Evaluierung und Evaluationsforschung: Begriffe, Modelle und Methoden

Helmut Kromrey
Freie Universität Berlin

Evaluating and Evaluation Research: Concepts, Models and Methods

Summary: Evaluation stands for varied concepts of systematical, data-based assessments, and within those frameworks for different empirical models and methods employed. After explicating the relevant terms, evaluation models, their theoretical underpinnings, their methodological and methodical problems as well as the requirements for application are critical discussed.

Keywords: Evaluation, evaluation-models, evaluation-application, evaluation-research

Zusammenfassung: Evaluation ist nicht nur ein vielfältig verwendeter Begriff. Er steht auch für verschiedenartige Konzepte von systematischer, datengestützter Bewertung; und in deren Rahmen wiederum wird auf unterschiedliche empirische Modelle und Methoden zurückgegriffen. Der vorliegende Text stellt nach einer einführenden Begriffsexplikation die häufigsten Evaluationsmodelle sowie die Voraussetzungen für ihre Anwendung vor und diskutiert die dabei zu bedenkenden methodologischen und methodischen Probleme.

Schlüsselbegriffe: Evaluation, Evaluationsmodelle, Evaluationsanwendung, Evaluationsforschung

Vorbemerkungen und sprachliche Vorklärungen

„Evaluation" ist gegenwärtig zu einem äußerst unscharfen Modewort geworden und wird von manchem lediglich als „wohlklingendes" Fremdwort für jede Form von Bewertung oder Beurteilung verwendet. Doch selbst wenn der Begriff klar definiert ist, bleibt eine gewisse Unschärfe, da dasselbe sprachliche Zeichen für unterschiedliche Typen von Referenzobjekten stehen kann (und steht). Nicht selten ist dies der Ausgangspunkt für Missverständnisse und Verwirrungen.

- Eine erste Gruppe von Referenzobjekten ist auf der symbolischen und gedanklichen Ebene angesiedelt. „Evaluation" steht für ein *spezifisches wissenschaftliches Denkmodell:* ein nachprüfbares Verfahren des Bewertens. Vor allem um dieses Denkmodell geht es, wenn mit methodologischem Akzent über Verfahren und Ansätze der Evaluation diskutiert wird.
- Die zweite Begriffsebene bezieht sich auf ein *spezifisches Handeln,* einen Prozess: auf *zielorientiertes Informationsmanagement.* Im allgemeinen Sinne gilt als Evaluation jede methodisch kontrollierte, verwertungs- und bewertungsorientierte Form des Sammelns, Auswertens und Verwertens von Informationen. Dabei ist es müßig, darüber zu streiten, ob das Erheben rein deskriptiver Daten über einen zu bewertenden Sachverhalt „schon" und das Ziehen von Schlussfolgerungen und Konsequenzen für diesen Sachverhalt „noch" zur Evaluation zählt.
- Und schließlich bezeichnet „Evaluation" auch noch etwas Punktuelles: das *Resultat* des Evaluationsprozesses, die Dokumentation der Wertaussagen in einem Evaluationsbericht oder -gutachten.

Der folgende Text ist vor allem auf der zweiten Begriffsebene – zielorientiertes Informationsmanagement – angesiedelt und behandelt

Psychologie in Erziehung und Unterricht, 2003, *50,* 11–26

auch das Thema „Methoden" in diesem Kontext. Diese Klarstellung ist erforderlich, da wir auf „Evaluation" und „Evaluierung" in den verschiedensten Diskussionskontexten stoßen – im Alltag ebenso wie in der Politik, in der Methodologie empirischer Wissenschaft ebenso wie im (spezifischeren) Zusammenhang der Umfrageforschung – und da selbstverständlich unterschiedliche Kontexte sich unterschiedliche Begriffsbedeutungen schaffen.

Der *alltägliche Sprachgebrauch* ist ausgesprochen unspezifisch: „Evaluation" bedeutet nichts weiter als „Bewertung". *Irgendetwas wird von irgendjemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet.* Derselbe Sachverhalt kann – wie die Alltagserfahrung – von verschiedenen Individuen sehr unterschiedlich bis gegensätzlich eingeschätzt und beurteilt werden.

In *politischen Argumentationen* sind die Begriffsverwendungen wesentlich spezifischer, unglücklicherweise aber außerordentlich vielfältig. Die Bezeichnung gilt für Effizienz*messungen* in ökonomischen Zusammenhängen ebenso wie für die von Sachverständigen vorgenommene *Analyse* der Funktionsfähigkeit von Organisationen (etwa: „Evaluation" wissenschaftlicher Einrichtungen). Auch die *beratende und moderierende Beteiligung* im Prozess der Entwicklung von Handlungsprogrammen mit dem Ziel ihrer Optimierung wird von diesem Begriff erfasst („formative" oder „responsive" Evaluation).

In der *empirischen Methodologie* meint „Evaluation" hingegen das *Design* für einen spezifischen Typ von Sozialforschung, der die Beschaffung von Informationen über Verlauf und Resultate eines (Handlungs- und Maßnahmen-)„Programms" mit explizit formulierten Zielen und Instrumenten zum Gegenstand hat. Evaluationsziele sind die wissenschaftliche Begleitung der Programm-Implementation und/oder die „Erfolgskontrolle" und „Wirkungsanalyse". Der Ansatz ist im Idealfall experimentell oder quasi-experimentell und strebt an, zu empirisch begründeten, intersubjektiv nachprüfbaren und somit *„objektivierten" Bewertungen* zu gelangen.

Schließlich wird auch im Zusammenhang „gewöhnlicher" *Umfrageforschung* von „Evaluation" gesprochen. Hier ist die Erhebung und Auswertung bewertender (also „evaluierender") Aussagen von Befragten gemeint, die in einem angebbaren Verhältnis zu dem zu evaluierenden „Gegenstand" stehen (etwa Kunden/Klienten, Betroffene, Teilnehmer von Bildungsveranstaltungen). Ermittelt werden durch Evaluationsumfragen *subjektive Werturteile* anhand explizit vorgegebener spezifischer Kriterien ebenso wie allgemeinere Zufriedenheits- oder Unzufriedenheits*äußerungen* oder auch Akzeptanz*informationen.* Ein spezifisches Evaluationsdesign existiert in diesem Fall nicht. Ins Auge fällt statt dessen die Analogie zur Meinungsforschung, mit dem einzigen Unterschied, dass nicht Meinungen, sondern Bewertungen und/oder Zufriedenheitseinschätzungen abgefragt werden.

Gemeinsam ist allen diesen Verwendungen, dass – im *Unterschied zum alltagssprachlichen Verständnis* – nicht „irgendetwas" evaluiert wird, sondern dass spezifizierte Sachverhalte, Programme, Maßnahmen, manchmal auch ganze Organisationen Gegenstand der Betrachtung sind. Zweitens nimmt nicht „irgendjemand" die Evaluation vor, sondern es sind Personen, die dazu in besonderer Weise befähigt erscheinen: wissenschaftliche „Sachverständige", methodische oder durch Praxiserfahrungen ausgewiesene „Experten", konkret „Betroffene". Drittens kommt das Urteil nicht nach „irgendwelchen" 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.

Voraussetzungen für ein erfolgreiches Evaluationsvorhaben: Präzisierungen, Rollendefinitionen und Kompetenzabklärungen

Präzisierungen zu jedem der oben genannten vier Aspekte (Gegenstand – Evaluator – Kriterien – Verfahren) sind in unterschiedlicher Weise möglich und kommen im Evaluationsalltag in unterschiedlichen Kombinationen vor. Soll ein Evaluationsprojekt nicht unkalkulierbaren Risiken des Scheiterns ausgesetzt sein, sind diese Präzisierungen im Vorfeld im Detail, verbindlich, nachvollziehbar und gut dokumentiert vorzunehmen.

Als relativ unproblematisch, möglicherweise gar als entbehrlich erscheint auf den ersten Blick die Präzisierung des „Gegenstands" der Evaluation. Er entspricht – so sollte man meinen – der Beschreibung des „Programms", dessen Implementation und Effektivität zu beurteilen ist (bzw. der spezifischen Maßnahme oder der Organisation etc., die im Fokus des Interesses steht). Mit der präzisen Beschreibung eines solchen Vorhabens/Sachverhaltes ist jedoch noch nicht der *„Gegenstand der Evaluation"* bezeichnet. Selbst wenn eine „umfassende Evaluation" (im Sinne von Rossi & Freeman, 1988 bzw. Rein, 1981) angestrebt würde, wäre doch noch (stark selektiv) zu entscheiden, welche Teilaspekte denn tatsächlich im Detail einer systematischen Beurteilung unterzogen werden sollen und welche allenfalls als Randbedingungen berücksichtigt werden könnten. Jede Evaluation wäre überfordert, wollte sie ein Programm, eine Einrichtung o. Ä. quasi „ganzheitlich" *zu ihrem Gegenstand machen.* Empirische Informationsgewinnung im Kontext von Evaluierung hat für konkrete Entscheidungszwecke zielgenaue Befunde zu liefern, die zudem für die Nutzer „relevant" zu sein haben; das heißt: von ihnen muss „etwas abhängen". Befunde, die zwar als „ganz interessant" aufgenommen werden, bei denen es

Tabelle 1: Evaluation – Begriffsdimensionen und Klärungsbedarf

alltäglicher Sprachgebrauch	**wissenschaftlicher Sprachgebrauch**	**Klärungsbedarf**
• Irgendetwas	• Programme, Maßnahmen, Organisation etc.	• Was ist das „Programm" und was sind seine Ziele? • Was ist der „Gegenstand" der Evaluierung? Was sind die Evaluationsziele?
wird	werden	
• von irgendjemand	• durch Personen, die zur Bewertung besonders befähigt sind,	Wer hat welche Funktionen/ Kompetenzen? • Informanten/Informationsquellen • Informationsbeschaffung und -aufbereitung • Evaluierende
• in irgendeiner Weise	• in einem objektivierten Verfahren	• Methoden und Verfahren der Informationsbeschaffung • Methoden und Verfahren des Bewertens • Legitimation zum Bewerten
• nach irgendwelchen Kriterien	• nach explizit auf den Sachverhalt bezogenen und begründeten Kriterien (und ggf. Standards) bewertet.	• Ziele (wessen Ziele?) • Kriterien • Standards
bewertet.		

aber für das Entscheidungshandeln keinen Unterschied ausmacht, ob sie so oder anders ausfallen, sind irrelevant.

Bei der Präzisierung des Evaluations-Gegenstands ist zudem zu unterscheiden zwischen *Merkmalen und Zielen des zu bewertenden Sachverhalts* (des Programms, des Entwicklungs-Vorhabens) auf der einen und den *Merkmalen und Zielen des Evaluations-Vorhabens* auf der anderen Seite. Soll das Evaluations-Vorhaben „nützlich" sein, d. h. bei den Nutzern der Befunde auf Akzeptanz stoßen, ist (selbstverständlich ebenfalls im Vorfeld) abzuklären, welche Personen, Gremien, Institutionen etc. als Nutzer vorgesehen sind, von welcher Art deren vorgesehene Nutzung sein soll und was deren Informationsbedarf ist. Patton (1997) – der in diesem Zusammenhang von „intended use by intended users" spricht – empfiehlt Planern und Durchführenden von Evaluations-Vorhaben, sich in der Programmdurchführung die handlungslogische Abfolge von Schritten oder Stufen („logical framework") zu vergegenwärtigen und diesen Stufen die entsprechenden evaluationsrelevanten Informationen zuzuordnen (S. 234ff., Tabelle 10.5). Dies beginnt auf der *Implementationsseite* mit den Programm-Inputs (bzw. auf der *Informationsbeschaffungsseite* mit Daten über Ausgaben, Personal, investierte Zeit), verläuft über die Implementations-Aktivitäten (bzw. deren monitoring), über die Beteiligten, die Zielgruppen, die weiteren Betroffenen und deren Reaktionen schließlich zu den bewirkten Veränderungen im Hinblick auf Kenntnisse, Einstellungen, Fertigkeiten sowie den daraus ggf. folgenden kurz-, mittel- und langfristigen Auswirkungen auf die Programm-Umwelt, auf geänderte Verhaltensweisen der Zielgruppen. An oberster Stelle der „Programmdesign-Hierarchie" steht schließlich das „eigentliche" Ziel, zu dem das Programm konzipiert und implementiert wurde, etwa Verbesserung der Chancengleichheit von Kindern aus benachteiligten Bevölkerungsgruppen bei einem bildungspolitischen Programm. In dieser Weise systematisch angegangen, entspräche die Präzisierung des „Evaluations-Gegenstands" einer Rekonstruktion der (impliziten) Programmtheorie und der für jede Hierarchiestufe vorgenommenen Zuordnung evaluationsrelevanter Informationen.

Ebenfalls auf den ersten Blick einfach erscheint die Einlösung des Klärungsbedarfs in der zweiten Zeile der obigen Tabelle *(Wer „evaluiert"?)*, so dass auch hier häufig der Fehler begangen wird, ein Projekt ohne eindeutige und verbindliche Absprachen über Funktionen und Zuständigkeiten der am Evaluations-Vorhaben Beteiligten zu beginnen. Dies kann zu vielfältigen Behinderungen der Arbeit führen; im ungünstigsten Fall kann es auch mit dem vollständigen Scheitern des Vorhabens enden. Die mit dem Evaluations-Vorhaben betrauten Personen können zum Gegenstand der Bewertung in unterschiedlichstem 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.

Auch wenn es keine Patentrezepte geben kann, sollte ein Rat auf jeden Fall beherzigt werden: Bei der Planung des Evaluationsvorhabens sind zumindest *drei Funktionen* analytisch klar voneinander zu trennen: *Informationsbeschaffung, Evaluierung, Ableitung von Konsequenzen* aus den Befunden. Zwischen den Beteiligten ist auszuhandeln und verbindlich festzulegen, wer welche Aufgaben übernimmt und wem welche Zuständigkeiten zugebilligt werden. Nur in seltenen Ausnahmefällen werden die Aufgaben und Kompetenzen für alle drei Funktionen „in einer Hand" liegen (können). Für Evaluationen im Rahmen von Organisationsentwicklungs-Vorhaben (wie z. B. der Evaluation von Studiengängen mit dem Ziel der Qualitätsentwicklung) empfiehlt sich eine Dreiteilung der Kompetenzen. Zum Bei-

spiel: Ein Team empirisch-methodisch ausgewiesener Forschungsexperten ist für die Informationsbeschaffung, -analyse und -präsentation zuständig; ein kleines Gremium von legitimierten Vertretern der beteiligten Gruppen (z. B. vom Fachbereichsrat eingesetzt) diskutiert auf dieser Basis Bewertungsalternativen und entwickelt Vorschläge und Empfehlungen; eine verantwortliche Instanz auf der Leitungsebene entscheidet, welche Konsequenzen für die Organisation zu ziehen sind und/oder handelt mit den Beteiligten konkrete Maßnahmenpläne/Zielvereinbarungen aus.

Nach diesen Klärungen bildet die *Festlegung der Bewertungskriterien* (letzte Zeile in Tabelle 1) den Abschluss der sozusagen (organisations- bzw. programm-)„politischen" Entscheidungen für das Evaluationsvorhaben. Notwendig sind auch in dieser Hinsicht eindeutige (und dokumentierte) Festlegungen im Vorfeld: Schließlich sollen die Aussagen der für die Bewertungen zuständigen Instanz (die „Evaluatoren" im engeren Sinne) nachvollziehbar, überprüfbar und kritisierbar sein. Denkbar ist wiederum ein ganzes Spektrum sehr unterschiedlicher Bewertungskriterien und -standards. Sie können sich auf die *Wirkungen* und Nebenwirkungen der Maßnahmen eines Programms beziehen, auf die Art und Effizienz der *Durchführung, auf die Eignung und Effektivität* der gewählten Maßnahmen für die Ziel-Erreichung und auf die Angemessenheit und *Legitimierbarkeit der Ziele* selbst. Die Kriterien können zudem aus unterschiedlicher Perspektive hergeleitet werden (Auftraggeber – Betroffene – Durchführende; ökonomische Effizienz – Nutzen für das Allgemeinwohl – Sozialverträglichkeit etc.).

Nicht mehr von „evaluationspolitischem", sondern von methodologischem Charakter sind die Entscheidungen, die sich auf die *Art und Weise der Durchführung des Evaluationsprojekts* beziehen. Hier liefert das Arsenal der Methodologie und Methodik der empirischen Sozialforschung eine bewährte Basis für die Entwicklung eines Designs, das die Nützlichkeit der Ergebnisse zu gewährleisten hat. Dennoch: „Musterlösungen" quasi aus dem „Kochbuch der Methodenlehre" existieren nicht, so dass immer „maßgeschneiderte" Lösungen gefunden werden müssen. Das Verfahren der Evaluierung 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 Curriculums für den Fremdsprachenunterricht in der Grundschule 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.

Was „ist" im empirisch-wissenschaftlichen Sinne „Evaluation"?

Wenn – wie im vorigen Abschnitt skizziert – „Gegenstand" der Evaluation im Prinzip alles sein kann, wenn das Spektrum der Evaluations-„Fragestellungen" oder „-Zwecke" praktisch unbegrenzt ist, wenn keine speziellen Methoden der Evaluation existieren, sondern auf das bekannte Arsenal der „gewöhnlichen" empirischen Sozialforschung zurückzugreifen ist, wenn es also (zusammen genommen) kein „Musterdesign für Evaluationen" geben kann (s. ausführlich Patton, 1997, S. 192ff.),

sondern je nach Konstellation von Gegenstand und Fragestellungen „maßgeschneiderte" Vorgehensweisen zu entwickeln und zu begründen sind – *was ist dann eigentlich „Evaluation"* als empirisch-wissenschaftliches Verfahren?

- Die einzig methodologisch sinnvolle Antwort kann nur lauten: Es handelt sich um 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 (empirisch-wissenschaftliche) „Evaluator" einzunehmen hat: Welche Zwecke soll der zu evaluierende Gegenstand für welche Zielgruppen erfüllen? Werden die Zwecke erreicht? Was muss ggf. verändert 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 als empirisch-wissenschaftliches Handeln ist es nicht, am Fall des zu evaluierenden Gegenstands die *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 mit den Funktionsansprüchen des zu evaluierenden Gegenstands in Konflikt, haben die wissenschaftlichen Ansprüche zurückzutreten und ist nach wissenschaftlich suboptimalen Lösungen zu suchen, die das Funktionsgefüge im sozialen Feld nicht beeinträchtigen.

Die Vielfalt von Evaluationen: eine grobe Klassifikation

Natürlich ist es wenig sinnvoll, ohne den Versuch eines Ordnungsschemas vor der geschilderten Variationsbreite von Evaluationen zu kapitulieren und lediglich zu sagen: „es kommt darauf an". In der Tat existierten eine Reihe von Versuchen, die Detailvielfalt auf eine überschaubare Zahl von Typen zu reduzieren. Unter methodologischer Perspektive besonders nützlich ist ein Vorschlag von Eleanor Chelimsky (1997, S. 100ff.), die drei „conceptual frameworks" unterscheidet:

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

Der Vorteil dieser Einteilung ist, dass jedes der drei „Paradigmen" eine je spezifische Affinität zu Designtypen, zur Logik bzw. „Theorie" der Evaluation sowie zu Methoden und Qualitätskriterien des Evaluationshandelns aufweist.

Das „Forschungsparadigma" der Evaluation

Für Universitätswissenschaftler gelten Evaluationsprojekte häufig als Chance und als Herausforderung, neben dem „eigentlichen" Evaluationszweck grundlagenwissenschaftliche Ziele zu verfolgen. Evaluation wird aus dieser Perspektive als angewandte Forschung verstanden, die sich mit der Wirksamkeit von sozialen Interventionen befasst. Ihr kommt die Rolle eines Bindeglieds zwischen Theorie und Praxis zu (Weiss, 1974). Alle Anlässe, Ak-

tionsprogramme zur Bewältigung sozialer oder bildungsrelevanter 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. Im Unterschied zu forschungsproduzierten Daten zeichnen sich Untersuchungen unmittelbar im sozialen Feld zudem 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 die beobachteten Veränderungen mit den vom Programm angestrebten Effekten (den Programmzielen) vergleicht. Dem Forschungsparadigma verpflichtete Evaluatoren werden versuchen, so weit wie möglich wissenschaftlichen Gütekriterien Geltung zu verschaffen und Designs zu realisieren, die methodisch unstrittige Zurechnungen von Effekten zu Elementen des Programms durch Kontrolle der relevanten Randbedingungen erlauben.

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 als eine Kontrollform administrativen Handelns vertreten und stehen damit neben Rechtmäßigkeits-Kontrolle (Gerichte), politischer Kontrolle (Parlamente) und Wirtschaftlichkeits-Kontrolle (Rechnungshöfe). Eine charakteristische Definition ist: „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, S. 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 vom Informationsbedarf der programmdurchführenden und/oder der finanzierenden Instanz ab. Gefordert werden häufig quantitative Informationen. Eine sehr gute Darstellung dieses Ansatzes findet sich in Eekhoff u. a. (1977).

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; vielmehr geht es hier darum, Konzepte und Vorstellungen zu entwickeln, um die Fähigkeit von Organisationen zur Problemwahrnehmung und -bewältigung zu stärken, um mitzuwirken, retrospektiv und prospektiv Problemfelder 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, S. 33) es ausdrückt – „Geburtshelfer" einer

Idee und ihrer Realisierung sein. Gelegentlich wird diese Konstellation auch als „offene" Evaluation bezeichnet und unterscheidet sich von den zuvor geschilderten „geschlossenen" Evaluationen, in denen Problem- und Fragestellungen, methodisches Vorgehen, Bewertungskriterien und die Zielgruppen der Evaluationsberichte von vornherein feststehen. In „offenen" Evaluationen ist 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, S. 268).

Die Funktion der Evaluation ist hier in erster Linie die eines Helfers und Beraters. Im Kontext gängiger Lehrbewertungen an Schulen und Hochschulen ist dieses Modell weniger einschlägig und wird deshalb im Folgenden nicht weiter vertieft.[1] Angemessen wäre es allenfalls bei der Entwicklung und dem Test neuer – z. B. internetgestützter – Lehr- und Lernformen. An Bedeutung könnte das Entwicklungsparadigma im Schul- und Hochschulkontext jedoch in dem Maße gewinnen, wie als follow up zur Evaluation die Entwicklung und Durchführung von Qualitätsentwicklungs-Konzepten in Zielvereinbarungen mit den evaluierten Fächern oder Schulen festgeschrieben wird.

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

Begriffsexplikation

Evaluatoren sehen sich einer großen Vielfalt von Aufgabenprofilen und Rahmenbedingungen gegenüber. Bei aller Vielfalt bleibt dennoch – zumindest für das Forschungs- und für 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. Diese Programmdimensionen 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 allerdings *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

[1] Ein konkret ausformuliertes Design für eine Evaluation dieses Typs präsentiert Hanne K. Krogstrup (1997). Es ist besonders auf komplexe 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. Wie schwierig ggf. ein solches Modell zu realisieren sein kann, schildern anschaulich A. Smith u. a. (1997), die erfahren mussten, wie in ihrem Projekt unterschiedliche und durch die Evaluatoren kaum vermittelbare Kulturen (die Autoren sprechen von „Welten") aufeinander prallten, so dass Lösungen für eine zumindest indirekte – nämlich über den „Puffer" Evaluatoren verlaufende – Kommunikation zwischen den „Welten" zu finden waren.

Veränderungen als Wirkungen des Programms identifiziert und gegenüber anderen Einflüssen abgegrenzt werden?
Selbstverständlich ist eine alles umfassende Evaluation in keinem Projekt realisierbar. Es müssen daher 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.

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").

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 bei *begleitender Evaluation* üblicherweise zugleich regelmäßige Rückkoppelungen von Ergebnissen in das Projekt vorgesehen sind, wirkt die Forschung 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.
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.*

Externe oder interne Evaluation: *Wo ist die Evaluation angesiedelt?*
Diese dritte – und für die Praxis wichtige – Unterscheidung berücksichtigt, 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 können zum einen in der Gefahr mangelnder Professionalität,

zum anderen im Hinblick auf die „Objektivität" der Resultate bestehen.
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.

Instanzen der Evaluierung: *Wer beurteilt nach welchen Kriterien?*
Unter diesem Gesichtspunkt wird nach der Herkunft der Evaluationskriterien und nach der Bewertungsinstanz gefragt.
Im „traditionellen" Fall stammen die Beurteilungskriterien aus dem zu evaluierenden Programm selbst. Seine Implementation sowie seine Wirkungen werden im Lichte seiner eigenen Ziele bewertet. Die Beurteilung wird vom Evaluationsteam vorgenommen, das 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. Dies ist jedoch im Falle von Pilotprojekten und Modellversuchen oft nicht vorhanden. 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 kennen gelernt 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 anhand dokumentierter und *nachvollziehbarer* Kriterien und Standards. Es sind vielmehr „Akzeptanzaussagen" von Personen, die in einer besonderen Beziehung (eben als Nutzer, als Betroffene) zum Untersuchungsgegenstand stehen und die zu ihren Urteilen aus ihrer je *individuellen* Perspektive gelangen. Folgerichtig wird diese Evaluationsstrategie als *Akzeptanzforschung* bezeichnet.

Methoden der Programmforschung: Das Feldexperiment als Referenzdesign
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 oben aufgeführten vier Variablenbereiche (Ziele – Maßnahmen – Effekte – Programmumwelt) mit empirischen Daten abzubilden (zu „messen") und miteinander zu verknüpfen. Wir-

kungs- 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, S. 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 (s. dazu Kromrey, 1995). 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. Drei dieser meist implizit gelassenen Voraussetzungen sind besonders hervorzuheben, da deren Erfüllung eine wesentliche Bedingung dafür ist, um das methodologische Forschungsprogramm empirischer Kausalanalysen überhaupt anwenden zu können:

- Bereits vor der Entwicklung des Forschungsdesigns muss Klarheit über die Untersuchungsziele bestehen. Für die Dauer der Datenerhebung dürfen sich weder die Untersuchungsziele noch die wesentlichen Randbedingungen des Untersuchungsgegenstandes in unvorhersehbarer Weise ändern.
- Ebenfalls vor der Entwicklung des Forschungsdesigns müssen 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 in dieser Hinsicht vielmehr erhebliche „Mängel" auf. Die von der empirischen Sozialforschung entwickelte Methodologie der Programmevaluation ist daher weniger ein Real- als ein Idealtyp.

Zu den idealtypischen Elementen der Programmevaluations-Methodologie gehört die Orientierung am Referenzdesign „Feldexperiment", das unter methodologischen Gesichtspunkten am ehesten in der Lage ist, die

o. g. anspruchsvolle analytische Aufgabe der differenziellen Zurechnung beobachteter Effekte auf die Programm-Maßnahmen zu lösen (für eine anschauliche Darstellung s. Frey & Frenz, 1982). Unter der Rahmenbedingung des „Primats der Praxis vor der Wissenschaft" ist insbesondere das Design eines „echten" Experiments mit randomisierten Versuchs- und Kontrollgruppen nicht realisierbar.

Aber auch weniger anspruchsvolle „quasi-experimentelle Anordnungen" sind nur selten zu verwirklichen. Die Anwendbarkeit der skizzierten Methodik der Programmforschung in Evaluations-Vorhaben kann daher eher als der Ausnahmefall gelten. Es muss also zu „Ersatzlösungen" gegriffen werden, die praktikabel erscheinen und dennoch hinreichend gültige Ergebnisse liefern.

Alternativen zum Experimentaldesign

Alternativen im Forschungsparadigma: Ex-post-facto-Design", theoriebasierte Evaluation

Dass der methodologische „Königsweg" der Evaluationsforschung, das Experimentaldesign, in der Evaluationspraxis als nahezu unrealisierbar gelten kann, bedeutet jedoch nicht, dass auf dessen Forschungs*logik* verzichtet werden müsste. Am nächsten steht dem „echten" Experiment naturgemäß das o. g. Quasi-Experiment, das so viele Elemente des klassischen Designs wie möglich zu realisieren versucht und für nicht realisierbare Design-Elemente methodisch kontrollierte Ersatzlösungen einführt (für Details s. Hellstern & Wollmann, 1983; Kromrey, 1987, 1995). So tritt etwa bei der Zusammenstellung strukturäquivalenter Versuchs- und Kontrollgruppen das Matching-Verfahren an die Stelle der zufälligen Zuweisung; oder die nicht mögliche Abschirmung von Störgrößen in der Informationsbeschaffungsphase wird ersetzt durch umfassende Erhebung relevanter potenzieller exogener Wirkungsfaktoren, um nachträglich in der Auswertungsphase die exogenen Einflüsse statistisch zu kontrollieren.

Mit letzterem Beispiel sind wir 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 die Evaluation 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 die Einteilung von Personen nach Programmnutzern bzw. -teilnehmern und Nichtnutzern bzw. Nicht-Teilnehmern (in Analogie zu Versuchs- und Kontrollgruppen), die empirische Klassifikation der Nutzer bzw. Nichtnutzer im Hinblick auf relevante demographische und Persönlichkeitsvariablen (in Analogie zur Bildung *äquivalenter* Gruppen) sowie die statistische Kontrolle exogener Einflüsse (in Analogie 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 kontrolliert zufälligen Zuweisung zur Experimental- bzw. Kontrollgruppe (Randomisierung) 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, im Vergleich zum echten Experiment, die

generelle Problematik der Messung sozialer Sachverhalte ein erheblich größeres Gewicht: Soll die Gültigkeit der Analyse-Resultate gesichert sein, müssen alle potenziellen 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, eine gegenstandsbezogene Handlungstheorie der Akteure, eine Theorie des Programmablaufs (für eine detaillierte Darstellung mit Beispielen s. Weiss, 1995, 1997). Die Bezeichnung „logisches Modell" wäre hier vielleicht treffender (vgl. Patton, 1997). Die Bezeichnung „theoriebasierte Evaluation" ist etwas irreführend, denn auch das Modell der Programmforschung ist „theoriebasiert": Methodische Voraussetzung für die Analyse ist auch hier ein in sich schlüssiges, einheitliches System von operationalisierbaren Hypothesen, das die theoretische Basis für die Planung des Programms, seine Implementation und die gezielte Messung der Effekte rekonstruieren soll.

Bei dem logischen Modell der Programmevaluation bleibt allerdings das zentrale Problem unberücksichtigt, dass häufig eine solche *einheitliche Programmtheorie* lediglich als Konstrukt des Forschers an das Programm herangetragen wird, um das Evaluationsdesign wissenschaftlich und methodologisch begründet entwickeln zu können. Für die Entscheidungen über das Programm selbst dürften dagegen die Überzeugungen der Planer der konkreten Maßnahmen entscheidend sein, also deren individuelle Vermutungen über die Notwendigkeit der Erreichung bestimmter Ziele und über die Eignung dafür einzusetzender Instrumente. Ebenso dürften die mit der Implementation betrauten Instanzen eigene – vielleicht sogar von den Planern abweichende – Vorstellungen darüber besitzen, wie die Maßnahmen im Detail unter 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 empirisch begründetes *handlungslogisches Rahmenkonzept* sein, das den von den Beteiligten vermuteten Prozess von den Maßnahmen über alle Zwischenschritte bis zu den Wirkungen skizziert.

Für eine theoriebasierte Evaluation stellt ein solches ablaufsorientiertes „logisches Modell" den „roten Faden" zur Verfügung, an dem entlang Detailinformationen über den gesamten Prozess aus der Perspektive der jeweiligen Akteure gesammelt werden. Durch die Orientierung an den Akteurstheorien wird vermieden, dass zwischen dem Einsatz eines Instruments und der Messung der Veränderungen im vorgesehenen Wirkungsfeld eine Blackbox verbleibt. So kann die Evaluation unmittelbar und konkret 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.

Alternativen im Kontrollparadigma: Indikatorenmodelle, Bewertung durch Betroffene

Beim Kontrollparadigma steht nicht das Interesse an der Gewinnung übergreifender und transferfähiger Erkenntnisse im Vordergrund, sondern die Beurteilung der Implementation und des Erfolgs eines bestimmten, konkreten Interventionsprogramms. Soweit es sich um ein Programm mit explizierten Ziel-Mittel-Relationen handelt, sind unter methodischem Gesichtspunkt selbstverständlich das Experiment bzw. seine Alternativen Quasi-Experiment oder Ex-post-facto-Design die geeignete Wahl. Allerdings steht nicht selten eine andere Thematik im Zentrum des Kontroll-Interesses, nämlich Qualitätssicherung und Qualitätsentwicklung – gerade im Falle zielgruppenbezogener Programme, wie etwa fortlaufend zu erbringende Humandienstleistungen durch eine Organisation oder Institution (z. B. Lehr- und Ausbildungsdienstleistungen von Schulen und Hochschulen).

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 (etwa lange Studienzeiten oder hohe Studienabbruchquoten) in erster Linie die vorgeblich schlechte Lehre verantwortlich gemacht und deren Qualitätsverbesserung eingefordert.

Im Kontrollparadigma gehört es somit 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, nämlich 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" ein mehrdimensionales Konstrukt ist, das von außen an den Sachverhalt zum Zwecke der Beurteilung herangetragen wird. Wenn 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 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.

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 Voraussetzung und Rah-

men für die Leistungserbringung sowie die nachfolgenden Ergebnisse, die die erbrachte Leistung bei den Adressaten bewirkt. Damit sind drei Qualitätsbereiche benannt sowie drei Felder (Input, Prozess, Outcome) 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 für *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). Dieser an sich nahe liegende und plausible Zusammenhang wird in nicht wenigen Hochschulprojekten zur Evaluation erstaunlicherweise übersehen.

Wenn man 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 (ggf. unter Einbeziehung der Beteiligten und Betroffenen; 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, oder sie werden von vornherein unter dem Gesichtspunkt leichter Messbarkeit ausgewählt. Nicht nur ist die Validität solcher Indikatoren oft zweifelhaft. Sie bergen auch die Gefahr der Fehlsteuerung, indem statt der gewünschten Qualität vor allem die leicht messbaren Sachverhalte optimiert werden.

Wenn – wie dargelegt – als letztliches Kriterium für den Erfolg der Dienstleistung der positive Effekt bei den Adressaten der Dienstleistung (Outcome) gelten soll, dann ist als Beurteilungsmaßstab für die Güte der Indikatoren die sog. „Kriteriumsvalidität" zu wählen; d. h. die Indikatoren in den Bereichen Struktur und Prozess sind nur in dem Maße valide, wie sie signifikante empirische Beziehungen zu Outcome-Indikatoren aufweisen.

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 bereits kurz angesprochen – die Bewertung durch die Adressaten bzw. Nutzer und/oder die Ermittlung ihrer Zufriedenheit gesetzt: Sie – so wird argumentiert – sind als die von dem zu evaluierenden Programm ganz konkret „Betroffenen" in der Lage, aus eigener Erfahrung auch dessen Qualität zuverlässig zu beurteilen. 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 häufig als empirisch falsch. Die per Umfrageforschung bei Nutzern oder Betroffenen erhobenen Antworten auf bewertende (also evaluative) Fragen haben nur bei Vorliegen sehr spezifischer Rahmenbedingungen den Status von „Evaluation" als methodisch kontrollierter, empirischer Qualitätsbewertung. Beispiele dafür liefert die „Lehrevaluation" an Hochschulen (Kromrey, 1996, 1999, 2001 und die Beiträge in diesem Heft). Ermittelt wird auf diese Weise zunächst die „Akzeptanz" (oder Nicht-Akzeptanz), auf die der beurteilte Sachverhalt bei den Befragten stößt; und diese hängt im Wesentlichen von Merkmalen der Befragten und nur relativ ge-

ring von Merkmalen des beurteilten Sachverhalts ab.

Um Missverständnissen vorzubeugen: Dies spricht *nicht* gegen die Befragung als zentrales empirisches Informationsinstrument in Evaluationsvorhaben. Natürlich sind per Befragung erhobene Daten eine außerordentlich wichtige, häufig auf andere Weise nicht zu beschaffende Grundlage *für* Evaluationen; sie sind lediglich noch nicht die Evaluation selbst (für eine Darstellung verschiedener Einsatzmöglichkeiten von Befragungen für Evaluationen im Hochschulbereich s. Kromrey, 2000). Und natürlich sind auch Akzeptanzaussagen keine unwesentliche Information, insbesondere nicht in solchen Dienstleistungsbereichen, in denen der Erfolg von der aktiven Partizipation der Adressaten abhängt (beispielsweise eben in Lehr-Lern-Prozessen).

Literatur

Beywl, W. (1991). Entwicklung und Perspektiven praxiszentrierter Evaluation. *Sozialwissenschaften und Berufspraxis, 14,* 265–279.

Chelimsky, E. (1997). Thoughts for a new evaluation society. „Keynote speech" at the UK Evaluation Society conference in London 1996. *Evaluation, 3,* 97–109.

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

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

Ehrlich, K. (1995). Auf dem Weg zu einem neuen Konzept wissenschaftlicher Begleitung. *Berufsbildung in Wissenschaft und Praxis, 24,* 32–37.

Frey, S. & Frenz, H.-G. (1982). Experiment und Quasi-Experiment im Feld. In J.-L. Patry (Hrsg.), *Feldforschung* (S. 229–258). Bern, Stuttgart: Huber.

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

Herman, S. E. (1997). Exploring the link between service quality and outcomes. Parents' assessments of family support programs. *Evaluation Review, 21,* 388–404.

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

Krogstrup, H. K. (1997). User participation in quality assessment. A dialogue and learning oriented evaluation method. *Evaluation, 3,* 205–224.

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

Kromrey, H. (1995). Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 15,* 313–335.

Kromrey, H. (1996). Qualitätsverbesserung in Lehre und Studium statt so genannter Lehrevaluation. Ein Plädoyer für gute Lehre und gegen schlechte Sozialforschung. *Zeitschrift für Pädagogische Psychologie, 10,* 153–166.

Kromrey, H. (1999). Von den Problemen anwendungsorientierter Sozialforschung und den Gefahren methodischer Halbbildung. *Sozialwissenschaften und Berufspraxis, 22,* 58–77.

Kromrey, H. (2000). Qualität und Evaluation im System Hochschule. In H. Stockmann (Hrsg.), *Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder* (S. 233–258). Leverkusen: Leske + Budrich.

Kromrey, H. (2001). Studierendenbefragungen als Evaluation der Lehre? Anforderungen an Methodik und Design. In U. Engel (Hrsg.), *Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre* (S. 11–47). Frankfurt/M.: Campus.

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

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

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

Smith, A., Preston, D., Buchanan, D. & Jordan, S. (1997). When two worlds collide. Conducting a management evaluation in a medical environment. *Evaluation, 3,* 49–68.

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

Weiss, C. H. (1995). Nothing is as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Conell, A. C. Kubisch, L. B. Schorr & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives* (pp. 65–92). Washington, DC: The Aspen Institute.

Weiss, C. H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21,* 501–524.

Prof. Dr. Helmut Kromrey
Freie Universität Berlin
Institut für Soziologie
Garystr. 55
D-14195 Berlin
Tel. (0 30) 83 85 76 18
Fax (0 30) 83 85 76 17
E-Mail: kromrey@zedat.fu-berlin.de