Evaluation – ein vielschichtiges Konzept Begriff und Methodik von Evaluierung und Evaluationsforschung. Empfehlungen für die Praxis

Helmut Kromrey

1 Vorbemerkung

Die in den vergangenen Jahren an Umfang zunehmende Diskussion um Evaluation und Evaluierung sieht sich nicht nur einer schwer durchschaubaren Vielfalt und Komplexität von damit in Verbindung gebrachten Methoden und Verfahren, von Forschungs- und Beratungsansätzen gegenüber. Sie wird weiter erschwert durch ein kaum begrenzbares Spektrum von „Gegenständen" der Evaluation sowie durch eine unüberschaubare Fülle von Fragestellungen. Sie wird in ihrem Ertrag eingeschränkt durch eine geradezu inflationäre Verwendung des Begriffs in extrem unterschiedlichen Bedeutungsvarianten. Der vorliegende Beitrag versucht, ein wenig Abhilfe dadurch zu schaffen, dass er die gängigen Begriffe, Konzepte und Verfahrensweisen skizziert und systematisiert, denen wir im Diskussionslabyrinth um die „Evaluation" begegnen.

2 Was „ist" Evaluation?

Verwirrung kann bereits dadurch entstehen, dass *das sprachliche Zeichen „Evaluation" für unterschiedliche Typen von Referenzobjekten* stehen kann (und steht).

Eine erste Gruppe von Referenzobjekten ist auf der symbolischen und gedanklichen Ebene angesiedelt. „Evaluation" steht einerseits als vermeintlich wohlklingendes Fremdwort für den (durchaus alltäglichen) *Begriff* „Bewerten" und/oder „Bewertung", andererseits für ein spezifisches (nicht mehr alltägliches) *Denkmodell*: ein nachprüfbares Verfahren des Bewertens. Vor allem um dieses Denkmodell geht es, wenn wir über die Methoden, Verfahren und Ansätze der Evaluation diskutieren.

Die zweite Begriffsebene bezieht sich auf ein *spezifisches Handeln*, einen Prozess: auf zielorientiertes Informationsmanagement. Im allgemeinsten 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.

Wer in ein Evaluationsprojekt involviert ist, hat es immer mit allen drei Begriffsebenen zu tun; und wem es nicht gelingt, sie in seinen Argumentationen trennscharf auseinander zu halten, der wird leicht in Diskussions-Sackgassen landen.

3 Unterschiedliche Kontexte schaffen sich unterschiedliche Begriffsbedeutungen

Dem Begriff „Evaluation" begegnen wir gegenwärtig in den verschiedensten Diskussionskontexten: im Alltag ebenso wie in der Politik, in der Methodologie empirischer Wissenschaft ebenso wie im (spezifischeren) Zusammenhang der Umfrageforschung. Ärgerlicherweise aber treffen wir – sobald wir den Kontext wechseln – hinter der selben Worthülse auf recht unterschiedliche Konzepte und Vorstellungen.

Der *alltägliche Sprachgebrauch* ist ausgesprochen unspezifisch; „Evaluation" bedeutet nichts weiter als „Bewertung": *Irgend etwas wird von irgend jemandem 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). Selbst 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 Informationsbeschaffung über Verlauf und Resultate eines (Handlungs- und Maßnahmen-)„Programms" mit explizit formulierten

Zielen und Instrumenten[1] 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.

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 „irgend etwas“ evaluiert wird, sondern dass spezifizierte Sachverhalte, 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.

4 Voraussetzungen für ein erfolgreiches Evaluationsvorhaben: Präzisierungen, Rollendefinitionen und Kompetenzklärungen

Präzisierungen zu jedem der im vorigen Abschnitt 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.

---

[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 (vgl. Hellstern/Wollmann 1983, S. 7)

Tabelle: Evaluation: Begriffsdimensionen und Klärungsbedarf

alltäglicher Sprachgebrauch	*wissenschaftlicher Sprachgebrauch*	*Präzisierungen*	*Klärungsbedarf*
■ Irgend etwas wird	■ Programme, Maßnahmen, Organisation etc. werden	existierend; in Planung/Entwicklung; bereits implementiert; Feldversuch/Pilotprojekt; Programmumfeld etc.	■ Was ist das „Programm“ und seine Ziele? ■ Was ist der „Gegenstand“ der Evaluierung? Was sind die Evaluationsziele?
■ von irgend jemand	■ durch Personen, die zur Bewertung besonders befähigt sind,	unabhängige Wissenschaftler, Auftragsforscher, im Programm Mitwirkende, externe Berater, engagierte Betroffene etc.	Wer hat welche Funktionen / Kompetenzen? ■ Informanten / Informationsquellen ■ Informationsbeschaffung und –aufbereitung ■ Evaluierende
■ in irgend einer Weise	■ in einem objektivierten Verfahren	Hearing, qualitative / quantitative Forschungslogik, experimentell / nicht-experimentell, formativ / summativ etc.	■ Methoden und Verfahren der Informationsbeschaffung ■ Methoden und Verfahren des Bewertens ■ Legitimation zum Bewerten
■ nach irgendwelchen Kriterien bewertet.	■ nach explizit auf den Sachverhalt bezogenen und begründeten Kriterien (und ggf. Standards) bewertet.	Zielerreichung / Effekte / Nebenwirkungen, Effizienz / Effektivität, Sozialverträglichkeit, Zielgruppenbezug etc.	■ Ziele (wessen Ziele?) ■ Kriterien ■ Standards

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). Zwar ist für *diesen* „Gegenstand" ein kaum vollständig aufzählbares Spektrum an möglichen Variationen denkbar: Der zu evaluierende Sachverhalt kann schon lange bestehen, sich gerade im Prozess der Realisierung befinden oder gar erst als Planungs- und Entwicklungsabsicht existieren; er kann sehr umfassend und abstrakt oder aber eng umgrenzt und konkret sein; er kann (im Sinne „experimenteller Politik") ein Pilotvorhaben sein, das in einem abgegrenzten (zumindest prinzipiell abgrenzbaren) Feld durchgeführt wird, oder aber eine Innovation, die sich in Konkurrenz zu bestehenden Angebotsalternativen behaupten soll.

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)[2] 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 – anders als im Kontext von Grundlagenforschung – 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 aber für das Entscheidungshandeln keinen Unterschied ausmacht, ob sie so oder anders ausfallen, sind irrelevant, sind Verschwendung von Evaluationsressourcen.

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. Bei Patton (1997) – der in diesem Zusammenhang von „intended use by intended users" spricht – findet sich als Empfehlung für Planer und Durchführende von

---

[2] Eine umfassende Evaluation bestünde danach 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" (vgl. Lösel/Nowack 1987, S. 57).

Evaluations-Vorhaben, sich die handlungslogische Abfolge von Schritten oder Stufen in der Programmdurchführung („logical framework“) zu vergegenwärtigen und diesen Stufen die entsprechenden evaluationsrelevanten Informationen zuzuordnen (vgl. 1997, S. 234 ff., Tabelle 10.5). Dies beginnt auf der Implementationsseite mit den Programm-Inputs (1) – bzw. auf der Informationsbeschaffungsseite mit Daten über Ausgaben, Personal, investierte Zeit –, verläuft über die Implementations-Aktivitäten (2) – bzw. deren monitoring – , über die Beteiligten, die Zielgruppen und weiteren Betroffenen (3), über deren Reaktionen auf diese Aktivitäten (4) schließlich zu den bewirkten Veränderungen im Hinblick auf Kenntnisse, Einstellungen, Fertigkeiten (5) sowie den daraus ggf. folgenden kurz-, mittel- und langfristigen Auswirkungen auf die Programm-Umwelt, auf geänderte Verhaltensweisen der Zielgruppen (6). An oberster Stelle der „Programmdesign-Hierarchie“ steht schließlich das „eigentliche“ Ziel (7) , zu dem das Programm konzipiert und implementiert wurde, etwa Sicherung der Marktposition bei einem Unternehmensprojekt oder Verbesserung der Chancengleichheit benachteiligter Bevölkerungsgruppen bei einem sozialpolitischen 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 (man hat wechselseitig kein Verständnis für die Ansprüche und Empfindlichkeiten der anderen Beteiligten, man begegnet sich mit Misstrauen, „bremst sich gegenseitig aus“); im ungünstigsten Fall kann es auch mit dem vollständigen Scheitern des Vorhabens enden. Die Bedeutung „vertrauensbildender Maßnahmen“ im Vorfeld darf auf keinen Fall unterschätzt werden; Pa tentrezepte existieren allerdings nicht. Das liegt schon allein daran, dass die mit dem Evaluations-Vorhaben betrauten Personen in unterschiedlichster Weise zum Gegenstand der Bewertung in Bezug stehen können: 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.

Ein Rat sollte aber 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. Allenfalls in seltenen Ausnahmefällen werden die Aufgaben und Kompetenzen für alle drei Funktionen „in einer Hand“ liegen (können); etwa im Falle eines zur Fremdevaluation eingesetzten

externen Gremiums, das Daten sammelt, Bewertungen vornimmt und Empfehlungen ableitet. Für Evaluationen im Rahmen von Organisationsentwicklungs-Vorhaben empfiehlt sich eher eine Dreiteilung auch der Kompetenzen. Zum Beispiel: Ein Team externer, empirisch-methodisch ausgewiesener Forschungsexperten ist zuständig für die Informationsbeschaffung, -analyse und –präsentation; ein kleines Gremium von legitimierten Vertretern der beteiligten Gruppen 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. Natürlich sind auch andere Kombinationen von Aufgaben und Zuständigkeiten möglich und – je nach faktischen Gegebenheiten – erfolgversprechend. Zu vermeiden ist lediglich, dass ohne ausdrückliche Legitimation ein „Evaluationsteam" eingesetzt wird, das mit diffusen und für die Beteiligten undurchschaubaren Zielen und Kompetenzen seine Tätigkeit aufnimmt.

Nach diesen Klärungen bildet die *Festlegung der Bewertungskriterien* (letzte Zeile der o. g. Tabelle) 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 soll die für die Bewertungen zuständige Instanz (die „Evaluatoren" im engeren Sinne) ihre Urteile nicht nach ad hoc zustande kommenden Kriterien und Maßstäben fällen, sondern ihre Aussagen sollen nachvollziehbar, überprüfbar und kritisierbar sein. Sofern die Klärungen zu den beiden erstgenannten Bereichen hinreichend eindeutig getroffen wurden, dürften an diesem Punkt keine unüberwindbaren Probleme mehr auftreten. Denkbar ist allerdings wiederum ein ganzes Spektrum sehr unterschiedlicher Bewertungskriterien und -standards. Sie können sich beziehen auf die *Wirkungen* und Nebenwirkungen der Maßnahmen eines Programms, auf die Art und Effizienz der *Durchführung*, auf die *Eignung* und *Effektivität* der gewählten Maßnahmen für die Zielerreichung, 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 Sozialarbeitskonzepts 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.

5 Was also 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[3], 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: Erfüllt der zu evaluierende Gegenstand den 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: 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.*

---

[3] Michael Quinn Patton (1997) listet in seinem einflussreichen Werk „Utilization-Focused Evaluation“ (vgl. 1997, S. 192 ff.) nicht weniger als 58 das Design mitbestimmende Zwecke auf; und er fügt hinzu, dass damit noch bei weitem nicht das gesamte Spektrum erfasst sei.

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 – im Unterschied zu üblicher wissenschaftlicher Forschung – 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 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 beeinträchtigen.

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 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 immerhin 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 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.

6 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 existieren eine Reihe von Versuchen, die Vielfalt im Detail auf eine überschaubare Zahl von Typen zu reduzieren. Für besonders nützlich halte ich einen Vorschlag von Eleanor Chelimsky (1997, 100 ff.), die drei „conceptual frameworks" unterscheidet:

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

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

6.1 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 (vgl. Weiss 1974, S. 11). Insbesondere staatliche Auftragsforschung eröffnet einen Weg, Zugang zu den internen Strukturen und Prozessen des politisch-administrativen Systems zu erhalten. 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.

6.2 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“ (vgl. 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 ab vom Informationsbedarf der programmdurchführenden und/oder der finanzierenden Instanz. Gefordert werden häufig quantitative Informationen. Eine sehr gute Darstellung dieses Ansatzes findet sich in Eekhoff u.a. (1977).

6.3 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. 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, S. 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

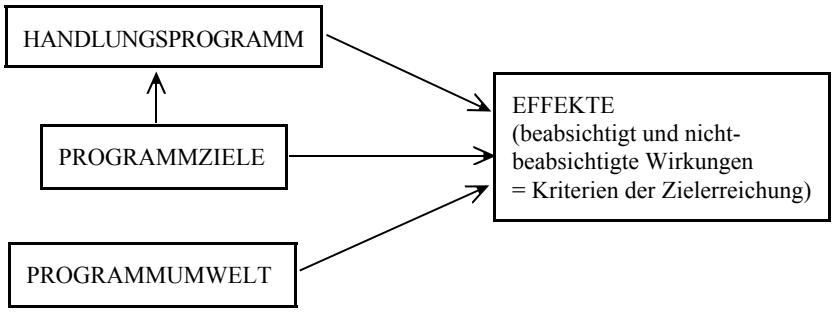
Eingangsphase einer Evaluation, aber auch die anschließenden Erhebungs-, Auswertungs-, Interpretations- und Berichtsarbeiten werden auf die Wünsche der Beteiligtengruppen abgestimmt“ (vgl. Beywl 1991, S. 268). Die Funktion der Evaluation ist hier in erster Linie die eines Helfers und Beraters.[4]

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

7.1 Begriffsexplikation

Es wurde bereits 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 irgend einer Weise bewertet* – durch eindeutige Präzisierungen absetzt (s.o., Abschnitt 2). Da diese Präzisierungen in unterschiedlicher Weise möglich und notwendig sind (s.o., Abschnitt 3), 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 (s.o., Abschnitt 4). 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.

Abbildung: Programmforschung



[4] Entsprechend findet sich manchmal auch die Charakterisierung als „Helfer- und Beratermodell der Evaluation“ (vgl. Abschnitt 7 dieses Beitrags).

Die drei in der Abbildung dargestellten Programmdimensionen (Ziele – Maßnahmen – Effekte) können – s.o.: „Evaluationsgegenstand" – 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. Schließ lich 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 ist – auch darauf wurde bereits hingewiesen – 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.

7.1.1 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")

7.1.2 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.

7.1.3 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.

7.1.4 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 Evaluationsteam, 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, 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.

7.2 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 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 (methodisch vergleichsweise einfachen) *deskriptiven* „Monitoring“ der Instrumentvariablen (Programm-Input), der exogenen Einflüsse und der Zielerreichungsgrade (Output).
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. 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. Die von der empirischen Sozialforschung entwickelte Methodologie der Programmevaluation ist daher weniger ein Real- als ein Idealtyp, an den anzunähern die Forscher sich je nach gegebener Situation bemühen werden.

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).

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).

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. Auch weniger anspruchsvolle „quasi-experimentelle Anordnungen", in denen Abweichungen vom echten Experiment durch alternative methodische Kontrollen ersetzt werden, sind nur selten realisierbar.

Auf die im einzelnen recht komplexen Details quasi-experimenteller Forschung sowie weiterer Methodenfragen soll in diesem Beitrag jedoch nicht näher eingegangen (für detaillierte Darstellungen vgl. Hellstern/Wollmann 1983; Kromrey 1987, 1988 und 1995). Wichtig ist mir an dieser Stelle lediglich zu betonen, dass die Anwendbarkeit der skizzierten Methodik der Programmforschung für Evaluations-Vorhaben nicht als der Regelfall, sondern eher als der Ausnahmefall gelten kann. Es muss also zu „Ersatzlösungen" gegriffen werden, die praktikabel erscheinen und dennoch hinreichend gültige Ergebnisse liefern.

7.3 Alternativen zum Experimentaldesign

7.3.1 Alternativen im Forschungsparadigma: „ex-post-facto-Design", theoriebasierte Evaluation

Als idealtypischer „Königsweg" der Evaluationsforschung (in angelsächsischen Texten auch als „Goldstandard" bezeichnet) gilt zwar – s.o. – 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 zufälligen Zuweisung; 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 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 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 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 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 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 (vgl. Weiss 1995, 1997). Die Bezeichnung „logisches Modell"

wäre vielleicht treffender (vgl. Patton 1997, S. 234 ff.: logical framework approach), zumal die Bezeichnung „theoriebasierte Evaluation" etwas irreführend ist, denn auch das Modell der Programmforschung ist „theoriebasiert": Methodische Voraussetzung für die Analyse ist ein in sich schlüssiges, einheitliches System von operationalisierbaren Hypothesen, 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 allerdings das 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. 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.

7.3.2 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. 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 z. B. fortlaufend zu erbringende 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 (etwa 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.

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 Ergebniseines Aushandlungsprozesses zwischen den Beteiligten,[5] 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.

---

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

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; 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.

Angesichts der Schwierigkeit und Aufwändigkeit solchen Vorgehens wird nicht selten eine einfachere Lösung gesucht und – vermeintlich – auch gefunden. An die Stelle methodisch kontrollierter Evaluation durch Forschung wird – wie oben (Abschnitt 6.1.4) 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 häufig als empirisch falsch (Beispiele dafür liefert die „Lehrevaluation" an Hochschulen; vgl. Kromrey 1996, 1999, 2001). 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 nicht in solchen Dienstleistungsbereichen, in denen der Erfolg von der aktiven Partizipation der Adressaten abhängt (beispielsweise eben in Lehr-Lern-Prozessen oder in der Familienhilfe oder generell in der Sozialarbeit).

7. Das Leitkonzept für das Entwicklungsparadigma der Evaluation: Das „Helfer- und Beratermodell“ der Evaluation

Das Konzept von Evaluation als Programmforschung ist – wie in Abschnitt 6.2 dargestellt – methodisch schwer realisierbar und muss von Voraussetzungen über den Untersuchungsgegenstand ausgehen, die nur selten hinreichend erfüllt sind. Auch das dem „Programm“-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, S. 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.[6]

Literatur:

---

[6] 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. 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.

Beywl, W., 1991: Entwicklung und Perspektiven praxiszentrierter Evaluation. In: Sozialwissen schaften und Berufspraxis, 14/3, 265-279.
Chelimsky, E., 1997: Thoughts for a new evaluation society. „Keynote speech“ at the UK Evaluati on Society conference in London 1996. In: Evaluation, 3/1, S. 97-109.
Donabedian, A., 1980: Explorations in quality assessment and monitoring: The definition of quality and approaches to its assessment, Ann Arbor, MI.
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, S. 32-37.
Frey, S.; Frenz, H.-G., 1982: Experiment und Quasi-Experiment im Feld. In: Patry, J.-L. (Hrsg.): Feldforschung. Bern, Stuttgart, S. 229-258.
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, S. 388-404.
Hübener, A.; Halberstadt, R., 1976: Erfolgskontrolle politischer Planung – Probleme und Ansätze in der Bundesrepublik Deutschland. Göttingen.
Krogstrup, H.K., 1997: User participation in quality assessment. A dialogue and learning oriented evaluation method. In: Evaluation, 3/2, S. 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, S. 478-499.
Kromrey, H., 1988: Akzeptanz- und Begleitforschung. Methodische Ansätze, Möglichkeiten und Grenzen. In: Massacommunicatie, 16/3, S. 221-242.
Kromrey, H., 1995: Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. In: ZSE Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, Jg. 15, H. 4, S. 313-335.
Kromrey, H., 1996: Qualitätsverbesserung in Lehre und Studium statt sogenannter Lehrevaluation. Ein Plädoyer für gute Lehre und gegen schlechte Sozialforschung. In: Zeitschrift für Päd agogische Psychologie, 10/3-4, S. 153-166.
Kromrey, H., 1999: Von den Problemen anwendungsorientierter Sozialforschung und den Gefahren methodischer Halbbildung. In: Sozialwissenschaften und Berufspraxis, Jg. 22, H. 1, S. 58-77.
Kromrey, H., 2001: Studierendenbefragungen als Evaluation der Lehre? Anforderungen an Methodik und Design. In: Engel, Uwe (Hrsg.): Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre. Frankfurt a. M., S. 11-47.
Lösel, F.; Nowack, W., 1987: Evaluationforschung. In: J. Schultz-Gambard (Hrsg.): Angewandte Sozialpsychologie. München, Weinheim, S. 57-87.
Patton, M.Q., 1997. Utilization-focused evaluation. 3rd ed, Thousand Oaks, CA, London.
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.
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, S. 49-68.

Weiss, C.H., 1974: Evaluierungsforschung. Methoden zur Einschätzung von sozialen Reform programmen. 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 S. 65-92.
Weiss, C.H., 1997: How can theory-based evaluation make greater headway? In: Evaluation Review, 21/4, S. 501-524.

Prof. Dr. Helmut Kromrey
Institut für Soziologie an der FU Berlin
Babelsberger Str. 14-16
10715 Berlin
Tel.: ++49.30.85002.230
Fax: ++49.30.85002.138
e-Mail: kromrey@zedat.fu-berlin.de
http://userpage.fu-berlin.de/~kromrey

Prof. Dr. Helmut Kromrey, Jahrgung 1940, berufliche Tätigkeiten in Industrie, Presse, Rundfunk und Hochschule. Abitur auf dem 2. Bildungsweg, Studium VWL und Soziologie in Köln; Promotion an der Universität Dortmund, Habilitation an der Ruhr-Universität Bochum. Seit 1994 Inhaber eines Lehrstuhls für Soziologie/Empirische Sozialforschung an der Freien Universität Berlin.