Evaluation von Lehre und Studium
Anforderungen an Methodik und Design

Helmut Kromrey

zur Veröffentlichung in:
Christiane Spiel (Hg.): Evaluation universitärer Lehre – Zwischen Qualitätsmanagement und Selbstzweck, Münster 2001, S. 21-60

Vorbemerkungen:

Würde man die hochschul- und wissenschaftspolitische Diskussion der vergangenen Jahre daraufhin durchsehen, welche Begriffe besonders häufig auftauchen, dann belegte "Evaluation" (bzw. Evaluierung) mit Sicherheit einen der Spitzenplätze. Versuchte man zu ergründen, was damit jeweils bezeichnet wird, stieße man zugleich auf eine nur schwer eingrenzbare Vielfalt von Begriffsverwendungen mit diversen dahinter stehenden Denk- und Handlungskonzepten, deren Gemeinsamkeit allenfalls als Leerformel ausdrückbar ist: *Irgend etwas wird von irgend jemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet.* Derart unklare Sprache führt zwangsläufig in ein Dilemma: Rationales Argumentieren erweist sich als unmöglich, solange sich bei den Beteiligten hinter gleichlautende sprachlichen Zeichen unterschiedliche Gedanken verbergen.

In der *politischen* Diskussion ist "Evaluation/Evaluierung" zu einem Allerweltswort geworden, das für *jede* Form von Bewertung/Bewerten stehen kann: von der gutachterlichen, datengestützten Beurteilung existierender Programme oder Einrichtungen bis zur subjektiven Vermutung über den künftigen Nutzen geplanter Maßnahmen oder Entwürfe.

Auch im *wissenschaftlichen* Kontext erweist sich die Situation als nicht viel klarer. Mit "Evaluation" kann – am einen Ende des Bedeutungsspektrums – eine systematische, theoretisch fundierte empirische Analyse zielgerichteter Aktivitäten (ein "Programm") unter dem Aspekt der "Erfolgskontrolle" gemeint sein. Es kann sich aber auch – am anderen Ende des Spektrums und sehr viel eingeschränkter – lediglich (in Analogie zur Meinungsforschung) um die empirische Erhebung und Analyse von per Befragung ermittelten subjektiven Urteilen einer irgendwie definierten Zielgruppe von Personen handeln

Um für den vorliegenden Beitrag der Gefahr von Mißverständnissen vorzubeugen, werden im nächsten Abschnitt zunächst verschiedene Konzepte empirisch-wissenschaftlicher Evaluation skizziert sowie die Möglichkeiten und Grenzen ihrer Realisierbarkeit angesprochen. Danach werden Evaluations-Modelle vorgestellt, die sich im Kontext der Institution Hochschule durchgesetzt haben. Ein dritter Teil beschäftigt sich mit den Möglichkeiten, Evaluation als Instrument zur Qualitätsentwicklung in Studium und Lehre einzusetzen. Den Abschluss bilden einige Beispiele der Nutzung von Befragungen zur Entwicklung von Lehrqualität.

1. Das Konzept "Evaluation": von empirischer Sozialforschung bis zu wissenschaftlicher Beratung

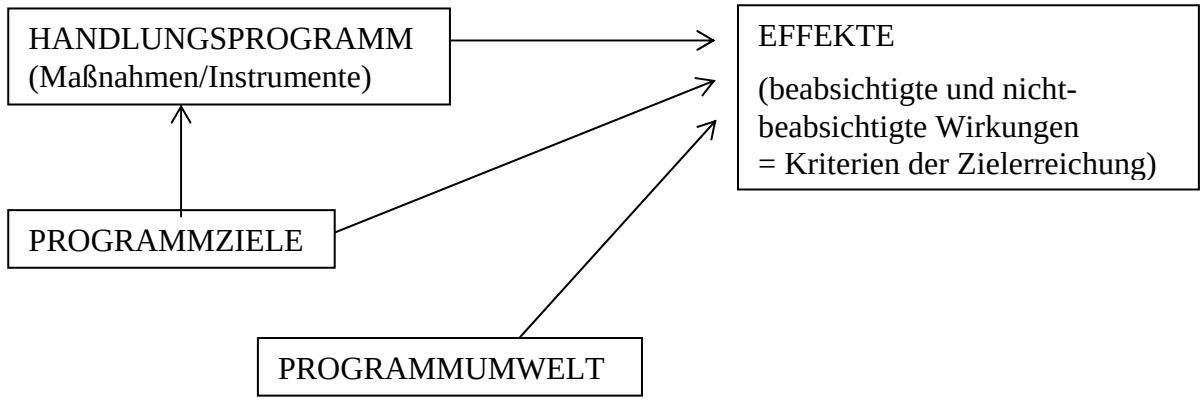
1.1 Evaluation als spezifisches Design angewandter Forschung

In der *Methodologie empirischer Sozialforschung* bezeichnet Evaluation das *Design* für einen spezifischen Forschungstyp, durchgeführt im idealtypischen Fall als Feldexperiment mit Kontrollgruppen. Wo die Voraussetzungen dafür nicht in vollem Maße erfüllt sind – und das ist überwiegend der Fall –, behilft man sich mit "Ersatzlösungen" für diejenigen Design-komponenten, die nicht idealtypisch realisiert werden können; erhalten bleibt aber in jedem Fall die generelle Orientierung an der Vorgehens- und Argumentationslogik des Experiments.

Zwingend notwendig sind für diesen Ansatz – in Absetzung zum Alltagssprachgebrauch oder auch zur politischen Diskussion (*s.* die Vorbemerkungen) – im Vorhinein zu leistende Präzisierungen *aller* mit einem Evaluationsvorhaben angesprochenen Aspekte: Nicht "i*rgend etwas"* soll beurteilt werden, sondern ein eindeutig zu definierender und empirisch abgrenzbarer Gegenstand oder Sachverhalt. Nicht "*irgend jemand"* ist für die Informationssammlung und –analyse zuständig, sondern eine mit der notwendigen Kompetenz versehene Evaluations-Instanz. Nicht "*nach irgendwelchen Kriterien"* und nicht *"in irgendeiner Weise"* kommen die Bewertungen zustande, sondern nach explizit auf den Sachverhalt bezogenen und begründeten Beurteilungskriterien und Standards sowie in einem objektivierten Verfahren im Rahmen eines im Detail geplanten Evaluationsdesigns. Da jedoch Präzisierungen zu den vier Aspekten (Gegenstand – Evaluationsinstanz – Kriterien – Verfahren) in unterschiedlicher Weise möglich sind und auch in unterschiedlichen Kombinationen vorkommen, sehen sich Evaluator(inn)en einer Vielfalt von Aufgabenprofilen und Rahmenbedingungen gegenüber, für die eine allgemein geltende "Evaluations-Methode" nicht existiert.

Bei aller Vielfalt bleibt dennoch allen Vorhaben gemeinsam, daß sie (mindestens) drei interdependente Dimensionen aufweisen – nämlich Ziele, Maßnahmenprogramm, Effekte – und daß sie (anders als in einem Forschungslabor) von Umgebungseinflüssen nicht abgeschirmt werden können.

Abbildung 1: Programm-Evaluation



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

können feststellbare Veränderungen als Wirkungen des Programms identifiziert und gegenüber anderen Einflüssen abgegrenzt werden?

Eine so umfassende Evaluation, wie sie nach dieser ersten groben Strukturierung des Aufgabenfeldes notwendig erscheint, ist selbstverständlich in keinem Projekt realisierbar. Es müssen Schwerpunkte entsprechend dem vorherrschenden Verwertungsinteresse gesetzt werden, damit das Vorhaben dennoch nutzbare Erkenntnisse liefert. 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.

1.1.1 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 dann die systematische und kontrollierte "Buchführung": Was passiert? Was wird wann und wie gemacht? ("monitoring")

1.1.2 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 "praxisrelevant". Andererseits ist es 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.

1.1.3 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 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.

1.1.4 Instanzen der Evaluierung: Wer beurteilt nach welchen Kriterien?

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

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

Ein solches Vorgehen verlangt relativ umfassendes theoretisches Wissen über die Struktur der Zusammenhänge zwischen Zielen, Maßnahmen, Wirkungen und Umwelteinflüssen, das jedoch– insbesondere bei Innovationsvorhaben – häufig nicht vorhanden ist. Hier behilft sich die Evaluation verschiedentlich 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 manchmal die *Befragung der Adressaten eines Programms* (Nutzer oder Betroffene) gewählt.

1.1.5 Methoden der Programmforschung

Die Methodologie der Programmforschung wurde im wesentlichen in den 70er und 80er Jahren entwickelt. Je nachdem, ob ein Evaluationsprojekt mehr in Richtung Wirkungsforschung oder mehr in Richtung Erfolgskontrolle tendiert, hat sich die Evaluation 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 orientieren 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 "Monitoring" der Instrumentvariablen (Programm-Input), der exogenen Einflüsse und der Zielerreichungsgrade (Output). Methodisch gesehen handelt es sich bei diesem dreifachen "Monitoring" somit um vergleichsweise einfache, *deskriptive* Forschungsaktivitäten.

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

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

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

- Analyse der Programmziele und ihrer Interdependenzen (Präzisierung eines Zielsystems einschließlich der Festlegung des angestrebten Zielniveaus) sowie Zuordnung der Instrumente zur Zielerreichung (Maßnahmen);
- 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. Bei näherem Hinsehen wird jedoch erkennbar, daß es von anspruchsvollen Voraussetzungen über den Gegenstand der Untersuchung wie auch von schwer einlösbaren Voraussetzungen bei den programm-durchführenden 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 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 sicherzustellen.

Im Normalfall der Begleitforschung zu Programm-Implementationen oder gar zu Modellversuchen neuer Techniken, neuer Schulformen, zur Erprobung alternativer Curricula oder Lernformen u.ä. ist keine einzige dieser Bedingungen voll erfüllt. Die Untersuchungssituation weist vielmehr in dieser Hinsicht erhebliche "Mängel" auf (ausführlicher dazu Kromrey 1988).

1.2 Alternativen zum Experimentaldesign: ex-post-facto-Analyse und theoriebasierte Evaluation

Als unbestrittener "Königsweg" der Programm-Evaluation gilt – wie bereits angesprochen – das Experimentaldesign, mit Einschränkungen noch das Quasi-Experiment, das so viele Elemente des klassischen Experiments wie möglich zu realisieren versucht und für nicht realisierbare Design-Elemente methodisch kontrollierte Ersatzlösungen einführt. So tritt etwa bei der Zusammenstellung strukturäquivalenter Versuchs- und Kontrollgruppen das matching-Verfahren an die Stelle der Randomisierung; oder die nicht mögliche Abschirmung von Störgrößen in der Informationsbeschaffungsphase wird ersetzt durch umfassende Erhebung relevanter potentieller exogener Wirkungsfaktoren, um nachträglich in der Auswertungsphase die exogenen Einflüsse statistisch zu kontrollieren (vgl. für einen zusammenfassenden Überblick Frey/Frenz 1982).

Mit letzterem Beispiel sind wird bereits auf halbem Wege, die Experimentallogik *in der Erhebungsphase* durch Experimentallogik *in der Auswertungsphase* zu simulieren. Wo ein Interventionsprogramm eine soziale Situation schafft, in der sich ein Feldexperiment verbietet, kann 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 Analogie zur *Abschirmung von Störgrößen*). Diese *nachträgliche* Anordnung der Informationen in einer Weise, als stammten die Daten aus einem Experiment, wird üblicherweise als "*ex-post-facto-Design*" bezeichnet.

Allerdings weist die ex-post-facto-Anordnung eine gravierende und prinzipiell nicht kontrollierbare Verletzung des Experimentalprinzips auf, nämlich das Problem der Selbstselektion der Teilnehmer/Nutzer. Auch das ausgefeilteste statistische Analysemodell kann kein Äquivalent zur kontrollierten Zuweisung zur Experimental- bzw. Kontrollgruppe

anbieten. Allenfalls kann versucht werden, diesen Mangel in der Feldphase dadurch zu mildern, daß Gründe für die Teilnahme oder Nicht-Teilnahme mit erhoben werden, um möglicherweise existierende systematische Unterschiede erkennen und abschätzen zu können. Darüber hinaus erhält die generelle Problematik der Messung sozialer Sachverhalte im Vergleich zum echten Experiment ein erheblich größeres Gewicht: Soll die Gültigkeit der Analyse-Resultate gesichert sein, müssen alle potentiellen exogenen Einflüsse und müssen alle relevanten Persönlichkeitsmerkmale nicht nur bekannt, sondern auch operationalisierbar sein und zuverlässig gemessen werden. Im echten Experiment entfällt diese Notwendigkeit dadurch, daß alle (bekannten und unbekannten) exogenen Einflußgrößen durch Randomisierung bei der Bildung von Experimental- und Kontrollgruppen neutralisiert werden.

Einen anderen Zugang zur Gewinnung detaillierten empirischen Wissens über das zu evaluierende Vorhaben wählt das Modell einer *"theoriebasierten Evaluation"* (theory-based evaluation). Gemeint ist hier mit dem Terminus "Theorie" allerdings nicht ein System hoch abstrakter, generalisierender, logisch verknüpfter Hypothesen mit im Idealfall räumlich und zeitlich uneingeschränktem Geltungsanspruch, sondern – ähnlich wie beim grounded-theory-Konzept – eine gegenstandsbezogene Theorie, eine Theorie des Programmablaufs (Weiss 1995, 1997). Die Bezeichnung "logisches Modell" wäre vielleicht treffender (vgl. Patton 1997, 234 ff.: logical framework approach), zumal die Bezeichnung "theoriebasierte Evaluation" etwas irreführend ist, denn auch das Modell der Programmforschung ist "theoriebasiert" (s.oben: 1.1.5). Sie benötigt, um Programmeffekte überhaupt erkennen und messen zu können, ein möglichst gut abgesichertes forschungsleitendes Wirkungsmodell, d.h. ein in sich schlüssiges, einheitliches System von operationalisierbaren Hypothesen, das die theoretische Basis für die Planung des Programms (Zuordnung von Maßnahmen/Instrumenten zu Programmzielen) und für dessen Implementation rekonstruiert und das die Begründung für die Zurechnung der gemessenen Effekte zu den durchgeführten Maßnahmen liefert.

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 deren jeweils eigene individuelle Vermutungen über die Notwendigkeit der Erreichung bestimmter Ziele und über 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 Evaluator(inn)en mehrere – im Idealfall sich ergänzende, manchmal 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.[1] Wo

[1] Während das Modell der Programmforschung (variablenorientiert) die *wirkungslogische* Beziehungsstruktur

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. [2]

1.3 Evaluation durch Indikatorenmodelle und Zielgruppenbefragung

Im Falle der Beurteilung eines Interventions- oder Dienstleistungsprogramms mit benennbaren Zielgruppen im Hinblick auf seine Implementation, seine Akzeptanz und seinen Erfolg wird häufig als Evaluationsmethode auch auf Input-Output-Modelle mit "objektiven" und "subjektiven" Indikatoren (im Idealfall als hinreichend lange Zeitreihen) zurückgegriffen. Dies gilt insbesondere, wenn das Verwertungsinteresse der Evaluation auf die Qualitätssicherung und Qualitätsentwicklung fortlaufend zu erbringender Humandienstleistungen gerichtet ist. Zwar wird mittlerweile weitgehend unbestritten der positive Effekt bei den Adressaten der Dienstleistung (outcome) als *letztliches Kriterium für den Erfolg* der Dienstleistung akzeptiert. Unabhängig davon gilt aber weiterhin die Annahme als fraglos gesichert, daß gute Servicequalität eine *weitgehende Gewähr für solchen Erfolg* sei. So wird z.B. in der Hochschulpolitik für wahrgenommene Mängel im universitär vermittelten Qualifikations-Output (z.B. lange Studienzeiten oder hohe Studienabbruchquoten) in erster Linie die vorgeblich schlechte Lehre verantwortlich gemacht und deren Qualitätsverbesserung eingefordert.

Somit gehört es in diesem Evaluationskonzept zu den ersten Aufgaben, die qualitätsrelevanten Dimensionen des Dienstleistungsangebots zu bestimmen und zu deren Beurteilung geeignete Qualitätsindikatoren zu begründen und zu operationalisieren – eine Aufgabe, mit der sich die Sozialwissenschaft im Rahmen der Sozialindikatorenbewegung seit Jahrzehnten befaßt hat. 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 *positiv*en 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

---

der Programmelemente einschließlich der Programmumwelt darstellt, handelt es sich hier um die (akteursorientierte) Skizze der *zeitlich-sachlogischen* Ablaufsstruktur des Programms.

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

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 der Definition, nämlich *Qualität als Grad der Übereinstimmung zwischen zuvor formulierten Kriterien und der tatsächlich erbrachten Leistung*. Diesen Prozeß bettet Donabedian 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. Zugleich 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).

Unterstellt man die sachliche Angemessenheit dieses dimensionalen Schemas, besteht die entscheidende Aufgabe der Evaluation darin, zu jeder der Dimensionen diejenigen Indikatoren zu bestimmen und zu operationalisieren, die dem konkret zu evaluierenden Programm angemessen sind. Dies kann nicht ohne Einbeziehung der Programmträger, des eigentlichen Dienstleistungspersonals sowie der Adressaten der Dienstleistung und ggf. weiterer Beteiligter und Betroffener geschehen (als Beispiel: Herman 1997). Des weiteren sind die Indikatoren als gültige Meßgrößen durch Formulierung von "Korrespondenzregeln" methodisch zu begründen; d.h. es ist nachzuweisen, daß sie "stellvertretend" die eigentlich interessierenden Dimensionen abbilden. Häufig genug geschieht dies entweder überhaupt nicht oder lediglich gestützt auf Vermutungen oder als Ergebnis eines Aushandlungsprozesses zwischen den Beteiligten,[3] oder sie werden von vornherein unter dem Gesichtspunkt leichter Messbarkeit ausgewählt. Nicht nur ist dann 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.[4]

Wenn – wie dargelegt – der positive Effekt bei den Adressaten der Dienstleistung (outcome) als letztliches Kriterium für den Erfolg der Dienstleistung gelten soll, dann ist als

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

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

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. Ein solches Indikatorensystem für das Donabedian-Modell wird in einem Artikel von Salzer u.a. (1997) vorgestellt und methodologisch grundlegend diskutiert.

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

Leider erweisen sich solche Vorstellungen als empirisch falsch (vgl. am Beispiel "Lehrevaluation" an Hochschulen: Kromrey 1994, 1996). Die per Umfrageforschung bei Nutzern oder Betroffenen erhobenen Antworten auf bewertende (also evaluative) Fragen haben nicht den Status von "Evaluation" als methodisch kontrollierter, empirischer Qualitätsbewertung. Ermittelt wird lediglich die "Akzeptanz" (oder Nicht-Akzeptanz), auf die der beurteilte Sachverhalt bei den Befragten stößt; und die hängt im wesentlichen ab von Merkmalen der Befragten und nur relativ gering von Merkmalen des beurteilten Sachverhalts. Natürlich sind auch Akzeptanzaussagen keine unwesentliche Information, insbesondere in solchen Dienstleistungsbereichen, in denen der Erfolg von der aktiven Partizipation der Adressaten abhängt. Akzeptanzaussagen geben Auskunft darüber, in welchem Ausmaß und unter welchen Bedingungen das Angebot "akzeptiert" (oder abgelehnt) wird, sowie darüber, welche Änderungen ggf. notwendig sind, um die "Akzeptanz" – nicht unbedingt das Produkt – zu verbessern. Die Akzeptanz eines Angebots kann jedoch sehr wohl eine wichtige *Voraussetzung für gute Ergebnisqualität* sein (z.B. gilt in der Didaktik die positive Einstellung zur Lehrperson und zu den Lehrinhalten unbestritten als eine notwendige – wenn auch noch nicht als eine hinreichende – Bedingung für erfolgreiches Lernen). Akzeptanz und/oder Zufriedenheit kann schließlich – wie im Donabedian-Modell – eine *Teildimension von outcome-Qualität* sein; dann nämlich, wenn die aktive Partizipation der Adressaten ein explizites Ziel des Programms ist. Aber selbst als Teildimension von Qualität kann sie nicht stellvertretend für das *gesamte* Qualitätskonzept stehen.

1.4 Evaluation und Beratung als Handlungszusammenhang

Bisher wurde Evaluation als *methodologisches Konzept* (Forschungsmodell der Evaluation) dargestellt. Doch selbst im Kontext angewandter Sozialwissenschaft steht der Begriff "Evaluation" nicht lediglich für Forschungshandeln, sondern umfasst nicht selten einen *weiter reichenden komplexen Handlungszusammenhang*, der sich neben der Forschung (dem Sammeln und Auswerten von Informationen über die Implementation und die Auswirkungen von Programmen) auch auf die wissenschaftliche Beratung oder gar die aktive Teilnahme an der Entwicklung, Implementation und Optimierung erstreckt. Zum Teil wird hierfür der Terminus "*offene Evaluation*" verwendet, im Unterschied zum eben skizzierten "geschlossenen" Konzept, bei dem für die Evaluation ein in der Umsetzung befindliches Programm, zumindest jedoch eine explizit ausformulierte Planung vorgegeben ist. Akteure in

"offenen" Evaluationsvorhaben können in die gesamte Programm-Historie eingebunden sein: 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. Zugleich darf ihr Blickwinkel nicht auf den Sachverhalt "Programm" beschränkt bleiben, sondern muss explizit auch die Beteiligten und Betroffenen – die "stakeholders", wie der schwer übersetzbare amerikanische Fachterminus lautet – einbeziehen. Die Aufgabe besteht nicht lediglich im Zurverfügung-Stellen von Informationen, sondern auch in deren Vermittlung, in zielgruppengerechter "Übersetzung", im Entwickeln von Umsetzungs-Ideen, im Moderieren, im Vermitteln zwischen unterschiedlichen Perspektiven. "Evaluation" wird so zu einem umfassenden Informations- und Prozessmanagement (für eine ausführlichere Darstellung dieser Konzepte s. Kromrey 1995a).

1.5 Gemeinsames: Was soll zu welchem Zweck durch wen auf welche Weise evaluiert werden?

So verschieden die skizzierten Evaluationskonzepte im Detail auch ausfallen mögen: Allen Ansätzen ist gemeinsam, dass einige (insbesondere nicht-methodologische) Fragen *vor* dem Anlaufen des Projekts eindeutig zu beantworten sind, sollen die Resultate nutzbringend verwertbar sein und soll das Vorhaben nicht zum reinen Selbstzweck, zum "Evaluations-Ritual" degenerieren.

An erster Stelle steht die *Klärung des Zwecks*. Ganz grob kann unterschieden werden zwischen Evaluationen mit dem Ziel der *Verbreiterung der Wissensbasis* in einem Handlungsfeld (um z.B. Interventionen besser planen und steuern, um Entscheidungen auf empirisch und theoretisch gesicherterer Basis treffen zu können) und Informationserhebungen zu *Kontrollzwecken* (die Verwendung von Ressourcen – insbesondere von öffentlichen Mitteln – hat sich anhand von Erfolgskriterien wie Effektivität, Effizienz und Akzeptanz zu legitimieren)[5]. Davon abzugrenzen ist schließlich noch die Evaluation zu *Entwicklungszwecken*. Eine vorherige bewußte Einordnung des Vorhabens innerhalb des Spektrums möglicher Verwertungsinteressen ist auch unter methodologischen Gesichtspunkten unabdingbar, da aus ihr extrem unterschiedliche Designtypen folgen; nicht ohne Grund wird in dieser Hinsicht zwischen dem Forschungs-, dem Kontroll- und dem Entwicklungsparadigma der Evaluation unterschieden.

Nahezu ebenso wichtig – wenn auch allenfalls analytisch von der Klärung des Zwecks zu trennen – ist die eindeutige *Definition und Abgrenzung des Gegenstands* der Evaluation. Der Blick kann sich richten auf das Handeln von Akteuren (z.B. von Lehrenden) innerhalb gegebener Rahmenbedingungen; das Interesse kann aber auch (oder statt dessen) der Angemessenheit der Organisationsstruktur und der Rahmenbedingungen selbst gelten. Gegenstand kann das Zielsystem einer Institution und dessen Umsetzung sein (etwa Studiengänge und Curricula eines Hochschulinstituts); es kann aber auch (oder statt dessen) das Verhalten der Nutzer – der "Klienten", z.B. der Studierenden – sein. Es versteht sich von selbst, dass für die Sicherstellung des Evaluationszwecks je nach "Gegenstand" ein jeweils weitgehend unterschiedlicher Satz von Informationen erforderlich ist.

Eine weitere klärungsbedürftige Frage lautet: *Wer evaluiert*? Sie wird häufig zu eng und zu ungenau beantwortet, was im Verlaufe des Vorhabens fast zwangsläufig zu Mißverständnissen und zu mangelnder Akzeptanz der Befunde führt. Die Frage spricht

[5] Evaluationen dieser Art werden argumentativ vertreten als eine weitere Kontrollform administrativen Handelns neben der Rechtmäßigkeitskontrolle durch Gerichte, der politischen Kontrolle durch Parlamente und der Wirtschaftlichkeitskontrolle durch Rechnungshöfe.

mehrere Teildimensionen an: (a) Wer führt das *Evaluationsprojekt* durch? (b) Wer ist die evaluierende Instanz, nimmt also die *Bewertungen* vor? (c) Wer ist zuständig für die *Verwertung* der Befunde? Erst wenn im Hinblick auf jeden dieser Aspekte klare Kompetenzzuschreibungen existieren, sind gute Voraussetzungen für eine nutzbringende Evaluation gegeben.

Zu (a): Das Evaluationsprojekt kann durch organisationseigenes Personal durchgeführt werden oder auf externen Sachverstand zurückgreifen; notwendige Bedingung für die Akzeptanz der Ergebnisse ist in jedem Fall die von allen Beteiligten anerkannte Methodenkompetenz.

Zu (b): Die Beurteilungen können anhand vorher vereinbarter Kriterien datenbasiert vom Evaluationsprojekt vorgenommen werden oder in der Zuständigkeit einer davon unabhängigen Instanz liegen (in der Hochschule kann dies beispielsweise eine eigens vom Fakultäts- oder Institutsrat eingesetzte Evaluationskommission sein). Sie können aber auch in die Hand von "Experten" gelegt werden (etwa externe Fachexperten oder auch Klienten bzw. in anderer Weise "Betroffene").

Zu (c): Für die Verwertung der Befunde schließlich sollten entscheidungsbefugte Instanzen zuständig und verantwortlich sein (in der Hochschule z.B. der Instituts- oder Fachbereichsrat oder die Institutsleitung bzw. das Dekanat). Auf jeden Fall ist – soll eine Evaluation nicht wirkungslos bleiben – explizit eine entsprechende Verantwortlichkeit von Beginn an festzuschreiben. Die Verwertung kann in der Aushandlung von Zielvereinbarungen münden, in denen die durchzuführenden Maßnahmen (und die dafür einzusetzenden Ressourcen) sowie Termine für deren Realisierung benannt werden, so dass eine spätere Erfolgskontrolle (ein "follow up") möglich ist.

Auf welche Weise das Evaluationsprojekt (im o.g. engeren Sinne)[6] durchzuführen ist, scheint nach der Festlegung der vorgenannten drei Punkte relativ problemlos und allein aus methodologischen Gesichtspunkten herleitbar. Da jedoch die Evaluation im "sozialen Feld" und unter Bedingungen abläuft, in denen der Vorrang für die wissenschaftlichen Ansprüche empirischer Forschung nicht (jedenfalls nicht vollständig) durchsetzbar ist, sind in der Regel eine Vielzahl von Kompromissen auszuhandeln, um die Akzeptanz des Verfahrens bei (möglichst) allen Beteiligten zu sichern: Welche Dokumente sind unter welchen Bedingungen zugänglich? Welche Statistiken existieren, und welche können für die Evaluation zusätzlich erstellt werden? Wer darf unter welchen Bedingungen mit welchen Instrumenten befragt werden? Zu welchen Situationen erhalten die Evaluator(inn)en unter welchen Bedingungen als Beobachter/innen Zugang? Wie ist der Datenschutz geregelt? Werden Ergebnisse veröffentlicht? Falls ja: in welcher Form? Hinter jeder dieser Fragen verbergen sich Fallstricke, und nur wenn sie für die Beteiligten akzeptabel geregelt werden, sind unverfälschte Informationen zu erwarten.

2. Was wird an Hochschulen unter dem Label "Lehrevaluation" getan?

2.1 Zur Unmöglichkeit von Evaluation als Erfolgskontrolle

Dass das Modell "Evaluation als Programmforschung" in der Hochschule flächendeckend nicht eingesetzt werden kann, ist offensichtlich. Die Universität kann – sowohl aus grundsätzlichen wie aus pragmatischen Gründen – nicht zum Experimentallabor

[6] Nämlich verstanden als angewandte Sozialforschung mit dem Ziel methodisch kontrollierten, verwertungs- und bewertungsorientierten Sammelns und Auswertens von Informationen.

umfunktioniert werden, nur um dadurch evaluierbar zu sein.[7] Aber auch aus weiteren methodologischen Gründen eignet sich diese Form der Evaluierung nicht.

So könnte beispielsweise das zu bewertende *Programm* der *Diplomstudiengang 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, außerdem Prüfungsordnung, Prüfungen und andere Leistungskontrollen in Betracht. Für die Messung der Zielerreichung böte sich der Zeitpunkt der Beendigung des Studiums (Diplomprüfung oder Studienabbruch) bei den einzelnen Studierenden an.

Ein erstes Bündel von Problemen ergäbe sich bereits bei der empirischen Beschreibung der für die Studierenden bis zum Examen relevant gewordenen Maßnahmen. 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 auch von ihren Inhalten häufig so stark variierend, dass 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 – falls überhaupt – 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 zu fragen, wie es um die Gültigkeit dieser Indikatoren bestellt wäre, wenn die Träger des zu evaluierenden Programms die Indikatorausprägungen selbst festlegten (nämlich in Prüfungen und Klausurbenotungen).

Ganz unmöglich schließlich wäre 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* können 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.

[7] Allenfalls Lehrexperimente *neben* dem laufenden Lehrbetrieb könnten in dieser Weise evaluiert werden; hier wären z.B. für die Vermittlung des herkömmlichen Stoffs Doppelangebote bei Realisierung unterschiedlicher Lernsettings denkbar.

2.2 Evaluation durch Umfrageforschung

Wenn Evaluation nach dem Modell der Programmforschung nicht möglich ist, liegt es nahe, das Fällen von Urteilen – also die *Tätigkeit des Evaluierens* – auf dafür geeignet erscheinende Dritte zu verlagern (auf Experten, auf Kunden, auf Betroffene) und die Funktion der Forschung auf das systematische Einholen und Auswerten solcher "Fremd-Evaluationen" zu beschränken. Dies wird in der Tat überall dort so gehandhabt, wo Lehrevaluation betrieben wird.

In besonders systematischer, formalisierter und nachprüfbarer Form geschieht dies in dem *Verfahren der zweistufigen (internen und externen) Evaluation*, wie es im Verbund Norddeutscher Universitäten und von der Zentralen Evaluationsagentur (ZEvA) in Niedersachsen angewendet wird und das auf Empfehlungen der Hochschulrektorenkonferenz (1995) und des Wissenschaftsrates (1996) beruht. Die so durchgeführte Evaluation erfasst als Gegenstand die Organisation und Durchführung der Lehre und des Studiums innerhalb einer Hochschuleinheit (Fakultät/Fachbereich oder Institut) und hat explizit 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 von der betreffenden Agentur vorbereitet wird, welche die Selbstbewertungsberichte an die Mitglieder der Peer-Group weiterleitet und, falls nötig, den zu evaluierenden Fachbereich um weitere Informationen bittet. 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., S. 11 f.; im selben Heft findet sich auf S. 15 f. ein Muster zur Gliederung des Lehrberichts sowie auf S. 17 f. ein Vorschlag für eine Gliederung des Evaluationsberichts).

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 (stakeholder-Modell). 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.

Es ist leicht nachvollziehbar, dass ein solches Evaluations- und Qualitätssicherungsmodell mit einem hohen Aufwand an Kosten, Zeit und Personal verbunden ist.

Wesentlich weniger anspruchsvoll – sowohl hinsichtlich des Verfahrens als auch hinsichtlich des Bemühens um Objektivierung – ist demgegenüber die weit verbreitete Strategie, die Lehrevaluation allein auf die *Befragung Studierender* zu stützen. Dem liegt offenbar die Gleichsetzung von Betroffen-Sein mit Expertentum zugrunde. Für die Beurteilung der Qualität der Lehre etwa lässt sich die folgende einfache (und auf den ersten Blick auch durchaus plausibel erscheinende) Argumentation rekonstruieren: “Ein aufwendiges Verfahren der Qualitätsbeurteilung durch Evaluations*forschung* 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” (s. oben: 1.3). Manche Lehrende gehen noch einen Schritt weiter und vertreten unter Verweis auf ”jahrzehntelang bewährte Praxis in den USA”[8] 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 aufwendige Auswertungsverfahren; Auszählungen und Durchschnittsberechnungen seien hinreichend. Ein Beispiel für diesen Typ von ”Einfach-Evaluation” ist das an der Freien Universität Berlin in regelmäßigen Abständen eingesetzte ”FU-Studienbarometer”, das für die standardisierte Beurteilung eines ganzen Studiengangs mit einer einzigen Fragebogenseite auskommt (s. Kromrey 1999, 62 ff.).

In dieser Form eingesetzt, sind mit einer Strategie der Erhebung studentischer Wahrnehmungen und Bewertungen *als* Evaluation von Studium und Lehre allerdings gleich mehrere *Fehlschlüsse* verbunden.

Im Unterschied zur Expertenevaluation anhand vorgegebener Kriterien und auf der Basis systematisch ausgewerteter Informationen mit anschließender kommunikativer Validierung (wie beim Verfahren der peer-Evaluation) sind die befragten Studierenden ”Alltags-

[8] Übersehen wird dabei, dass mit Studierendenbefragungen in US-Universitäten bewußt Akzeptanzmessung betrieben wird (schließlich sind die Studierenden ”Kunden” der Universität, nämlich Abnehmer einer Dienstleistung, die durch zum Teil hohe Studiengebühren die Institution mitfinanzieren). Für die Evaluation der Lehrenden durch die Hochschule ist allerdings die per Befragung ermittelte Akzeptanz nur *ein* Baustein in einem detaillierteren Bewertungsverfahren.

Evaluatoren": 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 ...). Die Bedeutung der im standardisierten Erhebungsbogen gegebenen Antworten ist so nicht mehr rekonstruierbar.

Werden *Globalaussagen* verglichen mit *differenziert erhobenen Beurteilungen* (unter Verwendung von Itemlisten, mit denen vor der Erhebung zusammenfassender Urteile zunächst Detail-Aspekte eingeschätzt werden), so zeigt sich, dass unter den Studierenden – grob zusammengefasst – zwei geradezu gegensätzliche Evaluierungsweisen existieren. Nahezu die Hälfte der Veranstaltungsteilnehmer urteilt so pauschal, dass in der Tat die Verwendung einfacher und kurzer Erhebungsinstrumente angemessen wäre: Die Tendenz der Einschätzungen auf *allen* Detail-Items einer Dimension stimmt überein mit dem zusammenfassenden Gesamturteil, d.h.: Man ist entweder in jeder Hinsicht zufrieden oder in jeder Hinsicht unzufrieden oder empfindet das Angebot durchweg als mittelmäßig. Die andere Hälfte der Befragten urteilt in den Details differenziert (man ist mit dem einen Teilaspekt zufrieden, mit dem anderen weniger, mit dem dritten unzufrieden) und bildet dann für die Gesamtbewertung der Dimension einen subjektiven Mittelwert. Bei diesem Teil der Studierenden gingen bei Verwendung kurzer "Alles-in-allem"-Fragebögen nicht nur wesentliche Informationen verloren, vielmehr würde dadurch auch der unzutreffende Eindruck einer einfachen, in sich widerspruchslosen Urteilsstruktur erweckt – methodologisch ausgedrückt: Es würde ein *Erhebungsartefakt* produziert.

Ähnlich 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 zeigt die komplexe Analyse 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 erst "Gestalt annehmen". Die isolierte Auszählung einzelner Variablen aber lässt solche Profile nicht 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 (s.o.), sondern insbesondere auch daran, dass es sich 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, produziert *Auswertungsartefakte*.

Fazit: Ein komplexer Sachverhalt kann angemessen nur durch hinreichend komplexe empirische Erhebungen valide abgebildet werden; und komplexe Interdependenzen im abzubildenden Sachverhalt werden erst durch hinreichend komplexe Analyseverfahren sichtbar. Die "ärgerlich komplizierte" soziale Realität wird nicht dadurch einfach, überschaubar und leicht handhabbar, dass man sich Scheuklappen anlegt. Lediglich die Gefahr, Wichtiges zu übersehen und fehlerhafte Entscheidungen zu treffen, wächst.

Der Verweis auf die o.g. Gefahren von Fehlschlüssen sollte allerdings nicht als Argument gegen die Verwendung von "Alltagsevaluationen" Betroffener missverstanden werden. Um diese als gültige Informationen nutzen zu können, muss jedoch im Zuge der Analyse das Kriteriensystem der Evaluierenden rekonstruiert werden. In aufwendigerer Form kann die Befragung von studentischen Veranstaltungsteilnehmern sogar als wertvolles Informationsinstrument zur Entwicklung von Lehrqualität genutzt werden. Auf ein Beispiel wird später (Abschnitt 4.1) noch eingegangen.

2.3 Umfrageforschung als Rückmeldung

Weitere Formen der Umfrageforschung in der Hochschule sind in Lehrveranstaltungen eingesetzte und von Didaktikern schon seit langem empfohlene (kürzere oder längere) Fragebögen als Instrument der *Rückmeldung an die Lehrperson*. Sie dienen nicht der Evaluation, sondern der *Kommunikation* über Lehre und sind vor allem in größeren Veranstaltungen hilfreich, in denen eine direkte Interaktion zwischen Lehrenden und Lernenden nicht mehr ohne weiteres möglich ist. Im Unterschied zu Befragungen *als* Evaluationsverfahren sind der Differenzierungsgrad und die methodische Qualität der Fragebögen ebenso wie die Form der Erhebung zweitrangig. Die Ergebnisse sollen der Lehrperson einen Eindruck von der Sichtweise der Teilnehmer vermitteln, *und* sie sollen der Ausgangspunkt für die Diskussion zwischen Lehrenden und Studierenden über die Lehre sein. Für diesen Zweck sind kurze Fragebögen mit durchaus auch groben Kategorien und mit zusammenfassend vorzunehmenden Bewertungen sogar von Vorteil: Sie bieten mehr Raum für die Interpretationsphantasie und damit auch mehr Ansatzpunkte für eine engagierte Diskussion. Eine lange Liste statistischer Werte, die den Eindruck erweckt, daran gebe es "nichts mehr zu deuteln", blockiert dagegen von vornherein jede Diskussion. Gute Dienste leisten z.B. an jeden Teilnehmer verteilte Blätter (DIN A 5 oder kleiner) mit folgendem Text:

Mini-Rückmeldung

Wenn Ihnen während des Seminarverlaufs etwas Anmerkenswertes auffällt, bitte sofort notieren (*Anmerkenswert* ist alles, was Sie stört, was Ihnen besonders gefällt, was Sie an Ideen für Veränderungen haben). Bitte immer nur eine Mitteilung pro Blatt!

Intensitätsskala: Bewertung:

0 1 2 3 4 5 + (oder) –

Eigentlich nicht so wichtig — Sehr bedeutsam

Regelmäßig eingesetzt und am Ende von Veranstaltungen eingesammelt, ist ein solches Mini-Rückmeldeinstrument selbst in großen Vorlesungen ein wertvoller Seismograph, der sofort anzeigt, was im Auditorium ansonsten vom Lehrenden unbemerkt abläuft. Wichtig ist jedoch bei jeder Form eingesetzter Rückmeldeinstrumente, dass sie tatsächlich für das Ingangsetzen einer Diskussion über die Lehre genutzt werden und nicht lediglich als eine lästige Pflicht erscheinen.

2.4 Andere Formen des Einsatzes von Befragungen und Erhebungen

Gegenstand der Befragung und Bewertung müssen nicht in jedem Fall Lehrveranstaltungen sein. Sinnvolle Fragestellungen können sich richten auf das Curriculum und die übergreifende *Studiensituation* im Fach (Institut, Fakultät/Fachbereich): Wie nehmen die Studierenden die durch Studien- und Prüfungsordnung vorgenommene Definition des Fachs wahr? Wie einleuchtend sind ihnen Struktur und Inhalte des Lehrangebots? Wird der Zusammenhang zwischen Lehre und Prüfungen als hinreichend erkannt? Wie wird die Betreuung empfunden? und vieles mehr. Ebenso sind aktuelle *Kenntnisse über die Studierenden* und ihre Art und Weise des Studierens für die Träger des Curriculums von Bedeutung: Unter welchen persönlichen Bedingungen und wie intensiv wird studiert (etwa Berufstätigkeit neben dem Studium, Anzahl der besuchten Veranstaltungen, zeitlicher Aufwand für das Studium)? Welche inhaltlichen Schwerpunkte setzen die Studierenden dort, wo sie Wahlmöglichkeiten haben? Welche Studierstile sind im Grund-, welche im Hauptstudium anzutreffen? Wie ist das Informationsverhalten der Studierenden? usw. In diesem Zusammenhang kann auch eine Vollerhebung der *Teilnehmerstruktur* in allen Veranstaltungen eines Semesters wichtige Informationen liefern: nicht nur darüber, ob und in welchem Maße eine Veranstaltung ihre definierte Zielgruppe tatsächlich erreicht, sondern auch darüber, ob und in welchem Ausmaß das Fach mit seinen Angeboten Dienstleistungen für andere Fächer liefert (für Nebenfächler, aber auch durch Teilnehmer anderer Fächer, die lediglich spezielle Angebote wahrnehmen und dort erbrachte Leistungen im eigenen Fach anerkennen lassen).

Befragungen müssen sich nicht lediglich an einen Querschnitt der aktuell Studierenden richten. Auch *spezifischere Auswahlen* können nützlich sein: Studienanfänger, Studierende

im Grundstudium vor der Zwischenprüfung, Studierende bei Beginn des Hauptstudiums, in der Examensphase. Darüber hinaus werden zunehmend *Absolventenbefragungen* durchgeführt, entweder zur ex-post-Evaluation des Studiums aus der späteren Perspektive von Berufstätigen und/oder als Verbleibstudien ehemaliger Studierender. Schließlich kommen auch *"Abnehmer"-Befragungen* vor, insbesondere mit dem Ziel, in potentiellen Berufsfeldern Profile von Anforderungen an das Qualifikationsprofil der Bewerber zu ermitteln. Nicht zuletzt – wenn auch (warum eigentlich?) ganz selten durchgeführt – könnten (und sollten) auch *die Lehrenden* eine Zielgruppe von Erhebungen sein. Eine Konfrontation der Wahrnehmung von Lehre und Lehrpersonen aus der Perspektive der Studierenden mit der Wahrnehmung der Studierenden und ihres Studienengagements durch die Lehrenden dürfte interessante Ergebnisse bringen.

3. Qualitätsentwicklung durch Evaluation

Evaluation von Studium und Lehre – wenn sie in der dem Gegenstand angemessenen Komplexität realisiert wird – ist zeit- und ressourcenaufwändig. Dieser Aufwand muss sich lohnen. Die Evaluation muss für die Beteiligten einen erkennbaren Nutzen bringen, soll sie auf die erforderliche Mitwirkungsbereitschaft treffen. Indes scheint eine Vielzahl von Evaluationen im Hochschulbereich eher punktuellen Charakter ohne nachfolgende Veränderungs-Konsequenzen zu haben, so dass deren Sinnhaftigkeit von den Beteiligten bezweifelt wird.

In der aktuellen hochschulpolitischen Diskussion aber wird von den Hochschulen gefordert, in kontrollierten Veränderungsprozessen die Qualität von Lehre und Studium zu verbessern. Dabei erscheinen Evaluationsverfahren als geeignetes Instrument, um Hinweise zu gewinnen, *wo* etwas verbesserungsbedürftig ist und *wie* es verbessert werden kann. Evaluation hätte in diesem Kontext also zunächst einmal die Funktion zu erfüllen, *Qualität* zu *"messen"*.[9] Selbst wenn man als letztliches Kriterium für die Leistungsqualität der Hochschule der "Qualifizierungserfolg" bei den Studierenden akzeptiert (s.oben: 1.3), bleibt doch die Aufgabe der "Messung" der *Qualität des Prozesses der Leistungserbringung sowie seiner Rahmenbedingungen* (hier also: der Qualität der Lehre und der Studienbedingungen) davon unberührt. Qualitätsentwicklung von Lehre und Studium wäre also abzubilden über – in der Sprache des Donabedian-Modells – Struktur- und Prozess-Indikatoren.

Für das Ziel Qualitätsentwicklung und/oder Qualitätssicherung ist allerdings allein mit dem Bereitstellen solcher Informationen noch nicht viel gewonnen. Informationen sind lediglich eine notwendige Voraussetzung dafür, gezielte Veränderungen dort in Gang zu setzen, wo der in Frage stehende Sachverhalt verbesserungsbedürftig und verbesserungsfähig erscheint. Somit wäre die Liste der im Abschnitt 1.5 aufgeführten Fragen um eine weitere zu ergänzen: Wer ist *Träger des Qualitätsentwicklungs-Vorhabens*? Anders formuliert: Wer ist verantwortlich dafür, dass die gelieferten Evaluations-Informationen in Handeln umgesetzt werden? Im allgemeinen wird die für die *Verwertung der Evaluationsbefunde* entscheidungsbefugte Instanz nicht auch das Tagesgeschäft der Umsetzung übernehmen können. Genauso wenig erfolgversprechend ist es, den Auftrag zur Qualitätsentwicklung

[9] Diskutiert werden in diesem Zusammenhang auch die Funktionen Kontrolle und Wettbewerb: 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 eine Art "Währung", die einen funktionierenden Wettbewerb (etwa um Reputation, aber auch um öffentliche Finanzmittel, um Forschungsförderung, sogar um besonders leistungswillige Studierende) ermöglichen und anregen könnte. Die wiederholt unternommenen Versuche, "Rankings" von Hochschulen, Hochschulfächern bis hin zu Lehrveranstaltungen zu erstellen, sind u.a. als Bemühung zu verstehen, Transparenz auf einem solchen Wettbewerbsmarkt zu schaffen.

pauschal an die einzelnen Akteuren im Prozess der Leistungserbringung (z.B. die Lehrenden und die Verwaltung) zu delegieren.

Eine mögliche Lösung könnte darin bestehen, das "Evaluationsprojekt" zu einem "Qualitätsentwicklungsprojekt" auszuweiten und mit den dafür notwendigen Ressourcen auszustatten. Durch die konkrete Aufgabe der Umsetzung empirisch gewonnener Befunde in Planungen und Empfehlungen wüchse die Einsicht in die Notwendigkeit differenzierten Vorgehens und entwickelte sich eine realistische Einschätzung des Nutzens unterschiedlicher Informations-Arten (pauschale Bewertungen als Ausgangspunkt für Meinungsbildungsprozesse und Diskussionen, "harte Fakten" und detaillierte Indikatoren als Basis für das Erkennen *konkreten* Veränderungsbedarfs und für die Ableitung *konkreter* Maßnahmen).

Nicht beantwortet ist aber allein mit der Installation eines Qualitätsprojektes noch nicht die zentrale Frage: Was *ist* eigentlich Qualität von Lehre und Studium? Eine *Qualität "alles in allem"* existiert offensichtlich nicht. Ein Sachverhalt kann zugleich in einer Hinsicht von ausgezeichneter Qualität, in anderer Hinsicht dagegen fehlerbehaftet sein. Es sind also verschiedene Aspekte *oder "Dimensionen" von Qualität* – in der Fachdiskussion des Qualitätsmanagements *"Kriterien"* genannt – zu unterscheiden. Zum anderen: Qualitätsaussagen sind Werturteile. Sollen sie intersubjektiv gefällt werden, sind anerkannte Vergleichsmaßstäbe – Fachausdruck: *"Standards"* – notwendig. Mit der Festlegung, aus wessen Perspektive Kriterien und Standards ausgewählt und formuliert werden, wird aber eine wesentliche Weichenstellung vorgenommen. Diese muss so legitimiert sein, dass alle Beteiligten sie akzeptieren können.

Soll also als Ausgangspunkt für den Entwicklungsprozess die Qualität des in Frage stehenden Sachverhalts "gemessen", d.h. intersubjektiv gültig abgebildet werden, ist das Qualitätskonzept in einer dem Gegenstand angemessenen Weise präzise zu definieren und sind die anzulegenden Kriterien und Standards durch geeignete, gültige Indikatoren zu operationalisieren.

Um es an einem einfachen Beispiel zu veranschaulichen: Zu beurteilen sei die Qualität von Autoreifen. Als *Qualitätskriterien* kämen wesentliche Eigenschaften des Objekts selbst in Frage. *Ein* Qualitätskriterium wäre etwa die Haltbarkeit des Produkts, gemessen an der Laufleistung in Kilometern; ein bei der Beurteilung anzulegender *Standard* könnte lauten: mindestens 30.000 km auf glatten Straßen. Andere Kriterien könnten sein: die Bodenhaftung (auf trockener sowie auf nasser Straße), die Sicherheit (bei Überbeanspruchung sowie bei Außeneinwirkung) u.ä. Auch dazu sind messbare Standards und zuverlässig durchführbare Qualitätstests relativ leicht definierbar.

Nicht so problemlos einlösbar ist diese Forderung beim "Sachverhalt Lehre", dessen Merkmale nicht als Eigenschaften des "Objekts" direkt ablesbar und in diesem Sinne "objektiv" messbar wären. Im Unterschied zu gegenständlichen Produkten – wie dem o.g. Autoreifen – ist Lehre eine *Dienstleistung*, deren Produkt (Lernservice für Studierende) sich erst in der Interaktion von Lehrenden und Lernenden herstellt. Bemühungen, die Qualität von Lehre kontextunabhängig verbindlich zu definieren, sind somit von vornherein zum Scheitern verurteilt. Qualität ist hier keine "*objektive*", dem Gegenstand (dem "*Objekt"*) zurechenbare, sondern eine relationale Eigenschaft. Wo dennoch der Versuch unternommen wird, Merkmale "guter Lehre" aufzulisten, setzt dieser – unabhängig vom Lehr-Inhalt – an der didaktischen Oberfläche an (Webler 1991, S. 246)[10]; und selbst da fällt es schwer, Einigkeit über einen Kriterienkatalog für "gute Didaktik" zu erzielen. Für Einführungsveranstaltungen mit Pflichtcharakter, in denen ein bei Studierenden eher unbeliebter Stoff vermittelt werden

[10] Oder es werden recht abstrakte und damit kaum intersubjektiv prüfbare "erfolgsrelevante Persönlichkeitsmerkmale der Lehrenden" genannt (ders., S. 247 f.).

soll, wird eine andere Didaktik angemessen sein als in Hauptstudienseminaren zu Spezialthemen mit ausschließlich freiwillig teilnehmenden und interessierten Studierenden oder als in Trainings zur Vermittlung fachübergreifender Schlüsselqualifikationen – um nur wenige unterschiedliche Lehr-Lern-Situationen zu benennen. Und welche Didaktik in diesen Situationen jeweils als angemessen gelten kann, dürfte von verschiedenen Lehrenden ebenso unterschiedlich eingeschätzt werden wie von Studierenden ohne oder mit Vorkenntnissen, ohne oder mit Leistungsmotivation, mit passiv-konsumierendem oder mit aktiv-entdeckendem Lernstil. Eine rein formale Definition – als Qualität der Darbietung – geht jedoch auch *prinzipiell* am Ziel der "Dienstleistung Lehre" vorbei. Lehre soll ja nicht stromlinienförmig nach Rezeptbuch abgespult werden, ihr Ziel ist auch nicht lediglich das Sich-Wohlfühlen oder die gute oder gar spannende Unterhaltung der Teilnehmer von Lehrveranstaltungen. Sie soll vielmehr Anregungen, Orientierung und – wo nötig – auch Anstöße zum aktiven Studieren geben. Ihr Ergebnis kann nicht in "Einschaltquoten" oder Zufriedenheits-Kennziffern gemessen werden.

Es bleibt nur der Ausweg *relativer* Qualitätsdefinitionen, wie dies in der Diskussion um Qualitätsentwicklung und Qualitätssicherung von Dienstleistungen geschieht. Für Ingenieurwissenschaftler liegt es nahe, auf Qualitätsdefinitionen aus der Industrie zurückzugreifen und sie analog auch für die Organisation Hochschule anzuwenden (z.B. Weule 1999). So findet sich etwa in der DIN/ISO-Norm 8402 eine inhalts- und ergebnisbezogene Definition: "Qualität ist die Beschaffenheit einer Einheit bezüglich ihrer Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen." Für welche Zwecke die Leistung geeignet sein soll, welche und wessen Erfordernisse festzulegen und vorauszusetzen sind, müsste demnach zunächst ermittelt werden, bevor Evaluation und Qualitätsentwicklung beginnen können. Qualität der Lehre – so ist bis jetzt zu resümieren – kann nicht adressatenunabhängig, sondern kann nur zielgruppenorientiert bestimmt und realisiert werden. Von Studienanfängern und Fortgeschrittenen, von gegenwärtig Studierenden und künftigen Absolventen, von Arbeitgebern und fachwissenschaftlicher community werden unterschiedliche, teils sogar gegensätzliche Erfordernisse geltend gemacht. Die Vorstellung von Lehre als Dienstleistung hat konsequenterweise zur Übernahme des oben bereits genannten Begriffs der Kundenorientierung in die Qualitätsdiskussion geführt – hier allerdings nicht in Analogie zum Wettbewerbsmarkt, sondern als Bezugspunkt für die Definition von Leistungsanforderungen. Soll Lehre ihrem Charakter als Dienstleistung gerecht werden, kann somit ihre Qualität und können Qualitätskriterien nicht extern (von wem auch immer) und auch nicht ein für allemal festgesetzt werden, sondern sie müssen den jeweiligen Gegebenheiten angepasst und – wo keine direkte Marktabstimmung durch Angebot und Nachfrage wirksam wird – zwischen den Beteiligten "ausgehandelt" werden. Dies findet seinen Niederschlag in einem weiteren, an den DIN/ISO-Normen orientierten Definitionsversuch: "Qualität ist die Erfüllung der gemeinsam (Kunde – Lieferant) vereinbarten Anforderungen – einschließlich der Erwartungen und Wünsche" (Rühl 1998, S. 22). Die Grundtendenz dieser Definition aus dem Produktionsbereich wird inhaltlich auch auf das Qualitätsmanagement von Dienstleistungen übertragen (DIN/ISO 9001 sowie 9004/2, wo als Anwendungsfall ausdrücklich u.a. auf die Wissenschaft verwiesen wird; ausführlicher dazu Stock 1994).

Wenn also Qualität von Dienstleistungen nicht absolut, sondern nur relativ definierbar ist – nämlich relativ zu den Adressaten (oder "Kunden") der Dienstleistung –, dann kann Qualität auch nicht (im Sinne strukturtreuer Abbildung einer Objekteigenschaft) "gemessen", sondern nur zwischen den Beteiligten "ausgehandelt" werden. Dementsprechend hat der Träger eines Qualitätsentwicklungs-Projekts zu entscheiden und zu begründen, *für welche Zielgruppe* die Dienstleistung optimiert werden soll. Dies bedeutet immer zugleich eine Entscheidung *gegen* andere potentielle Adressaten. Der Versuch, einem imaginären ‚Durchschnitt' heterogener

Zielgruppen mit heterogenen Bedürfnissen und Ansprüchen gerecht zu werden, führt nahezu zwangsläufig zu dem Resultat, dass die Leistung für keine Gruppe von großem Nutzen ist.

4. Qualität entwickeln, ohne Qualität zu "messen" – einige Beispiele

Im Abschnitt 2.1 war resümiert worden, im Kontext Studium und Lehre an der Hochschule sei das speziell für die Evaluierung entwickelte methodische Design – das Konzept der (experimentellen oder quasi-experimentellen) Programmforschung – nicht einsetzbar. Im Abschnitt 3 wurde des weiteren festgestellt, auch das "Messen" von Qualität als Aufgabe von Evaluation sei nicht einlösbar. Es wäre jedoch falsch, daraus die Empfehlung herzuleiten, in der Hochschule von dem Vorhaben generell Abstand zu nehmen. Angezeigt ist lediglich der Verzicht auf diese *Formen* von Evaluation. Es wurden ja auch bereits einige Beispiele skizziert, wie dennoch sinnvoll und erfolgversprechend auch in der Hochschule evaluiert werden kann (Kapitel 2), allerdings *nicht* verstanden als *spezifisches methodologisches Konzept*, sondern als empirische *Sozialforschung in einem spezifischen Verwertungskontext*. Damit ist der Evaluation explizit die Aufgabe zugewiesen, zur Entwicklung und Verbesserung von Qualität durch Bereitstellung einer differenzierten und qualitätsrelevanten Beratungs- und Entscheidungsbasis beizutragen. Evaluation ohne Konsequenzen ist nutzlos.

So klagen Künzel/Nickel/Zechlin, die über Erfahrungen aus einem Organisationsentwicklungsprojekt berichten: "Evaluieren ist schon schwierig genug, aber noch eine viel größere Herausforderung stellt die Veränderung der Realität an Hochschulen dar. Solange man nur Datenmaterial zur Qualität von Lehre und Forschung auf geduldigem Papier zusammenstellt und darüber in den Gremien redet, bleibt die Evaluation folgenlos und damit harmlos. Unbequem wird es erst, wenn aus dem vermeintlichen Datenfriedhof Konsequenzen auf der Handlungsebene gezogen werden. (...) Ob und was auf der Handlungsebene dann tatsächlich passiert, ist die Messlatte für den Erfolg einer Evaluation." (1999, S.105).

In ihrem Beitrag schildern die Autoren konkret den Ablauf eines solchen Entwicklungsprojekts – von der "Stärken-Schwächen-Analyse als Basis von Veränderungsprozessen" über die "Erstellung eines Handlungskatalogs", über "Zielplanung und Zielvereinbarung" bis zu Problemen in der "Leitungs- und Entscheidungsstruktur" (S.107-113). Mit Verweis darauf wird an dieser Stelle auf ein diesbezügliches Beispiel verzichtet. Statt dessen wird im folgenden etwas ausführlicher ein Anwendungsbeispiel auf der Mikroebene – die Aufgabe "Qualitätsentwicklung in Lehrveranstaltungen" – illustriert.

4.1 Beispiel: Zielgruppenorientierte Lehre

Es wurde bereits dargestellt, dass Lehrqualität sinnvoll nur relational – als Angemessenheit des Angebots (der Lehrenden) für definierte "Kunden" (Studierende) – entwickelbar ist. *Lehre* kann – wie sehr sie auch einer "best practice" didaktischer Kunst folgen mag – immer nur in begrenztem Ausmaß *Lernen* bewirken. Ob sie den Lernprozess und das Lernergebnis positiv beeinflusst oder erfolglos bleibt, hängt nicht in erster Linie von der didaktischen Qualität der Darbietung ab, sondern ist das Resultat der gesamten Lehr-Lern-Situation. Damit sind vielfältige – in wechselseitiger Beziehung stehende – Dimensionen angesprochen, u.a.

- das Lehr- bzw. Ausbildungsprogramm / der Studiengang
  (zu vermittelnde "Inhalte" – also Kenntnisse, Fertigkeiten und Fähigkeiten – und deren Stellenwert im Gesamtcurriculum)
- die "Lehrmittel"
  (Lehrpersonal, Medien, Aufgaben, Betreuung/Beratung)

- die "Lerner"
  (übergeordnete Studienziele, Lernziele in der Veranstaltung, Teilnahmegründe, Interesse/Motivation, Vorkenntnisse, Lern-Erfahrungen, Lern- und Arbeitsstile, Stellenwert des Studierens gegenüber anderen Tätigkeiten und Verpflichtungen, Zeitbudget zum Lernen für diese Veranstaltung)
- die materielle Lernumgebung
  (Lernort, Ausstattung, zeitliche/räumliche Flexibilität, Teilnehmerzahl)
- die soziale Lernumgebung
  (Einzel-/Gruppenlernen, Lern-"Klima", Stellenwert des Studiums im sonstigen Lebenskontext, Kommiliton(inn)en, Familie/Bekanntenkreis).

Nur ein kleiner Teil der genannten Dimensionen ist vom Lehrenden *gestaltbar*. Alle aber müssen für einen gelingenden Lernprozess *bekannt sein* und bei der Lehr-/Lernplanung *berücksichtigt werden*. D.h. die gestaltbaren Faktoren der Lernumwelt sind so auf die nicht-veränderbaren Faktoren abzustimmen, dass ein in sich stimmiges, situationsangepasstes Lernarrangement zustande kommt. Die Konsequenz aus dieser Forderung ist, dass jede Lehr-/Lernplanung immer zumindest ein Minimum an Vorab-Informationen verlangt.

Von diesen Überlegungen ausgehend, wurde (und wird) am Institut für Soziologie (IfS) der Freien Universität Berlin ein Projekt mit dem Ziel der Lehrqualitäts-Entwicklung für *interessierte* Studierende (als explizite Zielgruppe[11]) durchgeführt. Für wiederkehrende Lehrveranstaltungen, die zum Pflichtkanon des teilweise neu zu konzipierenden Diplom-studiengangs gehören, sollen für diese Zielgruppe in den jeweiligen Veranstaltungen "Lösungen nach Maß" gefunden werden. Dies erfordert hinreichende Informationen über die Teilnehmer; in Teilnehmerbefragungen können sie ermittelt werden.

Allerdings ist Qualitätsentwicklung und Qualitätssicherung nicht auf der Basis grober, globaler, vereinfachender Informationen möglich. Benötigt werden detaillierte Daten über die "Kunden", für die die Dienstleistung entwickelt werden soll; hier: Lernziele und Lern-voraussetzungen der Studierenden, Erwartungen, Ansprüche, natürlich deren Einschätzungen und Urteile. Auf all dies muss eine Lehrperson eingehen können, wenn sie eine Lehr-veranstaltung zielgruppenorientiert und adressatengerecht konzipieren und durchführen will. Dabei kann es sich herausstellen, dass die Erwartungen der Teilnehmer zueinander in Widerspruch stehen, nicht "unter einen Hut" zu bekommen sind, oder dass die Teilnehmer-erwartungen mit den Absichten und Zielen der Lehrperson oder der Lehrinstitution in Widerspruch stehen. In solchen Fällen kann (und soll) das Ergebnis in der Lehrsituation thematisiert werden, muss entschieden werden, welchen Erwartungen entsprochen werden kann und welchen nicht; dann haben auch die Teilnehmer die Chance, sich zu entscheiden, ob sie dennoch weiter teilnehmen wollen oder ob sie lieber wegbleiben.

Qualitätsentwicklung (und Qualitätssicherung) ist jedoch kein punktuelles, lediglich einmal stattfindendes Vorhaben, sondern zwangsläufig ein Prozess, der einige Zeit braucht und einige Durchläufe benötigt. Von daher ergibt sich: Es ist ein Konzept, das *nicht in jedem Semester in jeder Veranstaltung* verfolgt werden kann. Es eignet sich eher für ein auf Wiederholung angelegtes Lehrprogramm: Einführungsveranstaltungen im Grundstudium, regelmäßig wiederkehrende Bestandteile des Hauptstudiums-Curriculums. In anderen Situationen ist der Lehrprozess im Idealfall im direkten oder durch schriftliche

[11] Jede/r Lehrende wird die Erfahrung gemacht haben, dass ein nicht unerheblicher Teil der Immatrikulierten nicht aus intrinsischem Interesse am Fach studiert und daher auch in manchen Lehrveranstaltungen nur anwesend ist, um den formalen Anforderungen der Studienordnung zu genügen. Studierendenbefragungen bestätigen diesen Eindruck; von über 10.000 befragten Hörern in Vorlesungen mehrerer Fakultäten der Ruhr-Universität Bochum waren 55,6 % (!) ausschließlich deshalb in der Veranstaltung, weil es die Studien- oder Prüfungsordnung zwingend vorschrieb (s. auch weiter unten: Abschnitt 4.3). Es wäre wenig sinnvoll, wollte man "Lehrqualität" für diese Personengruppe optimieren.

Rückmeldeinstrumente unterstützten Diskurs (s.oben: 2.3) zwischen Lehrenden und Lernenden “auszuhandeln”.

Bestandteile des am IfS verfolgten Konzepts sind im Kern zwei schriftliche Befragungen (eine zu Beginn, eine abschließende gegen Ende des Semesters):

Die *Anfangsbefragung* der Teilnehmer (in der zweiten Semesterwoche) erhebt für jede Lehrveranstaltung folgende Schwerpunkte:

- Gründe für die Teilnahme an dieser Veranstaltung; Informationsquellen, aufgrund derer die Wahl getroffen wurde;
- Beschreibung der Studiensituation der Befragten: zeitliche Belastung durch Jobs und andere Verpflichtungen außerhalb des Studiums (etwa Familie, Ehrenämter), geplante zeitliche Investition in diese Veranstaltung, beabsichtigte Lernform (Einzelarbeit, Gruppe, Besuch ergänzender Veranstaltungen, Selbststudium);
- persönliche Lernziele, an die Veranstaltung geknüpfte Erwartungen und Befürchtungen;
- die persönliche "Didaktik-Theorie" der Teilnehmer (Einschätzung der Wichtigkeit bestimmter didaktischer Aspekte wie Diskussion, Skript, Medieneinsatz, Wiederholungen, Lernfortschrittstests etc.).

Schwerpunkte der *Abschlussbefragung* der Teilnehmer (in der vorletzten Veranstaltungswoche) sind:

- Wiederaufgreifen der Themenbereiche "Erwartungen / Befürchtungen" und abschließende (rückblickende) Beurteilung, ob sie sich bestätigt haben bzw. ob es besser oder schlechter gelaufen ist;
- Wiederaufgreifen des Themas "Lernformen": tatsächlich in die Veranstaltung investierter Arbeitsaufwand; Beeinträchtigung des Studiums durch außeruniversitäre Verpflichtungen;
- Wiederaufgreifen der didaktischen Aspekte und deren rückblickende Beurteilung für die abgelaufene Veranstaltung; Beurteilung der Person und des Auftretens der/des Lehrenden sowie des wahrgenommenen eigenen Lernerfolgs;
- Informationen über Teilnehmerfluktuation (eigener Wechsel der Lehrveranstaltung und dessen Gründe; Berichte über Gründe von Bekannten, die aus der Veranstaltung weggeblieben sind).

Die Anfangsbefragung wird bis zur vierten Semesterwoche ausgewertet und in die Veranstaltung rückgekoppelt (Diskussion mit den Teilnehmern). Auf der Basis der gewonnenen Informationen kann noch im laufenden Semester das Angebot und seine Darbietung an die Zielgruppe angepasst werden (bzw. bei nicht beeinflussbaren Rahmenbedingungen kann dies mit den Teilnehmern erörtert werden). Die Endbefragung (einschließlich der Evaluationen durch die Teilnehmer) ergibt Informationen für die längerfristige Planung des Veranstaltungstyps.

Geplant waren in diesem Projekt zusätzlich Gruppendiskussionen mit den Teilnehmern im Folgesemester, sobald die Gesamtauswertungen vorlagen. Diese kamen jedoch nicht zustande. Das Interesse der Studierenden, sich mit der Lehre auseinanderzusetzen, erweist sich schon bei der konkreten Rückmeldung in die laufende Veranstaltung als enttäuschend gering. Sich im Nachhinein noch einmal mit einer bereits absolvierten Veranstaltung abstrakt (mit dem Blick auf die künftige Gestaltung des Veranstaltungstyps für spätere Adressaten) auseinanderzusetzen – dazu fehlt offenbar die Motivation. Dagegen ist die Akzeptanz der nur punktuell auszufüllenden Rückmelde-Fragebögen bei den Studierenden relativ hoch: Mehr als 75 % plädieren für regelmäßige Wiederholungen solcher Befragungen. Allerdings wurde schon ab dem zweiten Semester auch vermehrt Unmut über die “lästige Befragerei” geäußert. Doch selbst wenn sich auf die Dauer - im Zuge einer “Veralltäglichung” des Ansatzes - die Beteiligung auf einem niedrigeren Niveau einpendelt, sollte das kein Grund zur Resignation sein. Wenn mit Bemühungen um eine qualitative Verbesserung des Studienservice Lehre nicht ein “Marktanteil” von 100 % erreichbar ist, reicht dies nicht als Argument für eine nur

mittelmäßige Qualität der Dienstleistung Lehre aus. Für die Zielgruppe der Studieninteressierten lohnt sich die Mühe.

Das gleiche Ziel – Qualitätsverbesserung der Lehre in Veranstaltungen durch einen formalisiert geförderten Dialog zwischen Lehrenden und Lernenden – verfolgt ein Vorhaben in einem Reformprojekt "Intensivstudium Psychologie" am Fachbereich Erziehungswissenschaft und Psychologie der Freien Universität Berlin. In diesem Konzept – "dialogische Evaluation" genannt – wird der Kommunikation zwischen den am Lernprozess Beteiligten ein noch höheres Gewicht beigemessen als im oben geschilderten Verfahren. Die Erfahrungen werden in einem jetzt vorliegenden Werkstattbericht dokumentiert (Knäuper, Kroeger u.a. 1999).

4.2 Beispiele evaluationsverwertbarer Informationen aus Befragungen in Lehrveranstaltungen

Erhebt man studentische Urteile in Lehrveranstaltungen und möchte die Bewertungen als Qualitätsindikatoren interpretieren (z.B. durch Berechnung von Mittelwerten pro Veranstaltung), dann wird man in der Regel mit einem "störenden" Resultat konfrontiert: Die Studierenden sind sich nicht einig. Was einer Gruppe von Teilnehmern als überzeugende Lehre erscheint, kritisieren andere als absolut untauglich; wieder andere urteilen "teils / teils" – und alles in derselben Veranstaltung, die doch "objektiv" für alle identisch ist (s. oben: 2.2 und im Detail Kromrey 1994/1995b). Analysiert man die Fragebögen im Hinblick auf die Teilnehmerstruktur in den Veranstaltungen, dann findet man parallel zur *Heterogenität der Urteile* eine entsprechende Vielfalt von Interessenlagen und Teilnahmegründen, von Erwartungen und Befürchtungen, von Zugehörigkeiten zu Studienphasen und Studiengängen, manchmal sogar zu Fachbereichen/Fakultäten. Die feststellbare Heterogenität wird nicht nur von manchen Lehrenden erheblich unterschätzt (sie haben schließlich ihre Veranstaltung für eine ganz bestimmte Zielgruppe eines bestimmten Studiengangs angekündigt). Auch studentische Interessenvertreter verfallen in den gleichen Fehler, wenn sie sich für *die* (vermeintlichen) Interessen *der* Studierenden einsetzen.

Möchte man die in solchen Kontexten erzielten Befragungsergebnisse *als* Evaluation nutzen, so ist diese *Heterogenität subjektiver Bestimmungsgründe der Urteilsfindung* (und natürlich der Urteile selbst) ein Indiz für die äußerst zweifelhafte Gültigkeit der Datenbasis. Dies ist jedoch nicht gleichbedeutend damit, dass die Informationen selbst von geringem Wert seien. Im Gegenteil: Sie informieren über wesentliche Randbedingungen für die Lehre und können für die inhaltliche und formale Gestaltung außerordentlich wichtige Anregungen geben, z.B.: Orientierung an vielfältigen Beispielen, bei Übungen Aufteilung in homogene Teilgruppen, bei Gruppenarbeit bewusste Diskussion zwischen unterschiedlichen (jeweils homogenen) Teilgruppen. Heterogenität macht die Lehre schwerer, kann aber durchaus auch positiv als bewusst eingesetztes didaktisches Prinzip genutzt werden. Auch für die Lehrangebots-*organisation* ist eine konkrete Schlussfolgerung naheliegend: Parallelangebote der betreffenden Veranstaltung mit jeweils unterschiedlicher Didaktik; etwa: eine Veranstaltung mit stärkerer Betonung der Eigenaktivität der Studierenden ("betreutes Selbstlernen"), eine andere im "herkömmlichen" Stil für rezeptives Lernen (Vorlesung mit Übung). Wo die personelle Kapazität nicht zur Duplizierung des Angebots ausreicht, kann eine solche Variation über die Semester verteilt stattfinden.

Ein weiterer Auswertungsbefund aus Lehrveranstaltungsbefragungen ist frappierend und wird selten berücksichtigt: der *Einfluss der studentischen "peers"* auf die Wahrnehmung und Beurteilung der Lehre, insbesondere derjenigen im selben Hörsaal. Wird eine Veranstaltung überwiegend von desinteressierten Studierenden besucht, werden *alle* Teilnehmer negativ beeinflusst. Im Gegenzug findet sich ein deutlicher "positiver Ansteckungseffekt", sofern die Mehrheit der Teilnehmer Interesse zeigt. Dies gilt für Urteile über die Lehrdarbietung ebenso wie für die Selbsteinschätzung des eigenen Lernerfolgs oder die Bereitschaft, sich mit dem

behandelten Stoff intensiver selbstständig zu beschäftigen (dazu im Detail Kromrey 1994). In Veranstaltungen mit negativer Grundstimmung hat die Lehrperson – wie Befragungen belegen – kaum Chancen, durch eigene Bemühungen "Interesse zu wecken". Eine genauere vergleichende Analyse über eine Vielzahl von Veranstaltungen hinweg zeigt jedoch: Eine ähnliche Wirkung wie das von Studierenden mitgebrachte eigene *Interesse* am Stoff hat die Einsicht in den *Nutzen* für das weitere Studium oder für den späteren Beruf. Und diese wiederum lässt sich im Kontext des Gesamtfachs durch ein in sich stimmiges und für die Studierenden nachvollziehbares Curriculum ebenso fördern wie durch Studienberatung und Orientierungsangebote über die Berufspraxis. Auch dies ein Beispiel dafür, dass die Feststellung einer vermeintlichen "Störgröße" zu einer wichtigen Information für die Planung werden kann.

4.3 Beispiele evaluationsverwertbarer Informationen aus Befragungen zur Studiensituation in Fachbereichen/Fakultäten

Wie in einzelnen Lehrveranstaltungen, so können Studierende auch in Fachbereichen / Fakultäten oder Studienfächern befragt werden: über ihre soziale Herkunft, ihre Gründe der Wahl von Studienfach und -ort, über Studienmotive, Arbeitsbelastung, Zeitbudget, Berufspläne, die Transparenz von Studienanforderungen, über Einschätzungen des Lehrangebots sowie der Betreuungs- und Prüfungssituation, über die wahrgenommene Abstimmung zwischen Lehr- und Prüfungsinhalten u.ä.m. Und wie in einzelnen Lehrveranstaltungen, ist auch in den meisten Fächern die große Heterogenität ein ins Auge fallendes Charakteristikum der Resultate.

So ergab etwa an der Fakultät für Sozialwissenschaft der Ruhr-Universität Bochum die Vollerhebung aller anwesenden Studierenden in den Lehrveranstaltungen einer Stichtagswoche die folgende Verteilung der Studienziele:

- 44 % mit Studienziel Diplom-Sozialwissenschaft,
- 14 % Lehramt mit sozialwissenschaftlichem Fach,
- 22 % sozialwissenschaftliches Nebenfach in Studiengängen anderer Fakultäten,
- 20 % Studierende anderer Fakultäten ohne sozialwissenschaftliches Fach (in manchen Lehrveranstaltungen machten die Nicht-Sozialwissenschaftler sogar die Mehrheit aus).

Die offizielle Studienplanung orientiert(e) sich – natürlich – weitestgehend am Diplomstudiengang Sozialwissenschaft, also an der Minderheit. Die Mehrheit der anwesenden Studierenden dagegen orientierte ihr Nachfrageverhalten an anderen Motiven und verfolgte anders gelagerte Ziele als das für die Angebotsplanung maßgebliche Leitbild.

Um bei derselben Fakultät zu bleiben: Selbst deren eigentliche Zielgruppe – also die Diplom-Studierenden – hatte alles andere als homogene Studienmotive.

Schon die Information über die Gründe für die Wahl des Fachs überraschte: Nur für eine Minderheit galt – wie sich herausstellte – dieses Fach als "erste Wahl"; für die Mehrheit war es eher eine Übergangs- oder eine Notlösung:

- Rund 30 % woll(t)en eigentlich ein anderes Fach studieren; 22 % konnten ihr Wunschstudium – vor allem Publizistik und Psychologie – wegen Zulassungsbeschränkungen nicht (sofort) realisieren und benutzten die Sozialwissenschaft als "Parkfach"; 8 % hatten andere Gründe.
- Weitere rund 30 % hatten zunächst ein anderes Fachstudium begonnen und es dann abgebrochen (Hauptgründe: zu schwer, zu arbeitsintensiv; Sozialwissenschaft bot sich in räumlicher Nähe als Unterschlupf an, um nicht die Universität ganz verlassen zu müssen.
- Nur rund 40 % hatten die Sozialwissenschaft von vornherein und mit der Absicht des Fachstudiums gewählt!

So kann es nicht verwundern, dass sich in der Gesamtheit der Studierenden die unterschiedlichsten Motivations-Profile repräsentiert fanden. Eine Clusteranalyse der Studiengründe, Abschlussziele und Berufsvorstellungen ergab unter den Befragten neun gut unterscheidbare Gruppen, die von einem breiten Studieninteresse mit Orientierung an Wissenschaft und Beruf über ein strikt berufsorientiertes Ausbildungsinteresse und dem Gegenstück Bildungsstudium bis zu Motivkonstellationen reicht wie Studium als Lebensstil oder Immatrikulation zur zeitlichen Überbrückung bis zum Wechsel auf einen Arbeitsplatz, für den ein Studienabschluss nicht erforderlich ist.[12]

Es dürfte unmittelbar erkennbar sein, dass eine Lehrangebotsplanung, die *allen* Studierendengruppen gleichzeitig gerecht werden könnte, schlechterdings unmöglich ist. Selbst wenn die nicht im eigentlichen Sinne studierwilligen Immatrikulierten außer Betracht bleiben, existieren höchst unterschiedliche bis gegensätzliche Anforderungsprofile. Will eine Fakultät / ein Fachbereich die von hochschulpolitischer Seite erhobene Forderung einlösen, Studiengänge zu konzipieren und zu realisieren, die innerhalb einer angebbaren Zahl von Semestern (Regelstudienzeit) "studierbar" sind, dann wird sie unter der gegenwärtigen Bedingung knapper Kassen nicht das gesamte Spektrum von Anforderungsprofilen der *Studierenden* (wobei dieser Begriff hier im wörtlichen Sinne gemeint ist, also die eher Studierunwilligen bewusst ausklammert) bedienen können. Das im vorigen Abschnitt (4.2) formulierte Postulat, dass die Entwicklung von Lehrqualität in einzelnen Lehrveranstaltungen jeweils nur zielgruppenorientiert erfolgen könne, gilt in gleicher Weise für das Curriculum und das Studienangebot eines ganzen Faches. Dies geschieht natürlich seit jeher, und zwar durch Orientierung an einem (expliziten oder zumindest impliziten) idealtypischen Leitbild: breites Studieninteresse auf Seiten der Nachfrager, Berufsqualifizierung durch Wissenschaft auf Seiten des Angebots. Befragungen wie die hier als Beispiel angeführten liefern allerdings empirisch fundiertes Wissen über die "Nachfrageseite" und schützen vor Fehlentscheidungen bei Planungen, die ausschließlich nach bestem Gutdünken "am grünen Tisch" erfolgen.

5. Literatur

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

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

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

Knäuper, Bärbel; Kroeger, Matthias und Studierende, 1999: Qualitätssicherung und –verbesserung im Intensivstudium Psychologie: Ein Werkstattbericht zur Lehrevaluation, Berlin: FU Studiengang Psychologie (der Bericht ist auf der Webseite http://userpage.fu-berlin.de/~sciencec/iStudium/ einsehbar).

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

[12] Ganz ähnlich war das Resultat in den anderen 12 Fakultäten, in denen entsprechende Erhebungen durchgeführt wurden. Ausnahmen bildeten lediglich stark "verschulte" Fächer (hier z.B. Bauingenieurwesen) oder solche mit einer "verbindlicheren" Fachidentität (hier z.B. Theologie).

Kromrey, Helmut, 1994: Wie erkennt man “gute Lehre”? Was studentische Vorlesungsbefragungen (nicht) aussagen. In: Empirische Pädagogik, Jg. 8, H. 2, S. 153-168

Kromrey, Helmut, 1995a: Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. In: ZSE Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, Jg. 15, Heft 4, S. 313-336

Kromrey, Helmut, 1995b: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: Peter Ph. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung, Münster, New York, 2. Aufl., S. 105-127

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

Kromrey , Helmut, 1999: Von den Problemen anwendungsorientierter Sozialforschung und den Gefahren methodischer Halbbildung. In: SuB Sozialwissenschaften und Berufspraxis, Jg. 22, H. 1, S. 58-77

Künzel, Ellen; Nickel, Sigrun; Zechlin, Lothar, 1999: Organisationsentwicklung an Hochschulen. Was geschieht mit den Evaluationsergebnissen? In: Hochschulrektorenkonferenz (Hg.): “Viel Lärm um nichts?” Evaluation von Studium und Lehre und ihre Folgen. Beiträge zur Hochschulpolitik 4/1999, Bonn: HRK, S.105-119

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, S. 7-13

Patton, Michael Q., 1997: Utilization-focused evaluation. 3rd ed., Thousand Oaks, CA, London

Rühl, Werner J., 1998: ISO 9000 – Erfahrungsbericht aus einem technischen Entwicklungszentrum. In: Hochschulrektorenkonferenz: Qualitätsmanagement in der Lehre. TQL 98. Beiträge zur Hochschulpolitik 5/1998, Bonn: HRK, S. 21-46

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

Stock, Wolfgang G., 1994: Wissenschaftsevaluation. Die Bewertung wissenschaftlicher Forschung und Lehre. ifo Diskussionsbeiträge 17, München: ifo Institut für Wirtschaftsforschung

Webler, Wolff-Dietrich, 1991: Kriterien für gute akademische Lehre. In: Das Hochschulwesen, Jg. 39, Heft 6, S. 243-249

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

Weule, Hartmut, 1999: Praktische Probleme der Qualitätssicherung an Hochschulen. In: Hochschulrektorenkonferenz (Hg.): Ein Schritt in die Zukunft. Qualitätssicherung im Hochschulbereich. Beiträge zur Hochschulpolitik 3/1999, Bonn: HRK, S. 45-54

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

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