zur Veröffentlichung in:
Zeitschrift für
Pädagogische Psychologie (1996)

HELMUT KROMREY

Qualitätsverbesserung in Lehre und Studium statt sogenannter Lehrevaluation

Ein Plädoyer für gute Lehre und gegen schlechte Sozialforschung

0. Vorbemerkungen

Das Image der Professoren als Lehrende an deutschen Hochschulen ist schlecht. "*Bodenlos* schlecht" sei die Lehre, heißt es sogar mancherorts (vgl. Mußgnug 1992). Ein journalistisches Forum, das dieses Image gern verbreitet, ist etwa der SPIEGEL, wo zum Beispiel zu lesen war, deutsche Hochschullehrer seien "unfähig, gut zu lehren". In den Universitäten treiben danach "Lehrflaschen" ihr Unwesen, "demotivieren ungezählte junge Menschen" und richten so "einen kaum wieder gutzumachenden Schaden an"? (Der Spiegel 1993, 86). Aber auch das Image der Studenten ist nicht das beste. Mängel zuhauf werden ihnen nachgesagt: Studierunfähigkeit, lediglich "dumpfes Absolvieren" formaler Anforderungen, große Wissenslücken, fehlende Sprach- und Kulturfähigkeiten (s. Stephan 1992; Deutscher Hochschulverband 1992; Otten 1992).

Welche Seite hat nun recht? Nach aller Alltagserfahrung ebenso wie nach empirischen Forschungsbefunden wohl beide - glücklicherweise aber nur für jeweils eine Minderheit unter den angegriffenen Personengruppen. Es ist deshalb nicht ein hoffnungs- und zweckloses Unternehmen, sich um bessere Qualität von Lehre *und* Studium zu bemühen. Bei solchem Bemühen wäre es allerdings a) ein grundlegend verkehrter Ansatz, sich mit der Evaluation *nur* "der Lehre" zu befassen und damit den Blickwinkel nur auf die *eine* Seite einzuschränken. Lehren und Studieren gehören zusammen, bedingen sich wechselseitig. Genauso ist es aber auch b) ein weiterer falscher Ansatz, von der "*Qualität*" von Lehre und/oder Studium *als Ziel* der Bemühungen auszugehen. *Ziel* muß vielmehr der "*Erfolg*" des Studiums sein. Qualitativ gute Lehre kann für die Erreichung dieses Ziels allenfalls *ein* Instrument sein; ob ein wirkungsvolles Instrument (wie von den Aktionsprogrammen verschiedener Länderregierungen als selbstverständlich unterstellt), ist nach vorliegenden Befunden der Hochschulforschung im In- und Ausland leider durchaus fraglich. So kommen etwa Abrami

u.a. nach Auswertung von Forschungsbefunden aus den USA und aus Kanada zu dem Fazit: "Instructors may have genuinely small effects on what students learn" (Abrami 1990, 221)

Vor diesem so skizzierten Hintergrund soll der folgende Beitrag zunächst (Teil 1) schlaglichtartig ein paar Ergebnisse aus den in letzter Zeit an deutschen Hochschulen durchgeführten Befragungen zur "Lehrevaluation" durch Studierende präsentieren und einige Probleme solcher "Lehrevaluationen" illustrieren. Danach (Teil 2) wird - angesichts der Komplexität des Gegenstands notwendigerweise bruchstückhaft - darauf eingegangen, was denn "Evaluation" eigentlich ist, wenn man sie als eine kontrollierte Forschungsaktivität - und nicht im rein alltagssprachlichen Sinne - versteht. Zugleich wird unter methodischen Gesichtspunkten geprüft, ob Teilnehmerumfragen überhaupt "Evaluation" sein können. Abschließend (Teil 3) wird skizziert, wie Teilnehmerumfragen in Lehrveranstaltungen wertvolle und durch andere Verfahren nicht ersetzbare Beiträge zu Bemühungen um Qualitätsverbesserungen in der Lehre leisten können.

1. Einige Ergebnisse studentischer "Lehrevaluationen"

Die bisherigen Umfragen unter Teilnehmern von Lehrveranstaltungen bringen sehr differenzierte Ergebnisse. Sie fallen sehr unterschiedlich aus, sind schwer unter einen Hut zu bringen; und dies nicht nur *zwischen* verschiedenen Lehrveranstaltungen, sondern ebenso innerhalb *jeder* (oder fast jeder) Veranstaltung, die eine bestimmte Teilnehmerzahl (ca. 20) überschreitet. Berichtet wird hier deshalb nur - in exemplarischer Absicht - über einige grob zusammenfassende Resultate (ausführlicher: Kromrey 1994, 1995a); und dies lediglich für ausgewählte Fragen und für einen Veranstaltungstyp: die Vorlesung. Zudem können die Ergebnisse im Rahmen dieses Beitrags nicht im Detail präsentiert und erläutert werden. Sie sollen vor allem Eindrücke von der Komplexität des Vorhabens vermitteln.

1.1 Ergebnisse I: Studentische Einschätzungen von Didaktik und Lehrperson

Die folgende grafische Darstellung, die wie ein Schnittmusterbogen für Hobby-Schneider wirkt, vermittelt einen ersten "Eindruck" von der Komplexität der Ergebnisse.

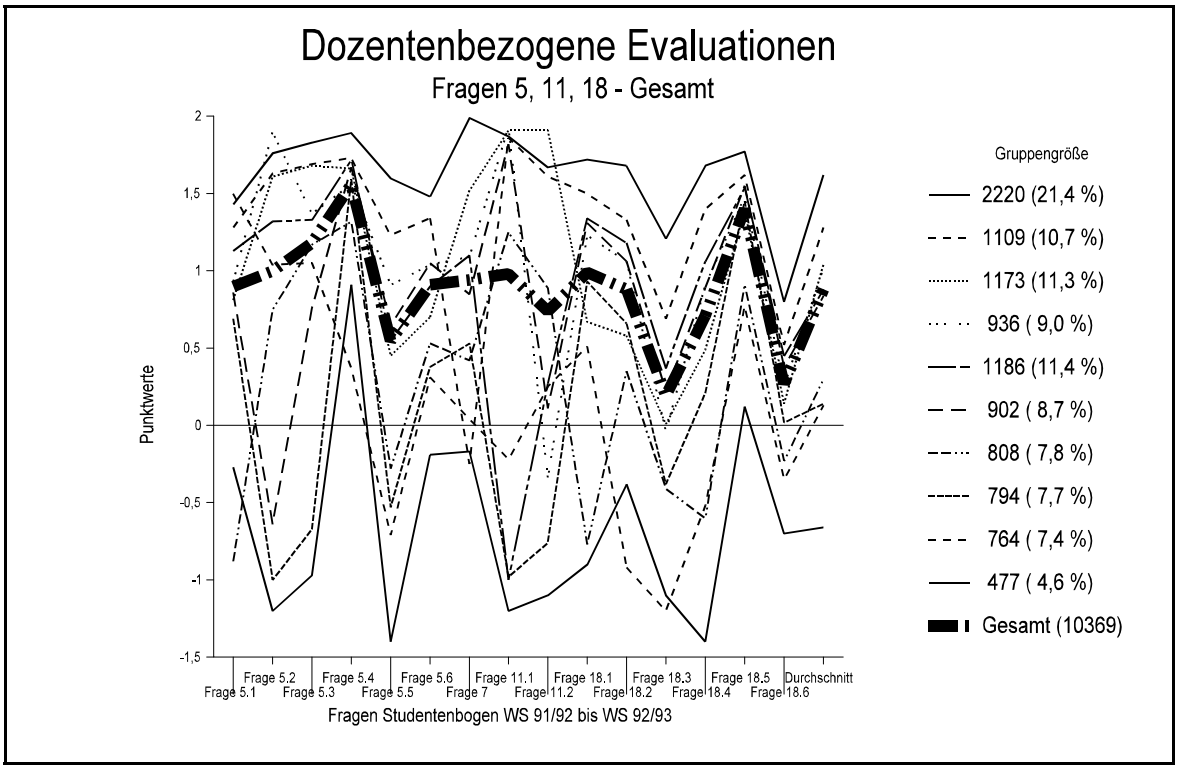

Figure 1 Gesamtübersicht (alle Urteilsprofile)

Dargestellt werden in dieser verwirrenden Grafik die Durchschnittswerte der Teilnehmer-Urteile zu einer Reihe von Items (bewertet auf einer 5-Punkte-Ratingskala von -2 bis +2), die sich auf das Lehrverhalten von Dozentinnen und Dozenten beziehen: Vortragsweise, Themenzentrierung, Verständlichkeit, Umfang und Schwierigkeitsgrad des Stoffs, Strukturierung, Klarheit der Lernanforderungen, Dozenten-Interesse, Orientierung an den Studierenden.

Betrachtet man - wie dies oft geschieht - lediglich die Mittelwerte aus den Angaben *aller* Befragten (breit-gestrichelt gezeichneter Kurvenverlauf), dann könnten die Hochschulen mit dem studentischen Urteil durchaus zufrieden

a) sprachlicher Vortrag des Stoffs:

Frage 5.1	(gelesen - frei vorgetragen):	0,9
Frage 5.2	(zu schnell - gerade richtiges Tempo):	1,0
Frage 5.3	(zu dicht/konzentriert - gerade richtig):	1,2
Frage 5.4	(zu eng am Thema - gerade richtig):	1,6
Frage 5.5	(unverständlich - gut verständlich):	0,6
Frage 5.6	(ungeläufige Fremdwörter - geläufige Wörter):	0,9
Frage 7	(Medieneinsatz: zu wenig - gerade richtig):	0,9

b) Schwierigkeitsgrad und Umfang des Stoffs:

Frage 11.1	(zu schwierig - gerade richtig):	1,0
Frage 11.2	(zu umfangreich - gerade richtig):	0,7

c) durch Dozent/in geschaffenes "Lernklima"

Frage 18.1	(Dozent/in gelangweilt - interessiert):	1,0
Frage 18.2	(keine Struktur erkennbar - Vorlesung klar gegliedert):	0,9
Frage 18.3	(Lernanforderungen unklar - klar):	0,2
Frage 18.4	(Dozent/in distanziert - ansprechbar):	0,7
Frage 18.5	(Eingehen auf Zwischenfragen: zu wenig - richtig):	1,4
Frage 18.6	(Eingehen auf Wünsche: zu wenig - richtig):	0,3

d) Durchschnitt über alle dozentenbezogenen Items: 0,9

"Im Durchschnitt" wird also die Lehre von den Veranstaltungsteilnehmern als "gut" (+1) wahrgenommen - mit zwei Abweichungen ins noch weiter Positive (Themenzentrierung und Grad des Eingehens auf Zwischenfragen: "gerade richtig") sowie ebenfalls zwei Abweichungen hin zum "befriedigend" (± 0) (Klarheit der Lernanforderungen, Berücksichtigung spezieller Wünsche von Studierenden). Entsprechend lauteten erste Reaktionen (auch in der Presse): Die Lehre ist "besser als ihr Ruf".

Bei differenzierterem Blick auf die Daten stellt sich allerdings heraus: Dieses Durchschnittsurteil ist ein statistisches Artefakt, das entsteht, wenn isoliert für die einzelnen Items Mittelwerte berechnet werden. Sucht man mit Hilfe geeigneter statistischer Verfahrern (z.B. Clusteranalyse) danach, wie groß die Gruppe der Studierenden ist, die in dieser Weise "durchschnittlich" urteilt (also ein Urteilsprofil aufweist, das dadurch charakterisiert ist, daß die Dozentenleistung auf fast allen Items in etwa mit +1 bewertet wird), dann zeigt sich: Ein solches "Cluster" existiert nicht. Statt dessen gibt es vielfältige Gruppen von - jeweils ähnlich urteilenden - Vorlesungsteilnehmern: ganz oben in der Grafik (Abb. 1) die "Fans", denen alles sehr gut erscheint (immerhin mehr als 20 %); ganz unten das entgegengesetzte Extrem: alles erscheint schlecht bis sehr schlecht (wenn auch nur bei knapp 5 %). Dazwischen liegen die Urteilsprofile von Gruppen Studierender, die - in jeweils unterschiedlicher Konstellation - teils positiv, teils negativ werten. Ein statistisch befriedigendes Ergebnis (hinreichend "homogene Cluster", d.h. hinreichend große Gleichartigkeit der Urteile *innerhalb* der jeweiligen Gruppen von Befragten bei zugleich möglichst deutlichen Unterschieden *zwischen* den Gruppen) stellt sich erst ein, wenn mindestens zehn Cluster gebildet werden (im einzelnen s. Tab. 1).

Tabelle 1: Bewertungsprofile 1

	Profile dozentenbezogener Evaluationen										GESAMT
	1	2	3	4	5	6	7	8	9	10	
Sprachl. Vortrag:											
Frage 5.1: freier Vortrag	1,43	1,28	0,81	0,93	1,13	0,86	-0,88	0,69	1,50	-0,28	0,90
Frage 5.2: Tempo	1,76	1,63	1,61	1,90	1,32	-0,64	0,75	-1,05	1,04	-1,17	1,00
Frage 5.3: Dichte	1,83	1,69	1,68	1,38	1,33	0,76	1,17	-0,67	1,06	-0,97	1,18
Frage 5.4: Themenbezug	1,89	1,73	1,66	1,56	1,72	1,67	1,32	1,59	0,37	0,91	1,56
Frage 5.5: Verständlichk.	1,60	1,23	0,45	0,91	0,55	0,65	-0,28	-0,53	-0,71	-1,39	0,55
Frage 5.6: Fremdwörter	1,48	1,34	0,70	1,04	0,90	1,05	0,53	0,38	0,31	-0,19	0,91
Frage 7: Medieneinsatz	1,99	-0,25	1,52	1,09	1,10	0,85	0,42	0,53	0,04	-0,17	0,94
Stoff:											
Frage 11.1: Schwierigk.	1,87	1,86	1,91	1,93	-1,05	1,82	1,26	-0,98	-0,23	-1,19	0,98
Frage 11.2: Umfang	1,67	1,61	1,91	-0,34	0,27	0,11	0,89	-0,76	0,25	-1,09	0,74
Dozentenverhalten:											
Frage 18.1: Interesse	1,72	1,50	0,67	1,23	1,34	1,31	-0,77	0,93	0,51	-0,90	0,99
Frage 18.2: Vorl.struktur	1,68	1,33	0,58	1,06	1,18	1,06	0,35	0,66	-0,92	-0,38	0,88
Frage 18.3: Lernanford.	1,21	0,69	0,01	-0,04	0,37	0,22	-0,41	-0,38	-1,18	-1,13	0,19
Frage 18.4: Zuwendung	1,68	1,40	0,49	0,91	1,06	0,89	-0,60	0,21	-0,52	-1,45	0,71
Frage 18.5: Fragen	1,77	1,62	1,39	1,48	1,54	1,55	0,91	1,45	0,77	0,12	1,40
Frage 18.6: Wünsche	0,80	0,52	0,14	0,25	0,45	0,38	-0,24	0,01	-0,35	-0,71	0,27
Durchschnittsbewertung	1,62	1,28	1,04	1,02	0,88	0,84	0,30	0,14	0,13	-0,66	0,88

Fazit: Die Uneinigkeit der Studierenden darüber, was als gutes Lehrverhalten empfunden wird, ist außerordentlich groß. Sie besteht darüber hinaus praktisch innerhalb jeder Lehrveranstaltung, d.h. derselbe Sachverhalt wirkt positiv auf die einen, negativ auf die anderen. Erfahrene Lehrpersonen werden darüber nicht überrascht sein. Der Befund dürfte mittlerweile auch in der Forschung nicht mehr umstritten sein; er stellt sich auch ein, wenn mit andersartigen Erhebungsverfahren gearbeitet wird (vgl. z.B. Schweer & Rosemann 1995).

An dieser Stelle scheint ein Exkurs zum Konzept "Qualität der Lehre" angebracht; denn sie soll durch die Evaluation (hier: durch studentische Qualitätsurteile) gemessen werden.

1.2 Was ist "Qualität"?

Für die oben vorgestellten Urteilsprofile könnte es durchaus plausibel sein anzunehmen, man könne sie auf einer Qualitätsdimension so anordnen, daß eine Rangordnung von den "Fans" über die im wesentlichen positiv Urteilenden bis hin zu den Generalkritikern aufstellbar wäre. Gelänge dies, wäre ein Index "Qualität der Lehre" (gemessen als subjektive Wahrnehmung der Veranstaltungsteilnehmer, gemittelt über eine Reihe von Einzel-Items) sinnvoll operationalisierbar.

Leider erweist es sich, daß dies lediglich für einen Teil der Befragten in empirisch gültiger Weise möglich ist. Nur vier der zehn Urteilsprofile (sie stehen für 45,1 % der Befragten) lassen sich widerspruchsfrei - d.h. ohne größere Überschneidungen der Kurvenverläufe - auf einer Positiv/Negativ-Dimension anordnen. Für sie gilt: Personen mit dem Urteilsprofil 1 bewerten das Lehrverhalten in der Tendenz eindeutig positiver als Personen mit dem Urteilsprofil 2. Diese 45 % der Befragten orientieren sich in ihren *Detail*-Evaluationen offenbar an einem *Global*urteil (Lehrqualität "alles in allem"). Sie urteilen innerhalb eines Schemas, das offenbar diejenigen Befürworter von Lehrevaluationen im Sinne haben, die kurze und damit leicht handhabbare Fragebögen empfehlen (darauf wird im Teil 2 näher eingegangen).

Bei den restlichen sechs Profilen handelt es sich demgegenüber um Urteilsdimensionen, die nicht bruchlos ordinalskalierbar sind. Hier überwiegen differenzierende Detail-Bewertungen: manches gut, manches mittelmäßig, manches schlecht.

Nun ist es eigentlich ja gerade letzteres, was sich diejenigen Lehrpersonen erhoffen, die nicht an Globalbewertungen, sondern an detaillierten Rückmeldungen interessiert sind: differenzierte Urteile über die einzelnen Aspekte ihrer Arbeit. *Aber* - leider wiederum "aber" -: Diese differenzierten Urteile scheinen wenig mit dem zu tun zu haben, was in der Lehrveranstaltung faktisch abläuft. In *ein und derselben* Veranstaltung kommt es vor, daß ein Teil der Studierenden *das* als positiv bewertet, was einem anderen Teil explizit negativ auffällt und umgekehrt. Würde also die Lehrperson ein Detail ändern, weil sie dort mit Kritik konfrontiert wurde, dann würde sie nach der Änderung wiederum mit Kritik konfrontiert werden: von denjenigen nämlich, die vorher die Lehre - so wie sie war - gut fanden. Und bei *den* 45 %, die sich ohnehin in ihren Detailbewertungen von einem Gesamteindruck ("alles-in-allem") leiten lassen, bewirkte die Änderung eines Details so gut wie gar nichts: Es würde weiterhin als gut oder weiterhin als schlecht eingestuft.

Die "ärgerlichen" Befunde zeigen: Was "Qualität" ist und was nicht, wird von unterschiedlichen Gruppen Studierender nach unterschiedlichen Kriterien eingeschätzt - eine Alltags-Selbstverständlichkeit, die erstaunlicherweise bei Anwendung empirischer Erhebungsmethoden und mehr noch beim Einsatz statistischer Auswertungsverfahren häufig in Vergessenheit gerät.

Nach diesem Exkurs zurück zu weiteren Umfragebefunden.

1.3 Ergebnisse II: Studentische Selbsteinschätzungen von Lernprozeß und Lernerfolg

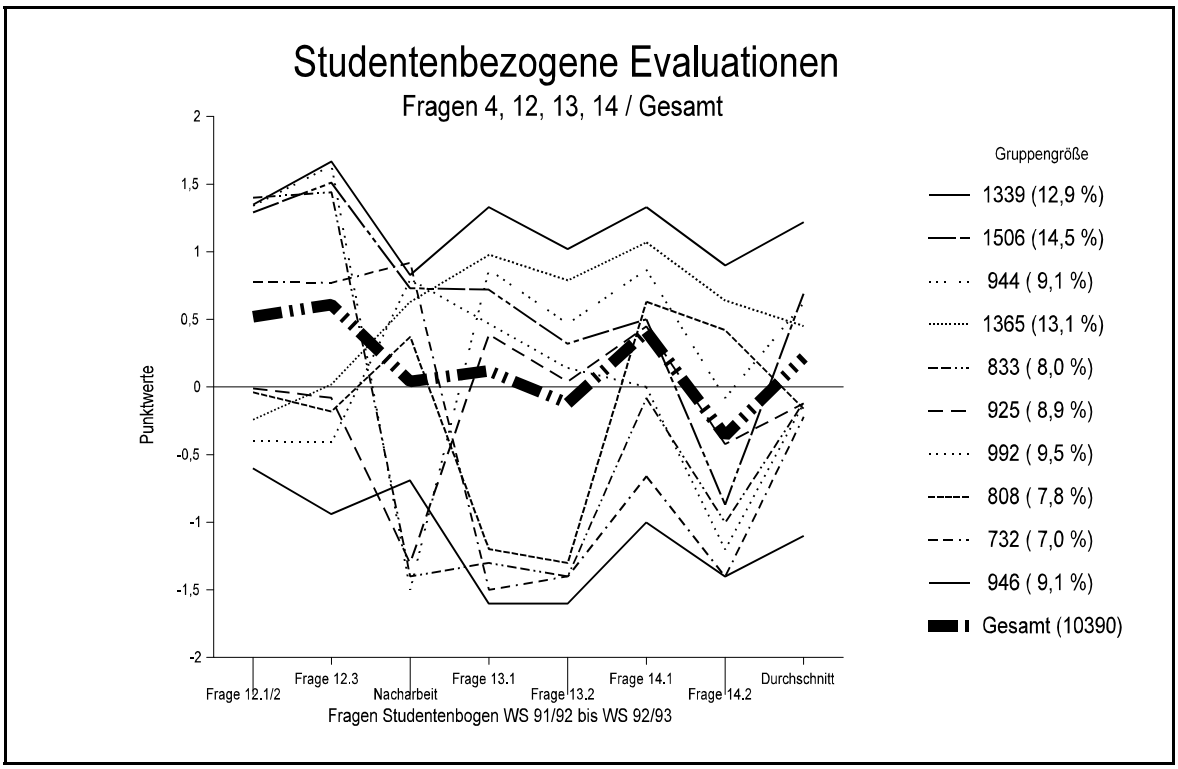

Figure 2 Gesamtübersicht (alle Urteilsprofile)

In Abb. 2 werden (analog zu Abb. 1) die Selbsteinschätzungen der Studierenden zu Lernsituation und Lernergebnissen dargestellt: Vorhandensein oder Fehlen erforderlicher Vorkenntnisse, Verständnis, Förderung von Interesse, Anregungen zur weiteren Befassung, eigene Nacharbeit, Lernerfolge. Das Durchschnittsprofil (berechnet als Mittelwert aller Antworten je Item) verläuft hier weniger positiv als hinsichtlich der Urteile über das Verhalten der Dozentinnen und Dozenten; es wäre als insgesamt "befriedigend" zu deuten (± 0):

Frage 12.1/2	(zuviel Vorkenntnisse vorausgesetzt - nicht vorausges.)	0,5
Frage 12.3	(Schwierigkeiten zu folgen: häufig - nie):	0,6
Frage 16.2	(Nacharbeit: nie - regelmäßig/vollständig):	0,0
Frage 13.1	(Interesse gefördert: nein - ja, sehr):	0,1
Frage 13.2	(Anregungen zur Weiterarbeit: nein - viele):	-0,1
Frage 14.1	(Zusammenhänge verdeutlicht: nein - sehr):	0,4
Frage 14.2	(wissenschaftl. Arbeiten gelernt: nein - sehr):	-0,2
Durchschnitt über alle lernprozeßbezogenen Items:		0,2

"Im Durchschnitt" wäre also *die Lehre* gar nicht so schlecht, *der Lernerfolg* dagegen eher mittelmäßig. Auch hier aber zeigt sich wieder: Dieses statistische "Durchschnittsprofil" existiert empirisch nicht. Die Situation stellt sich ähnlich dar wie bei den Beurteilungen der Lehrperson. Charakteristisch ist nicht die einheitliche Urteilstendenz, sondern die große Vielfalt unterschiedlicher Evaluationsprofile. Wieder finden sich auf der einen Seite des Spektrums Studierende, die alles positiv wahrnehmen, und auf der anderen Seite eine Gruppe, die durchgehend negativ urteilt. Wieder finden sich dazwischen Gruppen von Befragten mit in charakteristischer Weise unterschiedlichen Antwort-Konstellationen.

Auch hier öffnet eine Aufgliederung der Grafik den Blick auf interessante Zusammenhänge. Zu diesem Zweck seien die (Selbstbeurteilungs-)Profile nach einem aus der Perspektive der Didaktik wichtigen Kriterium geordnet: gute Lernvoraussetzungen versus schlechte Lernvoraussetzungen. Als "gute Lernvoraussetzung" soll gelten, wenn der/die Befragte über die in der Lehrveranstaltung geforderten Vorkenntnisse verfügte und er/sie keine Schwierigkeiten hatte, dem Lernstoff zu folgen. Entsprechend soll als "schlechte Lernvoraussetzung" gelten, wenn beides als problematisch empfunden wird (Vorkenntnisse fehlen, man beklagt daher Verständnisschwierigkeiten). Aus didaktischer Sicht wäre zu prognostizieren: Bei – in diesem Sinne - "guten" Lernvoraussetzungen wird sich tendenziell auch ein höherer Lernerfolg einstellen als bei "schlechten" Lernvoraussetzungen. Eine allseits anerkannte didaktische Regel lautet konsequenterweise, Lernende dort "abzuholen", wo sie sich befinden; d.h. von den Vorkenntnissen und dem Verständnis der Lernenden auszugehen und die Lehre darauf aufzubauen.

Leider vermitteln die Profilverläufe der Gruppen mit "guten Lernvoraussetzungen" den wenig erfreulichen Eindruck, daß sich offenbar der Lernerfolg dennoch nicht prognostizieren läßt. Es existieren in gleicher Weise Gruppen mit gutem, mit mittelmäßigem und mit schlechtem Lernerfolg. Gleiches gilt für den Fall "schlechter Lernvoraussetzungen". Am ehesten besteht ein Zusammenhang mit der berichteten eigenen Lernbemühung (hier illustriert am Item: Nacharbeit des Lernstoffs).

Die häufig verwendete einfache Gleichung "gute Lehre und gute Lernvoraussetzungen = gute Lernergebnisse" stimmt offenbar so nicht. Vielmehr dokumentiert sich in den Daten die - wiederum - Alltags-Selbstverständlichkeit, daß Lernen Arbeit bedeutet, und zwar nicht in

erster Linie Arbeit der Lehrenden, sondern Arbeit der Lernenden. Auch nach den Urteilen der Studierenden gilt: Lehre - auch gute Lehre - hat nur relativ geringe Chancen, gute Lern-Ergebnisse *zu bewirken*. Sie kann im Idealfall lernfördernder Service sein, der angenommen oder auch ignoriert werden kann. Umgekehrt gilt: Studierende sind bei schlechter Lehre nicht sogleich zum Mißerfolg verdammt. Sie können ein mangelhaftes Lehrangebot durch eigene Studienbemühungen kompensieren.

Natürlich ist es dennoch nicht gleichgültig, wie gut oder wie schlecht die Lehre ist. Natürlich ist es nicht müßig, sich mit der Frage der Lehrqualität zu befassen und sich um ihre Verbesserung zu bemühen, wo dies notwendig ist, bzw. Qualitätssicherung dort zu betreiben, wo gegenwärtig Lehre und Studienbedingungen (noch) zufriedenstellend sind. Wenn auch gute Lehre nicht Positives (automatisch) erzwingen kann, so hat doch schlechte Lehre starke "Chancen", Lernen zu behindern. Schlechter Lernservice macht es den Studierenden zwar nicht unmöglich, erfolgreich zu studieren; aber das Kompensieren-Müssen macht das Studieren schwerer als nötig, verlängert es, trägt dazu bei, daß nicht das volle Potential ausgeschöpft werden kann. Dennoch: Die Konzentration auf *Lehr*qualität allein reicht nicht; als notwendige Voraussetzung muß *Studien*qualität hinzukommen.

Leider erfreut sich bei vielen - bei Didaktikern, bei Studierenden, in der Öffentlichkeit - das Lernmodell "Nürnberger Trichter" immer noch großer Beliebtheit: Der Lehrende soll die Lehrinhalte in gut verdaulicher Form aufbereiten und sie schön verpackt und angenehm seinen Zielpersonen "eintrichtern". Und wenn dies in didaktisch korrekter Weise geschieht, dann haben die Objekte der Lehre das ihnen Vermittelte "gelernt"; und dann werden natürlich auch die so Belehrten zufrieden sein. Solche Übervereinfachung gilt nicht nur für die Hochschullehre; sie wird in gleicher Weise auch für das Feld beruflicher Weiterbildung beklagt (vgl. Volpert 1994). Man kann es nur immer wieder wiederholen: Der mögliche Effekt von Lehre wird hier weit überschätzt.

2. Was ist Evaluation?

Zu welchem Zweck wurden (und werden) Befragungen, die unter der Bezeichnung "Lehrevaluation" firmieren, durchgeführt?

Bisher (im Teil 1) wurden sie vor allem unter der Perspektive vorgestellt: *Was kann man daraus lernen?* Unter dieser Perspektive aber sind die Teilnehmerumfragen in Lehrveranstaltungen nicht zu einem modischen Renner in den letzten Jahren geworden, sondern unter der Perspektive: *"Evaluierung" von Lehrqualität*.

Evaluation heißt: Bewertung. Im Zusammenhang mit empirischer Sozialforschung heißt Evaluation: *methodisch kontrollierte* Bewertung im Sinne von *Messung von Qualität* (im Detail s. Kromrey 1995b sowie diesbezügliche Anmerkungen zu einer Studie von Rindermann/Amelang in Kromrey 1995c). Wenn nämlich die Qualität der Lehre in verschiedenen Veranstaltungen "gemessen" werden könnte, dann könnte man zugleich die *Güte der Lehre* der jeweiligen Lehrenden miteinander vergleichen; dann könnten die "Guten" gefördert und die "Schlechten" zum didaktischen Nachhilfeunterricht geschickt werden.

"Messen" aber bedeutet: das zu Messende "intersubjektiv" auf einer vorgegebenen Skala einordnen. Um nun die Qualität von Dienstleistungen intersubjektiv beurteilen - also "messen" - zu können, müssen einige ganz wesentliche Voraussetzungen erfüllt sein:

- Man benötigt präzise angebbare Ziele, die mit dieser Dienstleistung erreicht werden sollen (hier: Lehrziele).
- Man benötigt Kenntnisse über die angemessene Vorgehensweise zur Erreichung der Ziele (hier: Lernstoffauswahl, –aufbereitung, Didaktik).
- Man benötigt genaue Daten sowohl darüber, was tatsächlich abgelaufen ist (hier:Prozeß der Lehre) und welche Wirkungen das gehabt hat (hier: Lernergebnisse und deren Zurechnung zu den Ursachen: Wurden sie durch die Lehre bewirkt oder durch ein über- bzw. unterproportionales Lernen der Studierenden oder durch deren Vorkenntnisse?)
- Und man benötigt schließlich auch noch Kriterien, um einzuordnen, ab wann das, was man festgestellt hat, schlecht, zumindest mittelmäßíg oder schon gut oder gar sehr gut ist, wann also niedrige, mittelmäßige, hohe oder sehr hohe Qualität vorliegt.

Das alles wäre außerordentlich aufwendig und (falls überhaupt) nur individuell je Lehrveranstaltung einlösbar, jedenfalls nicht parallel zum laufenden Arbeitsprozeß von

Fakultäten und Instituten zu realisieren. In diesem Dilemma scheint sich ein Ausweg geradezu aufzudrängen: Die Studierenden sind die von der Lehre "Betroffenen"; sie können aus eigener Erfahrung sagen, wie es mit der Qualität der Lehre bestellt ist.

Um das Dilemma und seine "Lösung" an einem Beispiel aus dem Alltag zu illustrieren: Man kann ein hochkomplexes technisches Produkt wie ein Automobil hinsichtlich vieler Aspekte "objektiv" beurteilen, also evaluieren lassen (z.B. vom TÜV). Man kann es aber auch von den Autofahrern selbst beurteilen lassen, die ja täglich damit umgehen. Allerdings: Manche finden einen VW unerträglich und einen Mercedes ausgezeichnet; manche finden Mercedes viel zu pompös, aber Porsche ganz toll; andere wieder finden Porsche zu angeberisch, japanische Autos dagegen grundsolide und gut. .... Es dürfte kaum Einigkeit darüber herbeizuführen sein, in welche "tatsächliche" Rangordnung die einzelnen Produkte gesetzt werden müßten. Dies ist auch nicht erforderlich, weil sich jeder Automobilinteressierte am Markt ***das*** Produkt aussuchen kann, das seinen Wünschen und Vorstellungen in Relation zu seinen finanziellen Möglichkeiten - aber auch zu seinen Absichten, Geld in ein Automobil zu investieren - am besten entspricht.

In der Hochschule ist dies so nicht möglich: Zum einen ist das Angebot, aus dem Studierende in einem Studiengang wählen können, nicht annähernd so groß. Manches muß gewählt werden, selbst wenn man es überhaupt nicht möchte. Zum anderen sind auch die Chancen, in Relation zu den eigenen Absichten und Möglichkeiten Zeit und Arbeitsaufwand ins Studium zu investieren und dementsprechend Veranstaltungen mit höherem oder geringerem Anspruchsniveau zu wählen, gering (das Angebot ist nicht frei, sondern durch Studienordnungen reguliert). Dennoch (und gerade deshalb) möchten Hochschulpolitiker nach Möglichkeit ein Ranking von Lehrveranstaltungen und Lehrenden nach ihrer Qualität aufstellen können. Denn das Qualitätsniveau der Hochschulabschlüsse soll gerade ***möglichst wenig*** differieren, es soll vielmehr überall möglichst gleich sein, nämlich: möglichst hoch. Deshalb müssen Angebote und Anbieter mit geringerer Qualität erkannt und deshalb muß dort die Qualität angehoben werden. Das ist verständlich und legitim; denn universitäre Ausbildung ist keine am Markt angebotene Dienstleistung, sondern ein öffentliches Gut.

Zurück zur Realisierbarkeit von Lehrevaluationen. Für die gewünschte Verwendung von Teilnehmerumfragen als Instrument der Qualitätsbewertung gelten Rahmenbedingungen, die eine detaillierte Erhebung und Auswertung von Teilnehmerurteilen unangemessen erscheinen lassen: Der laufende Betrieb von Fakultäten und Instituten hat selbstverständlich Vorrang vor solchen Erhebungen und darf nicht über Gebühr gestört werden. Will man dennoch alle Lehrveranstaltungen einer Fakultät in ***einem*** Semester evaluieren, muß man sich auf Fragebögen beschränken, die nur wenige Fragen enthalten und somit in größerer Menge verteilt, ausgefüllt, erfaßt und ausgezählt werden können. Diese ***wenigen*** Fragen müssen zudem einen ganz zentralen Aspekt - den Lehr- bzw. Lernstoff der zu evaluierenden

Veranstaltungen - (da nicht standardisierbar) ausklammern. Und: Diese wenigen Fragen müssen von den Studierenden relativ grob zusammenfassende Urteile abverlangen, Urteile vom Typ "Wie zufrieden waren Sie mit...?" oder "Alles in allem gesehen: Wie beurteilen Sie...?", verbunden mit einer Einschätzungsskala (z.B. von -2 bis +2 oder in Anlehnung an die Prozentrechnung von 0 bis 100 oder ähnlich).

Zusammengefaßt:

1. Erhebung und Datenaufbereitung müssen mit wenig Mühe bei einer großen Zahl von Veranstaltungen und Befragten möglich sein.
2. Auch die Auswertung muß ohne großen Aufwand und in standardisierbarer Form möglich sein.
3. Die Ergebnisse müssen einen anschaulichen Vergleich zwischen verschiedenen Fragen und zwischen verschiedenen Veranstaltungen ermöglichen.
4. Die Auswertungsergebnisse müssen auch von statistischen Laien ohne weiteres rezipierbar sein.

Alle diese Kriterien scheinen in besonderer Weise von Durchschnittswerten (üblicherweise arithmetisches Mittel) erfüllt zu werden. Durchschnitte haben nicht nur den Vorteil, daß auch jeder nicht statistisch geschulte Betrachter in Anlehnung an seine Alltagserfahrung sie unmittelbar verstehen kann. Für sie spricht auch das weit verbreitete Argument, daß sich individuelle Abweichungen "im Durchschnitt ausgleichen"; d.h.: Die einzelnen Studierenden werden zwar individuell unterschiedlich (einige etwas negativer, andere etwas positiver) urteilen, aber "im Durchschnitt" ergibt sich daraus die allen gemeinsame Urteils-Tendenz. Selbst in ernsthaften Veröffentlichungen kann man lesen, daß ein aus mindestens 15 oder 20 Teilnehmerantworten berechneter Wert eine zuverlässigere Qualitätsmessung sei als die Bewertung durch Experten (z.B. Schmidt 1980).

2.1 Fehlschlüsse bei der Verwendung von Akzeptanzmessungen als Evaluation

Die mit diesem Typ von "Evaluation" zusammenhängende Problematik wurde im Teil 1 bereits illustriert; in diesem Abschnitt soll sie vertieft werden: methodische Probleme und Fallen, sofern man Teilnehmerumfragen als Instrument der Evaluation einsetzen möchte.

Die einfache - und im gegebenen Zusammenhang falsche - Gleichung, wie sie sich für die Vertreter der Durchschnitts-Berechnung aufdrängt, lautet:

> Durchschnittswert hoch = Lehre gut; und
> Durchschnittswert niedrig = Lehre schlecht.

Diese Gleichung ist aus mehreren Gründen falsch. *Der erste Fehler* ist inhaltlicher Art. Gleichgesetzt wird durchschnittliche Akzeptanz mit intersubjektiver Bewertung; d.h. es wird unterstellt, daß es sich bei der Erhebung der Akzeptanz der Lehre durch die Veranstaltungsteilnehmer um eine "Evaluation" im Sinne von Qualitätsmessung handelt. Anders ausgedrückt: Wenn bewertende (also "evaluierende") Fragen gestellt werden, dann nimmt jede befragte Person Bewertungen (also "Evaluationen") vor, deren individuell unterschiedliche Motivationen sich mittels Durchschnittsberechnung ausgleichen (also keinen verzerrenden Einfluß ausüben).

Differenzierte (multivariate) Auswertungen belegen das Gegenteil. Den weitaus größten Einfluß darauf, wie positiv oder negativ das Durchschnittsurteil ausfällt, übt ein Aspekt aus, der in Kurzfragebögen - wie oben ausgeführt - gänzlich außer acht gelassen werden muß: der Lernstoff der Veranstaltung und seine generelle Beliebtheit oder Unbeliebtheit im jeweiligen Studiengang. Gleichgültig, wonach im Detail gefragt wird: die Lehre von unbeliebtem Stoff wird tendenziell eher negativ, die Lehre von beliebten Studieninhalten tendenziell eher positiv bewertet.

Eine zweite zentrale Einflußgröße stellen die individuellen Besuchsgründe der einzelnen Veranstaltungsteilnehmer dar: Bei Pflichtteilnahme und "Klausurandrohung" wird tendenziell negativ, bei Wahlmöglichkeit, bei Freiwilligkeit der Teilnahme, gar bei bestehendem Eigeninteresse am Lerngegenstand wird tendenziell positiv geurteilt. Beides zusammengenommen - generelle Beliebtheit/Unbeliebtheit des Studienstoffs und individuelle Motivation - erklären rund 80 % der Unterschiede von Beurteilungs*durchschnitten* zwischen den Lehrveranstaltungen. Erst bei Konstanz von Studienstoff und Teilnehmerstruktur können Akzeptanzaussagen näherungsweise als Evaluationen interpretiert werden.

Der zweite Fehler ist statistisch-methodischer Art. Jeder Statistikstudent lernt in den

Anfangssemestern, daß (neben Ansprüchen an das Meßniveau der Daten, worauf hier nicht eingegangen wird) für eine sinnvolle Verwendung des arithmetischen Mittels als Maß für eine "zentrale Tendenz" einer Variablen *eine* Voraussetzungen auf jeden Fall erfüllt sein muß, nämlich: die Verteilung der Daten muß in der Tat eine "zentrale Tendenz" aufweisen. Im einzelnen: Die Variation der Einzelwerte darf nicht übermäßig groß sein (je höher die Variation, desto geringer der Informationswert des Durchschnitts); und die Daten müssen eine gewisse Konzentration um einen mittleren Wert aufweisen.

Beide Voraussetzungen sind für die Teilnehmerurteile in einem großen Teil der evaluierten Veranstaltungen nicht erfüllt. Vielmehr zeigt sich in fast jeder Veranstaltung eine große Unterschiedlichkeit der von den Teilnehmern abgegebenen Urteile (relative Einmütigkeit ist die seltene Ausnahme). Zum Teil kommt es sogar zu einer Polarisierung von Einschätzungen. Was einer relativ großen Gruppe von Teilnehmern als besonders gut und angemessen erscheint, wird von einer anderen - ebenfalls relativ großen - Gruppe als besonders negativ wahrgenommen. Der Durchschnitt "mittelmäßig" ist in solchen Fällen - das wurde bereits illustriert - ein reines Artefakt des statistischen Rechenmodells.

Der dritte Fehler besteht darin, die Antwort auf *ein* Beurteilungs-Item als unabhängig von den Antworten zu anderen Beurteilungs-Items zu betrachten. Dieser Fehler tritt in zwei Varianten als unzulässige Unterstellung einer für alle Befragten gleichen Beantwortungsstrategie auf. Die eine Variante unterstellt (zumeist implizit), daß die - unabhängigen - Beurteilungen für jeden einzelnen Aspekt unter Rückgriff auf ein gleichartiges Bewertungskriterium (hier z.B. auf eine latente Dimension "Qualität der Lehre") vorgenommen wird, so daß aus einer Batterie von Einzelitems zur Lehrqualität ein zusammenfassender Gesamtindex berechnet - oder als Ersatz für eine solche differenzierte Erfassung auch direkt ein zusammenfassendes Gesamturteil erhoben - werden kann. Die andere - häufig von Hochschuldidaktikern präferierte - Variante unterstellt, daß jede Beurteilung eines einzelnen Aspekts als eigenständige Aussage betrachtet und wörtlich genommen werden könne. Wird etwa der Medieneinsatz als unzureichend wahrgenommen, dann folgt daraus der didaktische Rat, den Medieneinsatz zu "verbessern", ihn also (isoliert von anderen Aspekten der Lehre) zu optimieren.

Beide Varianten treffen empirisch zu - allerdings nur für jeweils einen Teil der Befragten. Zu

Fehlinterpretationen kommt man daher zwangsläufig, wenn die Gesamtheit der Befragten (wie bei Mittelwertbildungen und bei der Konstruktion additiver Indices) "in einen Topf geworfen" werden.

Die erste Variante wurde in Abschnitt 1.2 schon behandelt: 45,1 % der Befragten orientieren sich in ihren Einschätzungen widerspruchsfrei an einer Qualitätsdimension. Für sie ist die Berechnung eines Gesamtindex als Mittelwert aller Einzelurteile empirisch sinnvoll. Gleiches gilt für das Erfragen eines pauschalen Gesamturteils (das erfragte Gesamturteil und der berechnete Indexwert korrelieren sehr hoch: r = 0,76 auf Individualdaten-Ebene sowie r = 0,91 auf der Ebene der Vorlesungsmittelwerte).

Die Mehrzahl der Befragten bewertet nach anderen Kriterien und kommt zu differenzierteren Urteilsprofilen; etwa: stoff- und darbietungsbezogene Aspekte positiv, auf die Lehr*person* und deren Verhalten bezogene Aspekte dagegen negativ (oder umgekehrt). Ein Durchschnittswert aus solchen Einzelevaluationen ist dann nicht mehr als gültiger Qualitäts-indikator interpretierbar, im Extremfall sogar vollständig informationsleer.

Am Beispiel von drei Urteilsprofilverläufen, die "im Durchschnitt" - gemessen am Gesamtindex - als gleichwertig einzuschätzen wären (nämlich als "mittelmäßig"), kann dies unmittelbar abgelesen werden.

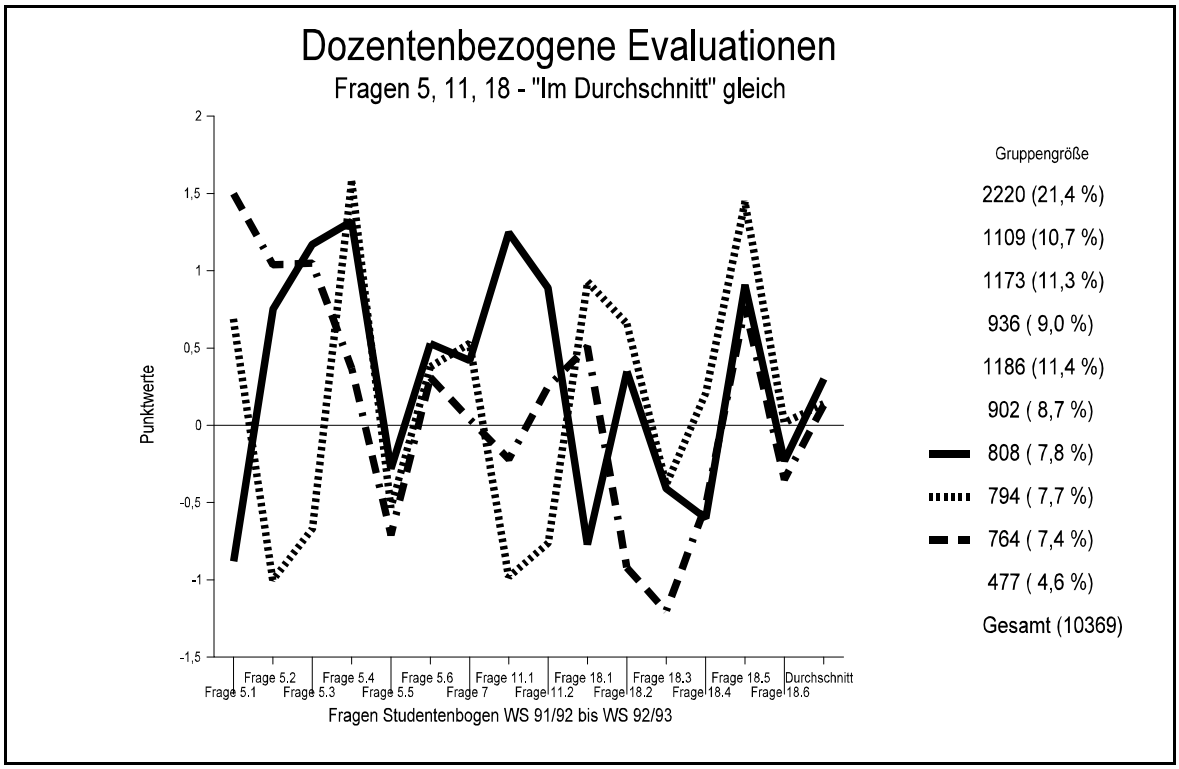

Figure 3 Beispiel: Urteilsprofile mit gleichem Durchschnittswert über alle Items

Profil 8 (n=794, 7,7 %; in Abb. 3 die quergestrichelte Linie) steht u.a. für folgende Detailurteile: frei vorgetragen, Vortrag zu schnell, zu dichte Informationsfülle, Stoff zu schwierig und zuviel, Dozent wirkt selbst interessiert. Profil 7 (n=808, 7,8 %; durchgezogene Linie) steht für in jedem dieser Aspekte gegenteilige Einschätzungen: Stoff abgelesen, Tempo und Dichte gerade richtig, Schwierigkeitsgrad und Menge des Stoffs gerade richtig, Dozent wirkt jedoch eher gelangweilt. Profil 9 (n=764, 7,4 %; unterbrochene Linie) liegt in den angesprochenen Details zum Teil zwischen den beiden genannten, weicht zum anderen Teil (freier Vortrag, fehlende Veranstaltungsstruktur, unklare Lernanforderungen) extrem ab. Relative Einmütigkeit besteht lediglich in Themenzentrierung, Verständlichkeit, Medieneinsatz und Eingehen auf Fragen. Dennoch erscheinen diese drei qualitativ völlig unterschiedlichen - zum Teil extrem gegensätzlichen - Beurteilungen der Lehre "im Durchschnitt" als identisch. Relativ unbestimmt stellt sich jedoch in diesen Gruppen das Ergebnis dar, wenn nicht ein Durchschnittsindex aus Detailurteilen ***berechnet***, sondern wenn stattdessen ***direkt*** eine zusammenfassende Bewertung abverlangt wird: Die von Befragten dieser Cluster selbst gebildete "Alles-in-allem"-Einschätzung fällt in ihrer Tendenz negativer aus und wird offenbar nach individuell unterschiedlichen Kriterien gebildet: Der Korrelationskoeffizient zwischen berechnetem und erfragtem Gesamturteil beträgt auf

Individualdatenebene lediglich r = 0,28 und erreicht auch für die Vorlesungsmittelwerte nur r = 0,33.

Zusammengefaßt ergeben sich damit als methodische Probleme von Akzeptanzbefragungen:

- Die isolierte Betrachtung einzelner Beurteilungs-Items führt in die Irre; notwendig ist die Ermittlung charakteristischer Urteilsprofile.
- Die empirisch zu ermittelnden Urteilsprofile sind nur zum Teil nach dem Kriterium "Qualität" zu ordnen; etwa mit gleichem Gewicht ergeben sich Bewertungskonstellationen, denen andere Maßstäbe zugrunde liegen.
- Auch global zusammenfassende Urteile (direkt als solche erhoben oder auch nachträglich aus Detailurteilen berechnet) sind methodisch nicht zulässig.
- Bei der Erhebung wie bei der Auswertung muß dem komplexen Sachverhalt von Lehren und Lernen methodisch Rechnung getragen werden; komplexe Sachverhalte können auch nur komplex abgebildet werden.

2.2 Lehr-Erfolg als Indikator für gute Lehre?

Angesichts der geschilderten Schwierigkeiten von Evaluationsumfragen drängt sich die Idee auf, den methodischen Problemen dadurch zu entgehen, daß anstelle von Bewertungs- und Akzeptanzmessungen versucht wird, Indikatoren für den Lehr-Erfolg zu erheben (Hier wird bewußt der Begriff Lehr-Erfolg, nicht Lern-Erfolg verwendet, da der Lern-Erfolg in hohem Maße von der Aktivität des Lernenden abhängt, nicht in erster Linie vom Lehrenden). Es ist bekannt und empirisch immer wieder bestätigt: Erfolgreiches Lernen hängt nicht unerheblich davon ab, ob der Lernende Interesse für das aufbringt, was er lernen soll, also: ob er ***motiviert*** ist zu lernen. Je höher die Lernmotivation, je höher das Interesse am Lernen und am Lerngegenstand, umso leichter fällt das Lernen, umso höher ist der Lernerfolg. Ebenso unstrittig ist daher der Rat an jede Lehrperson, die Lernenden zu "motivieren", ihr "Interesse zu wecken". Ein ***Lehr***-Erfolg wäre es also, wenn viele Teilnehmer der Lehrveranstaltungen bekundeten, die Lehrperson habe bei ihnen Interesse für das geweckt, was dort behandelt wurde. Die Vermutung liegt nahe, daß eine solche Auskunft ein gültigerer Indikator für die Qualität der Lehre sein müßte als die Erhebung von Zufriedenheit oder anderen subjektiven

Urteilen über die Didaktik und das Verhalten von Lehrpersonen.

Leider erweist sich auch hier die Angelegenheit empirisch als eher desillusionierend. "Geweckt werden" kann nun einmal nur etwas, was lediglich "schläft", also bereits latent vorhanden ist. "Motiviert werden" kann nur derjenige, bei dem ein Motiv im Prinzip schon vorhanden ist. Die Fähigkeit, etwas (noch) *nicht* Vorhandenes erst zu erschaffen, werden nur ganz wenige herausragende Künstler ihres Faches haben - die es natürlich auch unter Lehrenden gibt, aber bei weitem nicht so oft, wie rückblickend von Alt-Studierten über ihre früheren Lehrer berichtet wird. Diese mißliche Situation unterscheidet sich nicht im geringsten von anderen Bereichen des Lebens; selbst unter Künstlern (Maler, Musiker, Architekten, Schriftsteller ...) sind die wirklich großen Persönlichkeiten naturgemäß nur ganz selten vertreten. Vorherrschend sind die "ganz normalen" Handwerker, die ihre Arbeit mehr oder weniger sorgfältig erledigen. Das gilt ganz ähnlich auch für die Hochschule und ihre Lehrenden. Die "normalen Lehr-Handwerker" erfüllen weit überwiegend solide ihre Aufgaben, aber können selbstverständlich keine von ihnen verlangten Wunder vollbringen; sie können mit Erfolg im wesentlichen nur diejenigen erreichen, die bereits latent motiviert sind.

Tabelle 2:
Förderung von Interesse in der Vorlesung / Teilnahmemotivation

	TEILNAHMEMOTIVATION			
	extrins.	intrins. +Pflicht	intrins., ohne Pfl.	Zeile
INTERESSE GEFÖRDERT?	-1	0	+1	Total
-2 überhaupt nicht	1459 25,1%	168 5,1%	57 4,2%	1684 16,1%
-1	1213 20,9%	260 7,9%	99 7,4%	1572 15,1%
0 teils/teils	1537 26,5%	531 16,1%	200 14,9%	2268 21,7%
+1	1464 25,2%	1725 52,4%	445 33,1%	3634 34,8%
+2 ja, sehr	130 2,2%	608 18,5%	544 40,4%	1282 12,3%
Spalte Total	5803 55,6%	3292 31,5%	1345 12,9%	10440 100%

Tabelle **2** zeigt die Einschätzung der Veranstaltungsteilnehmer, ob der/die Lehrende sie für den Stoff hat interessieren können, aufgegliedert nach Teilnahmegründen der Studierenden: (-1) rein "extrinsische" Teilnahmemotivation (Pflichtveranstaltung, Klausur, kein Eigeninteresse), (0) neben "extrinsischer" auch "intrinsische" Motivation (neben Pflicht noch Eigeninteresse), (+1) rein "intrinsische" Motivation ohne Pflicht zur Teilnahme. "Intrinsische Motivation" wird hier bewußt sehr weit gefaßt; dazu zählt bereits *jeder* Teilnahmegrund, der nicht ausdrücklich nur Pflichtbesuch oder Klausur umfaßt, also etwa *Wahl*(pflicht), Einschätzung als "allgemein wichtig für das Studium", Besuch zur Wiederholung des Stoffs, natürlich auch persönliches Interesse. Vergleicht man das von den Teilnehmern berichtete "Fördern von Interesse" in den unterschiedlichen Motivationskategorien, so gelingt dies den Lehrenden in der Gruppe der nur "extrinsisch" Motivierten lediglich bei gut einem Viertel der Studierenden (27,4 %), jedoch bei fast drei Viertel der Studierenden aus den beiden Gruppen der "intrinsisch" Motivierten (70,9 % bzw. 73,5 %). Allerdings gilt für die Mehrheit (über 55 %) der in den Lehrveranstaltungen anzutreffenden Studierenden, daß sie ausschließlich aus dem einzigen Grunde dort sind, weil es für sie eine nicht zu umgehende Pflicht ist.

Noch schwieriger wird die Aufgabe des Förderns von Interesse, wenn die engere *Lernumwelt* der Studierenden dem Stoff gegenüber negativ eingestellt ist.

Tabelle 3:
Förderung von Interesse / vorherrschende Teilnehmermotivation in der Vorlesung

	VORHERRSCHENDE TEILNEHMERMOTIVATION					
	> 90 % extrins.	> 80 % extrins.	gemischt	> 70 % intrins.	> 75 % intrins.	Zeile
INTERESSE GEFÖRDERT?	-2	-1	0	1	2	Total
-2 überhaupt nicht	383 39,8%	721 25,8%	509 11,0%	69 4,4%	2 0,4%	1684 16,1%
-1	228 23,7%	520 18,6%	676 14,6%	130 8,4%	18 3,6%	1572 15,1%
0 teils/teils	190 19,7%	607 21,7%	1126 24,4%	277 17,8%	68 13,6%	2268 21,7%
+1	138 14,3%	783 28,0%	1794 38,8%	732 47,0%	187 37,5%	3634 34,8%
+2 ja, sehr	24 2,5%	168 6,0%	518 11,2%	348 22,4%	224 44,9%	1282 12,3%
Spalte Total	963 9,2%	2799 26,8%	4623 44,3%	1556 14,9%	499 4,8%	10440 100 %

Mehr als ein Drittel der Befragten (36 %) befanden sich in Lehrveranstaltungen, in denen mehr als 80 % der Anwesenden keinerlei Eigeninteresse an der Teilnahme zeigten. In solchen Umgebungen ist Lehr-Erfolg im obigen Sinne offenbar kaum noch erreichbar. Mit der Zunahme des Anteils von Vorlesungsteilnehmer(inne)n, die nur gezwungenermaßen anwesend sind, sinkt die Erfolgswahrscheinlichkeit des Förderns von Interesse bei *allen* Hörer(inne)n: Können in überwiegend intrinsisch motivierter Lernumgebung 82,4 % der Hörer interessiert werden, so gelingt dies den Lehrpersonen in reinen Pflichtveranstaltungen nur noch bei 16,8 % der Teilnehmer. Eine weitergehende tabellarische Aufgliederung (vgl. Kromrey 1994) belegt: Rein "extrinsische" Teilnehmer sind in solcher lernfeindlicher Lehrumwelt nur außerordentlich selten interessierbar. Sogar ursprünglich interessierte Studierende ("intrinsisch", keine Pflichtteilnahme) verlieren hier überwiegend ihre zuvor latent vorhandene Motivation.

Es zeigt sich wieder, was jeder erfahrene Lehrende weiß: Fehlt die Lernbereitschaft, ist jede Lehre ziemlich wirkungslos.

2.3 Konsequenzen aus den Erfahrungen mit Teilnehmerumfragen

Auf Ergebnisse wie die vorgestellten kann man in unterschiedlicher Weise reagieren. Man kann (1.) mit ihnen noch einmal illustrieren: Zur Evaluation im Sinne von methodisch kontrollierter Qualitätsmessung eignen sich Teilnehmerumfragen nur bedingt. Man kann (2.) damit eine resignative Einstellung legitimieren: Es hat ja doch alles keinen Zweck; die Umstände sind nun mal nicht zu ändern! Man kann aber auch (3.) daraus im Kleinen Konsequenzen ziehen, um wenigsten im eigenen Umfeld das Bestmögliche im Rahmen der gegebenen Umstände zu machen.

Eine wichtige Schlußfolgerung, die sich aufdrängt, ist allerdings: "Die *eine* gute Lehre" für die "typische Studentin" bzw. den "typischen Studenten" ist eine Idee ohne jeden Realitätsbezug. Lehre - wenn sie gut sein soll - muß zwangsläufig zielgruppen-orientiert sein, muß an den Lern- und Arbeitsweisen von Gruppen von Studierenden anknüpfen und an deren Interessen appellieren (und - das ist die Kehrseite der Medaille - zwangsläufig in Kauf nehmen, nicht zu gleicher Zeit die Interessen und Lernweisen *aller* Gruppen berücksichtigen zu können). Zielgruppenorientierte Lehre wird also nicht den Anspruch erheben, innerhalb

der Lehrveranstaltung bei *allen* (auch bei den absolut Desinteressierten) Interesse und Motivation "wecken" zu wollen.

Eine *weitere Schlußfolgerung* drängt sich auf: Auch gute Lehre ist noch nicht alles! Begleitend zu zielgruppen-orientierter Lehre ist gutes "Lehr-Marketing" vonnöten: Nicht erst *in* der Lehre Interesse "wecken", sondern bereits *vor* der Lehre durch rechtzeitige, gut zugängliche und hinreichende Information die potentiellen Interessenten für die Lehrveranstaltung gewinnen (etwa dadurch, daß in kommentierten Vorlesungsverzeichnissen neben dem Thema und dem Programm auch die vorgesehene Arbeitsform, das Anspruchsniveau, die vorausgesetzten Vorkenntnisse sowie nicht zuletzt die Beziehung zu anderen Studieninhalten, die Relevanz für den Studiengang und ggf. für Berufsfelder erläutert werden).

3. Teilnehmerbefragungen als Instrument für zielgruppenorientierte Qualitätsentwicklung in der Lehre

Im letzten Teil sei kurz ein Projekt vorgestellt, das sich die Lehrqualitäts-Entwicklung für interessierte Studierende im Rahmen eines teilweise neu zu konzipierenden Studiengangs am Institut für Soziologie der Freien Universität Berlin zum Ziel gesetzt hat.

Wenn der "normale Lehr-Handwerker" schon keine Lehr-Wunder vollbringen kann, so kann er doch lernen, "Lösungen nach Maß" zu entwickeln, d.h. sein Lernservice-Angebot auf die Zielgruppen in seiner Veranstaltung zuzuschneiden. Für eine "Lösung nach Maß" müssen allerdings die erforderlichen Informationen (hier: ausreichende Kenntnisse über die Zielgruppen der Lehre) zur Verfügung stehen. An dieser Überlegung setzt der Versuch an, Teilnehmerbefragungen in einen Kontext von Qualitätsentwicklung und Qualitätssicherung in der Lehre einzubeziehen: Teilnehmerbefragungen, um die benötigten Maße für "Lösungen nach Maß" zu erheben.

Qualitätsentwicklung und Qualitätssicherung ist Bestandteil jeder Produktentwicklung und Produktionskontrolle in der Industrie, wenn Produkte am Markt verkaufbar sein sollen. Sie ist zunehmend auch Bestandteil der Entwicklung und Kontrolle von Dienstleistungsangeboten, wenn diese sich mit ökonomischem Erfolg am Markt behaupten wollen. Es gibt dafür

mittlerweile Normen (ISO 9000 ff.), deren analoge Anwendung behördlich vorgeschrieben ist, wenn sich private Bildungseinrichtungen (etwa im Bereich Bildungsurlaub oder Weiterbildung) um öffentliche Aufträge bewerben. Die öffentlichen Bildungsinstitutionen können sich davon - angesichts überall knapper werdender Kassen - auf die Dauer nicht abkoppeln.

Qualitätsentwicklung und Qualitätssicherung ist jedoch nicht auf der Basis grober, globaler, vereinfachender Informationen möglich. Benötigt werden detaillierte Daten über die Zielgruppe(n), für die die Dienstleistung entwickelt werden soll; hier: über deren Lernziele und Lernvoraussetzungen, über deren Nutzungsverhalten (also: Lernverhalten), deren Erwartungen, Ansprüche und eben auch über deren Einschätzungen und Urteile. Auf all dies muß eine Lehrperson eingehen können, wenn sie eine Lehrveranstaltung 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 Teilnehmererwartungen 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; nur 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 allerdings 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 nur für ein auf Wiederholung angelegtes Lehrprogramm: Einführungsveranstaltungen im Grundstudium, regelmäßig wiederkehrende Bestandteile des Hauptstudium-Curriculums. In anderen Situationen ist der Lehrprozeß im Idealfall im Diskurs zwischen Lehrenden und Lernenden "auszuhandeln".

Bestandteile des Konzepts, wie es seit einigen Semestern erprobt wird, 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.).

Durchgeführt wird die Befragung von einer studentischen Hilfskraft; die jeweilige Lehrperson ist dabei nicht anwesend: Verteilung in der letzten halben Stunde, sofortiges Ausfüllen und Wiedereinsammeln. Ausgezählt werden die Antworten möglichst innerhalb einer Woche, so daß spätestens in der vierten Veranstaltungswoche die Ergebnisse von den Lehrenden an die Teilnehmer rückgekoppelt und mit ihnen diskutiert werden können. Der bisherige Eindruck dieser Ergebnisrückkoppelung ist, daß auf seiten der Lehrenden überraschenderweise die Bereitschaft, Probleme des Lehrens und Lernens öffentlich zu reflektieren, eher vorhanden ist als die Bereitschaft (oder Fähigkeit?) der Studierenden. Eine diesbezügliche Diskurs-Kultur muß offenbar erst (wieder) geschaffen werden. Auch dies ist vielleicht eine Rückmeldung an die Lehrenden: Möglicherweise muß es heute als ein fachübergreifendes Lehrziel gesehen werden, die für ein Studium unabdingbare Diskursbereitschaft erst neu anzuregen.

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 sei;

- 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 Durchführung geschieht nach dem Schema der Anfangsbefragung: Verteilung in der letzten halben Stunde, sofortiges Ausfüllen und Einsammeln. Eine kurzfristige Auszählung und Rückmeldung an die Lehrenden schafft die Voraussetzung dafür, daß noch am Schluß des Semesters ein Gespräch der beteiligten Lehrenden untereinander organisiert werden kann, in dem die Beteiligten ihre Einschätzungen des Nutzens der erhobenen studentischen Rückmeldungen sowie ihre Erfahrungen damit austauschen können.

Bis zum Beginn des Folgesemesters wird dann eine vergleichende Auswertung der Anfangs- und Endbefragung vorgenommen. Diese ist dadurch personenbezogen möglich, daß der Anfangs- und der Abschlußfragebogen von den Befragten mit einer (anonymen) Identifikationskennung versehen wird (d.h. für jeden, der vollständig geantwortet hat, können beide Bögen verknüpft werden, ohne daß Angaben zur Person erforderlich sind). Die Ergebnisse werden veranstaltungsweise den jeweiligen Lehrenden - und nur diesen - zur Verfügung gestellt. Veranstaltungs*übergreifende* Auswertungen können (und sollten!) darüber hinaus veröffentlicht werden.

Als Pendant zum Erfahrungsaustausch der beteiligten Lehrenden untereinander war geplant, zu Beginn des Folgesemesters Gruppendiskussionen mit Studierenden über deren Einschätzung des Vorhabens sowie über solche Themen zu führen, die in einer standardisierten Erhebung nicht angemessen behandelbar sind. Wie schon bei der Ergebnis-Rückmeldung innerhalb der Lehrveranstaltungen zeigte es sich jedoch, daß nur wenige Studierende für Themen der Lehre so engagiert sind, daß sie bereit wären, dafür außerhalb der "normalen" Studienaktivitäten zur Verfügung zu stehen.

Dagegen scheint die Akzeptanz der nur punktuell auszufüllenden Rückmelde-Fragebögen bei den Studierenden hoch zu sein: Mehr als 75 % plädieren für regelmäßige Wiederholungen solcher Befragungen. Allerdings wurde schon ab dem zweiten Semester auch vermehrt Unmut über die "lästige Befragerei" geäußert. Ob sich hier lediglich diejenigen explizit zu Wort melden, die der Aktion reserviert gegenüberstehen, oder ob darin bereits ein Abbröckeln der ursprünglich geäußerten Akzeptanz zu sehen ist, bleibt abzuwarten. Doch selbst wenn sich auf die Dauer - im Zuge einer "Veralltäglichung" des Ansatzes - die Beteiligung auf einem deutlich niedrigeren Niveau einpendeln sollte, ist dies kein Grund zur Resignation. Es ist kein Geheimnis, daß nicht alle Studierenden begierig sind, die Zeit ihres Studiums mit effektiver Arbeit zu füllen. Wenn also mit Bemühungen um eine qualitative Verbesserung des Studienservice "Lehre" nicht ein "Marktanteil" von 100 % erreichbar ist, ist dies kein Argument für eine nur mittelmäßige Qualität der Dienstleistung "Lehre". Für die Zielgruppe der Studieninteressierten lohnt sich die Mühe!

Literatur

Abrami, P. C. et al. (1990). Validity of Student Ratings of Instruction. What we know and what we do not. Journal of Educational Psychology, 82, 219-231.

Deutscher Hochschulverband (1992). Thesen zur Entlastung der Universitäten. Mitteilungen des Hochschulverbandes, 40, 59-60.

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

Kromrey, H. (1995a): Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: Mohler, P. (Hrsg.): Universität und Lehre. Münster/New York, 2. Aufl.: Waxmann, 105-128.

Kromrey, H. (1995b). Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeiten ihrer Realisierung. ZSE Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 15, 313-336.

Kromrey, H. (1995c). Buchbesprechung zu Rindermann, H. & Amelang, M. (1994). Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation, Heidelberg. Zeitschrift für Pädagogische Psychologie, 9, 221-224.

Mußgnug, R. (1992). Hochschulpolitik 2000. Mitteilungen des Hochschulverbandes, 40, 22-26.

Otten, K. (1992). Leserbrief. Mitteilungen des Hochschulverbandes, 40, 42.

Schmidt, J. (1980). Evaluation als Diagnose. HDZ-Dozentenkurs. Essen.

Schweer, M.K.W. & Rosemann, B. (1995). Qualität der Lehre: Bedingungsvariablen des studentischen Urteils. Zeitschrift für Pädagogische Psychologie, 9, 189-196.

DER SPIEGEL (1993). Welche Uni ist die beste? SPIEGEL-Rangliste der deutschen Hochschulen, Nr. 16, 80-101.

Stephan, R. (1992). Die Rache der Bildungsreform. Süddeutsche Zeitung vom 20.1.1992.

Volpert, W. (1994). Einleitung: Lernen am Arbeitsplatz. ZSE Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 14, 290-293.

(PAEDPSY.086 / Bearbeitungsstand 24.8.96)

Replik auf J.W.Diehl:

"Gute" oder "schlechte" Sozialforschung?

Einige notwendig scheinende Ergänzungen

Kein Text kann ausführlich genug sein, um Mißverständnisse bei den Leserinnen und Lesern vollständig auszuschließen. So scheinen also nach dem Kommentar von J.W. Diehl einige erläuternde Ergänzungen sinnvoll. Allerdings kann der folgende Versuch einer Klärung nicht auf alle angesprochenen Details eingehen; er beschränkt sich deshalb auf Antwortansätze zu drei Fragen:

1. Gibt es nur "gute Sozialforschung"?

Eine zentrale Botschaft des Diehl-Kommentars scheint zu sein: Bei den dem Autor bekannten Lehrevaluationen (einschließlich seiner eigenen) handelt es sich nicht um "schlechte Sozialforschung". Das ist eine erfreuliche Botschaft. Allerdings gibt es anderswo auch andere Eindrücke. Man schaue etwa in die vom Hochschulinformationssystem (HIS) in Hannover erstellte "Dokumentation Evalation der Lehre", in deren mir vorliegenden Teilen 1 und 2 (Stand 31.7.1992) 62 Vorhaben an deutschen Hochschulen kurz beschrieben werden. Die Durchsicht ergibt: Überwiegend wird die komplexe Aufgabe der "Evaluation" in - wie ich es nennen möchte - sehr "schlichter" Weise angegangen. Soweit Aussagen zur "statistischen Auswertung" gemacht werden, besteht diese in Durchschnittswertberechnungen pro Item und Lehrveranstaltung, üblicherweise in Form von Urteilsprofilen dargestellt. Oder nehmen wir detaillierter dokumentierte Projekte: Die "Prüf den Prof!!!"-Aktion des RCDS in Nürnberg, die vom Projekt pro Lehre der Freien Universität Berlin angebotene "Empirische Lehrveranstaltungsanalyse" (ELVA), die Untersuchung zum "Image von Lehrveranstaltungen und Professoren" am Lehrstuhl für Marketing der Universität Siegen verfahren so - um nur ein

paar Beispiele zu benennen. Selbst das ansonsten ausgesprochen differenziert vorgehende Modellprojekt "Evaluation der Lehre" an der Universität Mannheim (H.-D. Daniel) benutzt für die komprimierende Ergebnisdarstellung Durchschnittswerte pro Veranstaltung.

Im übrigen wird es nicht überraschen, wenn ich noch nachtrage: Natürlich handelt es sich auch bei den von mir geleiteten Erhebungen an der Ruhr-Universität Bochum (und jetzt an der FU Berlin) nicht um "schlechte Sozialforschung", die sich einer beliebig zusammengetragenen "Sammlung von Einzelitems" (Diehl) bediente und somit nur - rubbish in, rubbish out - zu wertlosem Ergebnismüll gekommen wäre. Die Absicht meines Textes "Qualitätsverbesserung in Lehre und Studium..." ist es allerdings nicht, die Qualität der eigenen Arbeit leuchten zu lassen, sondern auf Gefahren aufmerksam zu machen, die entstehen, wenn an sich gute und nützliche Verfahren in ungeeigneter Weise verwendet werden. Interessenten können bei mir gern die eingesetzten Erhebungsbögen zur Bewertung von Lehrveranstaltungen (gesonderte Versionen für Vorlesungen, Seminare, Übungen/Tutorien, jeweils für Studierende und für Dozentinnen/Dozenten), zur Evaluation von Studium und Lehre an Fakultäten bzw. Instituten, zur Erhebung von Studierstilen sowie zur Befragung von Studienanfängern und von Absolventen) anfordern.

2. Positive Urteile in "guten", negative Urteile in "schlechten" Lehrveranstaltungen?

Gleich zu Beginn meines Textes habe ich darauf hingewiesen (und dies im Fazit nach Tabelle 1 wiederholt), daß die Heterogenität der studentischen Wahrnehmung von Lehre nicht nur zwischen verschiedenen Lehrveranstaltungen festzustellen ist , sondern daß sie auch innerhalb (fast) jeder Veranstaltung auftritt. Bewertungsvarianz zwischen Veranstaltungen wäre natürlich für das Ziel einer hinreichend differenzierenden und trennscharfen Evaluation kein Nachteil, sondern geradezu eine notwendige Voraussetzung. Bezogen auf die festgestellten Urteilsprofile könnte man sich vorstellen - und dieser naheliegenden Erwartung folgt Diehl-, daß sich jeweils *vorlesungstypische* Konstellationen herausschälen, daß also in der einen Vorlesung zwei oder drei "benachbarte" Bewertungsprofile dominieren, in der anderen Vorlesung zwei oder drei andere Profile.

Diese Vermutung erweist sich jedoch als irrig: Läßt man sehr kleine Vorlesungen (weniger als 25 Hörer) außer Betracht, so existiert so gut wie keine (mindestens mittelgroße) Veranstaltung, in der bei den dozentenbezogenen oder bei den lernprozeßbezogenen Urteilsaspekten (vgl. Abb. 1 und 2) weniger als sechs verschiedene Beurteilungsmuster vertreten wären. Und selbst bei den kleinen Vorlesungen (10-24 Hörer; im allgemeinen Spezialvorlesungen im Hauptstudium mit homogenerer Hörerschaft) ist die (relativ) einheitliche Bewertung durch ihre Teilnehmer die seltene Ausnahme:

	dozentenbezogene Evaluationen (Abb. 1)						lernprozeßbezogene Evaluationen (Abb. 2)					
	unter 10 Hörer		10-24 Hörer		25 u.m. Hörer		unter 10 Hörer		10-24 Hörer		25 u.m. Hörer	
Urteils-profile pro Vorl.:	n	%	n	%	n	%	n	%	n	%	n	%
10	0	0	0	0	31	29,5	0	0	1	2,9	50	47,6
8-9	0	0	6	17,6	50	47,6	0	0	9	26,5	45	42,9
6-7	2	8,0	14	41,2	19	18,1	1	4,0	13	38,2	8	7,6
4-5	9	36,0	11	32,4	5	4,8	16	64,0	10	29,4	2	1,9
1-3	14	56,0	3	8,8	0	0	8	32,0	1	2,9	0	0
	25		34		105		25		34		105	

Diese Auszählung zeigt: In über Dreiviertel der mindestens mittelgroßen Vorlesungen finden sich unter den Hörern acht oder mehr verschiedene Dozentenbeurteilungs-Profile. Noch heterogener fallen die lernprozeßbezogenen Evaluationen aus: Mindestens acht verschiedene Urteilsmuster existieren hier in über 90 % der Veranstaltungen neben- bzw. gegeneinander; in fast jeder zweiten dieser Vorlesungen sind alle zehn Evaluationsprofile vertreten. Tendenziell nimmt die Heterogenität der Urteile zu mit der Größe der Hörerzahl, zusätzlich wenn es sich um Pflichtveranstaltungen ohne Wahlmöglichkeit zwischen Alternativen handelt. Noch uneinheitlicher wird die Situation, wenn an derselben Vorlesung Hörer unterschiedlicher Fächer teilnehmen. Mit anderen Worten: Derselbe "objektiv" gleiche Sachverhalt wird - für Psychologen wenig überraschend - von den beurteilenden Personen subjektiv umso unterschiedlicher wahrgenommen und bewertet, je vielfältiger deren Interessenstruktur ist.

3. Was ist "schlechte Sozialforschung"?

Die Antwort scheint - oberflächlich betrachtet - einfach: wenn die Methoden der Datenerhebung und -auswertung fehlerhaft angewendet werden. Das kommt zwar vor, aber in der Tat nicht sehr oft und ist auch für die Rezipienten solcher Forschung relativ leicht erkennbar.

Problematischer ist es, wenn unangemessene Modelle der Informationsbeschaffung und -aufbereitung gewählt werden. Um Fehler der Modellwahl handelt es sich,

- wenn Akzeptanzaussagen als Qualitätsurteile interpretiert werden (Akzeptanzurteile sind Aussagen über die auskunftgebende Person, nicht über den zu beurteilenden Sachverhalt),
- wenn per Umfrage erhobene Bewertungen mit empirischer Evaluation gleichgesetzt werden (Umfrageforschung - welche Fragen auch immer gestellt werden - ist von Design und Anspruch her etwas völlig anderes als Evaluationsforschung),
- wenn bei der Entwicklung von Meßskalen anstelle der Gültigkeit der verwendeten Indikatoren die Zuverlässigkeit des Meßvorgangs optimiert wird (selbst eine bis auf die Kommastelle stabile Messung nicht-gültiger Indikatoren bringt keine richtigen, sondern lediglich stabil falsche Ergebnisse),

- wenn komplexe Sachverhalte auf der Datenebene durch "grobe" Indikatoren übervereinfacht abgebildet werden oder ein interdependentes Beziehungsgeflecht auf univariate statistische Kennwerte reduziert wird (die empirische Realität wird nicht dadurch einfacher, daß man die Informationsbasis simplifiziert).

Ähnlich problematisch ist es aber auch, wenn ein durchaus anspruchsvolles, für einen bestimmten Typ von Fragestellungen hervorragend geeignetes Modell unreflektiert überall und immer eingesetzt wird (also auch für Fragestellungen, für die es ungeeignet ist). Das ist etwa der Fall bei dem allseits beliebten Modell der Faktorenanalyse, insbesondere in Kombination mit orthogonaler Rotation der Faktoren.

Für alle Formen von Modell-Fehlverwendungen gilt: Im günstigsten Fall führt das Vorgehen zu wenig informationshaltigen oder irrelevanten Befunden, im ungünstigsten Fall zu reinen Forschungsartefakten, die den Rezipienten vollständig in die Irre führen können. Für Beispiele lese man nach bei J. Kriz: Methodenkritik empirischer Sozialforschung.

Literatur:

Daniel, H.-D. (1995): Das Modellprojekt "Evaluation der Lehre" an der Universität Mannheim. In: Mohler, P. (Hrsg.): Universität und Lehre. Münster/New York, 2. Aufl.: Waxmann, 83-104.

Freter, H. (1992): Das Image von Lehrveranstaltungen und Professoren. Ergebnisse und Methodenprobleme aus Marketing-Sicht. Siegen: Ugh, Lehrstuhl für Marketing

HIS Hochschulinformationssystem (1992): Dokumentation Evaluation der Lehre, Teile 1 und 2. Hannover: HIS

Kriz, J. (1981): Methodenkritik empirischer Sozialforschung. Eine Problemanalyse sozialwissenschaftlicher Forschungspraxis. Stuttgart: Teubner

RCDS (1992): Prüf den Prof!!! Lehrstuhlbefragung am Betriebswirtschaftlichen Institut und am Institut für Wirtschaftsrecht der WiSo Nürnberg.