Prof. Dr. Helmut Kromrey

Freie Universität Berlin
Institut für Soziologie
Babelsberger Str. 14-16, 10715 Berlin
Tel. 030/85002230

Fallstricke bei der Implementations- und Wirkungsforschung sowie methodische Alternativen

Ausführlicher Text zum Vortrag
auf dem Europäischen Kongress
für Evaluation und Qualitätsmanagement in Sozialer Arbeit und Gesundheitswesen
am 3. September 1999 in Osnabrück

zur Veröffentlichung in
H. Müller-Kohlenberg, K. Münstermann (Hg.): Bewertung von Humandienstleistungen. Evaluation und Qualitätsmanagement in Sozialer Arbeit und Gesundheitswesen, Opladen 2000 : Leske+Budrich, S. 19-58

Berlin 1999

1. Vorbemerkung: Dilemmata einer Aufgabenkombination Evaluation, Innovationen, Dienstleistungen

Wer *Evaluation* als empirisch-wissenschaftliche Disziplin betreiben, wer ihren Zweck und ihre Methodik anderen vermitteln will, sieht sich einem mehrfachen Dilemma gegenüber. Dies ist *zum einen* bedingt durch eine mittlerweile geradezu inflationäre Verwendung des Begriffs, *zum anderen* durch einen "fließenden" Gegenstand, *drittens* durch eine unüberschaubare Vielfalt von Fragestellungen, Zwecken und Perspektiven und *schließlich* durch die Komplexität des Evaluationsprozesses selbst.

Zum *ersten* Problem:
Im alltäglichen Sprachgebrauch bedeutet "Evaluation" nichts weiter als "Bewertung": Irgend etwas wird von irgend jemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet. In anderen Kontexten sind die Begriffsverwendungen wesentlich spezifischer, unglücklicherweise aber außerordentlich vielfältig. Die Bezeichnung gilt für Effizienz*messungen* in ökonomischen Kontexten ebenso wie für die von Sachverständigen vorgenommene *Analyse* der Funktionsfähigkeit von Organisationen (etwa: "Evaluation" wissenschaftlicher Einrichtungen), sie umfaßt sowohl durch Umfragen ermittelte Zufriedenheits- oder Unzufriedenheits*äußerungen* (etwa: "Evaluation" von Fortbildungsveranstaltungen durch die Teilnehmer) als auch die Erfassung des Akzeptanz- und Abstimmungs*verhaltens* von Klienten- und Zielgruppen (etwa: Teilnehmerzahlen als Indikator der Güte einer Veranstaltung, Wahlresultate als "Evaluation" der Politik) und sogar die *beratende und moderierende Beteiligung* im Prozeß der Entwicklung von Handlungsprogrammen mit dem Ziel der Optimierung ihrer Nützlichkeit („formative" oder "responsive Evaluation").

Gemeinsam ist allen diesen Verwendungen, daß – im Unterschied zum alltagssprachlichen Verständnis – nicht „irgend etwas" evaluiert wird, sondern Programme, Maßnahmen, manchmal auch ganze Organisationen Gegenstand der Betrachtung sind. Zweitens nimmt nicht „irgend jemand" die Evaluation vor, sondern es sind Personen, die dazu in besonderer Weise befähigt erscheinen: „Sachverständige", methodische oder durch Praxiserfahrungen ausgewiesene „Experten", konkret „Betroffene". Drittens kommt das Urteil nicht nach „irgend welchen" Kriterien zustande, sondern diese müssen *explizit* auf den zu bewertenden Sachverhalt bezogen sein. Und schließlich darf bei einer systematischen Evaluation nicht „irgendwie" vorgegangen werden, sondern das Verfahren ist zu „objektivieren", d.h. im Detail zu planen und in einem „Evaluationsdesign" verbindlich für alle Beteiligten festzulegen.

Präzisierungen zu jedem dieser vier genannten Aspekte (Gegenstand – Evaluator – Kriterien – Verfahren) sind in unterschiedlicher Weise möglich und kommen im Evaluationsalltag in unterschiedlichen Kombinationen vor. So kann der zu evaluierende Sachverhalt schon lange bestehen oder aber gerade erst realisiert werden oder gar erst als Planungs- und Entwicklungsabsicht existieren; er kann sehr umfassend und abstrakt oder eng umgrenzt und konkret sein. Die mit der Evaluation betrauten Personen können in unterschiedlicher Weise zum Gegenstand der Bewertung in Bezug stehen: als außenstehende unabhängige Wissenschaftler, als Auftragsforscher für die Programmdurchführenden oder für eine Kontrollinstanz, als unmittelbar im Programm Mitwirkende oder als hinzugezogene externe Berater, als wenig engagierte Betroffene oder als organisierte Befürworter oder Gegner – um nur einige Varianten zu nennen. Die Bewertungskriterien können sich auf die *Wirkungen* und Nebenwirkungen der Maßnahmen des Programms beziehen, auf die Art und Effizienz der

Durchführung, auf die *Eignung* und Effektivität der gewählten Maßnahmen für die Ziel-Erreichung, auf die Angemessenheit und Legitimierbarkeit der Ziele selbst. Die Kriterien können aus unterschiedlicher Perspektive hergeleitet werden (Auftraggeber – Betroffene – Durchführende; ökonomische Effizienz – Nutzen für das Allgemeinwohl – Sozialverträglichkeit etc.). Das Verfahren der Evaluierung schließlich kann von der qualitativen oder der quantitativen Logik der Informationsgewinnung geprägt sein; das Forschungsdesign kann experimentell oder nicht-experimentell angelegt sein. Die Evaluationsaktivitäten können im Vorfeld, projektbegleitend oder im nachhinein unternommen werden; die Evaluation kann so angelegt sein, daß sie möglichst wenig Einfluß auf das laufende Programm ausübt (um "verzerrungsfreie" empirische Befunde zu gewährleisten), oder – im Gegenteil – so, daß jede gewonnene Information unmittelbar rückgekoppelt wird und somit direkte Konsequenzen für das Programm hat.

Hinzu kommt, daß zwischen den genannten vier Aspekten Wechselbeziehungen existieren. Die Evaluation eines noch in der Entwicklung und Erprobung befindlichen Sozialarbeits-konzepts in einem kommunalen sozialen Brennpunkt erfordert ein gänzlich anderes Design als etwa die Überprüfung, ob ein Bundesgesetz zum Anreiz von Investitionen im privaten innerstädtischen Wohnungsbestand zur Verbesserung der Wohnqualität "erfolgreich" ist, d.h. von den zuständigen Instanzen korrekt und effizient ausgeführt wird, die richtigen Zielgruppen erreicht und keine unerwünschten Nebeneffekte hervorruft.

Mehr als nur angedeutet ist in den genannten Beispielen bereits *das zweite Feld von Problemen*, mit denen es Evaluatoren zu tun haben:
Gegenstand der Evaluation kann praktisch alles sein – von einem bereits implementierten politischen Programm (etwa: Verbesserung der Gesundheits-Infrastruktur im ländlichen Raum) über Feldversuche zur Erprobung von Grundschulcurricula zur Aufklärung gegen Drogenkonsum, über Maßnahmen zur Verbesserung der Kommunikation zwischen Ärzten und den anderen medizinischen Berufsgruppen im Krankenhaus bis hin zur Entwicklung organisatorischer Innovationen (etwa: Konzipierung, Realisierung und Test eines Seminars zur EDV-Weiterbildung von Leitungspersonal). Evaluationsgegenstand sind nicht nur bereits realisierte (abgeschlossene) oder gegenwärtig durchgeführte (laufende) Maßnahmen, sondern ebenso auch noch in der Erprobung und Weiterentwicklung befindliche Programme; ausgeschlossen ist auch nicht die Ideenfindungs- und Konzipierungsphase neuer Politiken.

Dabei ist – *drittens* – das Spektrum der interessierenden *Evaluations-Fragestellungen* praktisch unbegrenzt; etwa (um nur einige zu nennen):

- Messung und Beurteilung der *Auswirkungen* von Programmen und Politiken;
- Ermittlung der *Effektivität und Effizienz* von Programmen, Projekten und eingesetzten Maßnahmen und Instrumenten;
- *Einblick gewinnen* in soziale Probleme sowie einen vergleichenden Überblick geben über gegenwärtige und vergangene Bemühungen zu ihrer Lösung;
- *Grundlagenwissen* bereitstellen über die interne Struktur und Arbeitsweise von Organisationen und Institutionen (etwa: Einrichtungen der häuslichen Pflege);
- Stärkung der „*Kundenorientierung*“ der öffentlichen Verwaltung; aber auch:
- Überprüfung der methodischen und fachlichen Güte bisher durchgeführter Evaluationen, also *Meta-Evaluation*;

- *Transfer* des aus bisherigen Evaluationsprojekten eines Politikfelds gewonnenen Handlungswissens in die gegenwärtige und künftige generelle Handlungspraxis (etwa durch Formulierung verbindlicher Standards).

Als Konsequenz folgt daraus *viertens* eine nicht auflösbare methodische Problematik für empirisch-wissenschaftlich zu betreibende Evaluationen:
Nicht nur sind die Aufgaben jeweils unterschiedlich und ist der Gegenstand vielgestaltig und fließend, sondern beides Aufgabe und Gegenstand der Evaluation ist jeweils "mitten im Leben" angesiedelt und entzieht sich der methodischen Kontrolle durch den Evaluator. Während bei anderen Typen wissenschaftlicher Forschung eine zu komplexe Fragestellung durch die Wahl eines vereinfachenden Modells auf das machbar Scheinende reduziert werden darf und der Untersuchungsgegenstand durch die Wahl eines geeigneten Forschungsdesigns zumindest teilweise gegen "störende" Umgebungseinflüsse abgeschirmt werden kann, ist beides in der Evaluationsforschung weder erlaubt noch möglich. Der zu evaluierende Gegenstand soll in seiner natürlichen Umwelt und in der ihm eigenen Komplexität analysiert, beurteilt und weiterentwickelt werden. Nicht einmal die Konstanz des Gegenstandes während der Dauer der Evaluationsstudie ist gewährleistet.

Zusammengefaßt bedeutet das für den empirisch-wissenschaftlichen Evaluator:
Es existieren *keine speziellen Methoden* der Evaluation; vielmehr ist aus dem gesamten Arsenal der empirischen Sozialforschung das für die spezifische Aufgabe Geeignete auszuwählen und an die jeweiligen Gegebenheiten anzupassen. Ebenso existiert *kein allgemeingültiges Evaluationsdesign*; vielmehr ist unter Rückgriff auf die Logik verschiedener Forschungsstrategien (quantitative und qualitative Ansätze) sowie Forschungspläne (Experimental- und Surveyforschung, Querschnitt- und Längsschnittstudien, Primär- und Sekundäranalysen etc.) ein für die jeweilige Aufgabe spezifisches, "maßgeschneidertes" Design zu entwerfen, das zudem so flexibel angelegt sein muß, daß es jederzeit ohne Funktionsverlust dem sich wandelnden Gegenstand angepaßt werden kann. Um es noch einmal hervorzuheben: Jede Erwartung, es könne einen allgemeinen und weitgehend verbindlichen methodologischen und/oder theoretischen Rahmen, eine Art Rezeptbuch für gute Evaluationen geben, ist eine Illusion. Michael Quinn Patton listet in seinem einflußreichen Werk „Utilization-Focused Evaluation" (Patton 1997, 192-194) nicht weniger als 58 das Design bestimmende Zwecke auf; und er fügt hinzu, daß damit noch bei weitem nicht das gesamte Spektrum erfaßt sei.

Was aber ist dann eigentlich "Evaluation" als empirisch-wissenschaftliches Verfahren? Es ist eine besondere Form angewandter Sozialwissenschaft (nicht nur Sozial*forschung*). Es ist eine *methodisch kontrollierte, verwertungs- und bewertungsorientierte Form des Sammelns und Auswertens von Informationen*.

Ihr Besonderes liegt nicht in der Methodik der Datengewinnung und liegt nicht in der Logik der Begründung und Absicherung der zu treffenden Aussagen. Das Besondere liegt vielmehr *zum einen* in der gewählten *Perspektive*, die der Evaluator einzunehmen hat: Erfüllt der zu evaluierende Gegenstand seinen ihm zugeschriebenen Zweck? Wie muß bzw. wie kann er ggf. verändert werden, damit er den vorgesehenen Zweck besser erfüllt? Bei noch in der Erprobung oder gar Konzipierung befindlichen Vorhaben auch: Welche Zwecke sollen überhaupt für welche Zielgruppen angestrebt werden? *Zur Evaluation wird empirische Wissenschaft somit nicht durch die Methode, sondern durch ein spezifisches Erkenntnis- und Verwertungsinteresse.*

Das Besondere liegt *zum anderen* in einer für die Wissenschaft ungewohnten Verschiebung von Rangordnungen, die sich im *Primat der Praxis* vor der Wissenschaft ausdrückt.

Vorrangiges Ziel der Evaluation – im Unterschied zu üblicher wissenschaftlicher Forschung – ist es nicht, am Fall des zu evaluierenden Gegenstands die wissenschaftliche *theoretische* Erkenntnis voranzutreiben, sondern wissenschaftliche Verfahren und Erkenntnisse *einzubringen*, um sie für den zu evaluierenden Gegenstand nutzbar zu machen. Wissenschaft liefert hier – ähnlich wie im Ingenieurwesen – *Handlungswissen* für die Praxis. Geraten wissenschaftlich-methodische Ansprüche einer möglichst objektiven Erkenntnisgewinnung (etwa methodische Kontrolle "störender" Umgebungseinflüsse) mit den Funktionsansprüchen des zu evaluierenden Gegenstands in Konflikt, haben die wissenschaftlichen Ansprüche zurückzutreten und ist nach – aus wissenschaftlicher Perspektive – suboptimalen Lösungen zu suchen, nach Lösungen jedenfalls, die das Funktionsgefüge im sozialen Feld nicht "stören".

Besonders massiv treten die angesprochenen Probleme in Erscheinung, wenn sich die Evaluationsaufgabe – was eher die Regel als der Ausnahmefall ist – auf *Innovationen* bezieht. Das ist leicht einsehbar, denn der Gegenstand, für den die Evaluation "maßgeschneidert" werden soll, existiert ja entweder noch gar nicht oder zumindest nicht in seiner endgültigen Form: Welcher „Gegenstand" also soll evaluiert werden? Ist unter solchen Bedingungen Evaluation überhaupt *sinnvoll möglich*? Ist sie nicht lediglich – so sieht es mancher „Praktiker" – eine modische, lästige und überflüssige Pflichtübung? Steht sie vielleicht bei dem beabsichtigten phantasievollen Vorstoß ins Neuland, beim Verfolgen neuer Ideen eher im Wege als daß sie förderlich und hilfreich wäre?

Diese Fragen sind – was ansonsten selten genug der Fall ist – eindeutig beantwortbar: Innovation wird durch Evaluation nicht behindert; im Gegenteil: Evaluation und Innovation sind wechselseitig aufeinander angewiesen. Ohne daß zumindest die Frage nach möglicherweise notwendigen Innovationen gestellt würde, wäre jede Evaluation in der Tat überflüssig. Und umgekehrt: Innovationen in Angriff zu nehmen, ohne die Situation, in der gehandelt werden soll, und ohne die Sachverhalte, auf die Innovationen abzielen sollen, kontrolliert und kontrollierbar einschätzen zu können, würde mit großer Wahrscheinlichkeit die Verschwendung von Geld, Arbeitsaufwand und Ressourcen bedeuten. Aber leider: Das Faktum, daß es sich beim Evaluationsgegenstand häufig um Innovationen handelt, macht die Evaluationsaufgabe in besonderem Maße schwierig.

Nun spricht aber das Rahmenthema dieses Kongresses noch einen dritten Grad von Schwierigkeiten an: die Evaluation von Humandienstleistungen, also eines nicht konkret faßbaren "Gegenstands". Im Falle der Güterproduktion lassen sich Effizienz und Effektivität des Produktionsprozesses sowie die Qualität des Produkts (output) – und das sind häufig an die Evaluation gerichtete zentrale Fragen – relativ leicht beurteilen. Haben wir es etwa mit der Herstellung von Automobilen zu tun, so kann man sich als Effektivitätskriterium die Fehlerfreiheit des Produktionsablaufs, als Effizienzkriterium die Stückkosten pro Automobil eines bestimmten Typs vorstellen. Und auch für den output selbst, das hergestellte Automobil, sind Qualitätskriterien sowie darauf zugeschnittene Tests relativ einfach zu definieren: Straßenlage, Bodenhaftung, Haltbarkeit der Teile, Kraftstoffverbrauch usw.

Wie aber sieht es bei Dienstleistungen, noch dazu bei Humandienstleistungen aus? Pfarrer Gohde sprach in seinem Einführungsreferat von der zwischenmenschlichen Dimension, deren Wirkungen nicht ausschließlich mit dem Begriff der organisatorischen Funktionalität erfaßbar sind, deren Gebrauchswert nicht darin besteht, daß sie verbraucht werden, sondern daß eine Austauschbeziehung zwischen Menschen im Kontext ihrer Lebenswert entsteht, deren besonderer Mehrwert einen nichtfiskalischen Charakter habe. Das klingt sehr abstrakt. Was also ist hier das Produkt? Was ist der Produktionsprozeß? Ist es der Aufbau und das

Vorhalten einer Dienstleistungsinfrastruktur (etwa Gesundheitsinfrastruktur im ländlichen Raum) oder die einzelne Dienstleistung selbst (etwa der Behandlungsfall im Krankenhaus oder in der Arztpraxis)? Oder interessiert nicht eher das, was durch die vorgehaltene und realisierte Dienstleistung bewirkt wird (outcome anstelle von output)? Und falls es sich um eine Dienstleistung handelt, die auf die Akzeptanz oder gar das aktive Mitwirken der Adressaten angewiesen ist (wie etwa bei Diensten im Bereich Bildung und Qualifizierung): Wer oder was ist dann eigentlich zu evaluieren? Im Bereich der Hochschule haben wir uns seit längerem damit herumzuschlagen – Stichwort: Evaluation von Lehre und Studium. Woran z.B. liegt es, wenn Studienzeiten länger werden und aus politischer Perspektive zu lang erscheinen? An den Lehrenden, an den Curricula, an der Ausstattung der Hochschulen, an den Studierenden, am Arbeitsmarkt, am gesellschaftlichen Anspruchsniveau für einen als angemessen erachteten Konsum-, Lebens- und Freizeitstandard der Studierenden schon während der Studienzeit? Noch ein Stück weiter gefragt: Ist es wirklich ein sinnvolles Ziel, Studienzeiten zu verkürzen (output), ohne danach zu fragen, was dann nach dem abgeschlossenen (kürzeren) Studium aus den Absolventen wird (outcome)?

Ich werde diese Fragen später an einem konkreten Beispiel illustrieren, einem US-amerikanischen Großprogramm zur Bekämpfung von Drogenmißbrauch unter Jugendlichen: Drug Abuse Resistance Education (D.A.R.E.), das darauf baute, durch gezielten Unterricht bereits in der Grundschule über Drogen und ihre Gefahren zu informieren und den Jugendlichen dadurch das kognitive Rüstzeug zu vermitteln, daß sie nicht unüberlegt in Drogenabhängigkeit geraten. In einem einjährigen Pflichtschulunterricht durch speziell ausgebildete und uniformiert auftretende Polizeibeamte sollte ihnen nicht nur Wissen vermittelt werden, sondern sollte ihre Entscheidungsfähigkeit geschult, sollten das Selbstvertrauen und die Fähigkeit gestärkt werden, sich ggf. auch dem Gruppendruck durch Gleichaltrige zu widersetzen. Was ist hier zu evaluieren und nach welchen Kriterien zu beurteilen: die flächendeckende Implementation des Programms, die Eignung der vermittelten Informationen und die Güte von Unterricht und Trainings, die Bereitschaft der Schüler zur engagierten Mitwirkung am Unterricht und ihre Bereitschaft, die Ziele des Programms als ihre Ziele zu übernehmen, oder der kurz-, mittel- und langfristige Effekt auf das faktische Drogenkonsum-Verhalten der Jugendlichen?

2. Die Vielfalt von Evaluationen: eine grobe Klassifikation

Angesichts der geschilderten Variationsbreite von Evaluationen existieren verständlicherweise eine Reihe von Versuchen, die Vielfalt im Detail auf eine überschaubare Zahl von Typen zu reduzieren. Ich greife hier auf einen Vorschlag von Eleanor Chelimsky (1997, 100 ff.) zurück, die drei „conceptual frameworks“ unterscheidet:

- Evaluation zur Verbreiterung der Wissensbasis (ich wähle dafür im folgenden den Begriff „Forschungsparadigma“ der Evaluation),
- Evaluation zu Kontrollzwecken (im folgenden das „Kontrollparadigma“) und
- Evaluation zu Entwicklungszwecken (im folgenden das „Entwicklungsparadigma“).

Für mein Vortragsthema hat diese Einteilung den Vorteil, daß jedes der drei „Paradigmen“ eine je spezifische Affinität zu Designtypen, zur Logik bzw. „Theorie“ der Evaluation, zu Methoden und Qualitätskriterien des Evaluationshandelns aufweist.

2.1 Das „Forschungsparadigma“ der Evaluation

Insbesondere für Universitätswissenschaftler gelten Evaluationsprojekte als Chance und als Herausforderung, neben dem „eigentlichen“ Evaluationszweck grundlagenwissenschaftliche Ziele zu verfolgen. Evaluation wird aus dieser Perspektive verstanden als angewandte Forschung, die sich mit der Wirksamkeit von sozialen Interventionen befaßt. Ihr kommt die Rolle eines Bindeglieds zwischen Theorie und Praxis zu (Weiss 1974, 11). Insbesondere staatliche Auftragsforschung eröffnet einen Weg, Zugang zu den internen Strukturen und Prozessen des politisch-administrativen Systems zu erhalten (Wollmann / Hellstern 1977, 456). Alle Anlässe, Aktionsprogramme zur Bewältigung sozialer Probleme zu implementieren, alle Situationskonstellationen, in denen durch neue gesetzliche Regelungen wichtige Randbedingungen geändert werden, alle Bemühungen, technische, organisatorische oder soziale Innovationen einzuführen, werfen zugleich sozialwissenschaftlich interessante Fragestellungen auf. Und im Unterschied zu forschungsproduzierten Daten zeichnen sich Untersuchungen unmittelbar im sozialen Feld durch einen ansonsten kaum erreichbaren Grad an externer Validität aus. Evaluationsforschung wird in erster Linie als Wirkungsforschung, die Evaluation selbst als wertneutrale technologische Aussage verstanden, die aus dem Vergleich von beobachteten Veränderungen mit den vom Programm angestrebten Effekten (den Programmzielen) besteht. Evaluatoren, die sich dem Forschungsparadigma verpflichtet fühlen, werden versuchen, wissenschaftlichen Gütekriterien so weit wie möglich Geltung zu verschaffen und Designs zu realisieren, die methodisch unstrittige Zurechnungen von Effekten zu Elementen des Programms durch Kontrolle der relevanten Randbedingungen erlauben. Es ist daher kaum ein Zufall, daß Beiträge zur Entwicklung einer allgemeinen Evaluationstheorie und -methodologie vor allem aus dem Kreis universitärer Evaluationsforscherinnen und -forscher geleistet wurden.

2.2 Das „Kontrollparadigma“ der Evaluation

Im Unterschied zur Wirkungsforschung versteht sich der zweite Typus von Evaluation als Beitrag zur Planungsrationalität durch Erfolgskontrolle des Programmhandelns. Planung, verstanden als Instrument zielgerichteten Handelns, um einen definierten Zweck zu erreichen, muß sich bestimmten Erfolgskriterien (Effektivität, Effizienz, Akzeptanz) unterwerfen. Evaluationen dieser Art werden argumentativ vertreten als eine weitere Kontrollform administrativen Handelns neben Rechtmäßigkeits-Kontrolle (Gerichte), politischer Kontrolle (Parlamente) und Wirtschaftlichkeits-Kontrolle (Rechnungshöfe). Eine charakteristische Definition: "Der Begriff Erfolgskontrolle impliziert ex-post-Kontrolle von Ausführung und Auswirkung von zu einem früheren Zeitpunkt geplanten Maßnahmen, und Erfolgskontrolle ist immer zugleich Problemanalyse für den nächsten Planungszyklus" (Hübener / Halberstadt 1976, 15). In welcher Weise der Erfolg kontrolliert wird und an welchen Kriterien der Erfolg gemessen wird, ob die Evaluation ihren Schwerpunkt auf output oder outcome des Programms legt oder auf dessen Implementation, hängt ab vom Informationsbedarf der programmdurchführenden und/oder der finanzierenden Instanz. Gefordert werden häufig quantitative Informationen.

2.3 Das „Entwicklungsparadigma“ der Evaluation

Im Vergleich zu den beiden vorhergehenden Klassen von Evaluationen sind Problemstellung und Erkenntnisinteresse bei diesem dritten Typus grundsätzlich anders gelagert. Am Beginn steht nicht ein bereits realisiertes oder in der Implementationsphase befindliches oder

zumindest ausformuliertes Programm;[1] vielmehr geht es darum, Konzepte und Vorstellungen zu entwickeln, die Fähigkeit von Organisationen zur Problemwahrnehmung und -bewältigung zu stärken, mitzuwirken retrospektiv und prospektiv Politikfelder zu strukturieren. Im Falle der Entwicklung und Erprobung von Programmen bedeutet dies: Die Evaluation ist in die gesamte Programm-Historie eingebunden, von der Aufarbeitung und Präzisierung von Problemwahrnehmungen und Zielvorstellungen über eine zunächst vage Programmidee, über die Entwicklung geeignet erscheinender Maßnahmen und deren Erprobung bis hin zu einem auf seine Güte und Eignung getesteten (endgültigen) Konzept. Evaluation unter solchen Bedingungen ist im wörtlichen Sinne "formativ", also programmgestaltend. Sie ist wesentlicher Bestandteil des Entwicklungsprozesses, in welchem ihr die Funktion der Qualitätsentwicklung und Qualitätssicherung zukommt. Sie kann sogar – wie Ehrlich (1995, 33) es ausdrückt – „Geburtshelfer" einer Idee und ihrer Realisierung sein. Gelegentlich wird diese Konstellation auch als „offene“ Evaluation bezeichnet, im Unterschied zu den zuvor geschilderten „geschlossenen“ Evaluationen, in denen Problem- und Fragestellungen, methodisches Vorgehen, Bewertungskriterien und die Zielgruppen der Evaluationsberichte von vornherein feststehen. Dagegen ist in „offenen“ Evaluationen nach einer Charakterisierung von Beywl „die Bestimmung der Feinziele, Fragestellungen, Hypothesen usw. zentrale Aufgabe des Evaluationsprozesses selbst. Der Evaluationsgegenstand ist lediglich vorläufig abgesteckt und wird im Fortgang der Untersuchung neu konturiert – je nach den Interessen der Organisationen, Gruppierungen oder Personen, die am Programm beteiligt sind. Besonders die Eingangsphase einer Evaluation, aber auch die anschließenden Erhebungs-, Auswertungs-, Interpretations- und Berichtsarbeiten werden auf die Wünsche der Beteiligtengruppen abgestimmt“ (Beywl 1991, 268). Die Funktion der Evaluation ist hier in erster Linie die eines Helfers und Beraters.[2]

2.4 Welches „Paradigma“ ist das zur Evaluation von Humandienstleistungen geeignetste?

Die Frage drängt sich auf – jedenfalls wenn man sich mit den Möglichkeiten von und den Anforderungen an die Evaluation von Humandienstleistungen befaßt -, ob sich nicht eines der drei skizzierten Paradigmen als besonders geeignet erweisen könnte. Schließlich handelt es sich hier um ein Feld mit besonderen Charakteristika: Humandienstleistungen erzeugen nicht ein konkretes „Produkt“, dessen Qualität mit einem Satz von Qualitätsindikatoren durch standardisierte Meßverfahren abgebildet werden kann; die erzielbaren Wirkungen sind in der Regel nicht unmittelbar gegenständlich sichtbar, oft sogar subjektiv schwer kommunizierbar (s. Müller-Kohlenberg 1997). Auch das professionelle Handeln ist – trotz aller Versuche zur Entwicklung verbindlicher Handlungsstandards – nur in geringem Maße standardisierbar; es bewegt sich in einem Feld der Interaktion mit den Klienten, in dem die komplexen Bedingungen des Einzelfalls darüber entscheiden, welches Handeln angemessen ist (Koditek 1997). In dieser Situation liegt es nahe, für das Entwicklungsparadigma in der Konkretisierung als Helfer- und Beratermodell der Evaluation zu plädieren, bis hin zur Vermittlung von Kompetenzen zur Selbstevaluation bei den im Humandienstleistungsbereich professionell Tätigen.

---

[1] Programme sind komplexe Handlungsmodelle, die auf die Erreichung bestimmter Ziele gerichtet sind, die auf bestimmten, den Zielen angemessen erscheinenden Handlungsstrategien beruhen und für deren Abwicklung finanzielle, personelle und sonstige Ressourcen bereitgestellt werden (Hellstern/Wollmann 1983, 7)

[2] Entsprechend findet sich manchmal auch die Charakterisierung als „Helfer- und Beratermodell der Evaluation“ (s. Abschnitt 4 dieses Beitrags).

Dennoch wäre eine Beschränkung auf dieses Paradigma kurzschlüssig. Je komplexer der Gegenstand der Evaluation, um so mehr besteht die methodologische Notwendigkeit der Triangulation durch Berücksichtigung unterschiedlicher Perspektiven. Damit meine ich nicht nur die Kombination verschiedenartiger Informationserhebungs*methoden* (z.B. qualitativ und quantitativ) und die Einbeziehung verschiedener *Akteursgruppen* (z.B. Projektträger – Geldgeber – Aufsichtsbehörden; Programmdurchführende – Mitarbeiter – Projektteam; Klienten – Adressaten – Nutzer – Betroffene; vgl. Müller-Kohlenberg 1997, 10 ff.), sondern eben auch die Wahl geeigneter, je nach Fragestellung durchaus unterschiedlicher Designtypen.

Thomas Koditek hat das für die sozialpädagogische Praxis und soziale Arbeit in einem Vortrag vor gut zwei Jahren deutlich herausgearbeitet (Koditek 1997). Darin verwies er *einerseits* auf den wachsenden „Innovations- und Legitimitätsdruck für die soziale Arbeit ... auf der Grundlage verschärfter externer Kostenvorgaben“ sowie auf die in diesem Zusammenhang geführte Debatte um neue Steuerungsmodelle, Qualitätssicherung und Qualitätsmanagement (a.a.O., 50). Auf der Seite der Evaluation entspricht dem natürlich das Kontrollparadigma.

Im Zuge der Forderungen nach kostensparender Reorganisation des öffentlichen Sektors insgesamt und der sozialpädagogischen Praxis im besonderen haben – so der *zweite* Argumentationsstrang von Koditek – „Organisations-, Management- und Personalentwicklungskonzepte auch in der sozialen Arbeit Konjunktur“. In dieser Debatte gehe es „nicht nur um die Entwicklung und Veränderung von Arbeitsbedingungen und professionellen Standards, sondern auch um die Beantwortung der zentralen Frage, wo die Kriterien in diesem Erneuerungsprozeß zu verorten seien“ (a.a.O.). Seine Antwort: „Der Weg dahin ist ... nur durch die jeweils konkret Beteiligten gestaltbar. Dies verpflichtet zu der Suche nach Verbindungen zwischen sozialpädagogischen Praxisforschern/Praxisforscherinnen und Praktikern/Praktikerinnen ..., die durch Prozesse der Interaktion und Kooperation erst geschaffen werden müssen. Es verpflichtet vor allem aber auch zu einer weitaus stärkeren und systematischen Beteiligung der Klienten (Klientenforschung) und zu weitgehender Integration ihrer Bedürfnisse in einen innovativen Gestaltungsprozeß“ (a.a.O., 52). Damit ist nun eindeutig das Helfer- und Beratermodell der Evaluation angesprochen.

Für diesen Gestaltungsprozeß aber – darauf weist Koditek *als drittes* ausdrücklich hin – ist gesichertes und praxisrelevantes Wissen notwendig. Aufgabe sozialpädagogischer Evaluation ist also ebenfalls die „Produktion von Grundlagenwissen in Bezug auf soziale Probleme sowie deren Verursachung“ und die „Erforschung der Lebenswelt potentieller oder realer Klientengruppen“. Damit ist nun auch das Forschungsmodell der Evaluation in die Pflicht genommen, denn – so Koditek – innovative Anstöße von Entwicklungsprozessen in sozialpädagogischen Institutionen haben „sozialpädagogische Praxisforschung bzw. die Evaluation sozialpädagogischer Praxis zur Voraussetzung“ (a.a.O., 53).

Mit der Wahl eines bestimmten Designtyps ist im übrigen – das wird häufig verwechselt – noch nicht entschieden, welche Methoden der Informationsbeschaffung und –analyse einzusetzen sind. Diese Entscheidung ist unabhängig davon so zu treffen, daß sie in bestmöglicher Weise den Bedingungen des Evaluationsfeldes und der beteiligten Akteure angepaßt sind. Selbstverständlich ist die Erhebung unmittelbar meßbarer Sachverhalte mittels standardisierter Verfahren und ist deren Auswertung mit dem Instrumentarium der Statistik vorzunehmen. Ebenso selbstverständlich sind in einem noch nicht durch gesichertes Wissen vorstrukturierbaren diffusen, komplexen Aktionsfeld offene Erhebungsverfahren sowie rekonstruktive Analysemethoden zu wählen. Darüber hinaus ist auch die Akzeptanz

bestimmter Forschungsstrategien bei den beteiligten Akteuren zu berücksichtigen. In der sozialen Arbeit, in der „weiche“ Verfahren auch die professionelle Tätigkeit bestimmen, sind „harte“ Forschungstechniken eher ein Störfaktor, der zu Widerständen führen kann. In der Schulmedizin oder in den Ingenieurwissenschaften, die ihre professionelle Tätigkeit und ihre professionellen Entscheidungen auf „harte“ Meßwerte stützen, stoßen dagegen rein qualitative Forschungstechniken auf Vorbehalte und auf eher geringe Akzeptanz. Ähnliches gilt tendenziell für die Entscheidungsebenen in Politik und Verwaltung.

2.4 Zur „Theorie der Evaluation“

Gefordert wird im akademischen Diskurs häufig eine generelle wissenschaftstheoretische und methodologische Fundierung angewandter Sozialwissenschaft, eine Theorie der sozialwissenschaftlichen Praxis – für unseren Kontext: eine „Theorie der Evaluation“. Gerade weil das Handlungsfeld so vielfältig ist und sich gegen Versuche der Standardisierung sperrt, werde ein Satz theoretischer Kategorien benötigt, anhand derer der konkrete Evaluationsfall präzise beschreibbar ist. Ohne einen solchen theoretischen Rahmen sei es sehr schwierig, die professionelle Qualität der Evaluation sowie die Aussagekraft und praktische Relevanz ihrer Ergebnisse einzuschätzen und ggf. zu kritisieren.

Shadish, Cook und Leviton (1991, 36-64) schlagen in diesem Zusammenhang eine Differenzierung nach fünf Dimensionen vor (Darstellung nach Rebien 1997, 441-444):

- *die soziale Intervention, auf die sich die Evaluation bezieht*: In welchem Kontext sozialer Veränderungen steht die Intervention, und wie wird sie von den Beteiligten und Betroffenen wahrgenommen? (Struktur und Funktion der Intervention, ihr externer Kontext, vorgesehener Ablauf)
- *die Konstruktion von Wissen durch die Evaluation*: Explikation der ontologischen, wissenschaftstheoretischen und methodischen Überlegungen des Evaluationsvorhabens, möglicher Bias des geplanten Vorgehens
- *Werte, Wertungen*: Wertungen sind in mehrfacher Hinsicht mit dem Evaluationsprozeß verbunden: als wirksam gewordene Prioritäten in politischen Entscheidungen, im Entstehungskontext der Evaluationsdaten, im programmbezogenen und Prioritäten setzenden Entscheidungssystem, in der Wertgeladenheit der Evaluationsdaten selbst, im spezifischen Nutzen der Daten für die Beteiligtengruppen
- *Nutzung, Verwertung*: Darstellung der potentiellen Nutzungsmöglichkeiten der Evaluation (instrumentell, konzeptionell, persuasiv) sowie Beschreibung des Zeithorizonts ihrer Verwertung (im Interventionsprozeß selbst, im politischen Prozeß, als zukünftiges Hintergrundwissen)
- *Praxis des Evaluationshandelns*: Unter Berücksichtigung gegebener Begrenztheit der zur Verfügung stehenden Zeit, Ressourcen und Qualifikationen haben Evaluatoren Designentscheidungen zu treffen und zu dokumentieren sowie damit verbundene trade-offs herauszuarbeiten.

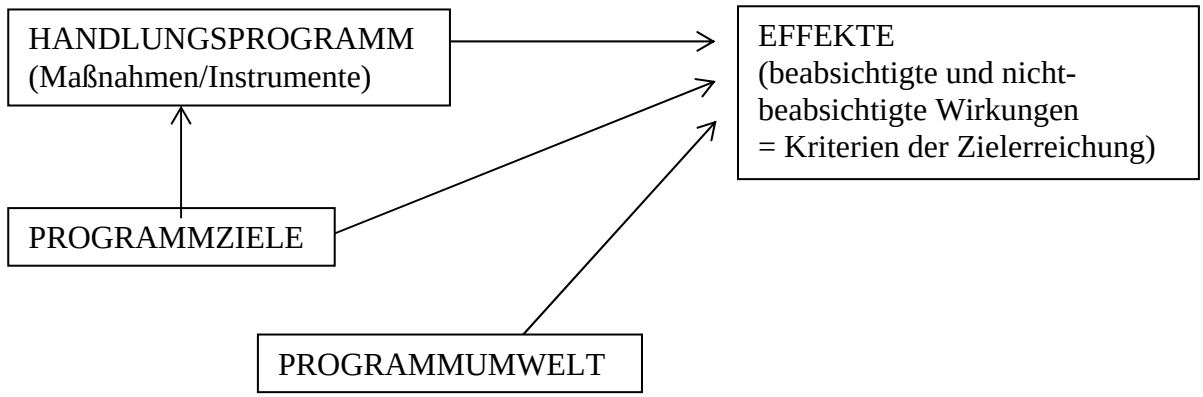
Wenn also Evaluationsvorhaben unter methodologischen und substanziellen Gesichtspunkten diskutierbar und kritisierbar sein sollen, wenn die mit dem Vorgehen verbundenen möglichen Fallstricke und Risiken erkennbar werden sollen, dann liefern die von Shadish, Cook und Leviton vorgeschlagenen Dimensionen ein nützliches Beschreibungs- und Argumentationsraster.

3. Das Leitkonzept für das Forschungs- und das Kontrollparadigma der Evaluation: Programmforschung

3.1 Begriffsexplikation

In den Vorbemerkungen habe ich darauf hingewiesen, daß die Fachsprache empirischer Wissenschaft sich vom unbestimmt-weiten, im Alltagssprachgebrauch und auch in der politischen Diskussion grassierenden Modebegriff ‚Evaluation' – *Irgend etwas wird von irgend jemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet* – durch eindeutige Präzisierungen absetzt. Da jedoch Präzisierungen zu den genannten vier Aspekten (Gegenstand – Evaluator – Kriterien – Verfahren) in unterschiedlicher Weise möglich sind und auch in unterschiedlichen Kombinationen vorkommen, sehen sich Evaluatoren einer solchen Vielfalt von Aufgabenprofilen und Rahmenbedingungen gegenüber, daß von einem vorherrschenden Evaluationsmodell und von einer *Methodik "der" Evaluation* nicht die Rede sein kann. Bei aller Vielfalt bleibt dennoch – zumindest für das Forschungs- und das Kontrollparadigma – allen Vorhaben gemeinsam, daß sie (mindestens) drei interdependente Dimensionen aufweisen – nämlich Ziele, Maßnahmenprogramm, Effekte – und daß sie (anders als in einem Forschungslabor) von Umgebungseinflüssen nicht abgeschirmt werden können.

Abbildung 1: Programmforschung



Die drei in der Abbildung dargestellten Programmdimensionen (Ziele – Maßnahmen – Effekte) können jeweils mehr oder weniger konkret oder abstrakt, mehr oder weniger festliegend oder variabel, mehr oder weniger ausformuliert oder nur implizit, mehr oder weniger offiziell oder informell sein. In jedem Fall aber orientieren die Beteiligten in dem zu evaluierenden Programm ihr Argumentieren und Handeln daran. Mit diesen drei Dimensionen muß sich daher auch jede Evaluation auseinandersetzen: Ungenaue Formulierungen von Zielen und Maßnahmen sind zu präzisieren und zu operationalisieren, implizit gelassene zu rekonstruieren, ungeordnete Ziele sind in einem Zielsystem zu ordnen, Zielkonflikte herauszuarbeiten. Ziele sind von Maßnahmen (als Instrumente zu deren Erreichung) abzugrenzen. Die Art und Weise der vorgesehenen Realisierung (Implementation) ist zu berücksichtigen und ggf. zu konkretisieren. Schließlich ist zu klären, was das Handlungsprogramm im Detail bewirken soll (und darüber hinaus bewirken kann): Welche

Veränderungen müssen in welcher Frist an welcher Stelle auftreten, damit die Ziele als erreicht gelten? Wie können sie festgestellt und gemessen werden? Wie können feststellbare Veränderungen als Wirkungen des Programms identifiziert und gegenüber anderen Einflüssen abgegrenzt werden?

Eine so umfassende Evaluation, wie sie nach dieser ersten groben Strukturierung des Aufgabenfeldes notwendig erscheint, ist in keinem Projekt realisierbar. Es müssen Schwerpunkte gesetzt werden. Hierzu sind vier zentrale Fragen zu beantworten:

- Was wird evaluiert? – Implementations- oder Wirkungsforschung
- Wann wird evaluiert? – Summative oder formative Evaluation
- Wo ist die Evaluation angesiedelt? – Externe oder interne Evaluation
- Wer beurteilt nach welchen Kriterien? – Instanzen der Evaluierung

Je nach deren Beantwortung lassen sich verschiedene Arten von Evaluation unterscheiden.

3.1.1 Implementations- oder Wirkungsforschung: Was wird evaluiert?

Die Unterscheidung bezieht sich hier auf den Gegenstand der Evaluation.

Stehen im Vordergrund die Effekte, die von den Maßnahmen eines Programms oder Projekts hervorgerufen werden, haben wir es mit *Wirkungsanalysen* (impact evaluations) zu tun. Im umfassendsten Fall kann sich das Bemühen darauf richten, möglichst alle, also nicht nur die intendierten Effekte (Zielvorgaben), sondern auch die unbeabsichtigten Konsequenzen und Nebenwirkungen – d.h. das gesamte "Wirkungsfeld" des Programms – zu erfassen.

Richtet sich der Blick nicht schwerpunktmäßig auf die Effekte, sondern steht die systematische Untersuchung der Planung, Durchsetzung und des Vollzugs im Vordergrund, spricht man von *Implementationsforschung*. Eine Hauptaufgabe der Evaluation ist die systematische und kontrollierte "Buchführung": Was passiert? Was wird wann und wie gemacht? (= "monitoring")

3.1.2 Summative oder formative Evaluation: Wann wird evaluiert?

Diese – ebenfalls gängige – Differenzierung bezieht sich auf den Zeitpunkt, an dem eine Evaluation ansetzt. Hier kann zwischen einer projektbegleitenden und einer abschließenden Evaluation unterschieden werden.

Da üblicherweise bei *begleitender Evaluation* zugleich regelmäßige Rückkoppelungen von Ergebnissen in das Projekt vorgesehen sind, hat die Forschung Konsequenzen für dessen Verlauf. Sie wirkt sozusagen programmgestaltend oder -formend. In einem solchen Fall spricht man deshalb von *"formativer" Evaluation*. Formative Evaluation ist definitionsgemäß besonders "praxisrelevant". Andererseits ist es besonders schwer, ihre Resultate im Sinne von Erfolgs- oder Wirkungskontrolle zu interpretieren, da die Forschung den Gegenstand der Bewertung selbst fortlaufend beeinflußt und verändert. Besonders geeignet ist sie dagegen als Instrument der Qualitätsentwicklung und/oder Qualitätssicherung. Anfangs- und Endpunkt einer formativen Evaluation sind methodisch nicht eindeutig definiert.

Eine erst gegen Ende oder gar nach Abschluß eines Projekts durchgeführte (oder erst dann zugänglich gemachte) Evaluation verzichtet explizit auf "projektformende" Effekte. Vielmehr gibt sie im Nachhinein ein zusammenfassendes Urteil, ein "Evaluationsgutachten" ab. Man

spricht hier von *"summativer" Evaluation*. Bei summativer Evaluation sind Anfang und Ende der Forschung klar definiert.

3.1.3 Externe oder interne Evaluation: Wo ist die Evaluation angesiedelt?

Diese dritte – und für die Praxis wichtige – Unterscheidung geschieht danach, wem die Evaluationsaufgabe übertragen wird.

In manchen Projekten ist die ständige Überprüfung und Ergebniskontrolle expliziter Bestandteil des Programms selbst. Die Informationssammlung und -einspeisung gehört als Instrument der Qualitätssicherung zum Entwicklungs- und Implementationskonzept. Da hiermit das eigene Personal des Projektträgers betraut wird, spricht man von *interner Evaluation*. Ihre Vorzüge werden darin gesehen, daß die Evaluation problemlosen Zugang zu allen notwendigen Informationen hat und während des gesamten Prozesses ständig "vor Ort" präsent ist. Probleme bestehen dagegen zum einen in der Gefahr mangelnder Professionalität, zum anderen im Hinblick auf die "Objektivität" der Resultate.

Werden dagegen die Dienste eines Forschungsinstituts oder außenstehender unabhängiger Forscher in Anspruch genommen, handelt es sich um *externe Evaluation*. Bei den meisten mit öffentlichen Mitteln geförderten Vorhaben ist eine externe wissenschaftliche Begleitung und/oder Begutachtung vorgeschrieben. Da es sich hierbei in der Regel um Forschungsexperten handelt, ist die notwendige Professionalität gewährleistet; und da die Evaluation ihre Arbeit nicht durch einen erfolgreichen Ablauf des zu begleitenden Projekts, sondern durch wissenschaftliche Standards zu legitimieren hat, kann auch von einem höheren Grad an Objektivität ausgegangen werden.

3.1.4 Instanzen der Evaluierung: Wer beurteilt nach welchen Kriterien?

Unter diesem Gesichtspunkt ist danach zu fragen, woher die Kriterien der Evaluation stammen und wer die Bewertungsinstanz ist.

Im "traditionellen" Fall stammen die Beurteilungskriterien aus dem zu evaluierenden Programm selbst. Seine Implementation sowie seine Wirkungen werden im Lichte seiner eigenen Ziele bewertet. Vorgenommen wird die Beurteilung vom Evaluationsforscher, der jedoch keine subjektiven Werturteile abgibt, sondern "*technologische Einschätzungen*" formuliert, die intersubjektiv nachprüfbar sein müssen (Vorher-nachher-Vergleich verbunden mit dem Vergleich des Soll-Zustands mit dem erreichten Ist-Zustand).

Ein solches Vorgehen verlangt relativ umfassendes theoretisches Wissen über die Struktur der Zusammenhänge zwischen Zielen, Maßnahmen, Wirkungen und Umwelteinflüssen, das jedoch gerade im Falle von Pilotprojekten und Modellversuchen nicht vorhanden ist. Hier behilft sich die Evaluation häufig damit, daß die eigentliche Bewertung auf *programm- und evaluationsexterne Instanzen* verlagert wird. Beispielsweise können Fachgutachten eingeholt werden. Oder es werden neutrale Experten befragt, die sich thematisch besonders intensiv mit projektrelevanten Themen befaßt haben oder die durch berufliche Erfahrungen mit ähnlich gelagerten Aufgaben ausgewiesen sind.

Als eine Variante des Verlagerns der Evaluierung auf eine programmexterne Instanz wird verschiedentlich die *Befragung der Adressaten eines Programms* (Nutzer oder Betroffene) favorisiert. Die Begründung fällt scheinbar leicht: Die Nutzer einer Dienstleistung, die Betroffenen einer Maßnahme sind die "eigentlichen" Experten. Sie haben den Gegenstand der Untersuchung aus eigener Erfahrung kennengelernt und wissen, wie er – bei ihnen – wirkt.

Bei den so erhobenen Urteilen handelt es sich allerdings weder um Bewertungen im Sinne "technologischer" Evaluationseinschätzung noch um Bewertungen neutraler Experten. Es sind vielmehr "Akzeptanzaussagen" von Personen, die in einer besonderen Beziehung (eben als Nutzer, als Betroffene) zum Untersuchungsgegenstand stehen. Folgerichtig wird diese Evaluationsstrategie als *Akzeptanzforschung* bezeichnet (s.u.: Abschnitt 3.3.2).

3.2 Methoden der Programmforschung

Die Methodologie der Programmforschung wurde im wesentlichen in den 70er und 80er Jahren entwickelt. Je nachdem, ob ein Evaluationsprojekt mehr in Richtung Wirkungsforschung oder mehr in Richtung Erfolgskontrolle tendiert, hat sich der Forscher zwar auf in der Gewichtung unterschiedliche Voraussetzungen und Anforderungen einzustellen. Gemeinsam bleibt aber allen Projekten die auf den ersten Blick simpel anmutende, praktisch jedoch kaum lösbare Aufgabe, die in Abb. 1 aufgeführten vier Variablenbereiche (Ziele – Maßnahmen – Effekte – Programmumwelt) mit empirischen Daten abzubilden (zu "messen") und miteinander zu verknüpfen. Wirkungs- und Erfolgskontrolle orientiert sich dabei am Modell der Kontrolle der "unabhängigen" bzw. "explikativen" Variablen (hier: Maßnahmen des Programms) und der Feststellung ihrer Effekte auf genau definierte "abhängige" Variablen (Zielerreichungs-Kriterien).

An Forschungsaufgaben folgen daraus:

- Messung der "unabhängigen Variablen", d.h.: das Handlungsprogramm mit seinen einzelnen Maßnahmen ist präzise zu erfassen;
- Identifizierung und Erfassung von Umwelt-Ereignissen und -Bedingungen, die ebenfalls auf die vom Programm angestrebte Zielsituation Einfluß nehmen könnten (exogene Einflüsse);
- Messung der "abhängigen Variablen", d.h.: das Wirkungsfeld (beabsichtigte und nicht-beabsichtigte Effekte) ist zu identifizieren, die Wirkungen sind anhand definierter Zielerreichungs-Kriterien (operationalisierter Ziele) zu messen.

Die Aufgabe der Datenerhebung besteht für die gesamte Dauer des Programmablaufs in einem – so Eekhoff u.a. 1977, 11ff. – "Monitoring" der Instrumentvariablen (Programm-Input), der exogenen Einflüsse und der Zielerreichungsgrade (Output). Methodisch gesehen handelt es sich bei diesem dreifachen "Monitoring" somit um vergleichsweise einfache, *deskriptive* Forschungsaktivitäten.

Wesentlich schwerer zu lösen ist die darauf folgende *analytische* Aufgabenstellung: Die festgestellten Veränderungen im Wirkungsfeld des Programms sind aufzubrechen

- in jene Teile, die den jeweiligen Maßnahmen als deren Wirkung zurechenbar sind,
- und in die verbleibenden Teile, die als Effekte exogener Einflüsse (Programmumwelt) zu gelten haben.

Die eigentliche "Erfolgskontrolle" oder "Evaluation" beinhaltet nach diesem Modell zwei Aspekte:

- Analyse der Programmziele und ihrer Interdependenzen (Präzisierung eines Zielsystems einschließlich der Festlegung des angestrebten Zielniveaus) sowie Zuordnung der Instrumente zur Zielerreichung (Maßnahmen des Programms);

- Vergleich der den einzelnen Maßnahmen zurechenbaren Effekte mit den angestrebten Zielniveaus.

Das damit skizzierte Modell einer kausalanalytisch angeleiteten Programmevaluations- und Wirkungsforschung wirkt in sich schlüssig und einleuchtend und scheint nur noch einer weiteren Differenzierung hinsichtlich der Methodik zu bedürfen (was im folgenden geschehen soll). Bei näherem Hinsehen allerdings wird erkennbar, daß es von anspruchsvollen Voraussetzungen über den Gegenstand der Untersuchung wie auch von Voraussetzungen bei den programmdurchführenden Instanzen und der Evaluation selbst ausgeht. Diese mögen zwar bei Vorhaben der Grundlagenforschung (vereinzelt) gegeben sein, sind jedoch in Programmforschungsprojekten wenig realitätsnah. Drei dieser meist implizit gelassenen Voraussetzungen sind besonders hervorzuheben, da deren Erfüllung eine wesentliche Bedingung dafür ist, das methodologische Forschungsprogramm empirischer Kausalanalysen überhaupt anwenden zu können:

- Vor der Entwicklung des Forschungsdesigns muß Klarheit über die Untersuchungsziele – bezogen auf einen definierbaren und empirisch abgrenzbaren Untersuchungsgegenstand – bestehen. Für die Dauer der Datenerhebung dürfen sich weder die Untersuchungsziele noch die wesentlichen Randbedingungen des Untersuchungsgegenstandes in unvorhersehbarer Weise ändern.
- Vor der Entwicklung des Forschungsdesigns müssen des weiteren begründete Vermutungen (Hypothesen) über die Struktur des Gegenstandes wie auch über Zusammenhänge und Beziehungen zwischen dessen wesentlichen Elementen existieren, nach Möglichkeit in Form empirisch bewährter Theorien. Erst auf ihrer Basis kann ein Gültigkeit beanspruchendes Indikatorenmodell konstruiert, können geeignete Meßinstrumente entwickelt, kann über problemangemessene Auswertungsverfahren entschieden werden.
- Der Forscher muß die Kontrolle über den Forschungsablauf haben, um die (interne und externe) Gültigkeit der Resultate weitestgehend sicherzustellen.

Im Normalfall der Begleitforschung zu Programm-Implementationen oder gar zu Modellversuchen neuer Techniken, neuer Schulformen, zur Erprobung alternativer Curricula oder Lernformen u.ä. ist keine einzige dieser Bedingungen voll erfüllt. Die Untersuchungssituation weist vielmehr in dieser Hinsicht erhebliche "Mängel" auf (ausführlicher dazu Kromrey 1988). Die im folgenden skizzierte Methodologie der Programmevaluation ist daher weniger ein Real- als ein Idealtyp, an den anzunähern der Forscher sich je nach gegebener Situation bemühen wird.

3.2.1 Ziel- und Maßnahmenanalyse

Nimmt man die gängigen Definitionen von "Interventionsprogramm" oder "Handlungsprogramm" beim Wort (vgl. Anmerkung 1), dann müßten die Kenntnisse, die man sich durch die Forschung erhofft, bei den Akteuren weitestgehend schon vorhanden sein: Neben einem widerspruchsfreien Zielsystem müßte zuverlässiges Praxiswissen existieren, um – auf der Basis von Daten über die gegebene Ausgangssituation – die erforderlichen Maßnahmen und Instrumente zur Erreichung der Zielsituation zu bestimmen. Solche Kenntnisse über Ziel-Mittel-Relationen müßten zudem technologisch verwertbar sein; d.h. die als strategisch wichtig erkannten Variablen müßten dem Eingriff der Programmdurchführenden zugänglich sein. Mayntz (1980, 4) weist jedoch mit Recht darauf hin, "daß nur im Ausnahmefall ein Programm zu Beginn des Implementationsprozesses als konkrete, faßbare Einheit vorliegt".

So sind die Ziele oft nicht eindeutig und nicht konkret, sondern vage und leerformelhaft formuliert. Das kann ganz bewußt im Prozeß der Programmaushandlung geschehen sein (um einen Kompromiß zwischen widerstreitenden Interessen zu ermöglichen oder um Konflikte zwischen Koalitionspartnern zu vermeiden). So wird die Aufgabe der Präzisierung aus der (politisch-öffentlichen) Zielfindungs- in die (weniger öffentliche) Implementierungsphase verschoben. Besonders problematisch ist dies, wenn am Programmvollzug mehrere Ebenen beteiligt sind (Bund – Länder – Kommunen – andere Träger). Auf jeder Ebene von Akteuren kann die Ausfüllung der Ziel-Leerformeln in unterschiedlicher Weise geschehen. Gleiches kann bei der Zuordnung von Maßnahmen (Instrumenten) der Fall sein, die der Zielerreichung dienen sollen. So können im Prinzip unterschiedliche Maßnahmen mit dem gleichen Programm vereinbar erscheinen. Eine mögliche Folge ist, daß bestimmte Träger ihre bereits auf Vorrat bestehenden "Schubladenprogramme" unter das neu beschlossene Programm subsumieren. In solchen Fällen entstehen Diskrepanzen zwischen offiziellen (manifesten) und verdeckten (latenten) Programmzielen. Außerdem können Ziele, die in ihrer vagen Formulierung als miteinander vereinbar erschienen, sich bei der Konkretisierung als im Widerspruch zueinander stehend erweisen. Des weiteren können zu Beginn gesetzte Ziele (selbst wenn sie präzise formuliert waren) sich im Laufe der Programmrealisierung ändern oder in ihrer Gewichtung verschieben, etwa weil sich wichtige Rahmenbedingungen für das Programm in nicht erwarteter Weise entwickelt haben. In manchen Programmen schließlich finden sich eher *Kataloge von Maßnahmen* statt eindeutiger Ziele, so daß unbestimmt bleibt, was mit dem Programm letztlich erreicht werden soll, welches also die angestrebten Effekte sind.

Mag es bei der *Formulierung von Programmzielen* im politischen Aushandlungsprozeß durchaus funktional sein, diese bewußt vage und mehrdeutig zu lassen, so darf die Interpretation empirischer Befunde dagegen nicht "aushandelbar" sein. Das bedeutet, daß ungenaue Zielformulierungen im Zuge der Designentwicklung konkretisiert werden müssen. Präzise Aussagen über die genannten Aspekte (Ziele – Instrumente – Ziel/Mittel-Relationen) sind schon rein forschungstechnisch unabdingbare Voraussetzung, um überhaupt eine Evaluation im hier verstandenen Sinne vornehmen zu können (darauf wurde bereits mehrfach hingewiesen). Der Evaluationsforscher ist somit im Falle "unvollständiger" Programme gezwungen, Lücken zu schließen und die notwendigen Präzisierungen vorzunehmen sowie im Falle widersprüchlicher Formulierungen Konflikte und Unverträglichkeiten zwischen den Zielen herauszuarbeiten und zu beseitigen – selbst auf die Gefahr hin, daß damit das Programm in Teilen zu einem "Konstrukt des Forschers" wird (Mayntz 1980, 4).[3]

Ein erster Ansatz ist eine eher "technische" Zielanalyse in der Absicht, die Liste der Programmziele zu komplettieren, ein hierarchisches Zielsystem zu konstruieren (ausgehend von Oberzielen über Haupt-, Teil- und Unterziele bis hin zu Indikatoren, die als Näherungskriterien den Grad der Zielerreichung zu messen gestatten) und die jeweils einzusetzenden Instrumente zuzuordnen (dazu und zum folgenden ausführlich Hellstern / Wollmann 1983, 11 ff.). Ausgangspunkt sind die vorhandenen Angaben im Programm; daneben sind relevante ergänzende Dokumente (Beratungsprotokolle, Grundsatz-Aussagen

[3] Dies impliziert ein häufig nicht gesehenes Wertproblem für die Programmforschung, nach deren Konzept die „Evaluierung“ kein (subjektives) Werturteil ist, sondern eine technologische Wertung: Vergleich vorgegebener Sollwerte (= Programmziele) mit empirisch erhobenen Daten über feststellbare Effekte. In dem Maße aber, wie bei der Rekonstruktion des Zielsystems als Wertbasis der Evaluation Ergänzungen vorzunehmen und Inkonsistenzen zu beseitigen waren, werden zwangsläufig Wertungen/Werturteile der Forscher Bestandteil der Wertbasis und die Evaluierung verliert ihren Status als lediglich technologisches Vergleichsurteil.

der beteiligten Organisationen/Parteien etc.) heranzuziehen sowie ggf. Beteiligte am *Entscheidungsprozeß* zu befragen. Ein weiterer Zugriff bietet sich über die Personen und Institutionen an, die für die *Implementation* verantwortlich sind (insbesondere wenn sich der Vollzug des Programms über mehrere Ebenen erstreckt). Eine Zielpräzisierung aus der Sicht der Beteiligten bietet bereits zu Beginn wichtige Anhaltspunkte, ob und in welcher Weise es im Verlaufe des Programmvollzugs zu Zielverschiebungen kommen dürfte. Bei umfassender Evaluation (comprehensive evaluation)[4] ist schließlich auch noch die Perspektive der *Zielgruppen des Programms* (Nutzer, Betroffene) bedeutsam. Sie erlaubt eine schon frühzeitige Aussage darüber, inwieweit das Programm die Bedürfnisse derer trifft, für die es konzipiert wurde, und insofern überhaupt die Mindestvoraussetzungen für einen "Erfolg" erfüllt.[5] Manche Programme bleiben hinsichtlich ihrer Zielpopulation so unbestimmt, daß eine Zielgruppenanalyse schon aus evaluationstechnischen Gründen unumgänglich ist: nämlich zur Bestimmung des potentiellen Wirkungsfeldes des Programms (Wer wird es in Anspruch nehmen? Wer wird von möglichen Aus- und Nebenwirkungen betroffen sein?).[6]

Das erstellte Zielsystem hat drei formalen Kriterien zu genügen: Konsistenz, Operationalisierbarkeit, Praktikabilität. Ein *inkonsistentes* System von Zielen (das also nicht in sich geschlossen und logisch widerspruchsfrei ist) kann nicht Grundlage für eine rationale Analyse und für die Bewertung eines Programms sein. *Nicht-operationalisierbare* Ziele sind nicht durch Daten abbildbar, können somit nicht Gegenstand empirischer Forschung sein. In Bezug auf den Zielinhalt sind Ziele operational, wenn sich gegenüber einer bestehenden Ausgangssituation die angestrebte Zielsituation genau herleiten und anhand geeigneter Indikatoren (Zielerreichungs-Kriterien) messen läßt. *Praktikabel* schließlich sind Ziele dann, wenn sie auf praktisches Handeln gerichtet sind und ihre Verwirklichung kontrolliert werden kann. Zielaussagen zu Sachverhalten, die im Rahmen des Programms nicht Gegenstand von planenden und handelnden Eingriffen sein können, sind für den Praktiker irrelevant;

---

[4] Eine umfassende Evaluation (Rossi/Freeman 1988, Rein 1981) bestünde in einer "systematischen Anwendung rationaler Methoden, um die Konzeptualisierung und Planung, Implementierung und Nützlichkeit eines sozialen Interventionsprogramms zu untersuchen". Sie beträfe "Fragen nach der Art, dem Ausmaß und der Verteilung des jeweiligen Problems, den Zielen und der Angemessenheit eines Programms, dem planmäßigen Ablauf der Intervention, dem Ausmaß, mit dem die beabsichtigten Änderungen bei der Zielpopulation erreicht werden, den Nebenwirkungen sowie der Nützlichkeit des Programms entsprechend Kosten-Effektivitäts- bzw. Kosten-Nutzen-Analysen" (Lösel/Nowack 1987, 57).

[55] Dieser Aspekt steht in besonderer Weise im Zentrum der „utilization focused evaluation“ (Patton 1997).

[6] Manche Autoren gehen noch ein Stück weiter und fordern, daß Evaluation nicht ausschließlich aus der Sicht und nach den Kriterien der politisch-administrativen Entscheidungsebene oder der Programmdurchführenden vorgenommen werden dürfe. Durch die Orientierung auf Verwaltungsinteressen würden der Sozialwissenschaft "Scheuklappen aufgezogen"; es würden alle jene Themenbereiche ausgeblendet, die nicht bereits von Politik und Planung als krisenhaft und zugleich auch als prinzipiell politisch-administrativ regelbar wahrgenommen worden sind. Solche Wirkungsanalysen seien zwar geeignet, "die Effektivität bei der staatlichen Bearbeitung bereits erkannter Probleme [zu] steigern. Die Belange bislang unberücksichtigter Interessen kann [sie] nur insoweit thematisieren, als sie im Brunnen nach den hineingefallenen Kindern forscht" (Häußermann / Siebel 1978, 493). Als methodisches Instrument, um Evaluation auch innovativ wirken zu lassen, schlägt Sjoberg eine – von ihm so genannte – "Gegensystemanalyse" (countersystem analysis) vor: Weder dürfe der Evaluationsforscher die herrschenden Systemkategorien akzeptieren noch unhinterfragt die Kategorien der Betroffenen übernehmen; vielmehr müsse er "alternative Ordnungen" formulieren, um Möglichkeiten zu erkennen, die gegenwärtige Situation zu "transzendieren" und für die Evaluation eine utopisch-denkbare Zielsituation als Vergleichsmaßstab zu erhalten (Sjoberg 1983, 81 ff.).

Forschungsbefunde zu Aspekten, die im Zuständigkeitsbereich der Programmdurchführer nicht veränderbar sind, tragen in diesem Kontext nicht zu praxisrelevantem Wissen bei und sind aus dieser Sicht "wertlos".

Eine wichtige Unterscheidung – die nicht in allen Programmen vorgenommen wird – ist die zwischen Zielen und Maßnahmen/Instrumenten. Ziele geben an, was erreicht werden soll. Instrumente sind die Hilfsmittel, die einzusetzen sind, um die Ziele zu erreichen. Diese eindeutig klingende Abgrenzung ist jedoch nicht so simpel, wie es den Anschein hat, insbesondere dann nicht, wenn mehrere Ebenen an der Durchführung beteiligt sind. Je nach Betrachtungsperspektive kann ein und derselbe Sachverhalt ein Ziel oder aber eine Maßnahme sein.

Als *Beispiel* seien aus einem (denkbaren) Programm zur Verbesserung der Wohnqualität in innerstädtischen Altbaugebieten die *Unterziele* "Lärmschutz" und "Energieeinsparung" herausgegriffen. Ein geeignetes *Instrument* könnte – "vor Ort", d.h. bei der Zielpopulation des Programms – der Einbau neuer Wohnungsfenster mit Mehrfachverglasung sein (Zielerreichungs-Kriterien: Verringerung des Geräuschpegels in den zur Straße gelegenen Räumen zur Hauptverkehrszeit, gemessen in Dezibel; Verringerung der Heizkosten, gemessen in DM); die Maßnahme wäre effektiv (sie wirkt) und zugleich effizient (sie wirkt gleichzeitig positiv auf beide Ziele). Die programmdurchführende Instanz – das zuständige kommunale Amt für Wohnungswesen – wird es dagegen als ihr "Ziel" betrachten, möglichst viele private Investoren (Wohnungseigentümer, Vermieter, Baugenossenschaften) dazu zu bewegen, den Einbau neuer Fenster vorzunehmen. Ihre "Instrumente" sind: Öffentlichkeitsarbeit, um private Investoren auf die Möglichkeit der Inanspruchnahme öffentlicher Förderungsmittel hinzuweisen, sowie möglichst schnelle und unbürokratische Bearbeitung von Förderungsanträgen (Zielerreichungs-Kriterien: bewilligtes Förderungsvolumen, durchschnittliche Bearbeitungsdauer der Anträge).[7] Für den Eigentümer von Gebäuden oder Wohnungen – durch seine Investition der eigentliche "Durchführende" des Programms – ist ein ganz anderes Ziel maßgeblich, das im Programm überhaupt nicht aufgeführt wird: die Wertsteigerung, zumindest die Substanzsicherung seines Kapitals, als Vermieter auch die Steigerung seiner Kapitalrendite.

Allerdings ist kein Evaluationsvorhaben so umfassend realisierbar, daß alle aus den verschiedenen Beteiligten-Perspektiven erstellbaren Ziel-Mittel-Systeme als alternative Maßstäbe an die Bewertung angelegt werden können. Somit ist die vorherige bewußte und begründete Entscheidung über den Verwendungszusammenhang notwendig: Wer ist Adressat der Evaluationsaussagen, und welchem Zweck sollen die Ergebnisse dienen? Offensichtlich kann also die Ziel-/Maßnahmen-Analyse niemals eine (interessenneutrale) Rekonstruktion und Präzisierung "des Programms" sein, sondern immer nur eine Perspektive, unter der das komplexe Gefüge Programm/Beteiligte/Umwelt betrachtet und untersucht wird. Selbst im Idealfall umfassender Evaluation können nur wenige ausgewählte Perspektiven evaluationsrelevant werden.

3.2.2 Konzipierung der Wirkungen (Modell des Wirkungsfeldes)

Mit der Ziel- und Maßnahmenanalyse ist zwar die Voraussetzung für die *Evaluierung*, nicht jedoch für die *Wirkungs*analyse geschaffen. Programmforschung erfolgt nicht in einer Laborsituation, in der jede einzelne Maßnahme isoliert auf ihre Effekte hin untersucht und in der die Wirkung der isolierten Maßnahme von allen übrigen Einflüssen ("Störgrößen") abgeschirmt werden könnte. Die Forschung hat es auch nicht mit einfachen Kausalketten zu tun (Maßnahme X_1 bewirkt über die intervenierenden Zwischenschritte Y_1 und Y_2 die Ver-

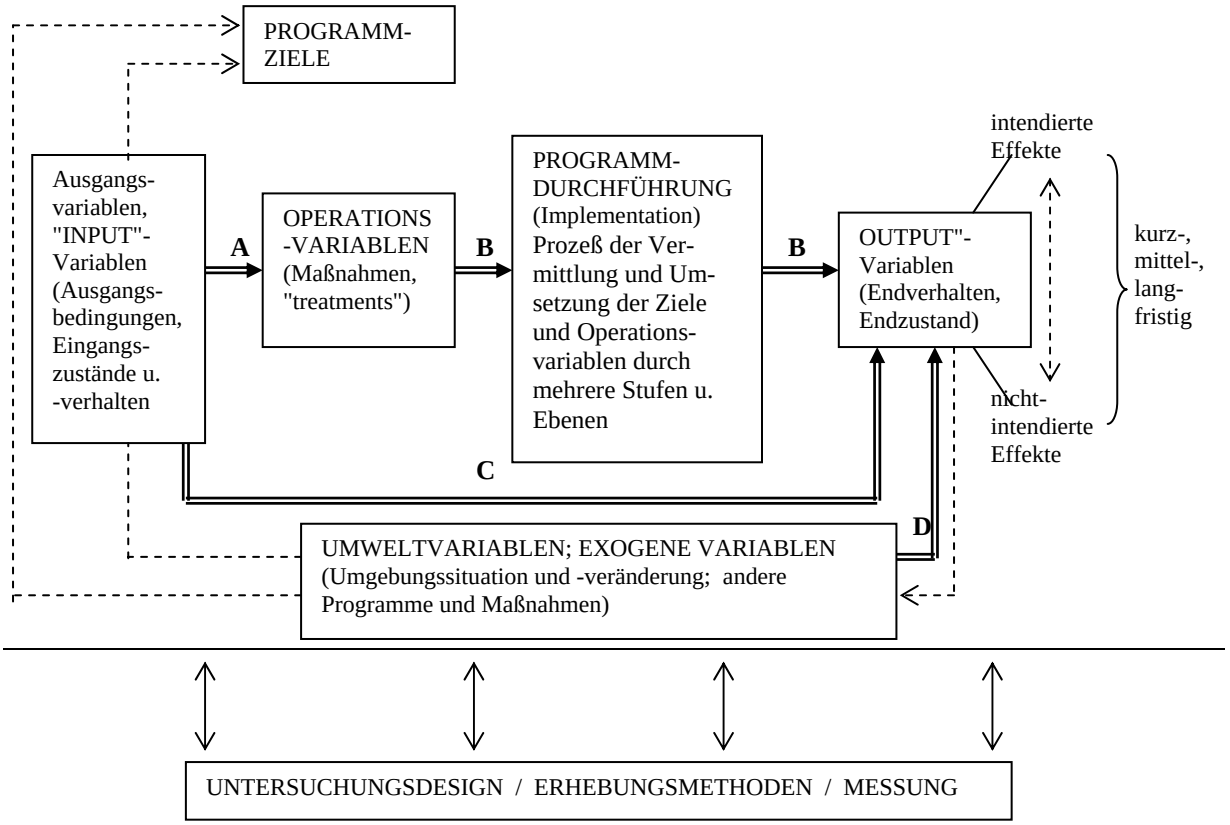
[7] Das Beispiel kann von Ebene zu Ebene mit wechselnden Ziel-Mittel-Konstellationen bis zu den politischen Parteien in den parlamentarischen Beschlußgremien weitergeführt werden.

änderung der Zielsituation von Z_0 nach Z_1), sondern mit einem komplexen *Wirkungsfeld*. Insbesondere wirkt eine Maßnahme nicht (trennscharf) nur auf ein Ziel und ist ein Ziel nicht (monokausal) nur durch eine Maßnahme erreichbar. Die Wirkungen treten nicht sämtlich zu gleicher Zeit ein; es ist zwischen kurz-, mittel- und langfristigen Effekten zu unterscheiden. Schließlich kann es neben den beabsichtigten auch zu ungeplanten Wirkungen und Wechselwirkungen kommen. Die Programmumwelt (Maßnahmen anderer Programme oder sozialer Akteure) können auf den Verlauf des zu evaluierenden Programms in erwünschter wie auch in unerwünschter Richtung Einfluß nehmen (und umgekehrt: das Programm kann seine Umwelt verändern).

Ziel des zu konzipierenden Wirkungsmodells ist es also, möglichst alle für die Beurteilung des Programmverlaufs relevanten (potentiellen) Wirkungen und Wirkungszusammenhänge auf gedanklicher Ebene vorab zu explizieren. Kriterien für die Entscheidung, was das "relevante Wirkungsfeld" sein soll, liefern die Programmziele. Benötigt werden aber darüber hinaus begründete Vermutungen und empirisch abgesicherte Theorien, die geeignet sind, die wechselseitigen Beziehungen zwischen Zielen, Maßnahmen und Umfeld zu antizipieren. Dieses Denkmodell eines Wirkungsfeldes – die formale Darstellung einer Theorie des Handlungsprogramms und seiner Einbettung in soziale Realität – ist die Basis für die Entwicklung eines "maßgeschneiderten" Forschungsdesigns. Dieses Design soll einerseits unter methodologischen Gesichtspunkten möglichst hohen Standards der empirischen Wissenschaft genügen, zugleich aber unter den durch das Programm gesetzten Rahmenbedingungen realisierbar sein. Jede ernstzunehmende Grundlagenforschung würde unter solchen Bedingungen versuchen, die Fragestellung so weit einzuschränken, daß möglichst alle wesentlichen (potentiellen) Einflußgrößen methodisch kontrolliert werden können. Evaluationsforschung darf genau diese Strategie – sollen die Ergebnisse ernstgenommen werden – nicht verfolgen, sondern muß möglichst viel Komplexität in ihrer Untersuchungsanlage mitberücksichtigen.

Das zu Beginn skizzierte Modell der Programmforschung stellt sich nach diesen Überlegungen nun wie folgt dar (wobei die methodischen Probleme der Indikatorenbildung und -messung in jedem der Variablenfelder sich nicht in anderer Weise stellen als bei jeder empirischen Untersuchung und daher hier nicht gesondert behandelt werden):

Abb. 2:
Grobes Variablenmodell einer Evaluationsstudie



Erläuterung: (die Pfeile symbolisieren die Wirkungsrichtung)

- ⇢ im allgemeinen nicht Gegenstand von Evaluationsstudien
- **A** ⟹ Maßnahmen orientieren sich am "Input" und an den Zielen
- **B** ⟹ Maßnahmen werden durchgeführt und haben "Wirkungen"
- **C** ⟹ "Reifungsprozesse" (Zustände ändern sich im Zeitablauf ohne Einwirkung von "Maßnahmen"
- **D** ⟹ externe Effekte (Wirkungen der Programmumwelt)

⇕ Zustände der Realität werden durch Meßinstrumente abgebildet ("Daten"); die verwendeten Untersuchungsmethoden sind jedoch niemals völlig neutral, sondern haben (insbesondere bei wiederholter Anwendung) Auswirkungen auf das Meßergebnis

Diese aus der Forschungslogik empirischer Theorie- bzw. Hypothesentests entlehnte Konzeption führt jedoch zu einem gewissen Paradoxon, wenn sie – wie hier – auf angewandte Forschung in einem Aufgabenfeld übertragen wird, das charakteristischerweise "Neuland" ist. Zu evaluierende Programme, sozialwissenschaftlich zu begleitende Modellvorhaben bewegen sich in einem Praxisfeld, in dem es gerade sowohl an theoretischem als auch an Erfahrungswissen mangelt (andernfalls wäre ihre Evaluierung überflüssig).[8] Bei der Konzipierung der Untersuchung sind dennoch die wesentlichen

[8] Außer Betracht bleibt bei der folgenden Darstellung die zusätzliche Schwierigkeit, daß je nach Programmtyp unterschiedliche Probleme auftreten. Bei einem explizit formulierten Programm mit feststehendem Katalog von Maßnahmen und Zuständigkeiten sind dies andere als bei einem Rahmenprogramm oder einem Projekt, das durch Setzen von „incentives“ die aktive Mitwirkung der Nutzer anregen und beeinflussen will. Die beiden letzteren Programmtypen leiden über die hier angesprochenen Probleme hinaus insbesondere an einer chronischen Unbestimmtheit des Wirkungsfeldes.

Wirkungszusammenhänge zu antizipieren, und zwar – anders als bei Forschungen zum Zwecke der Theorieentwicklung – nicht "versuchsweise", um dann empirisch auf ihre Haltbarkeit überprüft zu werden, sondern unmittelbar anwendungsbezogen. Die Befunde sollen anschließend nicht im "akademischen Elfenbeinturm" abgewogen werden, sondern sollen für das laufende Handlungsprogramm und für ggf. sich anschließende Entscheidungen Geltung haben. Das Paradoxon besteht nun in folgendem: Einerseits müssen unter *methodologischen* Gesichtspunkten empirisch bestätigte Theorien über die Struktur des Untersuchungsfeldes verlangt werden; sie sind als Basis für die Entwicklung eines auf Wirkungsmessungen angelegten Designs erforderlich. Andererseits existiert *logischerweise* bei "neuen" Untersuchungsfeldern das empirisch gesicherte Wissen noch nicht; es kann erst durch die noch durchzuführende Untersuchung gewonnen werden.

Dieses Dilemma ist nur durch einen Verstoß gegen die Methodologie traditioneller empirischer Forschung auflösbar. Den oben festgestellten "Mängeln" der Untersuchungssituation (fehlende Konstanz des Programms und der Rahmenbedingungen) kann nur mit einem analogen "Mangel" des Designs begegnet werden: Das zu Beginn formulierte Wirkungsmodell ebenso wie das darauf zugeschnittene Forschungsdesign sind "veränderungsoffen" anzulegen. Die Forschungslogik strukturtestender Verfahren (einschließlich standardisierter Erhebungsmethoden) ist zu ergänzen um Verfahren strukturentdeckender Forschung (etwa das von Glaser/Strauss entwickelte Konzept zur "Entdeckung einer gegenstandsbezogenen Theorie", vgl. Kromrey 1994a). Das zu entwickelnde und schrittweise zu optimierende Design hat jedoch nicht nur – wie bei jeder empirischen Datenerhebung – zu gewährleisten, daß die in Abb. 2 dargestellten Variablengruppen möglichst zuverlässig und verzerrungsfrei gemessen werden können, es muß darüber hinaus die Voraussetzungen für eine Kausalanalyse der Daten schaffen. Idealtypisch geeignet ist die (feld)experimentelle Vorgehensweise. Sie darf dementsprechend als der "Königsweg" der Programmforschung gelten, ist allerdings – wegen des Primats des Programms vor der Forschung – im Normalfall nur näherungsweise (in Form diverser quasi-experimenteller Ansätze) realisierbar.

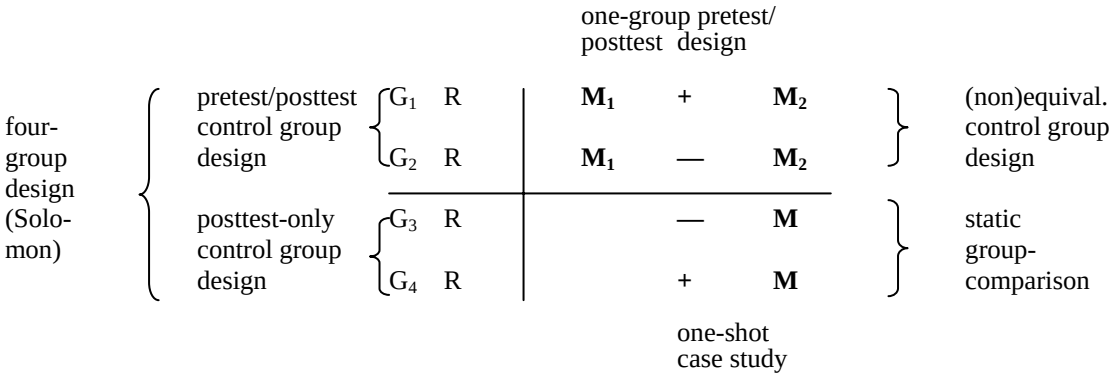
3.2.3 Das Feldexperiment als Referenzdesign für die Programmforschung

Das Design eines "echten" Experiments zeichnet sich dadurch aus, daß es mindestens die folgenden Merkmale aufweist:

- Es existiert eine Experimentalgruppe G_1, die dem experimentellen Stimulus X, dem "treatment" (hier: der auf ihre Auswirkungen zu untersuchenden Maßnahme), ausgesetzt wird.
- Es existiert eine in allen wesentlichen Merkmalen äquivalente Kontrollgruppe G_2, die dem experimentellen Stimulus nicht ausgesetzt wird, die also von der Maßnahme "verschont" bleibt.
- In beiden Gruppen werden vor dem Zeitpunkt des treatments und ausreichende Zeit danach die Ausprägungen der abhängigen Variablen (Merkmale, bei denen man Auswirkungen durch die Maßnahme erwartet) gemessen (M_1 und M_2).
- Stimmen vor dem treatment in der Experimental- und in der Kontrollgruppe die Verteilungen der abhängigen Variablen überein (und das sollten sie bei äquivalenten Kontrollgruppen), und sind nach dem treatment Unterschiede zwischen den Gruppen feststellbar, dann werden diese Unterschiede als Effekte des treatments (als Auswirkungen der Maßnahme) interpretiert.

Dieses Design kann noch um zwei weitere Gruppen (eine Experimental- und eine Kontrollgruppe, G_3 und G_4) erweitert werden, in denen man auf die Messung vor dem treatment verzichtet. Dadurch wird kontrolliert, ob nicht allein durch die Vorher-Messung schon Veränderungen in Gang gesetzt wurden (Versuchskaninchen-Effekt).

Abb. 3:
Designstrukturen bei Experimenten (nach Frey/Frenz 1982, 250)



Erläuterung:
G = Gruppe; R = Randomisierung; M = Messung; + = "treatment"

Für Untersuchungsgegenstände, bei denen in Bevölkerungs-Teilgruppen jeweils unterschiedliche Auswirkungen der gleichen Maßnahmen möglich sein könnten (z.B. alte Leute gegenüber Jugendlichen, Frauen gegenüber Männern, Familien mit Kleinkindern gegenüber älteren Ehepaaren usw.), wäre das Design auf eine größere Zahl von Experimental- und zugeordneten Kontrollgruppen auszuweiten (für jede relevante Bevölkerungs-Teilgruppe ein komplettes Experimentaldesign).

Als Prototyp des sozialwissenschaftlichen Experiments ist das (psychologische) Labor-Experiment anzusehen. Hierfür gilt, daß – im günstigsten Fall – die Auswirkungen möglichst aller Randbedingungen mit Ausnahme des experimentellen Stimulus bekannt sein sollten, so daß die Äquivalenz von Versuchs- und Kontrollgruppe auf der Basis empirisch bestätigter theoretischer Kenntnisse herstellbar ist. Die Zusammensetzung der Gruppen kann in einem solchen Fall gezielt vorgenommen werden, die möglichen Einflußgrößen sind gezielt kontrollierbar.

Diese anspruchsvolle Voraussetzung – Vorhandensein hinreichender empirisch bestätigter Kenntnisse, um voll kontrollierte Experimente durchzuführen – ist normalerweise nicht gegeben. Daher machen sich sozialwissenschaftliche Experimente den Vorteil des Zufallsprinzips zunutze, der darin besteht, auch (noch) unbekannte Merkmale und Faktoren mit angebbarer Wahrscheinlichkeit in einer nicht einseitig verzerrenden Weise zu repräsentieren. Allerdings wird das Zufallsprinzip nicht etwa auf die Auswahl der Experimentteilnehmer angewendet – hier ist man in der Regel auf Freiwillige angewiesen; man hat es selbst im echten Experimentaldesign (nicht nur in Quasi-Experimenten) in der Regel mit einer systematisch verzerrenden Teilnehmerstichprobe zu tun. Das Zufallsprinzip wird eine Stufe tiefer eingesetzt: Sobald eine genügende Anzahl von Experimentteilnehmern gefunden ist, werden diese nach den Merkmalen geschichtet, die für den Ausgang des Experiments als bedeutsam gelten – vielleicht Alter, Geschlecht, Bildung. Danach entscheidet ein

Zufallsverfahren, welche Personen aus jeder Schicht den Experimentalgruppen und welche den Kontrollgruppen zugewiesen werden (Randomisierung).

Auf diese Weise erreicht man zweierlei: zum einen die Bildung von äquivalenten Experimental- und Kontrollgruppen hinsichtlich der Schichtungsmerkmale, zum anderen durch das zufällige Zuweisen zur Experimental- oder Kontrollgruppe die Ausschaltung der Gefahr systematischer Ergebnisverzerrungen durch andere, dem Forscher vorab nicht bekannte Faktoren. Das Zufallsverfahren dient somit nicht dazu, eine möglichst weitgehende Identität zwischen Grundgesamtheit und verkleinertem Abbild (Repräsentativität einer Stichprobe) zu erreichen, sondern die möglichst weitgehende Identität zwischen Experimental- und Kontrollgruppen zu sichern.

Wenn diese gesichert ist, dann gilt: Die Differenz zwischen der Entwicklung der Zielvariablen in der Experimental- und in der Kontrollgruppe ist (da mit Ausnahme des "treatments" alle übrigen Situationsbedingungen und Einflüsse identisch sind) unmittelbar der getroffenen Maßnahme als Wirkung zurechenbar.

Es wurde bereits mehrfach darauf hingewiesen, daß die Evaluationsforschung in der unter methodischen Gesichtspunkten unangenehmen Situation ist, die Bedingungen der Untersuchung nur in beschränktem Maße festlegen und kontrollieren zu können. Vorrang vor der Forschung hat das Programm. Deshalb ist es praktisch niemals möglich, die Evaluation als "echtes (soziales) Experiment" zu konzipieren. Ersatzweise versucht man daher „quasi-experimentelle Anordnungen“ zu realisieren, in denen Abweichungen vom echten Experiment durch alternative methodische Kontrollen ersetzt werden.

Grundsätzliche Schwierigkeiten bereitet in der Regel insbesondere die Zusammenstellung "echter" Kontrollgruppen. So existiert z.B. im Falle von Programmen mit "Angebotscharakter" für die Zielpopulation ebenso wenig eine Pflicht zur aktiven Beteiligung (forschungsmethodisch ausgedrückt: in der Experimentalgruppe mitzuwirken) wie eine Möglichkeit des kontrollierten Ausschlusses (forschungsmethodisch: der Kontrollgruppe zugewiesen zu werden). Selbst wenn das zu evaluierende Programm letzteres zuließe (etwa durch regionale Begrenzung seiner Geltung), wäre dies unter ethischen Gesichtspunkten häufig nicht zu rechtfertigen (etwa bei sozialpolitischen Förderungsmaßnahmen).

Wenn aber die Inanspruchnahme von Maßnahmen oder Optionen eines Programms auf Freiwilligkeit beruht, lautet die Konsequenz für den Versuch der Realisierung eines Experimentaldesigns: Die Zuweisung zu Experimental- und Kontrollgruppen geschieht nicht per Randomisierung, sondern durch Selbstselektion der Teilnehmer. Damit erhält die Evaluation für die *Phase der empirischen Feldforschung* notwendigerweise lediglich nicht-äquivalente Kontrollgruppen (denn die Entscheidung für oder gegen das Programm geschieht nicht "zufällig", sondern ist systematisch mit der Lebens- und Persönlichkeitsstruktur der betreffenden Zielpersonen verknüpft). Um dennoch Aussagen mit angebbarem Grad an Gültigkeit zu gewinnen, muß die Forschung in der *Phase der Datenauswertung* versuchen, die Designmängel der Erhebungsphase durch geeignete statistische Analyseverfahren auszugleichen (für Möglichkeiten dazu s. Kromrey 1987). Ein einfacher Vergleich zwischen Experimental- und Kontrollgruppen (wie im echten Experiment) ist in keinem Falle hinreichend.

Besteht ein Programm nicht aus "Angeboten", sondern aus Maßnahmen, die vom Programmträger "durchgeführt" werden und die Auswirkungen für die gesamte Zielpopulation des Programms haben (etwa Einführung einer Geschwindigkeitsbeschränkung auf Bundesstraßen und Autobahnen, Einführung einer Öko-Steuer, flächendeckende Reform der gymnasialen Oberstufe), ist selbst die Möglichkeit der Bildung *nicht-äquivalenter*

Kontrollgruppen stark eingeschränkt. Nur selten wird es möglich sein, politisch-administrative Systeme oder Regionen zu finden, die in allen wesentlichen Hinsichten vergleichbar sind und in denen lediglich die jetzt zu untersuchende Veränderung nicht vorgenommen wurde. Im Normalfall wird sich die Analyse auf Vorher-Nachher-Vergleiche ohne Kontrollgruppen beschränken und umfassende Daten über mögliche "externe Einflüsse" erheben müssen, um im Zuge der Auswertung deren Effekte statistisch "kontrollieren" zu können.

Schließlich ist die Evaluation in einigen Fällen mit der Aufgabe konfrontiert, die Effekte von Programmen zu erforschen, die keine "Zielpopulation" (im Sinne von Personen, die Leistungen in Anspruch nehmen oder von Maßnahmen betroffen sind) aufweisen, sondern die das Verhalten von "Systemen" beeinflussen sollen: z.B. Eindämmen des Waldsterbens durch Maßnahmen zur Verringerung des CO_2-Ausstoßes; Verringerung der Arbeitslosigkeit in strukturschwachen Regionen durch Infrastruktur-Investitionen und Anreize zur Industrieansiedlung o.ä. In solchen Fällen existieren definitionsgemäß keinerlei "Kontrollgruppen". Hier sind in anderer Weise Referenzgrößen für die Zielvariablen zu finden, wie sie sich ohne die Maßnahmen des durchgeführten Programms (vermutlich) ergeben hätten. Solche Referenzwerte können in diesem Fall lediglich über Modellrechnungen oder Simulationsverfahren bestimmt werden: Nach Abschluß des Programms wird anhand der Daten über die Entwicklung aller relevanten Einflußgrößen, die nicht programmabhängig sind, ein Verlauf der Zielvariablen in einer hypothetischen Realität "ohne Programm" ermittelt (entsprechende Strategien sind bei Eekhoff u.a. 1977 nachzulesen).

Alle Untersuchungsansätze, die sich an der oben skizzierten grundlegenden Logik des Experiments orientieren, die aber dessen methodische Anforderungen nicht in vollem Umfang realisieren können, werden üblicherweise als "quasi-experimentell" bezeichnet. Je nach den "Mängeln der Forschungssituation" und je nach den gewählten methodischen Strategien zum bestmöglichen Ausgleich dieser Mängel ergibt sich eine Vielzahl quasi-experimenteller Designformen, die hier auch nicht ansatzweise darstellbar sind. Gute Überblicke bieten Frey/Frenz 1982, Lösel/Nowack 1987, Rossi/Freeman 1988, und natürlich die "Klassiker" dieses Ansatzes: Campbell 1983, Cook/Campbell 1979; für "handwerkliche" Fragen der Planung und Durchführung von Evaluationsprojekten s. Wottawa/Thierau 1990.

3.3 Alternativen zum Experimentaldesign

3.3.1 Alternativen im Forschungsparadigma: ex-post-facto-Design“, theoriebasierte Evaluation; Meta-Analyse

Als unbestrittener „Königsweg“ der Evaluationsforschung (in angelsächsischen Texten auch als „Goldstandard“ bezeichnet) gilt das Experimentaldesign, mit Einschränkungen noch das Quasi-Experiment, das so viele Elemente des klassischen Experiments wie möglich zu realisieren versucht und für nicht realisierbare Design-Elemente methodisch kontrollierte Ersatzlösungen einführt. So tritt etwa bei der Zusammenstellung strukturäquivalenter Versuchs- und Kontrollgruppen das matching-Verfahren an die Stelle der Randomisierung; oder die nicht mögliche Abschirmung von Störgrößen in der Informationsbeschaffungsphase wird ersetzt durch umfassende Erhebung relevanter potentieller exogener Wirkungsfaktoren, um nachträglich in der Auswertungsphase die exogenen Einflüsse statistisch zu kontrollieren.

Mit letzterem Beispiel sind wird bereits auf halbem Wege, die Experimentallogik in der Erhebungsphase durch *Experimentallogik in der Auswertungsphase* zu simulieren. Wo ein Interventionsprogramm eine soziale Situation schafft, in der sich ein Feldexperiment

verbietet, kann sich die Evaluation an dem in Abbildung 2 skizzierten Variablenmodell orientieren und eine möglichst vollständige Deskription des Programmverlaufs („monitoring“) anstreben; das heißt: Für alle untersuchungsrelevanten Variablen werden mit Hilfe des Instrumentariums der herkömmlichen empirischen Sozialforschung über die gesamte Laufzeit des Programms Daten erhoben. Erst im Nachhinein – im Zuge der Analyse – werden die Daten so gruppiert, daß Schlußfolgerungen wie bei einem Experiment möglich werden, also Einteilung von Personen nach Programmnutzern bzw. -teilnehmern und Nichtnutzern bzw. Nicht-Teilnehmern (in Analogie zu Versuchs- und Kontrollgruppen), empirische Klassifikation der Nutzer bzw. Nichtnutzer im Hinblick auf relevante demographische und Persönlichkeitsvariablen (in Analogie zur Bildung *äquivalenter* Gruppen) sowie statistische Kontrolle exogener Einflüsse (in Anologie zur Abschirmung von Störgrößen). Diese *nachträgliche* Anordnung der Informationen in einer Weise, als stammten die Daten aus einem Experiment, wird üblicherweise als „*ex-post-facto-Design*“ bezeichnet.

Allerdings weist die ex-post-facto-Anordnung eine gravierende und prinzipiell nicht kontrollierbare Verletzung des Experimentalprinzips auf, nämlich das Problem der Selbstselektion der Teilnehmer/Nutzer. Auch das ausgefeilteste statistische Analysemodell kann kein Äquivalent zur kontrollierten Zuweisung zur Experimental- bzw. Kontrollgruppe anbieten. Allenfalls kann versucht werden, diesen Mangel in der Feldphase dadurch zu mildern, daß Gründe für die Teilnahme oder Nicht-Teilnahme mit erhoben werden, um möglicherweise existierende systematische Unterschiede erkennen und abschätzen zu können. Darüber hinaus erhält die generelle Problematik der Messung sozialer Sachverhalte im Vergleich zum echten Experiment ein erheblich größeres Gewicht: Soll die Gültigkeit der Analyse-Resultate gesichert sein, müssen alle potentiellen exogenen Einflüsse und müssen alle relevanten Persönlichkeitsmerkmale nicht nur bekannt, sondern auch operationalisierbar sein und zuverlässig gemessen werden. Im echten Experiment entfällt diese Notwendigkeit dadurch, daß alle (bekannten und unbekannten) exogenen Einflußgrößen durch Randomisierung bei der Bildung von Experimental- und Kontrollgruppen neutralisiert werden.

Einen anderen Zugang zur Gewinnung detaillierten empirischen Wissens über das zu evaluierende Vorhaben wählt das Modell einer *„theoriebasierten Evaluation“* (theory-based evaluation). Gemeint ist hier mit dem Terminus „Theorie“ allerdings nicht ein System hoch abstrakter, generalisierender, logisch verknüpfter Hypothesen mit im Idealfall räumlich und zeitlich uneingeschränktem Geltungsanspruch, sondern – ähnlich wie beim grounded-theory-Konzept – eine gegenstandbezogene Theorie, eine Theorie des Programmablaufs (Weiss 1995, 1997). Die Bezeichnung „logisches Modell“ wäre vielleicht treffender (vgl. Patton 1997, 234 ff.: logical framework approach), zumal die Bezeichnung „theoriebasierte Evaluation“ etwas irreführend ist, denn auch das Modell der Programmforschung – wie oben vorgestellt – ist „theoriebasiert“(s. Abschnitt 3.2.2): Als forschungsleitendes Modell des Wirkungsfeldes versucht die Programmevaluation ein in sich schlüssiges, einheitliches System von operationalisierbaren Hypothesen zu formulieren, das die theoretische Basis für die Planung des Programms (Zuordnung von Maßnahmen/Instrumenten zu Programmzielen), für die Implementation und für die gezielte Messung der Effekte (Zurechnung der beobachteten Veränderungen zu den durchgeführten Maßnahmen) rekonstruieren soll.

Bei diesem Rationalmodell der Programmevaluation tritt nun das bereits genannte zentrale Problem auf, daß im allgemeinen eine solche einheitliche Programmtheorie als Grundlage rationaler Ziel- und Maßnahmenplanung nicht existiert, sondern ein Konstrukt des Forschers ist, das er an das Programm heranträgt, um sein Evaluationsdesign wissenschaftlich und methodologisch begründet entwickeln zu können. Faktisch dürften bei den Planern der

Maßnahmen ihre jeweils eigenen individuellen Vermutungen über die Notwendigkeit der Erreichung bestimmter Ziele und die Eignung dafür einzusetzender Instrumente für ihre Entscheidungen maßgebend sein. Ebenso dürften die mit der Implementation betrauten Instanzen eigene – vielleicht sogar von den Planern abweichende – Vorstellungen darüber besitzen, wie die Maßnahmen im Detail unter den jeweils gegebenen Randbedingungen zu organisieren und zu realisieren sind. Und schließlich werden auch die für den konkreten Alltagsbetrieb des Programms zuständigen Mitarbeiter sowie ggf. die Adressaten des Programms (soweit deren Akzeptanz und/oder Mitwirkung erforderlich ist) ihr Handeln von ihren jeweiligen Alltagstheorien leiten lassen.

Es existieren also im Normalfall unabhängig von den abstrahierenden theoretischen Vorstellungen der Evaluatoren mehrere – im Idealfall sich ergänzende, vielleicht aber auch in Konkurrenz stehende – Programmtheorien, die den Fortgang des Programms steuern und für dessen Erfolg oder Mißerfolg maßgeblich sind. Sie gilt es zu rekonstruieren und zum theoretischen Leitmodell der Evaluation zu systematisieren. Das Ergebnis könnte dann ein *handlungslogisches Rahmenkonzept* sein, in dem der von den Beteiligten vermutete Prozeß von den Maßnahmen über alle Zwischenschritte bis zu den Wirkungen skizziert ist.[9] Wo mehrere Wirkungsstränge denkbar sind, wären diese parallel darzustellen und ggf. zu vernetzen. Von einem solchen ablaufsorientierten „logischen Modell" angeleitet, kann die Evaluation Detailinformationen über den gesamten Prozeß aus der Perspektive der jeweiligen Akteure sammeln. Sie vermeidet es, zwischen dem Einsatz eines Instruments und der Messung der Veränderungen im vorgesehenen Wirkungsfeld eine black box zu belassen (wie dies etwa im Experimentaldesign geschieht). Sie kann nachzeichnen, an welcher Stelle ggf. der vermutete Prozeß von der Implementation über die Ingangsetzung von Wirkungsmechanismen bis zu den beabsichtigten Effekten von welchen Beteiligten auf welche Weise unterbrochen wurde, wo ggf. Auslöser für nicht-intendierte Effekte auftraten, an welchen Stellen und bei welchen Beteiligten Programmrevisionen angezeigt sind usw. Zudem kann eine so konzipierte Evaluation auf methodisch hoch anspruchsvolle, standardisierte, mit großem Kontrollaufwand durchzuführende und damit potentiell das Programm störende Datenerhebungen verzichten, da sie ihre Informationen jeweils ereignis- und akteursnah mit situationsangemessenen Instrumenten sammeln und direkt validieren kann.

Ich möchte mich zur Illustration auf eine Skizze beschränken und zu diesem Zweck auf das vorne bereits angesprochene US-amerikanische Programm D.A.R.E. (Drug Abuse Resistance Education) zurückgreifen.[10] Das mit außerordentlich hohem Aufwand durchgeführte Vorhaben – allein 1993 wurden 750 Millionen Dollar an öffentlichen Mitteln dafür bereitgestellt (McNeal/Hansen 1995, 141) – hatte seinen Ursprung 1983 in einer Zusammenarbeit zwischen Polizei und Schulverwaltung in Los Angeles. Ein grobes logisches Ablaufsmodell hat die folgende Struktur:

- *Akteure 1*: Polizei und Schulverwaltung in Los Angeles entwickeln ein Modellprojekt mit dem Programmziel, potentielle spätere Drogenkonsumenten bereits in frühem Jugendalter dagegen zu immunisieren (Theorie: Drogenmarkt als Denkmodell: Eingriff auf der Seite der „Nachfrage"; Persönlichkeitsentwicklung: Grundlagen für späteres Verhalten werden im Jugendalter gelegt; Wirkung von Unterricht: Wissen beeinflußt Einstellungen, geänderte Einstellungen beeinflussen langfristig das Verhalten). Instrumente/Ressourcen: Da verbreiteter Drogenkonsum unter Jugendlichen in der öffentlichen Diskussion als ein gesellschaftliches Problem akzeptiert ist, führen politische Bemühungen zur Bereitstellung öffentlicher Mittel.

[9] Während das Modell der Programmforschung in Abbildung 2 (variablenorientiert) die *wirkungslogische* Beziehungsstruktur der Programmelemente einschließlich der Programmumwelt darstellt, handelt es sich hier um die (akteursorientierte) Skizze der *zeitlich-sachlogischen* Ablaufsstruktur des Programms.

[10] Ein weiteres illustratives Beispiel für ein solches kleinschrittiges Programmwirkungsmodell ist bei Carol Weiss (1997, 503 ff.) zu finden.

- *Akteure 2*: Der Unterricht soll von Polizisten in Grundschulen gehalten werden; für diesen Unterricht sind geeignete Polizisten auszuwählen (Hypothesen über didaktische und soziale Voraussetzungen für den Anti-Drogen-Unterricht, z.B.: In der Grundschule sind die Jugendlichen noch eher formbar als in späteren Sozialisationsphasen; Unterricht durch Polizisten assoziiert Drogenkonsum mit einem Straftatbestand; Jugendliche akzeptieren junge Polizisten als Bezugspersonen eher als ältere).
- *Vorbereitende Aktivitäten 1*: Ein geeignetes Curriculum ist zu entwickeln; die ausgewählten Polizisten sind für ihren Schuleinsatz zu trainieren (kognitivistische Didaktiktheorien; Hypothesen über die Relevanz selektiver Wissensbestände für die beabsichtigte Beeinflussung von Einstellungen gegenüber Drogenkonsum: Nur harte Drogen? Soll auch Nikotin und Alkohol behandelt werden?).
- *Vorbereitende Aktivitäten 2*: Zuweisung von Polizisten zu Schulklassen, Unterrichtseinsatz in Uniform (Hypothesen über die Rolle des Vertrautseins mit dem räumlichen Kontext der Schüler, über die Wirkung des Auftretens und der Kleidung: Unterricht in Uniform verstärkt die Assoziation von Drogenkonsum mit Straftatbestand).
- *Programmaktivitäten: Durchführungsplanung und –organisation*: Kursdauer 10 Wochen, eine Stunde pro Woche innerhalb des regulären Unterrichts (Lernpsychologie und Didaktiktheorien: über angemessene Menge des Unterrichtsstoffs, über zur Wissensvermittlung und Wissensstabilisierung erforderliche Mindestdauer des Kurses, über maximale Dauer zur Vermeidung von Akzeptanzproblemen gegenüber dem neuen Unterrichtsfach).
- *Programmdurchführung/Beteiligungsverhalten*: Polizisten geben als externe Lehrende Unterricht in den Schulklassen (Hypothesen über Reaktionen der Schüler, über ihre Beteiligung, ihre Reaktionen; Vorannahmen über eventuell erforderliche flexible Anpassungen des Unterrichtsverhaltens an schulische Notwendigkeiten, über Unterstützung/Abwehrverhalten des regulären Lehrpersonals).
- Kurzfristige *Veränderungen von Wissen und Einstellungen*: Schüler erwerben Kenntnisse über Gefahren von Drogen sowie Fertigkeiten zum Fällen selbstverantwortlicher Entscheidungen (Hypothesen über die Auswirkung von Wissen auf Einstellungen, von Entscheidungsfähigkeit auf gesteigerte Selbstwerteinschätzung).
- Kurzfristige *Wirkung geänderter Einstellungen auf Verhalten*: Schüler gestalten ihr Leben bewußter, entwickeln – wenn notwendig – Widerstandskraft gegen Gruppendruck („rationalistische“ Handlungstheorie: Wissen und Fertigkeiten führen zu mehr Selbstbewußtsein, zu verbesserter Urteilsfähigkeit, zur Orientierung am eigenen Wohl: Drogen sind in diesem Zusammenhang negativ und werden gemieden).
- Geänderte *Einstellungen und Verhaltensänderungen bleiben langfristig erhalten* (Hypothesen über die Konstanz von erworbenen Persönlichkeitsmerkmalen; hier: rationales Entscheidungsverhalten und geänderte Einstellungen gegenüber Drogen im frühen Jugendalter verhindern den Drogenkonsum als Teenager).

Die hier nur in groben Zügen skizzierte Programmlogik (oder Programmtheorie) kann als ein Stufenmodell charakterisiert werden: Jede der Folgestufen kann erst erreicht werden, wenn davor liegenden Stufen erfolgreich durchlaufen wurden. Eine in dieser Weise „theoriebasierte“ Evaluation könnte programmbegleitend sofort feststellen, in welcher Stufe evtl. Probleme auftreten; sie könnte dadurch Informationen liefern, die den erfolgreichen Fortgang des Projekts unterstützen und sichern. Zugleich sind konkret gegenstandsbezogene Informationen eine geeignete empirische Basis für die Fortentwicklung sozialwissenschaftlichen Grundlagenwissens.

In besonderer Weise steht die Produktion von Grundlagenwissen natürlich im Vordergrund bei Vorhaben der *Meta-Evaluation* sowie der *Meta-Analyse von Evaluationsstudien*. Im ersten Fall geht es darum, durchgeführte Forschungsvorhaben zu evaluieren, einerseits um die Reichweite ihrer Aussagen zu überprüfen und ggf. (im Falle der Kumulation mehrerer Studien mit vergleichbarer Thematik) zu erhöhen, andererseits um methodologische Standards zu prüfen und weiter zu entwickeln (s. z.B. Marconi / Rudzinski 1995, McNeal / Hansen 1995).

Im zweiten Fall, der Meta-Analyse von Evaluationen, steht eindeutig die Prüfung der Geltung gewonnener Aussagen, ihre Differenzierung und Erweiterung an vorderster Stelle. Dies kann auf zweierlei Weise geschehen. Zum einen können vorliegende Studien zur gleichen Thematik (etwa Einzelprojekt-Evaluationen zum gleichen Programmtyp) gesammelt und ihre

Ergebnisse anhand eines vergleichenden Rasters kumuliert werden. Dies setzt voraus, daß die Untersuchungsdesigns der einzubeziehenden Studien weitgehend ähnlich und die Durchführung von hinreichender methodischer Qualität sind (etwa: nur Studien, die ein Experimentaldesign mit äquivalenten Kontrollgruppen realisiert haben). Die Güte der Ergebniskumulation ist dann mit abhängig von der Vollständigkeit der Erschließung der existierenden Untersuchungen sowie von der Stringenz der angewendeten Inklusions- bzw. Exklusionskriterien der Meta-Analyse (vgl. dazu Petrosino 1995). Eine andere Strategie ist, die recherchierten und dokumentierten Studien zu einem Themenkomplex als Objektmenge für eine Inhaltsanalyse zu nutzen. Designmerkmale können dann ebenso wie berichtete Befunde anhand eines Kategorienschemas codiert und im Auswertungsverfahren der Meta-Analyse zueinander in Beziehung gesetzt werden (vgl. Hellstern / Wollmann 1983).

3.3.2 Alternativen im Kontrollparadigma: Indiktorenmodelle, Bewertung durch Betroffene

Beim Kontrollparadigma steht, wie zu Beginn geschildert, nicht das Interesse an der Gewinnung übergreifender und transferfähiger Erkenntnisse im Vordergrund, sondern die Beurteilung der Implementation und des Erfolgs eines Interventionsprogramms. Soweit es sich um ein Programm mit explizierten Ziel-Mittel-Relationen handelt, sind unter methodischem Gesichtspunkt selbstverständlich das Experiment bzw. – wenn nicht realisierbar – seine Alternativen Quasi-Experiment bzw. ex-post-facto-Design die geeignete Wahl. Allerdings steht nicht selten eine andere Thematik im Zentrum des Kontroll-Interesses, nämlich Qualitätssicherung und Qualitätsentwicklung – auch und gerade im Falle fortlaufend zu erbringender Humandienstleistungen durch eine Organisation oder Institution. Zwar gilt inzwischen weitgehend unbestritten *der positive Effekt bei den Adressaten der Dienstleistung (outcome) als letztliches Kriterium für den Erfolg der Dienstleistung*. Doch ist zugleich die unerschütterliche Annahme weit verbreitet, daß gute Servicequalität eine weitgehende Gewähr für solchen Erfolg sei. So wird z.B. in der Hochschulpolitik für wahrgenommene Mängel im universitär vermittelten Qualifikations-Output (z.B. lange Studienzeiten oder hohe Studienabbruchquoten) in erster Linie die vorgeblich schlechte Lehre verantwortlich gemacht und deren Qualitätsverbesserung eingefordert.

Somit gehört es zu den ersten Aufgaben der Evaluation, die qualitätsrelevanten Dimensionen des Dienstleistungsangebots zu bestimmen und zu deren Beurteilung Qualitätsindikatoren zu begründen und zu operationalisieren – eine Aufgabe, mit der sich die Sozialwissenschaft im Rahmen der Sozialindikatorenbewegung seit Jahrzehnten befaßt. Hierbei wird die Evaluation gleich zu Beginn mit einem zentralen theoretischen und methodologischen Problem konfrontiert, der Unbestimmtheit des Begriffs „Qualität“. Je nachdem, auf welchen Aspekt der Dienstleistungserbringung sich der Blick richtet und aus welcher Perspektive der Sachverhalt betrachtet wird, kann Qualität etwas sehr Unterschiedliches bedeuten. Eine Durchsicht verschiedener Versuche der Annäherung an diese Thematik erweist sehr schnell, daß „Qualität“ keine Eigenschaft eines Sachverhalts (z.B. einer Dienstleistung) ist, sondern ein mehrdimensionales Konstrukt, das von außen an den Sachverhalt zum Zwecke der Beurteilung herangetragen wird. Wenn nun – wie oben angedeutet – die positiven Effekte bei den Adressaten einer Dienstleistung das eigentliche Kriterium der Qualitätsbeurteilung sein sollen, die Qualität der Dienstleistung jedoch aus unterschiedlichsten Gründen nicht an den Effekten auf die Adressaten abgelesen werden kann, dann erwächst daraus ein methodisches Problem, das ebenfalls schon in der Sozialindikatorenbewegung unter den Schlagworten subjektive versus objektive Indikatoren ausgiebig diskutiert worden ist. Dann muß entweder den Adressaten die Rolle der Evaluatoren zugeschoben werden, indem per mehr oder weniger

differenzierter Befragung ihre Beurteilung der Dienstleistung erhoben wird. Oder es müssen „objektive“ Qualitätsmerkmale der Dienstleistung und des Prozesses der Dienstleistungserbringung ermittelt werden, die auch „subjektive Bedeutung“ haben, die also in der Tat die Wahrscheinlichkeit positiver Effekte bei den Adressaten begründen können (Kromrey/Ollmann 1985).

Im Gesundheitswesen – und von dort ausgehend in anderen sozialen Dienstleistungsbereichen – ist der wohl bekannteste Ansatz das von Donabedian entworfene Qualitätskonzept (ausführlich in Donabedian 1980). Er stellt die Evaluation eines Prozesses in den Mittelpunkt seiner Definition, nämlich *Qualität als Grad der Übereinstimmung zwischen zuvor formulierten Kriterien und der tatsächlich erbrachten Leistung*. Diesen Prozeß bettet er ein in die Strukturen als Rahmenbedingungen für die Leistungserbringung sowie die Ergebnisse, die die erbrachte Leistung bei den Adressaten bewirkt. Damit sind drei Qualitätsbereiche benannt sowie drei Felder für die Auswahl und Operationalisierung qualitätsrelevanter Indikatoren abgegrenzt. Außerdem ist damit eine Wirkungshypothese impliziert: Die Strukturqualität (personelle, finanzielle und materielle Ressourcen, physische und organisatorische Rahmenbedingungen, physische und soziale Umwelt) ist die Bedingung der Möglichkeit von Prozeßqualität (Erbringung der Dienstleistung, Interaktionsbeziehung zwischen Anbieter und Klienten); diese wiederum ist eine Voraussetzung für Ergebnisqualität (Zustandsveränderung der Klienten im Hinblick auf den Zweck der Dienstleistung, Zufriedenheit der Klienten).

Die sachliche Angemessenheit dieses dimensionalen Schemas unterstellt, besteht die entscheidende Aufgabe der Evaluation darin, zu jeder der Dimensionen diejenigen Indikatoren zu bestimmen und zu operationalisieren, die dem konkret zu evaluierenden Programm angemessen sind. Dies kann nicht ohne Einbeziehung der Programmträger, des eigentlichen Dienstleistungspersonals sowie der Adressaten der Dienstleistung und ggf. weiterer Beteiligter und Betroffener geschehen (als Beispiel: Herman 1997). Des weiteren sind die Indikatoren als gültige Meßgrößen durch Formulierung von „Korrespondenzregeln“ methodisch zu begründen; d.h. es ist nachzuweisen, daß sie „stellvertretend“ die eigentlich interessierenden Dimensionen abbilden. Häufig genug geschieht dies entweder überhaupt nicht oder lediglich gestützt auf Vermutungen oder als Ergebnis eines Aushandlungsprozesses zwischen den Beteiligten,[11] oder sie werden von vornherein unter dem Gesichtspunkt leichter Meßbarkeit ausgewählt. Nicht nur ist die Validität solcher Indikatoren zweifelhaft (Wird damit wirklich die angezielte „Qualität“ gemessen?). Sie bergen auch die Gefahr der Fehlsteuerung, indem statt der gewünschten Qualität vor allem die leicht meßbaren Sachverhalte optimiert werden.[12]

Wenn – wie zu Beginn dargelegt – der positive Effekt bei den Adressaten der Dienstleistung (outcome) als letztliches Kriterium für den Erfolg der Dienstleistung gelten soll, dann ist als Beurteilungsmaßstab für die Güte der Indikatoren die sog. „Kriteriumsvalidität“ zu wählen;

[11] Die Entscheidung nach dem Konsensprinzip führt erfahrungsgemäß zur Einigung auf ein System von Indikatoren, dessen Anwendung am gegenwärtigen Zustand wenig bis gar nichts ändert.

[12] Selbst bei im Prinzip gültigen Indikatoren besteht das Dilemma, daß sie gültige Informationen nur so lange liefern, wie sie lediglich deskriptive Funktionen erfüllen, ihre Anwendung also ohne Konsequenzen bleibt. Andernfalls (wie etwa bei Verteilung von Haushaltsmitteln in Universitäten nach sog. Leistungs- und Belastungskriterien) wird jeder rational Handelnde versuchen, die Ausprägung der Indikatorwerte in seinem Sinne zu „optimieren“. Gilt beispielsweise der Anteil erfolgreicher Abschlüsse an der Zahl der Studierenden in einem Studiengang als ein „Leistungsindikator“, dann ist es unter Haushaltsgesichtspunkten rational, auch diejenigen zum Abschluß zu führen (unter „geeigneter“ Anpassung des Anspruchsniveaus), denen man „eigentlich“ die Annahme eines ihren Fähigkeiten entsprechenden Arbeitsplatzes ohne Fortführung des Studiums empfehlen sollte.

d.h. die Indikatoren in den Bereichen Struktur und Prozeß sind in dem Maße valide, wie sie signifikante empirische Beziehungen zu outcome-Indikatoren aufweisen. Dies folgt im Donabedian-Modell auch aus der kausalen Verknüpfung, die der Autor zwischen den Bereichen Struktur → Prozeß → Ergebnis postuliert. Eine nachweisbar gültige Messung von Qualität über Indikatoren hat somit stets aus einem theoretisch begründbaren und empirisch prüfbaren System von Indikatoren zu bestehen, in welchem zwischen Qualitätsindikatoren und Gültigkeitskontrollindikatoren („validators") unterschieden werden kann. Ein solches Indikatorensystem für das Donabedian-Modell wird in einem Artikel von Salzer u.a. (1997) vorgestellt und methodologisch grundlegend diskutiert.

Angesichts der Schwierigkeit und Aufwendigkeit solchen Vorgehens wird nicht selten eine einfachere Lösung gesucht und – vermeintlich – auch gefunden. An die Stelle methodisch kontrollierter Evaluation durch Forschung wird – wie oben (Abschnit 3.1.4) bereits kurz angesprochen – die Bewertung durch Betroffene und/oder die Ermittlung ihrer Zufriedenheit gesetzt: Man befrage die Adressaten und erhebe deren Bewertungen. Die Adressaten und Nutzer – so wird argumentiert – sind die von dem zu evaluierenden Programm ganz konkret „Betroffenen" und daher in der Lage, aus eigener Erfahrung auch dessen Qualität zuverlässig zu beurteilen. Sind die erbrachten Dienstleistungen „schlecht", so werden auch die Beurteilungen auf einer vorgegebenen Skala negativ ausfallen und umgekehrt. Befragt man eine hinreichend große Zahl von „Betroffenen" und berechnet pro Skala statistische Kennziffern (etwa Mittelwerte oder Prozentanteile), dann kommen – so die weitere Argumentation – individuelle Abweichungen der einzelnen Urteilenden darin nicht mehr zur Geltung. Erhofftes Fazit: Man erhält verläßliche Qualitätsindikatoren.

Leider erweisen sich solche Vorstellungen als empirisch falsch (vgl. am Beispiel „Lehrevaluation" an Hochschulen: Kromrey 1994b, 1996). Die per Umfrageforschung bei Nutzern oder Betroffenen erhobenen Antworten auf bewertende (also evaluative) Fragen haben nicht den Status von „Evaluation" als methodisch kontrollierter, empirischer Qualitätsbewertung. Ermittelt wird lediglich die „Akzeptanz" (oder Nicht-Akzeptanz), auf die der beurteilte Sachverhalt bei den Befragten stößt; und die hängt im wesentlichen ab von Merkmalen der Befragten und nur relativ gering von Merkmalen des beurteilten Sachverhalts. Natürlich sind auch Akzeptanzaussagen keine unwesentliche Information, insbesondere in solchen Dienstleistungsbereichen, in denen der Erfolg von der aktiven Partizipation der Adressaten abhängt (etwa in der Familienhilfe oder generell in der Sozialarbeit).[13] Akzeptanzaussagen geben Auskunft darüber, in welchem Ausmaß und unter welchen Bedingungen das neue Angebot, das neue Programm etc. 'akzeptiert' (oder abgelehnt) wird, sowie darüber, welche Änderungen ggf. notwendig sind, um die 'Akzeptanz' – nicht unbedingt das Produkt – zu verbessern. Akzeptanz und/oder Zufriedenheit kann auch – wie im Donabedian-Modell – eine Teildimension von outcome-Qualität sein; dann nämlich, wenn die aktive Partizipation der Adressaten ein explizites Ziel des Programms ist. Aber selbst als Teildimension von Qualität kann sie nicht stellvertretend für das *gesamte* Qualitätskonzept stehen.

4. Das Leitkonzept für das Entwicklungsparadigma der Evaluation: Das "Helfer- und Beratermodell" der Evaluation

Das Konzept von Evaluation als Programmforschung ist – wie in Abschnitt 3.2 dargestellt – methodisch schwer realisierbar und muß von Voraussetzungen über den Untersuchungsgegenstand ausgehen, die nur selten hinreichend erfüllt sind. Auch das dem "Programm"-

[13] Ein informationsreiches Akzeptanzforschungsprojekt stellt H. Müller-Kohlenberg in diesem Band vor.

Verständnis zugrunde liegende Leitbild rationaler Planung hat nicht mehr die gleiche Gültigkeit wie in den 1970er Jahren. Nach diesem Leitbild ist – ausgehend von sozialen Problemen, die systemimmanent lösbar erscheinen – auf der Basis einer Gegenüberstellung von Ist-Analyse und Soll-Zustand ein Handlungsprogramm zu entwerfen und zu implementieren; dieses ist begleitend und/oder abschließend auf seinen Erfolg zu überprüfen und erforderlichenfalls für die nächste Periode zu modifizieren. Für die Entwicklung und Erprobung innovativer Konzepte ist dieses Forschungsmodell außerordentlich unhandlich, in manchen Konstellationen auch überhaupt nicht realisierbar. Zunehmend werden in jüngerer Zeit empirische Informationen und sozialwissenschaftliches Know-how bereits bei der Entwicklung und Optimierung eines Programms sowie bei der Erkundung der Möglichkeiten seiner "Umsetzung" verlangt.

Gegenüber dem bisher dargestellten Konzept ergeben sich dadurch zwei grundlegende Unterschiede für die Funktion der Evaluation.

Zum einen steht *am Anfang nicht ein "fertiges" Programm*, dessen Implementierung und Wirksamkeit zu überprüfen ist. Vielmehr ist Evaluation in die gesamte Programm-Historie eingebunden: von der Aufarbeitung und Präzisierung von Problemwahrnehmungen und Zielvorstellungen über eine zunächst vage Programmidee, über die Entwicklung geeignet erscheinender Maßnahmen und deren Erprobung bis hin zu einem auf seine Güte und Eignung getesteten (endgültigen) Konzept. Evaluation unter solchen Bedingungen ist im wörtlichen Sinne "formativ". Sie ist wesentlicher Bestandteil des Entwicklungsprozesses, in welchem ihr die Funktion der Qualitätsentwicklung und Qualitätssicherung zukommt.

Zum zweiten kann der *Blickwinkel* der Evaluation in diesem Rahmen *nicht auf den Sachverhalt "Programm"* (Ziele – Maßnahmen – Effekte) *beschränkt* bleiben, sondern muß explizit auch die Beteiligten einbeziehen. Des weiteren reduziert sich die Programmumwelt nicht auf ein Bündel von "Störfaktoren", die es statistisch zu kontrollieren oder – im Experimentaldesign – zu eliminieren gilt. Vielmehr ist die Umwelt – neben dem System von Programmzielen – eine *wesentliche Referenzgröße* für die optimale Konzeptentwicklung. Bei der Entwicklungsaufgabe geht es nicht um einen abstrakten Katalog von Maßnahmen, der kontextunabhängig realisierbar und transferierbar sein soll; sondern die Aufgabe besteht in der optimalen Abstimmung von Zielen und Maßnahmen auf das vorgesehene Einsatzfeld.

Nicht von allen wird jedoch Evaluation im zuletzt skizzierten Kontext mit Forschung gleichgesetzt. Im exponiertesten Fall gilt *Evaluation als eine "Kunst"*, die "von Wissenschaft grundsätzlich verschieden" sei (Cronbach, zit. bei Ehrlich 1995, 35): Während in wissenschaftlich angelegten Vorhaben methodologische Standards und verallgemeinerbare Aussagen von ausschlaggebender Bedeutung seien, stehe für Evaluationsvorhaben das Interesse an nützlichen Informationen im Blickpunkt.

Methodisch verfährt Evaluation dieses Typus häufig ähnlich wie ein Forschungskonzept, das *Aktionsforschung* (Handlungsforschung, action research) genannt wird. Ihr *Ablauf ist iterativ, schleifenartig*, ist ein fortwährendes Fragenstellen, Antworten, Bewerten, Informieren und Aushandeln. Jede "Schleife" gliedert sich in drei Hauptphasen: Gegenstandsbestimmung, Informationssammlung, Ergebniseinspeisung. Der Zyklus ist entsprechend dem Programmfortschritt wiederholt zu durchlaufen.

Evaluatoren in diesem Konzept verstehen sich *als Moderatoren* im Diskurs der am Projekt beteiligten Gruppen (Informationssammler und -manager, "Übersetzer" unterschiedlicher Fachsprachen und Argumentationsmuster, Koordinatoren und Konfliktregulierer, Vermittler

von Fachwissen, Berater). Man kann daher mit Recht in diesem Zusammenhang von einem "Helfer- und Beratermodell" sprechen.

Evaluation dieses Typs – also *begleitende Beratung* – darf *auf keinen Fall mißverstanden* werden *als* die *"weichere" oder anspruchslosere Variante* im Vergleich zum Konzept der Programmforschung. Evaluatoren in der Funktion von Moderatoren und Beratern benötigen zunächst einmal alle im sozialwissenschaftlichen Studium üblicherweise vermittelten Kenntnisse und Fähigkeiten (insbesondere der *kompletten* empirischen Forschung: quantitative *und* qualitative Erhebungsmethoden, einfache *und* komplexe Verfahren der Datenanalyse, Daten- und Informationsverwaltung), darüber hinaus jedoch noch zusätzliche Qualifikationen, die nicht einfach "gelernt", sondern durch praktische Erfahrungen erworben werden müssen: interdisziplinäre Orientierung, Kommunikationsfähigkeit und Überzeugungskraft, wissenschaftlich-präzise und journalistisch-verständliche Sprache, Empathie, Phantasie, Moderationstechniken, Präsentations- und Vortragstechniken und manches mehr.

Ein konkret ausformuliertes Design für eine Evaluation dieses Typs – UPQA-Methode (User Participation in Quality Assessment) genannt – präsentiert Hanne K. Krogstrup (1997). Es ist – so die Autorin – besonders auf komplexe, schlecht strukturierte Problemstellungen in den Handlungsfeldern Soziales, Gesundheit und Bildung zugeschnitten und basiert methodisch auf dialogorientierten Formen der Interaktion zwischen den Akteuren im Feld sowie zwischen dem Feld und der Evaluation. Das Konzept verfolgt das Ziel, in prozeßbegleitender explorativer Forschung Anknüpfungspunkte für grundlegende Lernprozesse bei den Beteiligten im evaluierten Setting herauszuarbeiten und dadurch dauerhafte Kompetenzen für die Organisationsentwicklung zu schaffen (a.a.O., 205 ff.). Wie schwierig ggf. ein solches Modell zu realisieren sein kann, schildern anschaulich A. Smith u.a. (1997), die ein beteiligtenorientiertes Evaluationsvorhaben in einem größeren Krankenhaus durchführten. Die Forscher mußten erfahren, wie in ihrem Projekt unterschiedliche und durch die Evaluatoren kaum vermittelbare Kulturen (die Autoren sprechen von „Welten“) aufeinanderprallten, so daß Lösungen für eine zumindest indirekte – nämlich über den „Puffer“ Evaluatoren verlaufende – Kommunikation zwischen den „Welten“ gesucht werden mußten.

5. Literatur

Beywl, W. (1991): Entwicklung und Perspektiven praxiszentrierter Evaluation. In: Sozialwissenschaften und Berufspraxis, 14/3, 265-279

Beywl, W. (1999): Zielfindung und Zielklärung – ein Leitfaden. In: QS Materialien zur Qualitätssicherung in der Kinder- und Jugendhilfe, H. 21, Bonn (BMFSFJ)

Campbell, D. T. (1983): Reforms as experiments. In: Struening, E.L. / Brewer, M.B. (eds.): Handbook of evaluation research, Beverly Hills, London, 107-137 (zuerst in: The American Psychologist, 24/1969/4)

Chelimsky, E. (1997): Thoughts for a new evaluation society. „Keynote speech“ at the UK Evaluation Society conference in London 1996. In: Evaluation, 3/1, 97-109

Cook, Th.D.; Campbell, D.T. (1979): Quasi-experimentation. Design & analysis issues for field settings, Chicago

Donabedian, A. (1980): Explorations in quality assessment and monitoring: The definition of quality and approaches to its assessment, Ann Arbor, MI

Dukes, R.L.; Stein, J.A.; Ullman, J.B. (1997): Long-term impact of Drug Abuse Resistance Education (D.A.R.E.). In: Evaluation Review, 21/4, 483-500

Eekhoff, J.; Muthmann, R.; Sievert, O. (1977): Methoden und Möglichkeiten der Erfolgskontrolle städtischer Entwicklungsmaßnahmen, Bonn-Bad Godesberg, Schriftenreihe "Städtebauliche Forschung", Bd. 03.060

Ehrlich, K. (1995): Auf dem Weg zu einem neuen Konzept wissenschaftlicher Begleitung. In: Berufsbildung in Wissenschaft und Praxis, 24/1, 32-37

Frey, S.; Frenz, H.-G. (1982): Experiment und Quasi-Experiment im Feld. In: Patry, J.-L. (Hg.): Feldforschung, Bern, Stuttgart, 229-258

Häußermann, H.; Siebel, W. (1978): Thesen zur Soziologie der Stadt, Leviathan, 6/4, 484-500

Hellstern, G.-M.; Wollmann, H. (1983): Evaluierungsforschung. Ansätze und Methoden, dargestellt am Beispiel des Städtebaus, Basel, Stuttgart

Herman, S.E. (1997): Exploring the link between service quality and outcomes. Parents' assessments of family support programs. In: Evaluation Review, Vol. 21/3, 388-404

Hübener, A.; Halberstadt, R. (1976): Erfolgskontrolle politischer Planung – Probleme und Ansätze in der Bundesrepublik Deutschland, Göttingen

Koditek, Th. (1997): Voraussetzungen sozialpädagogischer Wirkungsforschung. In: QS Materialien zur Qualitätssicherung in der Kinder- und Jugendhilfe, H. 11, Bonn (BMFSFJ), 50-55

Krogstrup, H.K. (1997): User participation in quality assessment. A dialogue and learning oriented evaluation method. In: Evaluation, 3/2, 205-224

Kromrey, H. (1987): Zur Verallgemeinerbarkeit empirischer Befunde bei nicht-repräsentativen Stichproben. Ein Problem sozialwissenschaftlicher Begleitung von Modellversuchen und Pilotprojekten. In: Rundfunk und Fernsehen, 35/4, 478-499

Kromrey, H. (1988): Akzeptanz- und Begleitforschung. Methodische Ansätze, Möglichkeiten und Grenzen. In: Massacommunicatie, 16/3, 221-242

Kromrey, H. (1994a): Strategien des Informationsmanagements in der Sozialforschung. Ein Vergleich quantitativer und qualitativer Ansätze. In: Angewandte Sozialforschung, 18/3, 163-184

Kromrey, H. (1994b): Wie erkennt man „gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen. In: Empirische Pädagogik, 8/2, 153-168

Kromrey, H. (1996): Qualitätsverbesserung in Lehre und Studium statt sogenannter Lehrevaluation. Ein Plädoyer für gute Lehre und gegen schlechte Sozialforschung. In: Zeitschrift für Pädagogische Psychologie, 10/3-4, 153-166

Kromrey, H. (1998): Empirische Sozialforschung. Modelle und Methoden der Datenerhebung und Datenauswertung, 8. Aufl., Opladen: UTB

Kromrey, H.; Ollmann, R. (1985): Handlungsorientierungen und gebaute Umwelt. Zur subjektiven Bedeutung objektiver Indikatoren. In: Informationen zur Raumentwicklung, H. 5, 393-406

Lamnek, S. (1988/89): Qualitative Sozialforschung; Band 1: Methodologie; Band 2: Methoden und Techniken, München

Lösel, F.; Nowack, W. (1987): Evaluationforschung. In: J. Schultz-Gambard (Hg.): Angewandte Sozialpsychologie, München, Weinheim, 57-87

Marconi, K.M.; Rudzinski, K.A. (1995): A formative model to evaluate health services research. In: Evaluation Review, 19/5, 501-510

Mayntz, R. (1980): Die Entwicklung des analytischen Paradigmas der Implementations-forschung. In: dies. (Hg.): Implementation politischer Programme, Königstein/Ts., 1-17

McNeal Jr., R.B.; Hansen, W.B. (1995): An examination of strategies for gaining convergent validity in natural experiments. D.A.R.E. as an illustrative case study. In: Evaluation Review, 19/2, 141-158

Müller-Kohlenberg, H. (1997): Evaluation von sozialpädagogischen Maßnahmen aus unterschiedlicher Perspektive. In: QS Materialien zur Qualitätssicherung in der Kinder- und Jugendhilfe, H. 11, Bonn (BMFSFJ), 8-20

Patton, M.Q. (1997). Utilization-focused evaluation. 3rd ed., Thousand Oaks, CA, London

Petrosino, A.J. (1995): Specifying inclusion criteria for a meta-analysis. In: Evaluation Review, 19/3, 274-293

Rebien, C.C. (1997): Development assistance evaluation and the foundations of program evaluation. In: Evaluation Review, 21/4, 438-460

Rein, M. (1981): Comprehensive program evaluation. In: Levine, R.A.; Solomon, M.A.; Hellstern, G.-M.; Wollmann, H. (eds.): Evaluation research and practice, Beverly Hills, London

Rossi, P.H.; Freeman, H.E. (1988): Programmevaluation. Einführung in die Methoden angewandter Sozialforschung, Stuttgart

Salzer, M.S.; Nixon, C.T.; Schut, L.J.; Karver, M.S.; Bickman, L. (1997): Validating quality indicators. Quality as relationship between structure, process, and outcome. In: Evaluation Review, 21/3, 292-309

Shadish, W.R.; Cook, T.D.; Leviton, L.C. (1991): Foundations of program evaluation. Theories of practice, Newbury Park, CA

Sjoberg, G. (1983): Politics, ethics and evaluation research. In: Struening, Elmer L.; Brewer, M.B. (eds.): Handbook of evaluation research, Beverly Hills, London, 65-88

Smith, A.; Preston, D.; Buchanan, D.; Jordan, S. (1997): When two worlds collide. Conducting a management evaluation in a medical environment. In: Evaluation, 3/1, 49-68

Weiss, C.H. (1974): Evaluierungsforschung. Methoden zur Einschätzung von sozialen Reformprogrammen, Opladen 1974

Weiss, C.H. (1995): Nothing is as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Conell, J.P. et al. (eds.): New approaches to evaluating community initiatives, Washington, DC, 65-92

Weiss, C.H. (1997): How can theory-based evaluation make greater headway? In: Evaluation Review, 21/4, 501-524

Wottawa, H.; Thierau, H. (1990): Lehrbuch Evaluation, Bern, Stuttgart

Anschrift des Autors: Prof. Dr. Helmut Kromrey, Freie Universität Berlin, Institut für Soziologie, Babelsberger Str. 14-16, 10715 Berlin