Evaluation – ein Überblick

von Helmut Kromrey[1]

veröff. In: Heidrun Schöch (Hg.): Was ist Qualität. Die Entzauberung eines Mythos, Berlin 2005: Wissenschaftl. Verlag (Schriftenreihe Wandel und Kontinuität in Organisationen, Band 6), S. 31-85)

1 Prof. Dr. Helmut Kromrey, Kufsteiner Str. 12, D-10825 Berlin; eMail: kromrey@bds-soz.de; URL: www.profkromrey.de

1 Was „ist“ Evaluation?

1.1 Das Allerweltswort „Evaluation“

Der Begriff Evaluation ist zu einem schillernden Allerweltswort geworden, mit dem je nach Kontext sehr Unterschiedliches verbunden wird.

1.1.1 Verschiedene Referenzebenen

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. Um dieses Denkmodell geht es z.B., wenn über die Methoden, Verfahren und Ansätze der Evaluation diskutiert oder gestritten wird.

Die zweite Begriffsebene bezieht sich auf ein *spezifisches Handeln*: auf die Durchführung eines Evaluationsprojekts, auf zielorientiertes Informationsmanagement.

Und schließlich bezeichnet „Evaluation“ auch noch etwas Punktuelles: das *Resultat dieses 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.

1.1.2 Verschiedene Kontexte

Dem Begriff „Evaluation“ begegnen wir in den verschiedensten Diskussionskontexten: im Alltag ebenso wie in der Politik, in der Methodologie empirischer Wissenschaft ebenso wie im Zusammenhang der Umfrageforschung. Das wäre nicht weiter schlimm, wenn wir nicht – sobald wir den Kontext wechseln – hinter derselben Worthülse auf recht unterschiedliche Konzepte und Vorstellungen träfen.

Der *alltägliche Sprachgebrauch* ist ausgesprochen unspezifisch. Mit „Evaluation“ Bewertung gemeint sein: *Irgend etwas wird von irgend jemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet.* Die Konsequenz: Derselbe Sachverhalt kann von verschiedenen Individuen sehr unterschiedlich bis gegensätzlich eingeschätzt und beurteilt werden.

In *politischen Argumentationen* sind die Begriffsverwendungen zwar wesentlich spezifischer, zugleich aber *außerordentlich vielfältig*. Die Bezeichnung gilt für

Effizienzmessungen in ökonomischen Zusammenhängen; sie gilt für die von Sachverständigen vorgenommene Analyse der Funktionsfähigkeit von Organisationen; selbst die beratende und moderierende Beteiligung im Prozess der Entwicklung von Handlungsprogrammen mit dem Ziel ihrer Optimierung wird von diesem Begriff erfasst.

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 zum Gegenstand hat. Evaluationsziele sind die wissenschaftliche Begleitung der Programm-Implementation und/oder die „Erfolgskontrolle" und „Wirkungsanalyse".

Schließlich wird auch *im Zusammenhang „gewöhnlicher" Umfrageforschung* von „Evaluation" gesprochen. Gemeint ist hier die *Erhebung und Auswertung bewertender (also „evaluierender") Aussagen* von Befragten, die in einem angebbaren Verhältnis zu dem zu evaluierenden „Gegenstand" stehen (etwa Kunden/Klienten, Betroffene, Teilnehmer von Bildungsveranstaltungen). Ein spezifisches Evaluationsdesign existiert in diesem Fall nicht. Ins Auge fällt hier die Nähe zur Meinungsforschung .

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.

2 Was ist empirisch-wissenschaftliche Evaluation?

Wie noch auszuführen sein wird, kann „Gegenstand" der Evaluation im Prinzip alles sein und ist das Spektrum der Evaluations-„Fragestellungen" oder „Zwecke" praktisch unbegrenzt. Es existieren auch keine speziellen Methoden der Evaluation; vielmehr ist auf das bekannte Arsenal der „gewöhnlichen" empirischen Sozialforschung zurückzugreifen. Mit anderen Worten: Es kann kein „Musterdesign für Evaluationen" angeboten werden, sondern es sind immer „maßgeschneiderte" Vorgehensweisen zu entwickeln und zu begründen. Wenn dies so ist: Was ist dann eigentlich „Evaluation" als empirisch-wissenschaftliches Verfahren?

Um es vorwegzunehmen: Die methodologisch einzig sinnvolle Antwort kann nur lauten: Evaluation ist angewandte Sozialwissenschaft (nicht lediglich Sozialforschung). Evaluation als wissenschaftliches Vorgehen ist eine *methodisch kontrolliertes, verwertungs- und bewertungsorientiertes Sammeln und Verwerten 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 durch dieses spezifische 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 (obwohl auch dies nicht ausgeschlossen ist), sondern – umgekehrt - wissenschaftliche Verfahren und Erkenntnisse einzubringen, um sie für den zu evaluierenden Gegenstand nutzbar zu machen. Wissenschaft liefert hier 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 Lösungen zu suchen, die das Funktionsgefüge im sozialen Feld nicht beeinträchtigen.

2. 1 Die Vielfalt von Evaluationen: eine grobe Klassifikation

Natürlich wäre es wissenschaftlich wenig sinnvoll, ohne den Versuch eines Ordnungsschemas vor der potentiellen Variationsbreite von Evaluationen zu kapitulieren. Unter den Versuchen, die Vielfalt im Detail auf eine überschaubare Zahl von Typen zu reduzieren, ist ein Vorschlag von Eleanor Chelimsky (1997, 100 ff.), besonders nützlich. Sie unterscheidet drei „conceptual frameworks":

- Evaluation zur Verbreiterung der Wissensbasis (kurz: „Forschungsparadigma"),
- Evaluation zu Kontrollzwecken („Kontrollparadigma") und
- Evaluation zu Entwicklungszwecken („Entwicklungsparadigma").

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

2.1.1 Das „Forschungsparadigma" der Evaluation

Für anwendungsorientierte Universitätswissenschaftler sind Evaluationsprojekte eine Chance, neben dem eigentlichen Evaluationszweck auch grundlagen-wissenschaftliche Ziele zu verfolgen. Evaluation kommt aus dieser Perspektive

die Rolle eines Bindeglieds zwischen Theorie und Praxis zu (Weiss 1974, S. 11). Sie eröffnet einen Weg, Zugang zu den internen Strukturen und Prozessen z.B. des politisch-administrativen Systems zu erhalten. 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 diesem Ansatz in erster Linie als Wirkungsforschung, die Evaluierung 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. Daher ist es kein Zufall, dass Beiträge zur Entwicklung einer allgemeinen Evaluationstheorie und -methodologie vor allem aus dem Kreis universitärer Evaluationsforscherinnen und -forscher geleistet wurden.

2.1.2 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 objektivierbaren Erfolgskriterien (Effektivität, Effizienz, Akzeptanz) unterwerfen (ausführlicher in Kromrey 2003). Evaluationen dieser Art werden argumentativ vertreten als eine weitere Kontrollform administrativen Handelns neben Rechtmäßigkeits-Kontrolle (Gerichte), politischer Kontrolle (Parlamente) und Wirtschaftlichkeits-Kontrolle (Rechnungshöfe). 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 programm-durchführenden und/oder der finanzierenden Instanz. Gefordert werden häufig quantitative Informationen. [2]

2.1.3 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 „reinsten“ Fall ist Evaluation in die gesamte Programm-Historie eingebunden:

---

[2] Eine sehr gute Darstellung dieses Ansatzes findet sich in Eekhoff u.a. (1977).

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.

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

2.2.1 Begriffsexplikationen

Bei aller Vielfalt der bisher skizzierten Evaluationsbegriffe und –gegenstände sowie Aufgabenprofile bleibt dennoch – zumindest für das Forschungs- und das Kontrollparadigma – allen Vorhaben gemeinsam, dass sie (mindestens) drei interdependente Dimensionen aufweisen – nämlich Ziele, Maßnahmenprogramm, Effekte – und dass sie (anders als in einem Forschungslabor) von Umgebungseinflüssen nicht abgeschirmt werden können.

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

Eine Evaluation, die umfassend alle diese Punkte gleichgewichtig in Angriff nimmt, ist natürlich in keinem Projekt realisierbar. Es müssen Schwerpunkte gesetzt werden. Hierzu sind vier zentrale Fragen zu beantworten:

a) Was wird evaluiert? – Implementations- oder Wirkungsforschung

b) Wann wird evaluiert? – Summative oder formative Evaluation

c) Wo ist die Evaluation angesiedelt? – Externe oder interne Evaluation

d) Wer beurteilt nach welchen Kriterien? – Instanzen der Evaluierung

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

a) Stehen im Vordergrund des Evaluationsinteresses 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 zu erfassen. Steht dagegen die systematische Untersuchung der Planung, der Durchsetzung und des Vollzugs im Vordergrund, spricht man von *Implementationsforschung*. Eine Hauptaufgabe ist hier die systematische und kontrollierte „Buchführung": Was passiert? Was wird wann und wie gemacht? (= „monitoring")

b) Nach dem Zeitpunkt, an dem eine Evaluation ansetzt, 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; besonders geeignet ist sie 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 auf „projektformende" Effekte. Sie legt im Nachhinein ein zusammenfassendes Urteil, ein „Evaluationsgutachten" vor. Man spricht hier von *„summativer" Evaluation*.

c) Für die Evaluationspraxis und für die Akzeptanz der Resultate ist die Entscheidung wichtig, *wem die Evaluationsaufgabe übertragen wird*. In manchen Projekten ist die ständige (Zwischen-)Ergebniskontrolle expliziter Bestandteil des Programms selbst. Die Informationssammlung und -einspeisung gehört zum Konzept der Qualitätssicherung. Da dies durch das eigene Personal des Projektträgers geschieht, spricht man von *interner Evaluation*. Ihre Vorzüge werden darin gesehen, dass die Evaluation ständig „vor Ort" ist und problemlosen Zugang zu allen notwendigen Informationen hat. Werden dagegen die Dienste außenstehender unabhängiger Forscher in Anspruch genommen, handelt es sich um *externe Evaluation*. Ihre Vorzüge werden in der Professionalität der Evaluationsdurchführung (bei den Beauftragten handelt es sich in der Regel um Forschungsexperten) und in dem höheren Grad an Objektivität ihrer Resultate gesehen (die Evaluation legitimiert sich nicht durch einen erfolgreichen Ablauf des zu begleitenden Projekts, sondern durch wissenschaftliche Standards).

d) Von noch höherer Relevanz für die Akzeptanz des Evaluationsvorhabens ist schließlich die Entscheidung, *welche Instanz die letztlichen Bewertungen* (die „Evaluationen" im eigentlichen Sinne) *vornimmt* und woher die Kriterien stammen, nach denen diese Urteil gefällt werden. Im „traditionellen" Fall der *Programmevaluation* stammen die Beurteilungskriterien aus dem 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, 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 typischerweise nicht vorhanden ist, sondern durch die Evaluation erst geliefert werden soll (in Pilotprojekten werden schließlich Innovationen erprobt). Hier wird sich das Evaluationsprojekt häufig damit behelfen müssen, dass die *eigentliche Bewertung auf programm- und evaluationsexterne Instanzen verlagert* wird, z.B. durch Einholung von Fachgutachten oder Befragung neutraler Experten. Eine andere Variante des Verlagerns der Evaluierung auf eine programmexterne Instanz ist die Befragung der Adressaten eines Programms (Nutzer oder Betroffene). Dies ist nicht unproblematisch, denn bei den so erhobenen Urteilen handelt es sich weder um Bewertungen im Sinne „technologischer" Evaluationseinschätzung noch um Bewertungen neutraler, unabhängiger Experten. Erhoben werden vielmehr „Akzeptanzaussagen" von Personen, die in einer besonderen Beziehung (eben als Nutzer, als Betroffene) zum Untersuchungsgegenstand stehen. Korrekterweise dürfen daher solche Akzeptanzdaten noch nicht als die eigentlichen Evaluationen gelten, sondern als nutzungsnahe, „evaluationsrelevante" Informationen, die von einer legitimierten Bewertungsinstanz als empirische Basis für das zu treffende Evaluationsurteil genutzt werden können und sollen.

2.2.2 Methoden der Programmforschung: Das Feldexperiment als Referenzdesign

Forschungsmethodisch besteht das Konzept der Programmforschung darin, die auf den ersten Blick simpel anmutende Aufgabe zu lösen, die in Abschnitt 2.2.1 aufgeführten vier Variablenbereiche (Ziele – Maßnahmen – Effekte – Programmumwelt) mit empirischen Daten abzubilden (zu „messen") und miteinander zu verknüpfen.[3] 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

---

[3] Die methodisch komplexe Thematik kann hier nur kurz angerissen werden. Ausführlicher wird sie dargestellt in Kromrey 1995, Kap. 3.

„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" besteht nach diesem Modell aus den beiden Aspekten:

- Analyse der Programmziele und ihrer Interdependenzen sowie Zuordnung der Instrumente zur Zielerreichung (Maßnahmen des Programms) sowie
- Vergleich der den einzelnen Maßnahmen zurechenbaren Effekte mit den angestrebten Zielniveaus.

Das damit skizzierte Modell einer kausalanalytisch angeleiteten Programmevaluations- und Wirkungsforschung erscheint in sich schlüssig und einleuchtend. Bei näherem Hinsehen allerdings wird erkennbar, dass es von sehr 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 Programm-forschungsprojekten wenig realitätsnah. Um das methodologische Forschungsprogramm empirischer Kausalanalysen überhaupt anwenden zu können,

- muss *vor* der Entwicklung des Forschungsdesigns zum einen Klarheit über die Untersuchungsziele – bezogen auf einen definierbaren und empirisch abgrenzbaren Untersuchungsgegenstand – bestehen und dürfen sich für die Dauer der Datenerhebung weder die Untersuchungsziele noch die wesentlichen Randbedingungen des Untersuchungsgegenstandes in unvorhersehbarer Weise ändern;
- müssen des Weiteren begründete Vermutungen (Hypothesen) über die Struktur des Gegenstandes wie auch über Zusammenhänge und Beziehungen zwischen dessen wesentlichen Elementen existieren. Erst auf ihrer Basis kann ein Gültigkeit beanspruchendes Indikatorenmodell konstruiert, können geeignete Messinstrumente entwickelt, kann über problemangemessene Auswertungsverfahren entschieden werden.
- Und schließlich muss der Forscher die Kontrolle über den Forschungsablauf haben, um die (interne und externe) Gültigkeit der Resultate 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 jedoch 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.

Es wurde bereits 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,[4] so dass die Anwendbarkeit der skizzierten Methodik der Programmforschung für Evaluations-Vorhaben nicht als der Regelfall, sondern allenfalls als der Ausnahmefall gelten kann. Somit muss zu „Ersatzlösungen" gegriffen werden, die praktikabel erscheinen und dennoch hinreichend gültige Ergebnisse liefern.

2.3 Alternativen zum Experimentaldesign

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

Als idealtypischer „Königsweg" der Evaluationsforschung gilt zwar – s.o. – das Experimentaldesign. Der daran orientierte Realtyp ist 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. Beispielsweise tritt bei der Zusammenstellung strukturäquivalenter Versuchs- und Kontrollgruppen das matching-Verfahren an die Stelle der zufälligen Zuweisung („Randomisierung"); oder die nicht mögliche Abschirmung von Störgrößen in der Informationsbeschaffungsphase wird ersetzt durch umfassende Erhebung relevanter potentieller exogener Wirkungsfaktoren, um nachträglich in der Auswertungsphase die exogenen Einflüsse statistisch zu kontrollieren.

---

[4] Auch auf die Darstellung der Details experimenteller und quasi-experimenteller Forschung wird in diesem Überblick nicht näher eingegangen; s. dazu Frey/Frenz 1982 und Hellstern/Wollmann 1983.

An diese Überlegung knüpft eine Strategie an, die *Experimentallogik in der Erhebungsphase durch Experimentallogik in der Auswertungsphase zu simulieren*. Datengrundlage hierfür ist eine möglichst vollständige Deskription des Programmverlaufs („monitoring"); das heißt: Für alle untersuchungsrelevanten Variablen werden mit Hilfe des Instrumentariums der herkömmlichen empirischen Sozialforschung über die gesamte Laufzeit des Programms Informationen 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, 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.

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 insofern „theoriebasiert", als für die Analyse ein in sich schlüssiges, einheitliches System von operationalisierbaren Hypothesen, das als theoretische Basis für die Planung des Programms, für die Implementation und für die gezielte Messung der Effekte gelten kann, eine wesentliche methodische Voraussetzung ist.

Allerdings tritt bei diesem Rationalmodell der Programmevaluation tritt das Problem auf, dass eine solche einheitliche Programmtheorie als verbindliche Grundlage des Handelns der Programmdurchführenden im Allgemeinen faktisch nicht existiert, sondern ein Konstrukt des Forschers ist, um sein Evaluationsdesign wissenschaftlich und methodologisch begründet entwickeln zu können. In der Programmpraxis dürften vielmehr bei den Planern der Maßnahmen deren *jeweilige individuelle 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 ihr Handeln von ihren eigenen Alltagstheorien leiten lassen.

Es existieren also im Normalfall unabhängig von den theoretischen Vorstellungen der Evaluatoren mehrere – vielleicht sich ergänzende, vielleicht aber auch in Konkurrenz stehende – Programmtheorien, die den Fortgang des Programms steuern und für dessen Erfolg oder Misserfolg verantwortlich sind. Sie gilt es zu rekonstruieren und zum theoretischen Leitmodell der Evaluation zu systematisieren. Das Ergebnis könnten dann handlungslogische Rahmenkonzepte sein, in denen der von den Beteiligten vermutete Prozess von den Maßnahmen über alle Zwischenschritte bis zu den Wirkungen skizziert ist.

Von solchen ablaufsorientierten „logischen Modellen" angeleitet, kann die Evaluation Detailinformationen über den gesamten Prozess aus der Perspektive der jeweiligen Akteure sammeln. So vermeidet sie, zwischen dem Einsatz eines Instruments und der Messung der Veränderungen im vorgesehenen Wirkungsfeld eine black box zu belassen (wie dies 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.

2.3.2 Alternativen im Kontrollparadigma: Indiktorenmodelle, Bewertung durch Betroffene

Beim Kontrollparadigma geht es – wie bereits skizziert – nicht in erster Linie um die Gewinnung übergreifender und transferfähiger Erkenntnisse, sondern um die Beurteilung der Implementation und des Erfolgs eines Interventionsprogramms. Dabei steht nicht selten die Thematik Qualitätssicherung und Qualitäts-entwicklung im Zentrum des Kontroll-Interesses, insbesondere im Falle zielgruppenbezogener Programme, wie z.B. fortlaufend zu erbringende Human-dienstleistungen durch eine Organisation oder Institution. Zwar gilt letztlich der positive Effekt bei den Adressaten der Dienstleistung (outcome) als Kriterium für den Erfolg der Dienstleistung. Doch ist zugleich die Annahme weit verbreitet, dass gute Servicequalität eine weitgehende Gewähr für solchen Erfolg sei. 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.

Damit 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; denn „Qualität" ist keine Eigenschaft eines Sachverhalts (z.B. einer Dienstleistung), sondern ein Konstrukt, das von außen an den Sachverhalt zum Zwecke der Beurteilung herangetragen wird. Wenn nun die positiven Effekte bei den Adressaten einer Dienstleistung

das letztliche Kriterium der Qualitätsbeurteilung sein sollen, die Qualität der Dienstleistung jedoch aus unterschiedlichsten Gründen nicht an den Effekten auf die Adressaten abgelesen werden kann, dann steht die Evaluation vor einer schwer zu lösenden Aufgabe: Dann müssen entweder die Adressaten die Rolle der Evaluatoren übernehmen, oder es müssen „objektive" Qualitätsmerkmale der Dienstleistung und des Prozesses der Dienstleistungserbringung ermittelt werden, die auch „subjektive Bedeutung" haben, die also die Wahrscheinlichkeit positiver Effekte bei den Adressaten begründen können.

Das wohl bekannteste Beispiel für einen solchen Ansatz ist das von *Donabedian* für das Gesundheitswesen entworfene Qualitätskonzept (ausführlich in Donabedian 1980), das von dort ausgehend auch für andere soziale Dienstleistungsbereiche übernommen wurde. Er unterscheidet zwischen Strukturqualität (personelle, finanzielle und materielle Ressourcen, physische und organisatorische Rahmenbedingungen, physische und soziale Umwelt), Prozessqualität (Erbringung der Dienstleistung, Interaktionsbeziehung zwischen Anbieter und Klienten) und Ergebnisqualität (Zustandsveränderung der Klienten im Hinblick auf den Zweck der Dienstleistung, Zufriedenheit der Klienten).

Die sachliche Angemessenheit eines solchen dimensionalen Schemas unterstellt, sind dann 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. Die Güte der Evaluation mit Hilfe von Qualitätsindikatoren steht und fällt mit der Angemessenheit der gewählten Indikatoren. Wenn – wie dargelegt – der positive Effekt bei den Adressaten der Dienstleistung (outcome) als 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.

Angesichts der Schwierigkeit, die methodologischen Ansprüche des Indikatorenansatzes einzulösen, wird nicht selten – wie schon für Programmevaluation angesprochen (Abschnitt 2.2.1) – auf eine methodisch einfachere Alternative zurückgegriffen. An die Stelle methodisch kontrollierter Evaluation durch Forschung wird die Bewertung durch Betroffene und/oder die Ermittlung ihrer Zufriedenheit gesetzt. 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 sachverständig und 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 Resultat: Man erhält verlässliche Qualitätsaussagen, ohne zuvor ein theoretisch und empirisch abgesichertes Indikatorenmodell entwickeln zu müssen. Leider erweisen sich solche Vorstellungen häufig als empirisch falsch. Die per Befragung ermittelte „Akzeptanz" (oder Nicht-Akzeptanz) hängt nicht in erster Linie von der Qualität des zu beurteilenden Sachverhalts ab, sondern im Wesentlichen von Merkmalen der Befragten.

Das bedeutet nicht, dass in Nutzerbefragungen erhobene Akzeptanzaussagen wertlos seien! Ganz im Gegenteil liefern sie *für Zwecke der Evaluation* relevante Information, insbesondere 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). Sie sind aber nicht schon „*die Evaluation*".

2.4 Das Leitkonzept für das Entwicklungsparadigma der Evaluation: Das „Helfer- und Beratermodell"

Das Konzept von Evaluation als Programmforschung muss – wie dargestellt –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 einem zur Lösung anstehenden „Problem" – 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 Modell außerordentlich unhandlich, in manchen Konstellationen auch überhaupt nicht realisierbar. Empirische Informationen und sozialwissenschaftliches Know-how werden bereits bei der Entwicklung und Optimierung eines Programms sowie bei der Erkundung der Möglichkeiten seiner „Umsetzung" verlangt. Daraus ergeben sich für die Funktion der Evaluation drei grundlegende Unterschiede:

- Zum einen steht am Anfang nicht ein „fertiges" Programm, dessen Implementierung und Wirksamkeit zu überprüfen ist. Vielmehr ist Evaluation – wie in Abschnitt 2.1.3 skizziert – in die gesamte Programm-Historie eingebunden: beginnend mit der Aufarbeitung von Problemwahrnehmungen und Zielvorstellungen über die Entwicklung von Programmideen bis hin zu einem endgültigen Konzept.
- 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 gilt die Programmumwelt nicht lediglich als ein Bündel von „Störfaktoren", die es statistisch zu kontrollieren gilt. Vielmehr ist die Umwelt – neben den eigentlichen Programmzielen – eine explizite 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 dieser Typ von Evaluation 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 (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). Der Evaluation dieses Typs – also begleitende Beratung – wird man am ehesten mit der Bezeichnung „Helfer- und Beratermodell" gerecht. Man erliegt jedoch einem Irrtum, wenn man beratende Evaluation als die „weichere" oder gar anspruchslosere Variante im Vergleich zu Konzepten der Programm- und Implementationsforschung ansieht. Evaluatoren in der Funktion von Moderatoren und Beratern benötigen selbstverständlich alle im sozialwissenschaftlichen Studium üblicherweise vermittelten Kenntnisse und Fähigkeiten (insbesondere der kompletten empirischen Forschung: quantitative und qualitative Erhebungs- und Analysemethoden). Darüber hinaus jedoch sind zusätzliche Qualifikationen gefragt, die nicht einfach „gelernt" werden können, sondern die 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.

3 Evaluation – Wie macht man das?

Bisher wurde das Thema „Evaluation" unter begrifflich-theoretischen sowie unter methodologischen Gesichtspunkten dargestellt. Bei aller Vielfalt von Konzepten und Gegenständen standen im Hintergrund aller Überlegungen vier Aspekte, die bisher relativ abstrakt blieben: Evaluierung bezieht sich auf einen explizit benennbaren „*Gegenstand*"; sie wird – will sie intersubjektive Geltung beanspruchen - vorgenommen von einer *zur Bewertung legitimierten Instanz* anhand von gegenstands-angemessenen *Kriterien* und unter Einsatz geeigneter *Verfahren*.

Wer vor der Aufgabe steht, ein Evaluationsprojekt selbst durchzuführen, dem ist mit abstrakten Angaben jedoch noch nicht geholfen. Voraussetzung für die erfolgreiche Durchführung ist professionelles Projektmanagement, sind explizite

Präzisierungen, Rollendefinitionen und Kompetenzklärungen. Präzisierungen zu jedem der 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 nicht ad hoc, sondern schon im Vorfeld im Detail, verbindlich, nachvollziehbar und gut dokumentiert vorzunehmen. Die folgende Tabelle gibt einen Überblick über den Präzisierungs- und Klärungsbedarf:

Tabelle: Evaluation – Begriffsdimensionen und Klärungsbedarf

Dimensionen der Evaluation	mögliche Präzisierungen	Klärungsbedarf
Was wird evaluiert? Programme, Maßnahmen, Organisation etc.	existierende Einrichtung; in Planung/Entwicklung befindliches Projekt; bereits implementiert; Feldversuch/Pilotprojekt; Programmumfeld etc.	Was ist das „Programm“? Wie ist seine Struktur? Was ist der „eigentliche Gegenstand“ der Evaluierung? Was ist der Evaluationszweck? Wer sind die Nutzer?
Wer evaluiert? Personen, die zur Bewertung legitimiert bzw. 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
Wie wird evaluiert? objektiviertes, intersubjektiv nachprüfbares 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 welchen Kriterien? explizit auf den Sachverhalt bezogene und begründete Kriterien (ggf. konkretisiert durch Standards)	Zielerreichung / Effekte / Nebenwirkungen, Effizienz / Effektivität; Sozialverträglichkeit, Zielgruppenbezug etc.	Ziele (des Programms, der Evaluation), Kriterien (des Programms, der stakeholder), 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 ein Pilotvorhaben sein, das in einem abgegrenzten 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) angestrebt würde, wäre doch noch (stark selektiv) zu entscheiden, welche Teilaspekte denn tatsächlich im Detail einer systematischen Beurteilung unterzogen werden sollen. 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 auf der einen und den Merkmalen und Zielen des Evaluations-Vorhabens auf der anderen Seite. Soll das Evaluations-Vorhaben „nützlich" sein, d. h. sollen die Resultate 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 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 (1997, 234 ff.).

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). Im ungünstigsten Fall kann es auch mit dem vollständigen Scheitern des Vorhabens enden. Die Bedeutung „vertrauensbildender Maßnahmen" im Vorfeld kann gar nicht hoch genug eingeschätzt werden. Patentrezepte 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 Programm-durchführenden oder für eine Kontrollinstanz, als unmittelbar im Programm Mitwirkende oder als hinzugezogene externe Berater, als wenig engagierte Betroffene oder als organisierte Befürworter oder Gegner.

Von zentraler Bedeutung ist die klare Unterscheidung von mindestens drei Funktionen im „Projekt Evaluation": Informationsmanagement (Informations-beschaffung und –analyse), Evaluierung, Ableitung von Konsequenzen aus den Befunden. Es ist nicht selbstverständlich (sondern allenfalls der seltene Ausnahmefall), dass alle Projektfunktionen „in einer Hand" liegen, etwa im Falle eines zur Fremdevaluation eingesetzten externen Gremiums. Vielmehr ist zwischen den Beteiligten ausdrücklich auszuhandeln und verbindlich festzulegen, wer welche *Aufgaben* übernimmt und wem welche *Kompetenzen* zugebilligt werden. Für Evaluationen im Rahmen von Organisations-entwicklungs-Vorhaben empfiehlt sich eine Dreiteilung 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 Maßnahmen-pläne/Zielvereinbarungen aus. Natürlich sind auch andere Kombinationen von Aufgaben und Zuständigkeiten möglich und – je nach faktischen Gegebenheiten – Erfolg versprechend. 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 Tabelle) den Abschluss der (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-Kriterien und -Maßstäben fällen, sondern ihre Aussagen sollen nachvollziehbar, überprüfbar und kritisierbar sein. Sofern die Klärungen zu den beiden erstgenannten Bereichen (Gegenstand und Evaluationsinstanz) hinreichend eindeutig getroffen wurden, dürften an diesem Punkt größere Probleme nicht mehr auftreten. Denkbar ist allerdings wiederum ein ganzes Spektrum sehr unterschiedlicher Bewertungskriterien und -standards. Diese 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.).

Erst nachdem alle diese „evaluationspolitischen" Vorklärungen getroffen und dokumentiert sind, bekommt der Aspekt der „*Methoden*" Bedeutung (dem in

abstrakten Diskussionen über „Evaluation“ viel zu oft ein unangemessen großer Stellenwert zugeschrieben wird). Hier existieren die geringsten Konfliktpotentiale, liefert doch 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 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.

Dennoch: Diese Fragen sind vergleichsweise „objektiv“ zu beantworten und unter dem Gesichtspunkt der Nützlichkeit für die Erreichung der Ziele des Evaluationsprojekts entscheidbar. Unter diesem zuletzt genannten Gesichtspunkt ist übrigens dringend zu empfehlen, auf komplexe statistische Modelle der Datenanalyse zugunsten der unmittelbaren Nachvollziehbarkeit durch alle Beteiligten zu verzichten.

4 Evaluation und Qualitätsentwicklung in der Hochschule

4.1 Die Hochschule - ein besonders schwieriges Feld für die Evaluation

In den Abschniten 2.1.1 bis 2.1.3 wurden drei „conceptual frameworks“ für Evaluationen skizziert: das „Forschungs-“, das „Kontroll-“ und das „Entwicklungsparadigma“. Ein Evaluationsprojekt ist umso einfacher zu konzipieren, die Ziele, Bewertungskriterien und die Kompetenzen sind umso problemloser auszuhandeln, je eindeutiger ein Vorhaben einem dieser drei Paradigmen zuzuordnen ist.

Im Kontext der Qualitätsdiskussion im Bereich Hochschule schwingen jedoch – wenn Evaluationen eingefordert werden – fast immer die Erkenntnisinteressen aus allen drei „frameworks“ gleichzeitig mit. Natürlich möchte man neues empirisch abgesichertes Wissen darüber gewinnen, wovon erfolgreiches Lehren und Studieren abhängt und wie der Erfolg gefördert werden kann – insofern ist das Forschungsparadigma gefragt. Natürlich sollen zugleich Effektivität und Effizienz der Verwendung der in den Hochschulbereich fließenden öffentlichen Mittel kontrolliert werden, sollen die Hochschulen Rechenschaft über ihr Tun

ablegen – also ist auch das Kontrollparadigma angesprochen. Und ebenso natürlich soll Evaluation helfen, geeignete Maßnahmen zur Verbesserung der Qualität von Lehre und Studium zu konzipieren, zu implementieren und zu testen – womit schließlich das Entwicklungsparadigma zu seinem Recht kommt.

Trotz solcher unrealistisch hoher Erwartungshaltung wird dann aber nicht selten zugleich in aller Naivität gefordert, Evaluation müsse sich schnell, einfach, mit geringem Kosten- und Arbeitsaufwand realisieren lassen (denn Personalressourcen und Geld sind in den Hochschulen bekanntermaßen außerordentlich knapp). Außerdem darf die Evaluation den laufenden Betrieb nicht „stören" – schließlich ist das eigentliche Ziel der Hochschule die Sicherstellung eines geregelten Angebots für ein ordnungsgemäßes Studium und nicht dessen Evaluation.

Übersehen wird bei solchen Rufen nach simplen und belastungsfreien Verfahren die außerordentliche Komplexität potentieller Evaluationsgegenstände im System Hochschule: Es existiert weder ein präzise beschreibbares „Programm" mit klar definierten Zielen und ihnen zugeordneten Maßnahmen sowie eindeutig festgelegten Zielerreichungskriterien noch ein konkretes „Produkt", dessen Qualität mit einem Satz von Qualitätsindikatoren durch standardisierte Messverfahren abgebildet werden kann. Was ist also eigentlich der „Gegenstand" der Evaluation?

Zudem muss die Evaluation unmittelbar im aktiven Feld durchgeführt werden und kann – anders als etwa bei politischen Pilotprojekten – nicht einen Teilbereich abgrenzen und (zumindest teilweise) von der Umwelt isolieren. Nicht zuletzt ist sie dabei mit zahlreichen Akteuren mit je unterschiedlichen Zielen und Vorstellungen konfrontiert, deren Handeln sämtlich über Erfolg und Misserfolg des zu evaluierenden Programms wie auch der Evaluation selbst mitentscheidet. Auf welche Weise kann man diese „stakeholder" im Evaluationsprojekt berücksichtigen oder gar beteiligen?

Will Evaluation im System Hochschule dieser Komplexität gerecht werden, ist sie extrem zeit- und ressourcenaufwändig. Soll sie nicht lediglich Selbstzweck sein, sondern Veränderungen (Qualitätsverbesserungen) in Gang setzen, ist sie trotz des mit ihr verbundenen Aufwands bei allen Beteiligten auf aktive Akzeptanz, auf Mitwirkungsbereitschaft angewiesen. Damit diese erwartet werden darf, muss sich der Aufwand einer Mitwirkung lohnen: Die Evaluation muss für die Beteiligten einen erkennbaren Nutzen bringen. Akzeptanz ist darüber hinaus aber auch eine wesentliche Voraussetzung dafür, dass Evaluation überhaupt verwertbare Ergebnisse liefern kann. Es ist daher vorab zu klären und für alle Beteiligten erkennbar zu machen, zu welchem Zweck evaluiert werden soll, was mit den zu erhebenden und auszuwertenden Daten geschehen soll (s.o., Kap. 3). Evaluation darf nicht als „Evaluations-Ritual" erscheinen.

Eine geringe Akzeptanz ist auch dann zu erwarten, wenn Evaluation lediglich als Kontrollinstrument verwendet werden soll, um – seien es Lehrpersonen oder ganze Fächer – die „Guten" von den „Schlechten" zu sondern und daran Sanktionen zu knüpfen. Es wird oft unterschätzt, welches Spektrum an Möglichkeiten Evaluierte haben, kritische Informationen zu verschleiern und positive Informationen überdimensioniert in den Vordergrund zu rücken.

Weitgehend etabliert hat sich Evaluation als hochschul*internes* Steuerungsinstrument, zum Teil verknüpft mit „incentives" z.B. für gute Lehrorganisation und Forschungsleistungen. In manchen Bundesländern Deutschlands wird ein Teil der universitären Sachmittel „nach Leistungs- und Belastungskriterien" vergeben. Zu diesem Zweck ist – soll dies in der Universität routinemäßig und flächendeckend geschehen – ein Raster von möglichst wenigen Indikatoren erforderlich, die regelmäßig verfügbar sind und möglichst objektiviert Leistungen und Belastungen eines Fachs oder auch von kleineren Einheiten abbilden. Methodisch sind solche Verfahren nicht unstrittig: Indikatoren können immer nur einen Ausschnitt aus dem gesamten Problemfeld abbilden und auch dies immer nur mit zweifelhafter Gültigkeit (s.o., Abschn. 2.3.2). Des Weiteren besteht die Möglichkeit (und damit die Gefahr), lediglich die durch Indikatoren abgebildeten Bereiche zu „optimieren" und anderes zu vernachlässigen; ganz abgesehen von der Möglichkeit der Umdefinition von Kriterien, um „bessere" Ergebnisse zu erzielen (mehr Studienabschlüsse in kürzerer Zeit kann man auch dadurch erreichen, dass man das Anspruchsniveau senkt). Die Erfahrungen der Sozialindikator-Bewegung haben gezeigt: Indikatorensysteme liefern nur so lange gültige Resultate, wie sie lediglich zu Deskriptions- und Erklärungszwecken (allenfalls auch noch als prognostisches Frühwarnsystem) genutzt werden. Sobald an die Indikatorenwerte Sanktionen für diejenigen geknüpft werden, die die Ausprägungen durch ihr Handeln beeinflussen können, verlieren sie ihre neutrale Informationsfunktion.

Einen etwas anderen Zungenschlag erfährt die Diskussion um Evaluation als Steuerungsinstrument im Kontext der Forderung nach stärkerer Wettbewerbsorientierung der Hochschulen: „Auch ein Hochschulsystem, das staatlich globalgesteuert, aber zunehmend von Wettbewerb und Profilbildung gekennzeichnet ist, muss sich Marktgesetzlichkeiten stärker öffnen. Auch wenn sie nicht auf Gewinnerzielung hin orientiert sind, müssen Hochschulen sich in mancher Hinsicht wie Unternehmen verhalten lernen. Das heißt unter anderem, bei der Planung und Ausgestaltung von Lehrangeboten rascher auf Nachfrageänderungen zu reagieren und auch Studierende als ‚Kunden' ernster zu nehmen" (Landfried 1999, 10). Evaluation schafft in diesem Zusammenhang „ein Stück Markt-Ersatz, eine Art Quasi-Wettbewerb" (ders., 11). Mit wem die Hochschulen über das Medium Evaluation in welcher Form um welche knappen Ressourcen konkurrieren, bleibt allerdings ebenso unbeantwortet wie die Frage, auf wessen Nachfrageänderungen – und dann in welcher Weise – rascher zu reagieren sei. Auch die „Kundenrolle" von Studierenden bleibt diffus.

An der Schnittstelle von Kontroll- und Wettbewerbsargumentation schließlich finden wir die Vorstellung von Evaluation als Instrument globaler Qualitäts-„Messung". Wenn es gelänge, die Qualität der Leistungen der Institution Hochschule und ihrer Gliederungen umfassend, detailliert, gültig und zuverlässig zu messen, dann stünde damit einerseits ein „objektives" Kontrollinstrument zur Verfügung; andererseits existierte in Gestalt der Qualitätsmaße auch eine Art „Währung", die für einen funktionierenden Wettbewerb (etwa um Reputation, aber auch um öffentliche Finanzmittel, um Forschungsförderung, sogar um besonders leistungswillige Studierende) notwendig scheint. Die wiederholt unternommenen Versuche, Rankings von Hochschulen und Hochschulfächern bis hin zu Lehrveranstaltungen zu erstellen,

sind u.a. auch als Bemühung zu verstehen, Transparenz auf einem solchen Wettbewerbsmarkt zu schaffen.

Doch selbst wenn es gelänge, die methodischen Probleme von Qualitätsmessung durch Indikatoren zu lösen, wäre für das Ziel Qualitätsentwicklung und/oder Qualitätssicherung allein mit dem Bereitstellen solcher Informationen noch nicht viel gewonnen. Informationen sind allenfalls die notwendige (aber noch nicht hinreichende) Voraussetzung dafür, gezielte Veränderungen dort in Gang zu setzen, wo der evaluierte Sachverhalt verbesserungsbedürftig und verbesserungsfähig erscheint. Damit stoßen wir auf zwei Fragen, deren Zusammenhang häufig nicht beachtet wird. Wer ist Träger der Evaluation? Und wer ist Träger des Qualitätsentwicklungs-Vorhabens? (s.o., Kap. 3). Anders formuliert: Wer ist verantwortlich dafür, dass die gelieferten Evaluations-Informationen in Handeln umgesetzt werden? Nur in Ausnahmefällen wird dies auch der Träger des Evaluationsvorhabens sein (= „interne Evaluation"). Wo Qualitätsentwicklungs-Akteure und Evaluationsinstanz sich unterscheiden (= „externe Evaluation"), ist ein auf gegenseitigem Vertrauen basierendes Verhältnis beider Instanzen die Voraussetzung sowohl für gültige Evaluationsresultate (keine Unterdrückung „problematischer" Informationen, zuverlässige „Schwachstellen"-Analyse) wie für gelingende Umsetzung der Resultate in Maßnahmen zur Qualitätsentwicklung (Formulierung konstruktiver und realisierbarer Empfehlungen durch die Evaluation, Zusicherung der Vertraulichkeit erlangter interner Kenntnisse, Veröffentlichung nur im gegenseitigen Einvernehmen).

4.2 Teilnehmerbefragungen – der problematische „Königsweg" für die Lehrevaluation

Im vorigen Abschnitt wurden einige der Schwierigkeiten von Evaluationen im „besonders schwierigen" Feld Hochschule angesprochen: Bei dem zu bewertenden Leistungsspektrum haben wir es nicht mit einem konkret fassbaren „Gegenstand" zu tun – anders als im Falle der Güterproduktion, wo sich Effizienz und Effektivität des Produktionsprozesses sowie die Qualität des Produkts (output) relativ leicht beurteilen und in standardisierter Form messen lassen. Stattdessen geht es um die Bereitstellung von Dienstleistungen – bzw. noch eingeschränkter: von Humandienstleistungen –, die die aktive Mitwirkung der Klienten (hier: der Studierenden) voraussetzen, sollen sie einen „Erfolg" bewirken, also „Qualität" aufweisen, effizient und effektiv sein. Was ist in diesem Fall das „Produkt"? Was ist der „Produktionsprozess"? Ist es das Vorhalten einer Dienstleistungs-Infrastruktur (geregeltes Lehrangebot und die Informationen darüber in kommentierten Vorlesungsverzeichnissen, Personal für Beratungen und Prüfungen, PC-Räume und Bibliotheken) oder die einzelne Dienstleistung selbst (die Lehrveranstaltung, Prüfung, das Beratungsgespräch)? Oder interessiert eher, was durch die vorgehaltene und realisierte Dienstleistung bewirkt wird (outcome anstelle von output)? Schließlich: Wenn es sich – wie im Falle von Lehre und Studium – um eine Dienstleistung handelt, die auf die Akzeptanz und das aktive Mitwirken der Adressaten angewiesen ist: Wer oder was ist dann eigentlich zu evaluieren – der Anbieter, der Nachfrager oder beide? Und nicht zuletzt: Wer evaluiert wen?

Vergegenwärtigen wir uns zur Illustration die evaluationsrelevanten Fragen aus der Sicht des Konzepts der Programmforschung. Das zu bewertende Programm könnte z.B. der Studiengang eines Fachs sein. Als Ziele kämen die an die Studierenden zu vermittelnden Qualifikationen, als Maßnahmen Studienordnung, Studienverlaufspläne, Lehrveranstaltungen, Studieninfrastruktur sowie Betreuung und Beratung durch das Lehrpersonal in Betracht, außerdem Prüfungsordnung, Prüfungen und andere Leistungskontrollen. Für die Messung der Zielerreichung böte sich der Zeitpunkt der Beendigung des Studiums (Examen oder Studienabbruch) bei den einzelnen Studierenden an.

Mit einem ersten Bündel von Schwierigkeiten wären wir bereits bei der empirischen Beschreibung der für die Studierenden bis zum Examen relevant gewordenen Maßnahmen konfrontiert. Studienordnung, Studienverlaufspläne und Prüfungsordnung wären für alle im Verlaufe ihres Studiums konstant und somit (im Hinblick auf Unterschiede in den erworbenen Qualifikationen) ohne Wirkung. Lehrveranstaltungen dagegen – mit Ausnahme einiger standardisierter Vorlesungen und Übungen insbesondere im Grundstudium – sind schon in ihren Inhalten häufig so stark variierend, dass hier zusätzlich zu den im Prinzip statistisch kontrollierbaren Unterschieden der studentischen Veranstaltungsauswahl (feststellbar etwa durch Auswertung der Studienbücher) eine zusätzliche Variation in nicht kontrollierbarem Ausmaß hinzukäme. Beratung, Betreuung und Prüfungen schließlich ergeben sich in Interaktionen zwischen einzelnen Studierenden und einzelnen Mitgliedern des Lehrpersonals und wären bei Studienabschluss überhaupt nicht mehr rekonstruierbar.

Als ähnlich problematisch erwiese sich die Erfolgsmessung. Die im Studium zu vermittelnden Qualifikationen sind üblicherweise in den Studiengangsdokumenten (Studien- und Prüfungsordnung) nur sehr vage definiert. Ersatzweise kämen die in Klausuren und Prüfungen erbrachten Leistungen der Absolventen (gemessen in den erzielten Noten) in Betracht. Diese wären allerdings keine direkten Maße der Qualifikationen, sondern lediglich Indikatoren für eine Teilmenge von ihnen. Erfolge/Misserfolge auf anderen Dimensionen blieben unerkannt. Außerdem wäre die Gültigkeit dieser Indikatoren fraglich, wenn die Träger des zu evaluierenden Programms die Indikatorausprägungen selbst festlegten (nämlich in Prüfungen und Klausurbenotungen).

Als ganz unmöglich schließlich erwiese sich die Zurechnung der Beiträge einzelner Maßnahmen zum festgestellten Studienerfolg der jeweiligen Absolventen. In welcher Weise das Studium verläuft sowie ob und in welchem Ausmaß es erfolgreich abgeschlossen wird, hängt nach allen vorliegenden empirischen Erkenntnissen aus der Bildungsforschung in hohem Maße von Merkmalen in der Individualsphäre der Studierenden ab: wie Lebenssituation, Interesse und Leistungsmotivation, Studienstil und -intensität. Die von den Trägern des Studiengangs beeinflussbaren Gegebenheiten – Studieninfrastruktur, Lehre und Betreuung – können lediglich (wenn sie von schlechter Qualität sind) das Studium erschweren oder (bei guter Qualität) erleichtern. Den individuellen Erfolg bewirken sie nicht. Um also den relativen (fördernden oder hemmenden) Beitrag der angebotenen Maßnahmen zum Studienerfolg abschätzen zu können, müsste zunächst der individuelle Eigenbeitrag des jeweiligen Studierenden

bekannt sein – eine, wie leicht einsehbar, völlig unrealistische Anforderung, deren Nichterfüllbarkeit in diesem Bereich jede Evaluation im Sinne von Zielerreichungskontrolle prinzipiell unmöglich macht.

Selbst wenn wir „lediglich" einen Teilaspekt des universitären Alltags – die Qualität der Lehre – durch empirische Daten evaluieren wollen, sehen wir uns einem prinzipiellen Dilemma gegenüber:

Soll die Bewertung für ein ganzes Fach geschehen, dann ist dies in den einzelnen Lehrveranstaltungen nur in abstrahierender Weise, nämlich losgelöst von jedem Inhalt der Lehre möglich. Lehr- und Ausbildungsziele ohne Bezug zu den Inhalten aber existieren nicht; der Evaluation eines Programms (hier: Lehrangebot des Fachs) ohne jede Zielorientierung aber fehlen die Evaluationskriterien. Übrig bleiben dann lediglich Oberflächenmerkmale wie Teilnehmerzahl, Größe der Lehrräume, Medieneinsatz, Bereitstellung von Skripten, didaktische Aufbereitung des Stoffs u.ä. Ohne Berücksichtigung der Inhalte wird Lehre nicht daraufhin bewertet, wie gut sie *ist*, sondern wie gut sie aussieht.

Betrachten wir dagegen die einzelne Lehrveranstaltung als „Programm", so könnten zwar im Prinzip alle relevanten Merkmale wie Ziele, Instrumente, Effekte konkretisiert und ermittelt werden. Aber die parallele Evaluation einer Vielzahl solcher „Programme" (z.B. alle Lehrveranstaltungen eines Fachs in einem Semester) wäre nur mit einem unvertretbar hohen Aufwand zu realisieren. Außerdem müsste sich – will man nicht auf der Ebene isolierter Evaluationen einzelner Lehrveranstaltungen verbleiben, sondern eine Bewertung „der Lehre" im betreffenden Fach vornehmen – eine Meta-Analyse in Form einer Clusterevaluation anschließen.

Angesichts dieser unlösbaren Probleme einer Anwendung des Konzepts der Programmevaluation sowie angesichts der prinzipiellen Strittigkeit jedes Indikatorenmodells als Evaluationsbasis überrascht es nicht, wenn auf das Instrument „Betroffenenbefragung" als Rezept zurückgegriffen wird, um den geschilderten „gordischen Knoten" zu durchschlagen. Befragungen erscheinen erheblich weniger anspruchsvoll – sowohl hinsichtlich des Aufwands der Durchführung als auch hinsichtlich der Strategie der Objektivierung.

Für diese Wahl scheinen auch gute Argumente zu sprechen (s.o., Abschnitt 2.3.2), die sich für Lehrevaluationen in etwa so zusammenfassen lassen: „Ein aufwändiges Verfahren der Qualitätsbeurteilung durch Evaluationsforschung ist entbehrlich. Mit den Studierenden verfügt die Hochschule bereits über die Experten, die die Lehre aus erster Hand – als tagtäglich von ihr Betroffene – fundiert und zuverlässig beurteilen können. Deren Wahrnehmungen und Bewertungen brauchen nur in standardisierter Form erhoben und pro Lehrveranstaltung in geeigneter Form ausgewertet zu werden, um aussagekräftige Qualitätsindikatoren zu erhalten."

Manche gehen noch einen Schritt weiter und vertreten unter Verweis auf „jahrzehntelang bewährte Praxis in den USA" die Auffassung, hierzu werde nicht einmal ein detailliertes Instrumentarium benötigt. Vielmehr reichten kurze und damit schnell ausfüllbare Fragebögen aus, in denen von den Studierenden auf wenigen zentralen Dimensionen (typischerweise Didaktik, Angemessenheit von

Stoffmenge und Schwierigkeitsgrad, Auftreten der Lehrperson und soziales Klima, Lernerfolgseinschätzung) zusammenfassende Bewertungen erbeten werden. Studierende seien durchaus kompetent, solche Urteile zu fällen, wird – vermeintlich studentenfreundlich – argumentiert. Damit erübrigten sich zugleich auch komplexe Auswertungsverfahren; Auszählungen und Durchschnittsberechnungen seien hinreichend.

Leider treffen diese optimistischen Annahmen weitgehend nicht zu. In dieser Form eingesetzt, vielmehr sind mit einer solchen Einfachstrategie der Erhebung studentischer Wahrnehmungen und Bewertungen *als Evaluation* von Studium und Lehre allerdings fast zwangsläufig Fehlschlüsse verbunden.[5]

Im Unterschied zur Evaluation durch Experten anhand vorgegebener Kriterien und auf der Basis systematisch ausgewerteter Informationen sind die befragten Studierenden „Alltags-Evaluatoren", das heißt: Jeder einzelne von ihnen bewertet *irgend etwas* (was er mit dem in der Frage angesprochenen Sachverhalt ad hoc assoziiert) *irgendwie* („alles in allem" oder „aus aktueller Erfahrung" oder „mit Blick auf das Wesentliche" oder ...) *unter irgendwelchen Gesichtspunkten* (Nutzen für sein Studium oder vermuteter Nutzen für den angestrebten Beruf oder aktuelles persönliches Interesse oder abstrakt-verallgemeinertes Interesse der Studierenden oder ...).

Das heißt nicht, dass Studierende die Qualität der Dienstleistung Lehjre nicht aus ihrer Sicht zutreffend beurteilen könnten, sondern dass sie je nach individueller Studiensituation unterschiedliche (subjektiv rationale) Kriterien anwenden. Sofern deren individuelle Studiensituation und die angewendeten Kriterien nicht mit erhoben werden (was in „handhabbaren" Fragebögen nicht möglich ist), ist aber die Bedeutung gegebenen Antworten nicht mehr rekonstruierbar.

In gleicher Weise problematisch ist die Empfehlung „einfacher Auswertungen", insbesondere in Form isolierter Auszählungen der Antworten auf die einzelnen Fragen und/oder durch Berechnung von Mittelwerten. Auch hier zeigen komplexe Analysen differenziert erhobener studentischer Bewertungen die Unangemessenheit solchen Vorgehens: Zum einen werden von den Befragten die Einschätzungen hinsichtlich der verschiedenen Dimensionen und Teildimensionen des Evaluationsgegenstands (z.B. Lehrveranstaltung oder Lehrperson) nicht unabhängig voneinander vorgenommen, sondern sie stehen – selbstverständlich – in einem subjektiv sinnvollen Zusammenhang. Daraus folgt, dass sich die Einzelurteile jedes Befragten zu einem für seine Wahrnehmung typischen Urteilsprofil verbinden und dadurch sozusagen „Gestalt annehmen" (im Detail: Kromrey 1994). Die isolierte Auszählung einzelner Variablen lässt solche Profile gar nicht erst sichtbar werden. Zum anderen sind sich die Teilnehmer ein und derselben zu evaluierenden Veranstaltung – eigentlich ebenfalls selbstverständlich – in ihren Beurteilungen nicht einig. Das liegt nicht nur daran, dass ihnen für ihre „Alltags-Evaluationen" keine intersubjektiven Vergleichsstandards vorgegeben wurden, sondern es dokumentiert, dass es sich

---

[5] Im vorliegenden Text kann dies nur kurz angerissen werden; ausführlich dazu Kromrey 1999.

bei den Befragten nicht um austauschbare Exemplare der Gattung Studierende handelt, sondern um Individuen: mit unterschiedlichen Sozialisationserfahrungen und von daher unterschiedlichen Vorkenntnissen, Interessen und Lernstilen, mit unterschiedlichen Präferenzen und Sympathien/Antipathien für die Lehrperson, mit unterschiedlichen Standorten in ihrem Studiengang, mit unterschiedlicher Einschätzung der Brauchbarkeit ihres Studiums und des zu Lernenden für das Leben außerhalb der Hochschule usw. Das heißt: Die Gesichtspunkte, unter denen beurteilt wird, sind sehr verschiedenartig; sie müssen demgemäß – wenn der Fragebogen ernsthaft und kompetent ausgefüllt wird – zu unterschiedlichen Urteilen führen. Die Berechnung von Mittelwerten, die die studentischen Individualurteile zu Qualitätskennziffern der Teilnehmer kondensieren, unterdrückt den eigentlichen Informationsgehalt der erhobenen Daten und produziert Auswertungsartefakte.

Fazit: Ein komplexer Sachverhalt kann angemessen auch nur durch hinreichend komplexe empirische Erhebungen valide abgebildet werden; und komplexe Interdependenzen im abzubildenden Sachverhalt werden erst durch hinreichend komplexe Analyseverfahren sichtbar. Um es ausdrücklich hervorzuheben: Der Hinweis auf die o.g. Gefahren von Fehlschlüssen bei unangemessenem Einsatz des Instruments Teilnehmerbefragung sollte nicht als Argument gegen die Verwendung von „Alltagsevaluationen" Betroffener missverstanden werden. Um diese jedoch als gültige Informationen nutzen zu können, muss im Zuge der Analyse das Kriteriensystem der Evaluierenden rekonstruiert werden. So erst werden Akzeptanzaussagen zu wichtigen Informationen. Bei Einhaltung der einschlägigen methodologischen Qualitätsstandards (aber nur dann) ist die Befragung von studentischen Veranstaltungsteilnehmern ein wertvolles Informationsinstrument zur Entwicklung von Lehrqualität.

Zu beobachten ist jedoch leider häufig ein falscher und unprofessioneller Einsatz der Methode. Entgegen einem gängigen Vorurteil („Fragen stellen kann jeder") ist nämlich standardisiertes Befragen ein sehr problematisches Instrument der Informationsgewinnung. Wie in jedem Methodenlehrbuch nachzulesen ist, gelten sowohl hinsichtlich des Erhebungsgegenstandes wie auch der Forschungskontaktsituation schon für jede „herkömmliche" standardisierte Befragung sehr anspruchsvolle Anforderungen. Noch schwieriger wird es, wenn Befragung als Instrument des „Messens" eingesetzt werden soll. In diesem Fall müssen an formalen Voraussetzungen erfüllt sein:

- der „Gegenstand" (das Objekt) der Beurteilung ist eindeutig definiert;
- das zu messende „Merkmal" (Variable) ist eindeutig definiert und operationalisiert;
- eine „Mess-Skala" (Vergleichsmaßstab) existiert und ist eindeutig definiert (incl. eindeutiger Skalen-Endpunkte sowie unterscheidbarer Abstufungen zwischen ihnen);
- die Befragten sind in der Lage, den „Gegenstand" intersubjektiv übereinstimmend zu identifizieren, das zu messende „Merkmal" intersubjektiv übereinstimmend zu erkennen (identische Semantik + identische Perspektive) und die „Mess-Skala" in intersubjektiv übereinstimmender Weise darauf anzuwenden (= Messung durch

Vergleich der Merkmalsausprägung eines Objekts mit einem Vergleichsmaßstab, Übersetzung dieses Vergleichs in Skalenpunkte).

Wollen wir etwas so Abstraktes wie *Qualität* messen (also „evaluieren"), kommen noch weitere Probleme hinzu: Verschiedene Befragte urteilen nicht nur auf den vorgegebenen Kriterien unterschiedlich (d.h. es gibt individuelle Variation; die ließe sich durch geeignete statistische Auswertung „ausmitteln", sofern es nicht zu gruppenspezifisch „systematischer" Variation kommt), sondern sie urteilen mit Bezugnahme auf unterschiedliche Ziele, Bedürfnisse und Voraussetzungen.

Bezogen auf die Lehre als unterstützender Service für Lernende heißt dies: Verschiedene Studierende benötigen unterschiedliche Arten von Unterstützung durch das Lehrangebot und als Konsequenz daraus – wenn die Lehrqualität *für sie* „gut" sein soll – ein auf ihre Bedürfnisse zugeschnittenes unterschiedliches Lehrangebot. Wird aus den erhobenen Bewertungen der Befragten diese Unterschiedlichkeit statistisch „ausgemittelt" (also entsubjektiviert), wird das End-Resultat nicht „besser" (im Sinne von richtiger), sondern „schlechter" (denn es verliert jeden Informationswert).

Von der Befragungsmethodologie her gesehen ergeben sich daraus zwei weitere unabdingbare Anforderungen an die Qualität des Befragungsprojekts (das gilt im übrigen nicht nur für Befragungen mit dem Ziel der „Evaluierung", sondern auch für Befragungen mit dem Ziel der Erhebung von Akzeptanz und von anderen qualitätsrelevanten Informationen):

- Neben dem „Kriterium" der Beurteilung („Qualitäts-Merkmal") ist ein „Standard" der Beurteilung zu definieren: Wann ist etwas „gut", wann „schlecht", wann „mittelmäßig"? etc.
- Dieser „Standard" ist von allen Evaluierenden in intersubjektiv übereinstimmender Weise anzuwenden.

4.3 Die gegenwärtige „best practice": Das Modell mehrstufiger Evaluation

Evaluation nach dem methodologischen Modell der Programmforschung stellt – darauf wurde mehrfach hingewiesen – sehr hohe Ansprüche, die schwer einlösbar sind. Im Hochschulkontext sind die Voraussetzungen für die Anwendung praktisch niemals gegeben. Das (alleinige) Einholen von Urteilen über Lehre und Studium per Befragung Studierender ist – auch bei Verwendung differenzierter Erhebungsbögen – kein vollwertiger Ersatz. Es liefert zwar für die Evaluation nützliche und wichtige Informationen. Deren subjektive Einfärbung ist allerdings auch durch statistische Methoden nicht eliminierbar. Gleiches gilt für Befragungen anderer Beteiligtengruppen (z.B. Lehrende).

Um dennoch Evaluation mit dem Anspruch auf intersubjektive Geltung im Hochschulbereich zu ermöglichen, hat sich mittlerweile ein mehrstufiges Verfahren weitgehend etabliert, das die Komponenten Selbstevaluation (durch die zu bewertende Einheit) sowie externe Validierung und Fremdevaluation (durch unabhängige, von außen hinzugezogene „peers") kombiniert. Ist das Verfahren Bestandteil eines Qualitätsentwicklungsprozesses, wird das Modell

noch ergänzt um den Schritt des Aushandelns von Schlussfolgerungen und Konsequenzen („follow up“).

Beispiele in Deutschland sind die Evaluationsverfahren des Verbunds Norddeutscher Universitäten und der Zentralen Evaluationsagentur (ZEvA) in Niedersachsen, die auf Empfehlungen der Hochschulrektorenkonferenz (1995) und des Wissenschaftsrates (1996) beruhen. Die so durchgeführte Evaluation erfasst die Organisation und Durchführung der Lehre und des Studiums innerhalb einer Hochschuleinheit (Fakultät/Fachbereich oder Institut), hat also nicht die Bewertung einzelner Lehrveranstaltungen zum Ziel. Die Hauptelemente des Verfahrens sind (s. HRK 1998):

- „Der Lehrbericht eines Fachbereichs / einer Fakultät als kontinuierliche Sammlung von Basisdaten und Leistungsindikatoren.
- Die interne Evaluation (...), die von einer internen Arbeitsgruppe vorbereitet wird und auf der Analyse der in den Lehrberichten erfassten Daten und auf Interviews mit Studenten und Personal basiert. Sie führt zu einem kritisch-abwägenden Bericht über die Selbsteinschätzung der erreichten Resultate im Hinblick auf die selbstgesteckten Ziele; sie enthält eine Beschreibung möglicher Hindernisse und Defizite sowie von Maßnahmen zu ihrer Beseitigung, Vorschläge für die Kontrolle und Verbesserung der Qualität der Lehre und die Verteilung von Mitteln für Forschung und Lehre. (...)
- Der Vor-Ort-Besuch der Sachverständigen (Peers), (…). Der in der Regel zweitägige Vor-Ort-Besuch schließt Gespräche mit der Universitätsleitung, dem Dekan und den Lehrenden und Studierenden ein, (...).
- Der Evaluations-Bericht der Peers schließt eine kritische Würdigung der internen Evaluation und ihrer tatsächlichen Bedeutung als Mittel der Qualitätssicherung ein, weist auf Probleme hin und gibt Hinweise auf mögliche Lösungen. Vor der Veröffentlichung des Abschlussberichts erhält der evaluierte Fachbereich Gelegenheit, den vorläufigen Bericht zu bearbeiten, um Irrtümer und Missverständnisse zu korrigieren. Dies findet im Rahmen einer gemeinsamen Sitzung statt, an der die Mitglieder der Sachverständigengruppe (Peer-Group), Vertreter der evaluierten Einrichtung und der Evaluationsagentur teilnehmen. (...)
- Das „follow up“ umfasst eine Vereinbarung bzw. einen Vertrag zwischen dem Fachbereich und der Universitätsleitung über zu ergreifende Maßnahmen zur Verbesserung von Lehre und Studium, zur Optimierung der Ergebnisse bzw. zur Sicherstellung bestimmter zu erreichender Standards innerhalb eines definierten Zeitraums. (...)“ (a.a.O., 11 f.).

Die Evaluierung geschieht in diesem Modell – wie ersichtlich – nicht durch die Umfrageforschung, wohl aber (unter anderem) mit Umfragen, und wird ergänzt um andere Erhebungen sowie um Daten aus der Hochschulstatistik und um Beobachtung und Diskussion. Für die Evaluation dieses Typs erfüllt die empirische Forschung und deren Methodik nicht die Funktion einer Instanz der Qualitätsentscheidung mittels „objektiver“ Daten. Vielmehr finden wir hier ein Beispiel für das Prinzip der „Objektivierung durch Verfahren“. Die Sicherung der

Intersubjektivität der Ergebnisse wird durch ein darauf zugeschnittenes Verfahrensmodell angestrebt: Die Einbeziehung aller Beteiligten und Betroffenen in den Prozess soll gewährleisten, dass das für den Zweck der Evaluation relevante Informationsspektrum erfasst wird. Die Gültigkeit der Ergebnisse, wie sie der Evaluationsbericht dokumentiert, wird durch die Möglichkeit zur Korrektur sowie durch eine gemeinsame Abschlussdiskussion zwischen Evaluatoren und Evaluierten angestrebt (kommunikative Validierung). Damit die Evaluation nicht ins Leere läuft, sondern Anstöße zu Qualitätsverbesserungen gibt, mündet das Verfahren in konkrete Zielvereinbarungen (Festlegung nachprüfbarer Maßnahmen mit expliziten Terminen für die Realisierung). Und um es nicht bei einem einmaligen Anstoß bewenden zu lassen, sondern einen Prozess kontinuierlicher Qualitätsverbesserung in Gang zu setzen, sind schließlich in regelmäßigen Abständen (von mehreren Jahren) „follow ups“ vorgesehen.

Literatur

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

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

Eekhoff, Johann; Muthmann, Rainer; Sievert, Olaf; Werth, Gerhard; Zahl, Jost, 1977: Methoden und Möglichkeiten der Erfolgskontrolle städtischer Entwicklungsmaßnahmen, Bonn-Bad Godesberg: Schriftenreihe "Städtebauliche Forschung" 03.060

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

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

Hellstern, Gerd-Michael; Wollmann, Hellmut (1983): Evaluierungsforschung. Ansätze und Methoden, dargestellt am Beispiel des Städtebaus, Basel, Stuttgart

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

HRK Hochschulrektorenkonferenz (Hg.), 1998: Evaluation. Sachstandsbericht zur Qualitätsbewertung und Qualitätsentwicklung in deutschen Hochschulen. Dokumente & Informationen 1/1998, Bonn: HRK

Kromrey, Helmut (1994): Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung, Münster (2. Aufl. 1995): Waxmann, 105-128

Kromrey, Helmut, 1995: Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. In: ZSE Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, Jg. 15, H. 4, 313-335

Kromrey, Helmut, 1999: Von den Problemen anwendungsorientierter Sozialforschung und den Gefahren methodischer Halbbildung. In: Sozialwissenschaften und Berufspraxis, Jg. 22, H. 1, 40-57

Kromrey, Helmut, 2003: Evaluation in Wissenschaft und Gesellschaft. In: ZfEv Zeitschrift für Evaluation, Jg. 2, H. 1, 2003, 93-116

Landfried, Klaus, 1999: Qualitätssicherung als Aufgabe wettbewerblicher Hochschulen. In: HRK (Hg.): Ein Schritt in die Zukunft. Qualitätssicherung im Hochschulbereich. Beiträge zur Hochschulpolitik 3/1999, Bonn: HRK, 7-13

Patton, Michael Quinn, 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, Peter H. / Freeman, Howared E. (1988): Programmevaluation. Einführung in die Methoden angewandter Sozialforschung, Stuttgart

Weiss, Carol H., 1974: Evaluierungsforschung. Methoden zur Einschätzung von sozialen Reformprogrammen, Opladen 1974

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

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

Wissenschaftsrat 1996: Empfehlungen zur Stärkung der Lehre in den Hochschulen durch Evaluation. In: ders.: Empfehlungen und Stellungnahmen 1996, Band I, Köln