammenstellung der Ergebnisse engen Abstimmungsprozessen. Methodische Aspekte der inhaltlichen Durchführung von Evaluationsvorhaben – zu denen die Operationalisierung der zentralen Fragestellungen, die Instrumentenentwicklung, die Erhebung sowie die Auswertung zählen – spielen in der Entwicklungsphase des Evaluationsdesigns im Evaluationsprozess eine hervorgehobene Rolle.

Neben den Vorteilen, die ein Leitfaden für die Programmevaluation mit sich bringen kann, sollte an dieser Stelle auch auf **mögliche Restriktionen** bei der Anwendung aufmerksam gemacht werden. Der Leitfaden eignet sich primär für die Anwendung bei Programmevaluationen, die von externen Evaluatoren (nicht der Organisation, die das Programm initiieren, verantworten oder zugehörigen Einheiten) durchgeführt werden. Bei folgenden Aspekten ist die Anwendung der einzelnen Arbeitsschritte im Leitfaden individuell zu überprüfen:

- *Externe Vorgaben:* Programme können externen Vorgaben, Rechtsvorschriften oder Gesetzen unterliegen, die bei der Evaluation berücksichtigt werden müssen. Als Beispiel kann der Hochschulbereich genannt werden. Studiengänge (im gewissen Sinne Programme nach der hier getroffenen Definition) müssen so konzipiert werden, dass u.a. die Strukturvorgaben der Kultusministerkonferenz, Akkreditierungsgrundsätze sowie die Ländergesetzgebungen eingehalten wer-den. Nicht immer kann das methodisch am sinnvollsten erachtete Evaluationsverfahren zum Zuge kommen.
- *Interne vs. externe Evaluation:* Die Planung und Durchführung von Evaluationsstudien orientiert sich an Organisationsstrukturen und -zugehörigkeiten. Wird mit der Durchführung einer Evaluationsstudie eine externe Organisation beauftragt, sind die Zielsetzungen ggf. andere, als dies bei internen Evaluationen der Fall ist. Es erscheint sinnvoll, mit Wirkungsevaluationen von sozialen Programmen externe Dienstleister zu beauftragen; programmbegleitende Evaluationen zum Zweck der Kontrolle und Weiterentwicklung (z.B. der Curricula, Inhalte) dagegen intern zu betreiben. Die Kommunikation zwischen Programm-beteiligten und Evaluatoren kann bei internen Evaluationen von dem im Leitfaden vorgeschlagenen Ablauf abweichen.
- *Zeit- und Budgetfragen:* Der erarbeitete Leitfaden schildert einen Evaluations-ablauf, der von Programmverantwortlichen und Evaluatoren als kontinuierliches Entwicklungsprojekt verstanden werden kann. Bereits angesprochen wurde die Schwierigkeit, mit der Evaluatoren umgehen müssen, wenn Wirkungsevaluationen zu einem Zeitpunkt angesetzt sind, nachdem

das Programm schon gestartet ist (siehe Projekt *frühstart*). Eine weitere Herausforderung für die Konzeption einer Evaluationsstudie ist es, wenn die Umsetzung eines Evaluationsdesigns mit den zur Verfügung stehenden Mitteln nicht realisierbar ist. Im konkreten Fall von „Spielend lernen" wurden daher nur Teile des gesamten Programms evaluiert.

Bei Konfrontation des Evaluators mit den geschilderten Restriktionen empfiehlt sich ein pragmatischer Umgang bei der Suche nach Lösungsmöglichkeiten. Auch unter Wahrung methodischer Standards bietet der Leitfaden Evaluatoren und Programmverantwortlichen genügend Anpassungsspielraum. Der Leitfaden, als zentrales Ergebnis dieser Arbeit, soll Evaluatoren in einer an methodischen Aspekten orientierten, jedoch **situationsbezogen pragmatischen anzupassenden Vorgehensweise**, eine Hilfestellung leisten. Der Fokus wurde dabei auf eine starke Nutzungsorientierung bezüglich der Verwendung der Evaluationsergebnisse gelegt sowie auf Formen der direkten Einbindung von Programmverantwortlichen in den Prozess der Evaluation. Der Leitfaden ist flexibel gestaltet, so dass er gleichermaßen für programmbegleitende als auch für Wirkungsevaluationen eingesetzt werden kann.

9. Literaturverzeichnis

Alkin M. C., Christie C. A. (2004): *An evaluation theory tree*, in: Alkin M. C. (Hrsg.): Evaluation Roots. Thousand Oaks, CA: Sage.

Artelt C., Stanat P. (2010): *Leistungen von Schülerinnen und Schülern im internationalen Vergleich – Die PISA Studie*, in: Spiel C. (Hrsg.): Bildungspsychologie. Göttingen: Hogrefe, S. 352-356.

Astbury B., Leeuw F. L. (2010): *Unpacking blackboxes: mechanisms and theory building in evaluation*. American Journal of Evaluation, 31, (3), S. 363-381.

Balk M. (2000): *Evaluation von Lehrveranstaltungen. Die Wirkung von Evaluationsrückmeldungen*. Peter Lang GmbH, Frankfurt am Main.

Barker Bausell R. (1986): *A Practical Guide to Conducting Empirical Research*. New York, NY : Harper & Row.

Baumert J., Klieme E., Neubrand M. (Hrsg.) (2001): *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske+Budrich.

Becher G., Kuhlmann S. (Hrsg.) (1995): *Evaluation of Technology Policy Programs in Germany*. Boston, MA: Kluwer Academic Publishers.

Bentler P., Woodward J.A. (1978): *A Head Start reavaluation: positive effects are not yet demostrable*. Evaluation Quaterly, S. 493-510.

Berger R., Holler-Zittlau I., Dux W. (2004): Untersuchungen zum Sprachstand vierjähriger Vorschulkinder. Aktuelle phoniatrisch-pädaudiologische Aspekte, in: Bd.12. Berger R, Holler-Zittlau I. (Hrsg.): Ergebnisse einer Folgeuntersuchung der im Jahre 2003 ermittelten sprachauffälligen Vorschulkinder. 21. wissenschaftliche Jahrestagung der DGPP 2004. http://www.egms.de/en/meetings/dgpp2004/04dgpp68.

Berger R., Holler-Zittlau I., Dux W. (2006): *Marburger Sprach-Screening für 4- bis 6-jährige Kinder (MSS): Ein Sprachprüfverfahren für Kindergarten und Schule*. Persen Verlag.

Bergmann M. M., Cattacin S., Läubli-Loud M. (1998): *Evaluators Evaluating Evaluators: Peer-assessment and Training Opportunities in Switzerland*. Working paper 1/98. resop - université de genève

Beywl W. (1988): *Zur Weiterentwicklung der Evaluationsmethodologie. Grundlegung, Konzeption und Anwendung eines Modells der responsiven Evaluation*. Frankfurt a. M.: Peter Lang Verlagsgruppe.

Beywl W., Speer S., Kehr J. (2003): *Wirkungsorientierte Evaluation im Rahmen der Armuts- und Reichtumsberichterstattung*. Perspektivstudie. Im Auftrag des Bundesministeriums für Gesundheit und Soziale Sicherung (BMGS). Köln. http://www.univation.org/download/Evaluation_der_Armuts-_und_Reichtumsberichterstattung.pdf.

Bloom B. (1964): *Stability and Change in Human Characteristics*. New York, Wiley.

Blossfeld H.-P., Roßbach H.-G., von Maurice J. (Hrsg.) (2011): *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)*. Springer Verlag.

Bonate Peter L. (2000): *Analysis of Pretest-Posttest Designs*. Chapman & Hall/CRC.

Borrmann A., Stockmann, R. (2009): *Evaluation in der deutschen Entwicklungszusammenarbeit*. Band 1: Systemanalyse. Band 2: Fallstudien. Waxmann Verlag. Münster, New York, Berlin, München.

Bortz J., Döring N. (2003): *Forschungsmethoden und Evaluation*. 2. Auflage. Berlin, Heidelberg, New York: Springer.

Boruch R. F. (1997): *Randomized experiments for planning and evaluation: A practical guide*. (Applied Social Research Methods Series, Volume 44.) Thousand Oaks, CA: Sage Publications.

Böttcher W., Holtappels H. G., Brohm M. (2006): *Evaluation im Bildungswesen: Eine Einführung in Grundlagen und Praxisbeispiele*. Weinheim und München. Juventa Verlag.

Brandt T. (2009): *Evaluation in Deutschland. Professionalisierungsstand und -perspektiven*. Münster: Waxmann.

Bronfenbrenner U. (1976): *Is Early Intervention Effective?* Teachers College Record 76, no. 2, S. 279-303.

Buschhoff C. (2009): *Evaluation von Verwaltungsmodernisierung. Empirische Ergebnisse auf Grundlage der Binnenmodernisierung in einer Landesverwaltung*. Peter Lang GmbH, Frankfurt am Main.

Bussmann W. (1997): *Evaluationen und intensive demokratische Beteiligung: Ergänzung oder Ersatz?*, in: Swiss Political Science Review 3(2), S. 1-101.

Bussmann W., Klöti U., Knoepfel P. (Hrsg.) (1997): *Einführung in die Politikevaluation*. Basel, Fankfurt am Main, Helbing & Lichtenhahn.

Caliendo M., V. Steiner (2005): *Aktive Arbeitsmarktpolitik in Deutschland: Bestandsaufnahme und Bewertung der mikroökonomischen Evaluationsergebnisse*. Zeitschrift für Arbeitsmarktforschung, 38 (2/3), 396–418.

Campbell D. T., Stanley, J. C. (1966): *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Campbell D. T. (1969): *Reform as Experiments*, in: American Psychologist. Jg. 24, H. 4, S. 409-429.

Campbell D. T., Erlebacher A. E. (1970): *How regression artifacts can mistakenly make compensatory education programs look harmful*, in: Hellmuth J. (Hrsg.): The Disadvantaged Child: Vol. 3, Compensatory education: A national debate. New York: Brunner/Mazel. S. 185-210.

Campbell D. T. (1974): *Evolutionary Epistemology*, in: The philosophy of Karl R. Popper, in: Schilpp P. A. (Hrsg.) LaSalle, IL: Open Court, S. 412–463.

Campbell D. T (1975): *Qualitative Knowing in Action Research*. Journal of Social Issues.

Campbell D. T., Boruch R. F. (1975): *Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects*, In: Bennett C. A., Lumsdaine A. A. (Hrsg.): Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. New York: Academic Press, S. 195-296.

Campbell D. T. (1988): *Methodology and epistemology for social science*. Chicago, Univ. of Chicago Press.

Campbell D. T., Cook T. E., Shadish W. (2001): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing.

Chambers B., Cheung A., Slavin R., Smith D., Laurenzano M. (2010): *Effective Early Childhood Education Programs: A Systematic Review*. Best Evidence Encyclopedia (BEE). http://www.bestevidence.org/word/early_child_ed_Sep_22_2010.pdf.

Chelimsky E. (1997): *The coming Transformations in Evaluation*, In: Chelimsky E., Shadish W.R. (Hrsg.): Evaluation for the 21st century: a handbook. Sage Publishers.

Chen H. T. (1990): *Theory-Driven Evaluations*. Sage Publishers.

Chen H. T., Rossi P. H. (1983): *Evaluating with sense: The Theory-Driven Approach*, in: Evaluation Review. Jg. 7, S. 283-302.

Chen H. T., Rossi P. H. (1987): *The theory-driven approach to validity*, in: Evaluation and Program Planning, 10, S. 95–103.

Chen H.-T. (2004): *The roots of theory-driven evaluation. Current views and origins*, in: Alkin M. (Hrsg.): Evaluation roots. Tracing theorists' views and influences. Thousand Oaks, CA: Sage Publications. S. 132–152.

Chen H.-T. (2005): *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage Publications.

Chen H.-T. (2012): *Theory-driven evaluation: Conceptual framework, application and advancement*. In: Strobl R. et al. (Hrsg.): Evaluation von Programmen und Projekten für eine demokratische Kultur, Springer Fachmedien Wiesbaden 2012, S.17-26.

Cicirelli V. G. (1969): *The impact of Head Start. An evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*. Athens: Ohio University and Westinghouse Learning Corporation.

Cohen D. K., Garet M. S. (1975): *Reforming educational policy with applied research*, in: Harvard Educational Review, 43(1), S. 17-43.

Cohen J. (1988): *Statistical Power Analysis for the Behavioral Sciences*. 2. Aufl., Hillsdale: Lawrence Erlbaum Associates.

Cook T. D. (1978): *Introduction*, in: Cook T. D., DelRosario M. L., Hennigan K. M., Mark M. M., Trochim W. M. K. (Hrsg.): Evaluation Studies Review Annual (Vol. 3). Beverly Hills, CA: Sage Publications.

Cook T. D., Campbell D. T. (1979): *Quasi-Experimentation. Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cook T. D., Matt G. E. (1990): *Theorien der Programmevaluation*, in: Koch U., Wittmann W. W. (Hrsg.): Evaluationsforschung: Bewertungsgrundlage von Sozial- und Gesundheitsprogrammen. Berlin u. a.: Springer.

Cook T. D. (2003): *Why have educational evaluators chosen not to do randomized experiments?*, in: Annals of American Academy of Political and Social Science, 589, 114-149.

Cook T. D. (2006): *Describing what is Special about the Role of Experiments in Contemporary Educational Research: Putting the "Gold Standard" Rhetoric into Perspective*, in: Journal of MultiDisciplinary Evaluation, Number 6.

Coryn Ch., Noakes L., Westine C., Schroter D. (2011): *A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009*. In: American Journal of Evaluation, 32(2), Sage Publications, S.199-226. http://www.wmich.edu/evalphd/wp-content/uploads/2011/04/A-Systematic-Review-of-Theory-Driven-Evaluation-Practice-from-1990-to-2009.pdf.

Cronbach, L.J. (1963): *Educational Psychology*. Revised Edition, Harcourt, Brace and World, New York.

Cronbach L. J. (1972): *Evaluation zur Verbesserung von Curricula*, in: Wulf C. (Hrsg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : R. Piper & Co. Verlag, S. 41-59.

Cronbach L. J., Ambron S. R., Dornbusch S. M., Hess R. D., Hornik R. C., Phillips D. C., Walker D. F., Weiner S. S. (1980): *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Cronbach L. J. (1982): *Designing Evaluations of Educational and Social Programs*. San Francisco u.a.: Jossey-Bass.

Daniel K. (2008): *Managementprozesse und Performance: Ein Konzept zur reifegradbezogenen Verbesserung des Managementhandels (Unternehmensführung & Controlling)*. Verlag: Gabler Edition Wissenschaft.

Datta L. (1979): *Another Spring and Other Hopes: Some Findings from National Evaluations of Project Head Start*, in: Zigler E., Valentine J. (Hrsg.): Project Head Start: A Legacy of the War on Poverty, New York: Free Press.

DeGEval – Gesellschaft für Evaluation e.V. (2002): *Standards für Evaluation*. Köln.

DeGEval – Gesellschaft für Evaluation e.V. (2008): *Standards für Evaluation*. 4. unveränderte Auflage. Mainz: 10-13. http://www.degeval.de/images/stories/Publikationen/DeGEval_-_Standards.pdf.

Derlien, H.-U. (1990): *Genesis and Structure of Evaluation Efforts in Comparative Perspective*, in: Rist R. C. (Hrsg.): Program Evaluation and the Management of Govern-ment. New Brunswick: Transaction Publishers, S. 146-176.

Derlien, H.-U. (1997): *Die Entwicklung der Evaluationen im internationalen Kontext*, in: Bussmann W., Klöti U., Knoepfel P. (Hrsg.): Einführung in die Politikevaluation. Basel, S. 4-12.

Deming W. E. (1982): *Out of the Crisis*. Massachusetts Institute of Technology, Cambridge.

Descy P., Tessaring M. (2006): *Der Wert des Lernens Evaluation und Wirkung von Bildung und Ausbildung*. Dritter Bericht zum aktuellen Stand der Berufsbildungsforschung in Europa. Synthesebericht. Cedefop Reference series; 66. Luxemburg: Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaften.

Dollmann J., Kristen C. (2010): Herkunftssprache als Ressource für den Schulerfolg? - Das Beispiel türkischer Grundschulkinder, in: Zeitschrift für Pädagogik; 55. Beiheft, S. 123-146.

Doran G. T. (1981): *There's a S.M.A.R.T. way to write management's goals and objectives*, in: Management Review, Volume 70, Issue 11(AMA FORUM), S. 35-36.

Dugard P., Todman J. (1995): A*nalysis of pre-test post-test control group designs in educational research*. Educational Psychology, 15, 181-198.

Einsiedler W., Kirschhock E.-M. (2003): *Forschungsergebnisse zur phonologischen Bewusstheit*, in: Grundschule, 35, S. 55-57.

Einsiedler W., Frank A., Kirschhock E.-M., Martschinke S., Treinies G. (2002): *Der Einfluss verschiedener Unterrichtsmethoden auf die phonologische Bewusstheit sowie auf Lese- und Rechtschreibleistungen im 1. Schuljahr*. In: Psychologie in Erziehung und Unterricht, S. 194-209.

Ellsworth J. (1998): *Inspiring Delusions: Reflections on Head Start's Enduring Popularity*, in: Ellsworth J., Ames L. (Hrsg.): Critical Perspectives on Project Head Start. Albany: State University of New York.

Ellsworth J., Ames L. J. (Hrsg.) (1998): *Critical Perspectives on Project Head Start: Revisioning the Hope and Challenge*. New York: State University of New York Press.

Esser H. (2006): *Sprache und Integration. Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten*, Frankfurt a. Main/New York: Campus Verlag.

Esser H. (2008): Assimilation, ethnische Schichtung oder selektive Akkulturation? Neuere Theorien der Eingliederung von Migranten und das Modell der intergenerationalen Integration. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft, S. 81-107.

Esser H. (2010): *Integration, ethnische Vielfalt und moderne Gesellschaft*, in: Wienand J., Wienand C. (Hrsg.) Die kulturelle Integration Europas. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 143-169.

Farran D. (1990): *Effects of intervention with disadvantaged and disabled children*, in: Meisels S., Shonkoff J. (Hrsg.): Handbook of early childhood intervention, Cambridge University Press.

Fend H. (1977): *Wissenschaftssoziologische Perspektiven für eine Analyse der Begleitforschung im Rahmen von Modellversuchen*, in Mitter W., Weishaupt H. (Hrsg.): Theoretische und praktische Ansatzpunkte einer Evaluation von Begleituntersuchungen Weinheim: Beltz, S. 48-83.

Filsinger, D. (2008): *Bedingungen erfolgreicher Integration. Integrationsmonitoring und Evaluation*. Friedrich Ebert Stiftung; Gesprächskreis Migration und Integration.

Flick, U. (2006): *Qualitative Evaluationsforschung: Konzepte - Methoden - Umsetzung*. Rororo-Verlag.

Funnell S. C. (2000): *Developing and Using a Program Theory Matrix for Program Evaluation and Performance Monitoring*, in: Rogers P. J., Petrosino A. J., Hacsi T., Huebner T. A. (Hrsg.): Program Theory in Evaluation: Challenges and Opportunities. New Directions for Evaluation, no. 87. San Francisco: Jossey-Bass.

Giel S. (2013): *Theoriebasierte Evaluation. Konzepte und methodische Umsetzungen*. Waxmann Verlag (Münster/New York/München/Berlin).

Gill M., Turbin V. (1999): *Evaluating 'Realistic Evaluation': Evidence from a Study of CCTV, in:* Painter K., Tilley, N (Hrsg.): Surveillance of Public Space: CCTV, Street Lighting and Crime Prevention. Crime Prevention Studies Volume 10. Monsey, N.Y: Criminal Justice Press.

Gray S. W. (1974): *Children from Three to Ten: The Early Training Project*, in: Ryan S. (Hrsg.): A Report on Longitudinal Evaluations of Preschool Programs. DHEW Publication No. 76-30024. Washington D.C., S. 61-68.

Guba E. G. (1969): *The Failure of Educational Evaluation*, in: Educational Technology, V9, No.5, S. 29-38, auch in: Weiss C. H. (Hrsg.) (1972): *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn' and Bacon Inc., S. 250-266.

Guba E. G., Lincoln Y. S. (1981): *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco, CA: Jossey-Bass.

Guba E. G., Lincoln Y. S. (1989): *Fourth Generation Evaluation*. Newbury Park, London, New Dehli: Sage Publishers.

Hagen, T., Spermann A. (2004): *Hartz-Gesetze: Methodische Ansätze zu einer Evaluierung*. ZEW Wirtschaftsanalysen, Bd. 74, Baden-Baden.

Harrington J. (1991): *Business Process Improvement. A Breakthrough Strategy for Total Quality, Productivity, and Competitiveness*, New York u.a..

Haubrich K., Lüders C. (2003): *Evaluation – hohe Erwartungen und ungeklärte Fragen*. Deutsches Jugendinstitut e.V. (DJI). http://www.gesis.org/fileadmin/upload/dienstleistung/fachinformationen/servicepublikationen/sofid/Fachbeitraege/Jugend_2004-1.pdf.

Haubrich K., Lüders C. (2004): *Evaluation – mehr als ein Modewort?*, in: Recht der Jugend und des Bildungswesens, 3/2004, S. 316–337.

Haug S. (2005): *Familienstand, Schulbildung und Erwerbstätigkeit. Eine Analyse der ethnischen und geschlechtsspezifischen Ungleichheiten*, in: Haug S., Diehl C. (Hrsg.): Aspekte der Integration. Eingliederungsmuster und Lebenssituation italienisch- und türkischstämmiger junger Erwachsener in Deutschland, Wiesbaden (Band 35 der Schriftenreihe des Bundesinstituts für Bevölkerungsforschung), S. 51-75.

Haug S. (2008): *Sprachliche Integration von Migranten in Deutschland*. Working Paper 14 der Forschungsgruppe des Bundesamtes für Migration und Flüchtlinge. http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/WorkingPapers/wp14-sprachliche-integraion.pdf?__blob=publicationFile.

Heckmann F., Wunderlich T. (2001): *Integration fördern: Das Modellprojekt ‚Integrationskurse'*. Unveröffentlichter Antragstext, efms.

Heckmann F., Wunderlich T., Worbs S., Lederer H. W. (2001): *Integrationspolitische Aspekte einer gesteuerten Zuwanderung*. Gutachten für die interministerielle Arbeitsgruppe der Bayerischen Staatsregierung. München.

Held J., Bibouche S., Schork C., Dirr F. (2007): *Kommunale Integrationsprojekte mit Migranten – Eine subjektorientierte Evaluation*. Stuttgart: Landesstiftung BdWü.

Hellstern G.-M., Wollmann H. (Hrsg.) (1984): *Handbuch zur Evaluierungsforschung*. Band 1, Opladen: Westdeutscher Verlag.

Helmstädter W. (2009): *Ländervergleich – Evaluation Arbeitsmarktpolitik*, in: Beywl T., Widmer, T., Fabian, C. (Hrsg.): Evaluation. Ein systematisches Handbuch. Wiesbaden, S. 148 – 153.

Hitzler R., Honer A. (Hrsg.): *Sozialwissenschaftliche Hermeneutik. Eine Einführung*, Opladen 1997.

Hoffmann N, Polotzek S., Roos, J., Schöler H. (2008): *Sprachförderung im Vorschulalter – Evaluation dreier Sprachförderkonzepte*. In: Diskurs Kindheits- und Jugendforschung Heft 3-2008, S. 291-300. http://www01.ph-heidelberg.de/wp/schoeler/Datein/hofmann-ua_disk308.pdf.

Horton C. (2006): *Evaluating early care and education programs: A review of research methods and findings*. Report prepared for the National Early Childhood Accountability Task Force. http://www.erikson.edu/wp-content/uploads/ECEreport.pdf.

House E. R. (1983): *Philosophy of Evaluation*. San Francisco u.a. Jossey Bass.

House E. R. (1993): *Professional Evaluation, Social Impact and Political Consequences*. Newbury Park: Sage.

Höhne T. (2005): *Evaluation als Wissens- und Machtform*. Studien- und Forschungsberichte der Universität Gießen. http://geb.uni-giessen.de/geb/volltexte/2005/2105/index.html.

Kalter F., Granato N., Kristen C. (2011): *Die strukturelle Assimilation der zweiten Migrantengeneration in Deutschland: Eine Zerlegung gegenwärtiger Trends*, in:

Rolf Becker (Ed.) Integration durch Bildung: Bildungserwerb von jungen Migranten in Deutschland. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 257-289.

Kardorff E. von (2006): *Zur gesellschaftlichen Bedeutung und Entwicklung (qualitativer) Evaluationsforschung, i*n: Flick U. (Hrsg.): Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen, Reinbek: Rowohlt, S. 63-91.

Kazi M.A.F. (2001): *Realist Evaluation for Practice*. Keynote Address Thirteenth National Symposium on Doctoral Research in Social Work. https://kb.osu.edu/dspace/bitstream/handle/1811/36914/1/13_Kazi_paper.pdf.

Kennedy S.D. (1980): *Final Report of the Housing Allowance Demand Experiment*. Abt Associates Inc., Cambridge, Mass.

Kershaw D., Fair J. (1976): *The New Jersey Income-Maintenance Experiment*, in: Vol. 1. New York: Academic Press.

Kirschhock E.-M., Martschinke S., Treinies G., Einsiedler, W. (2002): *Vergleich von Unterrichtsmethoden zum Schriftspracherwerb mit Ergebnissen zum Lesen und Rechtschreiben im 1. und 2. Schuljahr*. Nr. 100. Nürnberg: Arbeiten aus dem Institut für Grundschulforschung. Erschienen in: Empirische Pädagogik.

Kiziak T., Kreuter V., Klingholz R. (2011): *Dem Nachwuchs eine Sprache geben. Was frühkindliche Sprachförderung leisten kann*. Diskussionspapier. Berliner Institut für Bevölkerung und Entwicklung. http://www.berlininsttut.org/fileadmin/user_upload/Veroeffentlichungen/DP_Sprachfoerderung/Sprachfoerderung_online.pdf.

Koch U., Wittmann W. W. (Hrsg.) (1990): *Evaluationsforschung. Bewertungsgrundlage von Sozial- und Gesundheitsprogrammen*. Berlin u.a.: Springer Verlag.

Kracht A., Rothweiler M. (2003): *Diagnostische Fragen zur kindlichen Grammatikentwicklung im Kontext von Mehrsprachigkeit*, in: Warzecha B. (Hrsg.): Heterogenität macht Schule. Beiträge aus sonderpädagogischer und interkultureller Perspektive. Münster: Waxmann. S. 189-204.

Kristen C., Granato N. (2004): *Bildungsinvestitionen in Migrantenfamilien*, in: I-MIS-Beiträge Heft 23/2004. Themenheft: Migration – Integration – Bildung – Grundfragen und Problembereiche. Für den Rat für Migration herausgegeben von Klaus J. Bade und Michael Bommes, S. 123-178.

Kromrey H. (2001): *Evaluation - ein vielschichtiges Konzept. Begriff und Methodik von Evaluierung und Evaluationsforschung. Empfehlungen für die Praxis*. Sozialwissenschaften und Berufspraxis, 24. Jg., Heft 2/2001.

Kromrey H. (2002): *Empirische Sozialforschung*. Lucius & Lucius Verlagsgesellschaft mbH Stuttgart.

Kromrey H. (2005): *Evaluation – Ein Überblick*, in: Schöch H. (Hrsg.): Was ist Qualität. Die Entzauberung eines Mythos, Berlin 2005: Wissenschaftl. Verlag (Schriftenreihe Wandel und Kontinuität in Organisationen, Band 6), S. 31-85.

Kromrey H. (2008): *Begleitforschung und Evaluation - fast das Gleiche, und doch etwas Anderes!* In: Michaela Glaser, Silke Schuster (Hg.): Evaluation präventiver Praxis gegen Rechtsextremismus. Positionen, Konzepte und Erfahrungen. Leipzig 2008: DJI, 113-135.

Kuhlmann S. (2005): *Selbstevaluation durch Leistungsvergleiche in deutschen Kommunen*, in: Zeitschrift für Evaluation, Heft 1, S. 7-28.

Lachenmann G. (1977): *Evaluierungsforschung – historische Hintergründe, sozialpolitische Zusammenhänge und wissenschaftliche Einordnung*, in: Kantowsky D. (Hrsg.): Evaluierungsforschung und -praxis in der Entwicklungshilfe. Zürich. S. 25-87.

Lamnek S. (1995): *Qualitative Sozialforschung*. Lehrbuch. Beltz.

Landerl K., Wimmer H. (1994): *Phonologische Bewußtheit als Prädiktor für Lese- und Schreibfertigkeiten in der Grundschule*, in: Zeitschrift für Pädagogische Psychologie, 8, S. 153-164.

Lange E. (1983): *Zur Entwicklung und Methodik der Evaluationsforschung in der Bundesrepublik Deutschland*, in: Zeitschrift für Soziologie 12 (3/1983), S. 253-270.

Leeuw F. L., Toulemonde J., Brouwers A. (1999): *Evaluation Activities in Europe: A Quick Scan of the Market in 1998*. Evaluation Sage Publications. S. 487-496.

Lenk H. (1999): *Wissenschaftstheoretische Bemerkungen zum Theoriebegriff und zu theoretischen Begriffen*. Gorokhov V.(Hrsg.): Jahrbuch des Deutsch-Russischen Kollegs, 2000, S. 169-199.

Leseman P. P. M. (2002): E*arly childhood education and care for children from low-income or minority backgrounds*, OECD.

LeVine R. A., Campbell D. T. (1972): *Ethnocentrism: Theories of Conflict, Ethnic Attitudes, and Group Behavior*. Wiley, New York.

Lisker A. (2010): *Sprachstandsfeststellung und Sprachförderung im Kindergarten sowie beim Übergang in die Schule*. Expertise im Auftrag des Deutschen Jugendinstituts. München: Deutsches Jugendinstitut. http://www.dji.de/bibs/Expertise_Sprachstandserhebung_Lisker_2010.pdf.

Lisker A. (2011): *Additive Maßnahmen zur vorschulischen Sprachförderung in den Bundesländern*. Expertise im Auftrag des Deutschen Jugendinstituts. München: Deutsches Jugendinstitut. http://www.dji.de/bibs/Expertise_Sprachfoerderung_Liser_2011.pdf.

Ludwig J., Phillips D. A. (2007): *The Benefits and Costs of Head Start*. Society for Research on Child Development, Social Policy Report. Volume XXI, Number 3. http://www.nber.org/papers/w12973.pdf.

Maddaus G. F., Stufflebeam D. L., Scriven M. S. (1983): *Program evaluation: A historical overview*, in: Maddaus G. F., Scriven M. S., Stufflebeam D. L. (Hrsg.): Evaluation Models, Boston: Kluwer-Nijhoff, S. 3-22.

Malorny C., Hummel Th. (2011): *Total Quality Management. Tipps für die Einführung*. 4. Aufl., Hanser Fachbuch.

Martschinke S., Kammermeyer G., King M., Forster M. (2005): *Diagnose und Förderung im Schriftspracherwerb*. Auer Verlag, Donauwörth.

Marx H., Jansen H., Mannhaupt G., Skowronek H., Näslund J. C., Schneider W. (1993): *Prediction of difficulties in reading and spelling on the basis of the Bielefeld screening*, in: Grimm H., Skowronek H (Hrsg.): Language acquisition problems and reading disorders aspects of diagnosis and intervention. Berlin: Walter de Gruyter.

Mayring P. (2002): *Einführung in die qualitative Sozialforschung. Eine Anleitung zu qualitativem Denken*. Psychologie Verlags Union, München, 5. Auflage, Beltz Studium, Weinheim.

McVicker Hunt J. (1961): *Intelligence and Experience*, New York, Roland Press.

Mertens D. M. (2000): *Institutionalizing Evaluation in the United States of America*, in: Stockmann R. (Hrsg.): Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder, Opladen: Leske+Budrich, S. 41-56.

Meyer W. (2002): *Was ist Evaluation?* Saarbrücken: Centrum für Evaluation, CEval-Arbeitspapiere. http://www.ceval.de/typo3/fileadmin/user_upload/PDFs/workpaper5.pdf.

Moosbrugger H., Schweizer K. (2002): *Evaluationsforschung in der Psychologie*, in: Zeitschrift für Evaluation, Heft 1/2002, S. 19-37.

Nachtigall C., Suhl U. (2002): *Der Regressionseffekt. Mythos und Wirklichkeit*, Methevalreport, 2002. http://www.metheval.uni-jena.de/materialien/reports/report_2002_02.pdf.

Nauck B., Kohlmann A., Diefenbach H. (1997): *Familiäre Netzwerke, intergenerative Transmission und Assimilationsprozesse bei türkischen Migrantenfamilien*, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 49, S. 477 - 499.

Naumann J., Artelt C., Schneider W., Stanat P. (2010): *Lesekompetenz von PISA 2000 bis PISA 2009,* in: Klieme E., Artelt C., Hartig J., Jude N., Köller O., Prenzel M., Schneider W., Stanat, P. (Hrsg.): PISA 2009. Bilanz nach einem Jahrzehnt. Münster/ New York/ München/ Berlin: Waxmann.

Oakley A. (1998): *Experimentation and social interventions: a forgotten but important history*, in: British Medical Journal, 317, S. 1239–1242.

OECD (2001a): *Knowledge and skills for life. First result from PISA 2000*. Paris: OECD.

OECD (2001b): *Starting strong: early childhood education and care*. Paris: OECD.

OECD (2004): The choice of tools for enhancing policy impact: Evaluation and review, OECD, Paris.

Osterloh M., Frost J. (2006): *Prozessmanagement als Kernkompetenz. Wie Sie Business Reengineering strategisch nutzen können*. 5., überarbeitete Auflage, Wiesbaden: Gabler Verlag.

Owen J. M. (2007): *Program Evaluation: Forms and Approaches*. Third Edition. New York: The Guilford Press.

Patton M. Q. (1978): *Utilization-focused evaluation*. Beverly Hills: Sage Publications.

Patton M. Q. (1981): *Creative evaluation*. Beverly Hills: Sage Publications.

Patton, M. Q. (1998): *Utilization focussed Evaluation*. 4. Aufl., London u.a, Sage Publishers.

Patton, M. Q. (2002): *Qualitative research and evaluation methods*. 3. Auflage. Thousand Oaks, CA: Sage.

Patton M. Q. (2010): Developmental Evaluation Applying Complexity Concepts to Enhance Innovation and Use. Guilford Press.

Pawson R., Tilley N. (1997): *Realistic Evaluation*. London, UK: Sage.

Pedersen L, Rieper O (2008): Is realist evaluation a realistic approach for complex reforms? In: Evaluation. 14(3). S. 271–293.

Polotzek S., Hofmann N., Roos J., Schöler H. (2008): *Wirkungen der vorschulischen Sprachförderungen in Mannheim und Heidelberg auf die schulischen Leistungen am Ende der 1. Klasse*. EVAS-Arbeitsbericht Nr. 4.

Posavac E. J., Carey R. G. (1992): *Program Evaluation: Methods and Case Studies*. Englewood Cliffs, NJ: Prentice Hall.

Prenzel M., Artelt C., Baumert J., Blum W., Hammann M., Klieme E., Pekrun R. (Hrsg.) (2007): *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie – Zusammenfassung*, PISA Konsortium Deutschland, Online: http://pisa.ipn.uni-kiel.de/zusammenfassung_PISA2006.pdf

Rasch B., Friese M., Hofmann W., Naumann E. (2006): *Quantitative Methoden*. Band 2, (2. Auflage). Heidelberg: Springer.

Reeves T., Hedberg J. (2003): *Interactive learning systems evaluation*. Englewood Cliffs, New Jersey: Educational Technology Publications.

Rogers P. (2008): Using programme theory to evaluate complicated and complex aspects of interventions. In: Evaluation. 14(1), S. 29–48.

Rogers, P. J., Petrosino A., Hübner T. A., Hacsi T. A. (2000): *Program theory evaluation: Practice, promise, and problems*, in: Rogers, P. J., Petrosino A., Hübner T. A., Hacsi T. A. (Hrsg.): Program Theory in Evaluation: Challenges and opportunities, New Directions for Evaluation, Jossey-Bass, S. 5-13.

Roos J., Schöler H. (2007): *Sprachentwicklungsdiagnostik mittels standardisierter Tests*, in: Schöler H., Welling A. (Hrsg.): Handbuch der Sonderpädagogik, Band 1 Sonderpädagogik der Sprache. Göttingen: Hogrefe. S. 531-550.

Roos J., Polotzek S., Schöler H. (2010a): *Evaluationsstudie zur Sprachförderung von Vorschulkindern*. Wissenschaftliche Begleitung der Sprachfördermaßnahmen im Programm „Sag' mal was – Sprachförderung für Vorschulkinder". Abschlussbericht.

Roos J., Polotzek S., Schöler H. (2010b): *EVAS Evaluationsstudie zur Sprachförderung von Vorschulkindern. Abschlussbericht der Wissenschaftlichen Begleitung der Sprachfördermaßnahmen im Programm „Sag' mal was – Sprachförderung für Vorschulkinder". Unmittelbare und längerfristige Wirkungen von Sprachförderungen in Mannheim und Heidelberg*. http://www.sagmalwas-bw.de/media/WiBe%201/pdf/EVAS_Abschlussbericht_Januar2010.pdf.

Roos J., Gasteiger-Klicpera B., Kucharz D., Knapp W., Schöler H. (2011): *Forschungsdesiderate*, in: Baden-Württemberg Stiftung (Hrsg.): Sag' mal was – Sprachförderung für Vorschulkinder. Tübingen: Narr Francke Attempto.

Rossi, P. H., Freeman H. E., Hofmann G. (1988): *Programm-Evaluation: Einführung in die Methoden angewandter Sozialforschung*. Stuttgart: Enke.

Rossi, P. H., Freeman H. E., Howard E. (1999): *Evaluation. A Systematic Approach*. 6 Aufl. Thousand Oaks u.a.: Sage.

Rother N. (2008): *Das Integrationspanel. Ergebnisse zur Integration von Teilnehmern zu Beginn ihres Integrationskurses*, Working Paper Nr. 19, Nürnberg: Bundes-

amt für Migration und Flüchtlinge.http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/WorkingPapers/wp19-Integrationspanel.pdf?__blob=publicationFile.

Rother N. (2009): *Das Integrationspanel. Entwicklungen von alltagsrelevanten Sprachfertigkeiten und Sprachkompetenzen der Integrationskursteilnehmer während des Kurses*, Working Paper Nr. 23, Nürnberg: Bundesamt für Migration und Flüchtlinge, Online: http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/WorkingPapers/wp23-integrationspanel2.pdf?__blob=publicationFile.

Schlösser E. (2007): *Wir verstehen uns gut: Spielerisch Deutsch lernen". Methoden und Bausteine zur Sprachförderung für deutsche und zugewanderte Kinder als Integrationsbeitrag in Kindergarten und Grundschule*. Ökotopia Verlag.

Schweinhart L. J., Barnes H. V., Weikart D. P. (with Barnett W. S. and Epstein A. S.) (1993): *Significant benefits: The High/Scope Perry Preschool Study through age 27*. Ypsilanti, MI: High/Scope Press.

Scriven M. (1972): *Die Methodologie der Evaluation*. In: Wulf, C. (Hrsg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper.

Scriven, M. (1980): *The Logic of Evaluation*. California: Edgepress.

Scriven M. (1986): *New frontiers of evaluation*, in: Evaluation Practice, 7, 7-44.

Scriven M. (1991): *Evaluation Thesaurus*. Newbury Park u.a.: Sage.

Shadish W. R., Cook T. D., Leviton L. C. (1991): *Foundations of Program Evaluation*. Newbury Park, Sage Publishers.

Shadish W., Cook T. D., Campbell D. T. (2001): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning.

Shewhart W A. (1931): Economic control of quality of manufactured product. New York: Van Nostrand.

Smith M. F. (1989): *Evaluability Assessment: A Practical Approach*. Dordrecht: Kluwer.

Smith M. F. (2005): *Evaluability Assessment*, in: Mathison S. (Hrsg.): Encyclopedia of Evaluation. Sage Publications, Thousand Oaks.

Stake R. E. (1975): *Program Evaluation. Particulary Responsive Evaluation*. Center for Instructional Research and Curriculum Evaluation. University of Illinois at Urbana-Champaign. Occational Paper Series.

Stake R., Easley J. (1978): *Case studies in science education*. Urbana, IL: Center for Instructional Research and Evaluation.

Stamm, M. (2003): *Evaluation und ihre Folgen für die Bildung: eine unterschätzte pädagogische Herausforderung*. Münster, New York, München, Berlin. Waxmann.

Stamm M. (2008): *Evaluation: Wirksame Wege zur Nutzung – Wege zur wirksamen Nutzung*, in: Ant M. et al. (Hrsg.): Nachhaltiger Mehrwert von Evaluation. Bielefeld: Bertelsmann. S. 145-158.

Stern E. (2004): *Philosophies and types of evaluation research*, in: Descy P., Tessaring M. (Hrsg.): The foundations of evaluation and impact research. Third report on vocational training research in Europe: Background Report. Luxemburg: Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaften, Cedefop Reference Series, 58.

Stockamnn R. (Hrsg.) (2000): *Evaluierungsforschung. Grundlagen und ausgewählte Forschungsfelder*. 1. Auflage, Münster, Waxmann Verlag.

Stockmann R. (Hrsg.) (2006): *Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder*. Münster, Waxmann Verlag.

Stockmann R., Caspari A. (1998): *Ex-post Evaluation als Instrument der Qualitätsentwicklung in der Entwicklungszusammenarbeit*. Gutachten im Auftrag der Deutschen Gesellschaft für Technische Zusammenarbeit (GTZ).

Stockmann R., Meyer W. (2014): *Evaluation. Eine Einführung*. Opladen: Budrich.

Stufflebeam D. L., Madaus G. F., Kellaghan T. (Hrsg.) (2000): *Evaluation models - viewpoints on educational and human services evaluation*. Boston: Kluwer Academic Publisher Group.

Stufflebeam D., Shinkfield A. (2007): *Evaluation Theory, Models, and Applications*. San Francisco: Jossey-Bass.

Suchman E. (1967): *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage.

Thoening, J. C. (2000): *Evaluation as usable knowledge for public management reforms*, in: Evaluation, Vol. 6/2, S. 217-231.

Tietgens H. (1984): *Institutionelle Aspekte der Erwachsenenbildung*, in: Schmitz E., Tietgens H. (Hrsg.): Erwachsenenbildung. Enzyklopädie Erziehungswissenschaft, Bd. 11. Stuttgart.

Tietze W. et al. (1998): Wie gut sind unsere Kindergärten? Eine Untersuchung zur pädagogischen Qualität in deutschen Kindergärten. Neuwied, Berlin: Luchterhand.

Toepel K., Tissen G. (2000): *Stand und Perspektiven der Evaluation in Deutschland. Eine Einführung*. Vierteljahrshefte zur Wirtschaftsforschung. 69. Jahrgang, Heft 3/2000, S. 347–349.

Toulemonde J. (2000): *Evaluation Culture(s) in Europe: Differences and Convergence between National Practices*. Vierteljahreshefte zur Wirtschaftsforschung. 69. Jahrgang, Heft 3/2000, S. 350–357.

Tulodziecki G. (1982): *Zur Bedeutung von Erhebung, Experiment und Evaluation für die Unterrichtswissenschaft*, in: Unterrichtswissenschaft 10 (1982), S. 364 – 377.

Tyler R. W. (1932): *Service Studies in Higher Education*. The Ohio state university.

Tyler R. W. (1949): *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.

Van der Knaap P. (2006): *Responsive Evaluation and Performance Management. Overcoming the Downsides of Policy Objectives and Performance Indicators*, in: Evaluation 12 (3). Sage Publications.

Wagner R., Torgesen J. (1987): *The nature of phonological processing and its causal role in the acquisition of reading skills*, in: Psychological Bulletin, 101, S. 192-212.

Weiss C. H. (Hrsg.) (1972): *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn' and Bacon Inc.

Weiss, C. H. (1972): *Evaluation Research*. Englewood Cliffs, NJ: Prentice Hall.

Weiss C. H. (1974): *Evaluierungsforschung. Methoden zur Einschätzung von sozialen Reformprogrammen*. Opladen.

Weiss C.H. (1976): *Using research in the policy process: potential and constraints*. Policy Studies Journal, Bd. 4, S. 224-228.

Weiss C. H. (1998): *Have we learned anything new about the use of evaluation?*, in: American Journal of Evaluation, Vol. 19/1, S. 21-33.

Weiss C. H. (2007): *Theory-based evaluation: Reflections ten years on: Theory-based evaluation: Past, present, and future*. New Directions of Evaluation. Volume 2007, Issue 114, Wiley Pub.

White S., Phillips D. (2001): *Designing Head Start: Roles played by developmental psychologists*, in: Featherman D. L., Vinovskis M. A. (Hrsg.): Social science and policymaking: A search for relevance in the twentieth century. Ann Arbor: University of Michigan Press.

Wholey J. S. (1977): *Evaluability Assessment*, in Rutman L. (Hrsg.): Evaluation Research Methods: a Basic Guide, Beverly Hills: Sage, S. 41-56.

Wholey, J. S. (1984): *Evaluierung – Grundlage und Voraussetzung für leistungsfähige Programme*, in: Hellstern G.-M., Wollmann H. (Hrsg.): Handbuch zur Evaluierungsforschung. Bd. I. Opladen: Westdeutscher Verlag.

Wholey J. S., Scanlon J. W., Duffy H. G., Fukumoto J. F., Vogt L. M. (1970): *Federal Evaluation Policy*, Washington D.C.: The Urban Institute.

Wholey J. S., Strosberg M. A. (1983): *Evaluability assessment: From theory to practice in the Department of Health and Human Services*. In: Public Adminis-tration Review, 43, S. 66-71.

Wholey J. S., Hatry H. P., Newcomer K. E. (2004): *Handbook of Practical Program Evaluation*. 2nd Edition. Jossey-Bass.

Widmer T. (2000): *Qualität der Evaluation*, in: Stockmann R. (Hrsg.): Evaluationsforschung. Opladen, S. 77-102.

Widmer T., Beywl W., Fabian C. (2009): *Evaluation. Ein systematisches Handbuch*. Verlag für Sozialwissenschaften. Wiesbaden.

Wild J. (1982): *Grundlagen der Unternehmensplanung*. Opladen, Westdt. Verlag.

Wilner D., Walkley R., Pinkerton T., Tayback M. (1962): *Housing Environment and Family Life: A Longitudinal Study of the Effects of Housing on Morbidity and Mental Health*. Baltimore, MD: John Hopkins University Press.

Wolf K. M., Stanat P., Wendt W. (2010): *EkoS – Evaluation der kompensatorischen Sprachförderung: Erster Zwischenbericht*. Berlin, AB Empirische Bildungsforschung der Freien Universität. Verfügbar unter: http://www.isq-bb.de/uploads/media/ekos-bericht-1-endfassung.pdf.

Wolf R. (2006): *Spielend Lernen – Zwischenbericht nach dem zweiten Projektjahr*. Unveröffentlichter Projektbericht, Bamberg.

Wollmann H. (2004) *Evaluation und Verwaltungspolitik. Konzepte und Praxis in Deutschland und im internationalen Kontext*, in: Stockmann, R. (Hrsg.): Evaluationsforschung. 2. Aufl., Opladen: Leske + Budrich, S. 205 –232.

Wollmann H. (2005a): *Evaluation*, in: Akademie für Raumforschung und Landesplanung (Hrsg.), Handwörterbuch der Raumordnung, Hannover: ARL, S. 274-280. http://amor.cms.hu-berlin.de/~h0598bce/docs/HW-2005-Evaluation.doc.

Wollmann H. (2005b): *Evaluierung von Verwaltungsmoderisierung*, in: Blanke, B. et al. (Hrsg.): Handbuch zur Verwaltungsreform, 3.Aufl., Wiesbaden: VS Verlag, S. 502-510.

Wollmann H., Kuhlmann, S. (2011): *Evaluierung von Verwaltungsmodernisierung*, in; Blanke, B. et al. (Hrsg.): Handbuch zur Verwaltungsreform, 4. Aufl., VS Verlag Wiesbaden, S. 563-571.

Wottawa H., Thierau H. (1998): *Lehrbuch Evaluation*. (2., vollst. überarb. Aufl.). Bern: Verlag Hans Huber.

Wu P., Campbell D. T. (1996): *Extending Latent Variable LISREL Analyses of the 1969 Westinghouse Head Start Evaluation to Blacks and Full Year Whites*, in: Evaluation and Program Planning, V19 Nr.3, S. 183-191.

Wulf C. (Hrsg.) (1972): *Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen*. München R. Piper & Co. Verlag.

Youker B. W., Ingraham A. (2014): *Goal-Free Evaluation: An Orientation for Foundations' Evaluations*. In: The Foundation Review: Vol. 5: Iss. 4, Article 7.

Zigler E., Styfco S., Gilman E. (1993): *The National Head Start Program for Disadvantaged Preschoolers,* in: Zigler E., Styfco S. (Hrsg.): Head Start and beyond: a national plan for extended childhood intervention. New Haven: Yale University Press.

Zigler E. F., Muenchow S. (1992): *Head Start: The inside story of America's most successful educational experiment*. New York: Basic Books.

Zimmermann E. (1972): *Das Experiment in den Sozialwissenschaften*. Stuttgart. Teubner-Verlag.

University of Bamberg Press

Die Arbeit verfolgt das Ziel, ein prozessorientiertes Rahmenmodell für die Tätigkeit von Evaluatoren in der Praxis zu entwickeln. Im Hauptteil der Arbeit werden zunächst in einem historisch-systematischen Vorgehen die Methoden einer Auswahl von Evaluationstheoretikern vorgestellt und hinsichtlich ihrer Relevanz für die Evaluation von sozialen Programmen diskutiert. Neben den Erkenntnissen aus der Evaluationstheorie wird die Anwendung von Evaluationsmethoden in drei Evaluationsstudien des Autors kritisch reflektiert. Die auf diese Weise gewonnen Erkenntnisse aus Theorie und Praxis werden schließlich in einem Rahmenmodell für die Durchführung von Evaluationsstudien aufbereitet. Das Rahmenmodell soll Evaluatoren bei der Gestaltung und Durchführung von Evaluationsstudien zu sozialen Programmen in verschiedenen Etappen und Phasen unterstützen.

ISBN: 978-3-86309-474-4

www.uni-bamberg.de/ubp