Helmut Kromrey

EVALUATION DER LEHRE DURCH UMFRAGEFORSCHUNG?
Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern[1]

veröff in:
Universität und Lehre (hrsg. von Peter Ph. Mohler)
Münster (Waxmann-Verlag)

0. Vorbemerkungen

Die Qualität von Lehre und Studium ist - mit Recht - ein Dauerthema in der hochschulpolitischen Diskussion. Lange Studienzeiten, hohe Fachwechsler- und Abbrecherquoten, überfüllte Hörsäle zu Semesterbeginn, nur noch teilweise gefüllte Hörsäle in der zweiten Semesterhälfte, Klagen über Desorientierung von Studenten im Hinblick auf Studienanforderungen und effektive Studiengestaltung sowie über unzureichende Studien- und Prüfungsleistungen auf der einen Seite, über ausufernde Stoffpläne, über mangelndes Engagement von Lehrenden, schlechte Betreuung und lustlose Lehre auf der anderen Seite sind aus dem aktuellen Hochschulalltag nicht wegzudiskutieren. Bemühungen um eine Verbesserung von Studium und Lehre sind in einer solchen Situation nicht nur (selbst)verständlich, sondern in jeder Hinsicht unterstützenswert. Naheliegend und kaum von der Hand zu weisen ist es auch, wenn die zuletzt genannten Aspekte (Stoffpläne, Lehre, Betreuung) als - zumindest *eine* - Ursache, die zuerst genannten Aspekte (Studiendauer und -leistungen) dagegen als deren Folge angesehen werden. "Aktionsprogramme" zur Qualität der Lehre, initiiert von den für die Hochschulen zuständigen Ministerien auf Länder- und Bundesebene, sind da nur die logische Konsequenz. Auch die Einbettung systematischer Evaluation von Lehrqualität in solche Programme ist prinzipiell zu begrüßen; schließlich sollen die Aktionen und Maßnahmen auf unbestreitbaren, empirisch abgesicherten Erkenntnissen basieren und nicht von gängigen Vorurteilen diktiert sein.

Hier endlich ist Skepsis angesagt: Kann die empirische Forschung solche "abgesicherten" oder gar "unbestreitbaren" Erkenntnisse überhaupt bereitstellen? Und falls prinzipiell ja: Wie muß das dafür geeignete methodische Verfahren aussehen?

1. Ist "Qualität der Lehre" in einzelnen Lehrveranstaltungen "meßbar"?

Konkret wird die Thematik "Qualität von Lehre und Studium" im allgemeinen zugespitzt auf die Frage nach der "guten Lehre", und diese wiederum wird ohne längeres Nachdenken bezogen auf die einzelne Lehrveranstaltung, meist sogar noch weiter eingeschränkt auf den einzelnen Dozenten bzw. die einzelne Dozentin und "seine" bzw. "ihre" Lehre. So liest man etwa im SPIEGEL 16/1993, S. 86: "Ein Großteil der Hochschullehrer in Deutschland ist unfähig, gut zu lehren"; bzw. (ebda.): "Lehrflaschen demotivieren ungezählte junge Menschen und richten damit einen kaum wieder gutzumachenden Schaden an."

Wie aber - so ist die zwangsläufig an solche Vorwürfe sich anschließende Frage - kann man sie denn erkennen, die "gute Lehre"? Schon aus unserer Alltagserfahrung wissen wir, daß wir etwas erst einmal *kennen* müssen, um es auch *er*kennen zu können; erst recht, um *anderen* sagen zu können, wie und woran *sie* es erkennen sollen.

Was also *ist* "gute Lehre"? Wiederum wissen wir aus unserer Alltagserfahrung, daß diese Frage nicht eindeutig beantwortbar ist, weil es *die* (eine) gute Lehre nicht gibt, gar nicht geben kann. Schon deshalb können wir "Qualität der Lehre" erst dann beurteilen, wenn wir zuvor festlegen: *Was* soll *für wen* und *zu welchem Zweck* gelehrt werden? D.h.: Was sind die Lehr- und die Lernziele?

Sollen beispielsweise

- die für das erfolgreiche Studium eines Fachs von den Lehrenden für notwendig gehaltenen und bei den Studierenden überwiegend ungeliebten Basiskenntnisse und Basisfertigkeiten vermittelt werden

- *und* muß man zudem davon ausgehen, daß es nennenswerte Lücken in den studentischen Vorkenntnissen gibt,
- dann ist Lehre ganz anders zu konzipieren, als wenn Studierende im Hauptstudium in einem sie interessierenden Teilgebiet zum kreativen Umgang mit wissenschaftlichen Erkenntnissen oder gar zum Finden *neuer* Erkenntnisse angeleitet werden sollen.

Was für den ersten Fall gute Lehre ist, wäre für den zweiten Fall extrem schlechte Lehre - und umgekehrt.

Nun macht aber eine solch differenzierende Betrachtungsweise die Angelegenheit ärgerlich kompliziert und ist zudem *dann* völlig unpraktikabel, wenn man ein standardisiertes Steuerungsinstrument zur Verbesserung der Studiensituation im Sinn hat. Ist man nämlich überzeugt, die "Effizienz des Studiums" durch eine - wie auch immer verstandene - bessere "Qualität der Lehre" erhöhen zu können, dann sucht man natürlich nicht nach differenzierten, sondern nach möglichst generellen Kriterien für "Qualität", nach Kriterien, die gerade stoff- und fach- und lernziel*un*abhängig gelten sollen. Man benötigt allgemeine (und zwar möglichst wenige) Vergleichsmaßstäbe, die "gut" von "schlecht" und die "gut" von "besser" zu unterscheiden erlauben, ohne dabei auf die Besonderheiten einzelner Fächer und einzelner Lehrveranstaltungen Rücksicht nehmen zu müssen.

Was lag in dieser Situation näher, als zum einen auf Ansätze der Hochschuldidaktik aus den 70er Jahren zurückzugreifen -studentische Lehrveranstaltungskritik- und zum anderen auf den Usus an US-amerikanischen Hochschulen zu verweisen: Einschätzung der Lehre und der Lehrenden durch Studierende anhand kurzer Fragebögen mit relativ globalen (also: inhaltsunabhängigen) Beurteilungen. Die Begründung hierfür fällt -scheinbar- leicht: Die Studierenden sind die von der Lehre ganz konkret "Betroffenen"; als solche können sie aus eigener Erfahrung auch deren Qualität zuverlässig beurteilen. Übersehen wird bei dieser oberflächlichen Übernahme von Konzepten, daß die Kontexte ihres Einsatzes ganz andere waren bzw. sind als bei ihrer heute beabsichtigten Verwendung als Qualitätsmaßstab.

Bei der studentischen Lehrveranstaltungskritik der 70er Jahre ging es um ein Instrument der Rückmeldung an den jeweiligen Dozenten von *seinen* Hörern über *seine* Lehrveranstaltung. Angestrebt war eine Rückspiegelung in *denselben* Kontext; es ging keinesfalls um einen *Vergleich zwischen* verschiedenen Veranstaltungen oder zwischen verschiedenen Lehrenden. Dafür - also: zur Rückmeldung - eignet sich dieses Instrument in der Tat hervorragend (und zwar in unterschiedlichen Graden der Standardisierung je nach der Größe der Teilnehmerzahl - und immer auch auf den *Inhalt* der Lehre bezogen).

Bei den US-amerikanischen Befragungen geht es um Akzeptanzerhebungen; die zentrale Frage ist, ob die Lehre bei den Studierenden "gut ankommt". Dies ist ein verständliches Informationsinteresse, wenn Hochschulen sich über Studiengebühren finanzieren müssen. Daß etwas "gut ankommt", daß etwas auf hohe Akzeptanz stößt, sagt aber nicht unbedingt auch etwas über die Qualität dessen aus, was beurteilt werden soll, sondern es sagt in erster Linie etwas über den Urteilenden und seine Präferenzen aus. Die Urteilskriterien können von Person zu Person und von Bewertungsgegenstand zu Bewertungsgegenstand ganz unterschiedlich sein. Beispielsweise - und auch das wissen wir aus unserer Alltagserfahrung - sind die Kriterien, nach denen wir etwas als das "kleinere Übel" einschätzen, ganz andere als die, nach denen wir unter mehreren guten Alternativen die beste aussuchen.

Warum sollte das alles im Hörsaal oder im Seminarraum plötzlich ganz anders sein? Auch bei der Bewertung von Lehre durch die "von Lehre Betroffenen" kommt es entscheidend an auf das individuelle Wechselspiel

- von Interesse der Studierenden und Anforderungen des Studienplans,
- von studentischem Lernverhalten und Lehre,
- von studentischen Vorkenntnissen und Anforderungen des Lernstoffs,
- von studentischer Arbeitsfähigkeit und den Rahmenbedingungen für das Lernen in der Universität.

Von diesen Gegebenheiten und manchen anderen hängt es im konkreten Fall ab, wie das individuelle Urteil ausfällt. Das heißt: Das Urteil des "Rezipienten" bezieht sich nicht allein auf den "Sachverhalt Lehre", sondern immer auf das gesamte, komplexe Beziehungsfeld, in dem unter anderem - aber eben: unter anderem - *auch* die jeweilige Lehrveranstaltung ihren Platz hat. Die entscheidende Frage ist also: Hat für die "Qualität des Studiums" der Teilaspekt "Lehre *in der jeweiligen Lehrveranstaltung*" wirklich den hohen Stellenwert, der ihm in den Programmen zur Qualität der Lehre zugeschrieben wird? Falls nein (und die vorliegenden Befunde der Lehrevaluationen nicht nur an deutschen Hochschulen deuten darauf hin), wäre es wenig erfolgversprechend, sich *allein* auf diesen Bereich zu konzentrieren.

2. Evaluation der Lehre oder Selbstevaluation der Studierenden?

Die zuletzt aufgeworfene Frage ist noch um einen wichtigen Schritt weiterzuführen: Sind studentische Evaluationen - insbesondere, wenn sie sich nicht auf eher deskriptive Aussagen über Details[2] beziehen, sondern "zusammenfassenden" Charakter haben - wirklich (bzw. zumindest in erster Linie) Urteile über die Lehrleistung oder vielleicht eher Aussagen über den Urteilenden selbst, sozusagen "Selbstbezichtigungen". Die Antwort fällt, wie am folgenden Beispiel illustriert wird, weit schwerer als von Verfechtern studentischer *Lehrkritik als Evaluation* unterstellt wird.[3]

Es dürfte unbestritten sein, daß sich jemand einer Aufgabe dann lieber stellt, wenn er sich freiwillig dafür entschieden hat und wenn die Aufgabe in sein eigenes Interessengebiet fällt. Bezogen auf die Teilnahme an und die Mitarbeit in Lehrveranstaltungen: Es dürfte einen Unterschied ausmachen, ob Studierende eine Lehrveranstaltung ausschließlich gezwungenermaßen besuchen (z.B. nur, weil am Ende des Semesters die Klausur droht), oder ob sie aus eigenem Antrieb mitarbeiten; bzw. in psychologischer Formulierung: ob die Teilnahmemotivation extrinsisch oder intrinsisch ist.

Welche Auswirkungen hat nun (1.) die unterschiedliche Teilnahme-Motivation auf das Urteil der einzelnen Studierenden über die Qualität der ihnen gebotenen Lehre? Und welche Auswirkungen hat es (2.) auf die Beurteilung einer gesamten Lehrveranstaltung, wenn sich in ihr überwiegend intrinsisch oder überwiegend extrinsisch motivierte Studierende befinden?

Die folgenden beiden Tabellen geben eine empirische Antwort. Stellvertretend für andere Urteilsdimensionen wird zunächst die Globalbeurteilung der Lehrveranstaltung insgesamt[4] herausgegriffen und einem Index gegenübergestellt, der die Studierenden anhand ihrer Teilnahme-Motivation grob in drei Gruppen einteilt:[5]

Tabelle 1: S04 Frage 4: Erwartungen erfüllt?
by INTRINS Intrinsische Motivation

	INTRINS			
Count	*			
Col Pct	*extrins.	intrins.	intrins.,	
	*	+ Pfl.	o.Pfl.	Row
	* -1	* 0	* +1	* Total
S04	3	3	3	1
-	* 1735	* 373	* 102	* 2210
negative Urteile	* 30,7%	* 11,5%	* 7,7%	* 21,6%
	3	3	3	1
0	* 1551	* 619	* 195	* 2365
	* 27,4%	* 19,0%	* 14,6%	* 23,1%
	3	3	3	1
+	* 2370	* 2262	* 1036	* 5668
positive Urteile	* 41,9%	* 69,5%	* 77,7%	* 55,3%
	2	2	2	-
Column	5656	3254	1333	10243
Total	55,2%	31,8%	13,0%	100,0%

Man erkennt: Über alle Fakultäten hinweg[6] bewerten Studierende, die den Vorlesungsbesuch als (lästige) Pflicht absolvieren, die von ihnen besuchte Veranstaltung deutlich schlechter als andere Studierende. Nur gut 40 % der Pflicht-Hörer fällen positive Urteile (Punktwert +1: 34%: +2: 8%), fast ein Drittel äußert sich explizit negativ. Auffallend ist die Größe dieser Studierendengruppe: Gut

55 % beträgt der Anteil der Befragten, der nur deshalb die Veranstaltung besuchte, weil am Ende eine Klausur oder eine andere Prüfung "drohte" (weitere Besuchsgründe wurden von ihnen nicht angekreuzt). Ist bei den beurteilenden Studierenden die Konstellation entgegengesetzt (kein Klausur- oder Prüfungsdruck, stattdessen Eigeninteresse oder Wahlveranstaltung), dann fallen auch die Urteile tendenziell entgegengesetzt aus: Mehr als drei Viertel loben die Veranstaltung, nur knapp 8 % äußern Kritik. Sehr ähnlich ist das Ergebnis, wenn extrinsische und intrinsische Motivation parallel bestehen ("intrins. + Pflicht"), wenn also Prüfungs-/Klausurpflicht durch Eigeninteresse der Teilnehmer *ergänzt* wird.

Nun verteilen sich jedoch die Teilnahmemotive der Hörer nicht gleichmäßig über alle Lehrveranstaltungen. Vielmehr gibt es Vorlesungen, die weit überwiegend von Nur-Pflicht-Hörern besucht werden, und solche, in denen die Interessierten in der Mehrzahl sind. Allein schon diese Teilnehmerstruktur bestimmt zu einem außerordentlich hohen Maß, ob das "durchschnittliche" Urteil über die "evaluierte" Lehrveranstaltung außerordentlich gut oder außerordentlich schlecht ausfällt.

Tabelle 2: S04 Frage 4: Erwartungen erfüllt?
by TEILNMOT Teilnehmermotivation in der Vorlesung[7]

S04 Erwartungen	TEILNMOT	> 90 % extrins.	> 80 % extrins.	gemischt > 70 %	> 75 % intrins.	intrins.	Row Total
überh.nicht erf.	-2	151	257	135	20	1	564
		16,1	9,4	3,0	1,3	,2	5,5
	-1	266	658	604	107	11	1646
		28,3	24,0	13,4	6,9	2,2	16,1
	0	266	787	1033	238	41	2365
		28,3	28,7	22,9	15,4	8,2	23,1
	+1	216	860	2006	720	172	3974
		23,0	31,4	44,4	46,5	34,5	38,8
Erwartungen voll erfüllt	+2	40	177	739	464	274	1694
		4,3	6,5	16,4	30,0	54,9	16,5
Column Total		939	2739	4517	1549	499	10243
		9,2	26,7	44,1	15,1	4,9	100,0

Vergleicht man die beiden Extremsituationen für Lehrende - Vorlesungen mit fast nur gezwungenermaßen teilnehmenden Studenten gegenüber Vorlesungen mit weit überwiegend interessierten Teilnehmern -, dann wird überdeutlich: In einem Setting mit (fast) nur Pflichthörern hat (in allen Fakultäten) die Dozentin oder der Dozent wenig Chancen, zu "guten Noten" für ihre/seine Lehre zu kommen. Nur 27 % der Urteile fallen positiv aus, über 44 % äußern explizit Kritik. Ist dagegen die große Majorität der Studierenden interessiert, kann der Lehrende wenig verkehrt machen: Den ungefähr 90 % positiven Urteilen stehen lediglich 2 % kritischer Bewertungen gegenüber.

Übrigens ergeben sich die Werte in Tab. 2 nicht allein aus der Kumulation des je individuellen Zusammenhangs zwischen Motivation und Bewertungsrichtung, wie er in Tab. 1 aufgezeigt wird. Hinzu kommt (aus der psychologischen Gruppenforschung bekannt) der Einfluß der engeren sozialen Umgebung auf das individuelle Urteil: In einer als weitgehend homogen wahrgenommenen Hörerschaft orientiert sich der Studierende an der Meinung seiner Kommilitonen. D.h. rein *ex*trinsisch motivierte Hörer urteilen in einer überwiegend interessierten Hörerschaft eher positiv; *in*trinsisch motivierte Hörer dagegen urteilen in einer überwiegend desinteressierten Hörerschaft eher negativ.

Zum Beispiel: Die Nur-Pflicht-Hörer insgesamt äußern zu 30,7 % Kritik und zu 41,9 % Lob (vgl. Tab. 1); in einer extrinsisch majorisierten Umgebung steigt bei ihnen der Anteil negativ Urteilender auf 47,5 % und sinkt der

Anteil positiver Urteile auf 23,9 %. Analoges gilt für intrinsisch motivierte Hörer. Laut Tab. 1 äußern sie sich insgesamt zu 77,7 % positiv und zu 7,7 % negativ. In einer positiv eingestellten Umgebung jedoch verteilen sie zu 88,9 % Lob und nur zu 2,6 % Tadel; in einer überwiegend negativ eingestellten Umgebung sinkt demgegenüber der Anteil positiver Urteile auf 52 % und steigt der Anteil negativer Bewertungen auf 18,7 %. Pointiert formuliert: Wer das Glück hat, vor interessierten Hörern zu lehren, wird dafür hoch gelobt. Wer die schwerere Aufgabe hat, Pflichtstoff vor wenig interessierten Studierenden zu vermitteln, erntet dafür Kritik.

Ähnliche Effekte lassen sich für (fast) beliebige andere Merkmale aufzeigen.

Um zur Eingangsfrage dieses Abschnittes zurückzukommen (Evaluation der gebotenen Lehrleistung oder Selbstbeurteilung des Urteilenden?): Beide Aspekte sind in den Fragebogenantworten konfundiert, und zwar - je nachdem, um welche Frage und um welche Vorlesung es sich handelt - in variierendem Ausmaß. Die Auswertung der Lehrveranstaltungs-Umfragedaten mit dem Ziel "Evaluation" verlangt daher komplexe, quasi-experimentell angelegte Analyseverfahren.[8]

3. Sind Durchschnittswerte pro Lehrveranstaltung ein gültiger Qualitätsindikator?

Verfolgt man die Präsentation der Ergebnisse studentischer Veranstaltungsumfragen an deutschen Hochschulen, so fällt auf, daß meist relativ unkritisch je Beurteilungs-Item arithmetische Mittel pro Veranstaltung berechnet und als Qualitätsindikator interpretiert werden. Eine durchaus verständliche Strategie; das Verfahren besticht durch die Einfachheit der Handhabung und die leichte Vermittelbarkeit seiner Ergebnisse auch gegenüber Nicht-Statistikern.

Die implizite Begründung dieses Vorgehens scheint in Überlegungen zu liegen, wie sie etwa im nordrhein-westfälischen Aktionsprogramm Qualität der Lehre explizit zum Ausdruck gebracht werden, wo es zur "Beurteilung der Lehrveranstaltungen und damit auch der Lehrenden" durch Studenten heißt: "Wertungen sind hierbei einerseits unvermeidlich, andererseits aber durchaus aussagekräftig, wenn sie in großer Zahl übereinstimmen." (MWF NW 1991, 124). Korrekterweise wird hier die (relativ große) Übereinstimmung der Teilnehmerurteile innerhalb der Lehrveranstaltungen als Bedingung genannt. Wäre sie erfüllt, d.h. gäbe es eine mehrheitlich übereinstimmende Beurteilung, so wäre sicher auch die weitere (ergänzende) Unterstellung zu rechtfertigen, daß die verbleibenden individuellen Abweichungen keine ins Gewicht fallende Verzerrung verursachen, sondern sich bei hinreichend großer Zahl von Befragten ausgleichen. Diese methodische Unterstellung, deren Zutreffen eine wesentliche Bedingung für die Anwendbarkeit einfacher Erhebungs- und einfacher Auswertungsverfahren wäre, erweist sich jedoch als falsch. Ganz im Gegenteil muß als Regelfall die außerordentlich große Heterogenität der Teilnehmerurteile konstatiert werden; die erhoffte (relativ große) Übereinstimmung findet sich allenfalls als seltene Ausnahme in den Daten wieder.

Dennoch: Selbst die festzustellende, überraschend hohe Variation der Antworten in (fast) jeder Vorlesung spricht allein immer noch nicht prinzipiell gegen eine Verwendung von Mittelwerten (gegebenenfalls des Medians anstelle des arithmetischen Mittels), solange wenigstens eine andere, weniger restriktive Bedingung erfüllt ist, nämlich daß in Veranstaltungen mit "guter Lehre" die Urteile der Teilnehmer (trotz Variation im Detail) *in ihrer Tendenz* relativ positiver, in Veranstaltungen mit "schlechter Lehre" *in ihrer Tendenz* relativ negativer ausfallen. Vorstellbar wäre dann zumindest eine grobe Klassifikation in (wenige) unbestritten "gute" Veranstaltungen auf der einen Seite sowie (wenige) unbestritten "schlechte" Veranstaltungen auf der anderen Seite[9] und dazwischen die große Masse von im Detail umstrittenen, jedoch in der Tendenz als "eher positiv" bis "eher negativ" einordenbaren Vorlesungen.

Die (implizite) Annahme für diesen Fall wäre, daß sich die Beurteiler zwar hinsichtlich der *Ausprägungen* ihrer Einschätzungen unterscheiden, daß sie ihre Urteile aber dennoch anhand der gleichen Evaluations-*Dimensionen* bilden. Träfe dies zu, dann müßten Studierende mit ähnlichem Gesamturteil sich auch in den Detaileinschätzungen ähneln, d.h. dann würden Studierende mit negativer Gesamtevaluation auch in den Details negativer evaluieren als Studierende mit positivem Gesamturteil. Auch diese Annahme erweist sich jedoch, wie eine differenzierte Datenanalyse zeigt, als Irrtum. Der Regelfall ist nicht nur eine außerordentlich große "Inter-Befragten-Heterogenität" (*Teilnehmer derselben Veranstaltung* beurteilen diese in großer Vielfalt und aus offensichtlich sehr

unterschiedlichen Perspektiven); zusätzlich zeigt sich eine überraschend starke "Intra-Befragten-Heterogenität" (Unterschiede und Widersprüchlichkeiten zwischen Detail- und Globalurteilen *beim einzelnen Befragten*). Mit anderen Worten: *Dieselbe* Form der Lehre ist für einen Teil der Hörer akzeptabel und wird als lernfördernd empfunden, für einen anderen Teil ist sie unakzeptabel und lernhemmend, von weiteren Hörergruppen schließlich wird sie in einigen Aspekten als positiv, in anderen als negativ wahrgenommen.

Dies soll im folgenden anhand von zwei Clusteranalysen illustriert werden - zum einen mit Aussagen zur "Lehre im engeren Sinne" (Vortrags- und Darstellungsweisen, Lernstoff, Dozentenverhalten), zum anderen mit eher studienbezogenen Urteilen (Vorkenntnisse, Förderung von Interesse und Arbeitsbereitschaft, Lernerfolgseinschätzung).[10]

3.1 Evaluation der "Lehre im engeren Sinne": Bewertungsprofile I

Kommen wir kurz zurück zur bereits diskutierten Unterstellung einer (weitgehenden) Homogenität der Urteile der Vorlesungsteilnehmer/innen. Ihr Zutreffen würde bedeuten, daß sich für jede einzelne Lehrveranstaltung ein je typisches "Bewertungsprofil" herausbildet. Zum Beispiel: Die Mehrheit der Evaluierenden in einer Veranstaltung fände die Lehre des Dozenten auf den Urteilsskalen A bis X ziemlich gut, in einer anderen Veranstaltung ziemlich schlecht; oder auch: auf der Urteilsskala A gut bis sehr gut, auf der Skala B mittelmäßig, auf der Skala C sehr gut bis ausgezeichnet, auf der Skala D ziemlich schlecht etc. Von diesem "Bewertungsprofil" gäbe es zwar individuell mehr oder weniger große Abweichungen, diese wären jedoch nicht von systematischer Art. In diesem Fall wäre die Berechnung von Durchschnittswerten je Urteilsskala ein nicht nur einfaches, sondern auch angemessenes Auswertungsverfahren.

Auswertungsmethodisch schwieriger ist dagegen die - faktisch vorzufindende - Situation, daß die Teilnehmer-Urteile in systematischer Weise heterogen sind, so daß sich in den einzelnen Veranstaltungen "konkurrierende" Urteilsprofile herausbilden. Dann nämlich tritt der Fall ein, daß z.B. ein Teil der Evaluierenden in derselben Veranstaltung die Lehre des Dozenten auf den Skalen A bis D etc. durchweg positiv, ein anderer Teil dieselbe Lehre durchweg negativ, ein dritter Teil durchweg mittelmäßig, ein vierter auf den Skalen A und B positive, auf den Skalen C und D negative Einschätzungen abgibt usw. Das heißt, es gibt (intern homogene) Teilnehmer-Gruppen, deren Mitglieder sich jeweils *untereinander* in ihren Qualitätsurteilen relativ einig sind, die sich jedoch von *anderen* (wiederum intern homogenen) Teilnehmer-Gruppen deutlich unterscheiden. In diesem Fall ist die Berechnung von Skalenmittelwerten pro Vorlesung ein nicht nur unzweckmäßiges, sondern ein fehlerhaftes und in die Irre führendes Verfahren. Im Extrem kann der Fall eintreten, daß ein aus Skalenmittelwerten gebildetes "Durchschnittsprofil" ein reines statistisches Artefakt darstellt, d.h. daß es keine einzige nennenswert große Gruppe von Studierenden gibt, die auf allen Skalen "durchschnittliche" (jedenfalls nah am Durchschnitt liegende) Werte angekreuzt hat.

Ein aus den Mittelwerten pro Bewertungsskala konstruiertes "Durchschnittsprofil" ergibt im vorliegenden Fall für die *Lehre im engeren Sinne* die folgende Konstellation:

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

Der sich aufdrängende Eindruck ist bei diesem Vorgehen geradezu überwältigend positiv: Die Lehre scheint hiernach "im Durchschnitt" in fast jeder Hinsicht "gut" bis "sehr gut" zu sein; mit einer Tendenz zu "befriedigend" in gerade nur zwei Aspekten: nicht immer hinreichende Klarheit über die gestellten Lernanforderungen und zum Teil mangelndes Eingehen auf studentische Wünsche. Das Problem ist nur: Ein solches "Durchschnittsprofil"' existiert empirisch unter den Vorlesungsteilnehmern nicht! Es handelt sich in der Tat um ein statistisches Artefakt. Stattdessen existieren unter den Hörern der Vorlesungen eine ganze Reihe qualitativ sehr verschiedenartiger Urteilskonstellationen, die bei graphischer Darstellung (Abb. 1) fast wie ein Schnittmusterbogen für Hobbyschneider/innen wirken:[11]

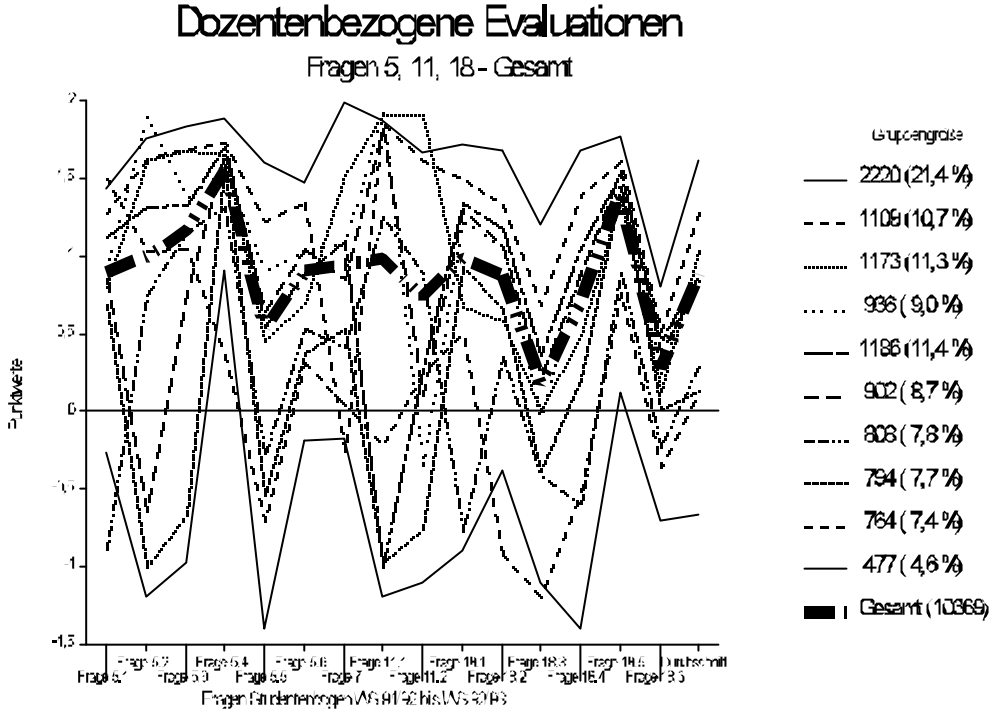

Figure 1 Gesamtübersicht

Genau betrachtet, erfüllt sich die Erwartung, daß studentische Teilnehmerurteile "in großer Zahl übereinstimmen", nur für drei der gestellten dozentenbezogenen Fragen:

- Tendenziell kritisch - sieht man von den besonders positiven Urteilsprofilen 1 und 2 ab - äußern sich die Studierenden darüber, daß nicht immer ganz klar ist, welche Lernanforderungen an sie gestellt werden (Frage 18.3), d.h. was sie denn eigentlich lernen sollen (dies gilt - wenig überraschend - für Studierende der Anfangssemester stärker als für höhere Semester).

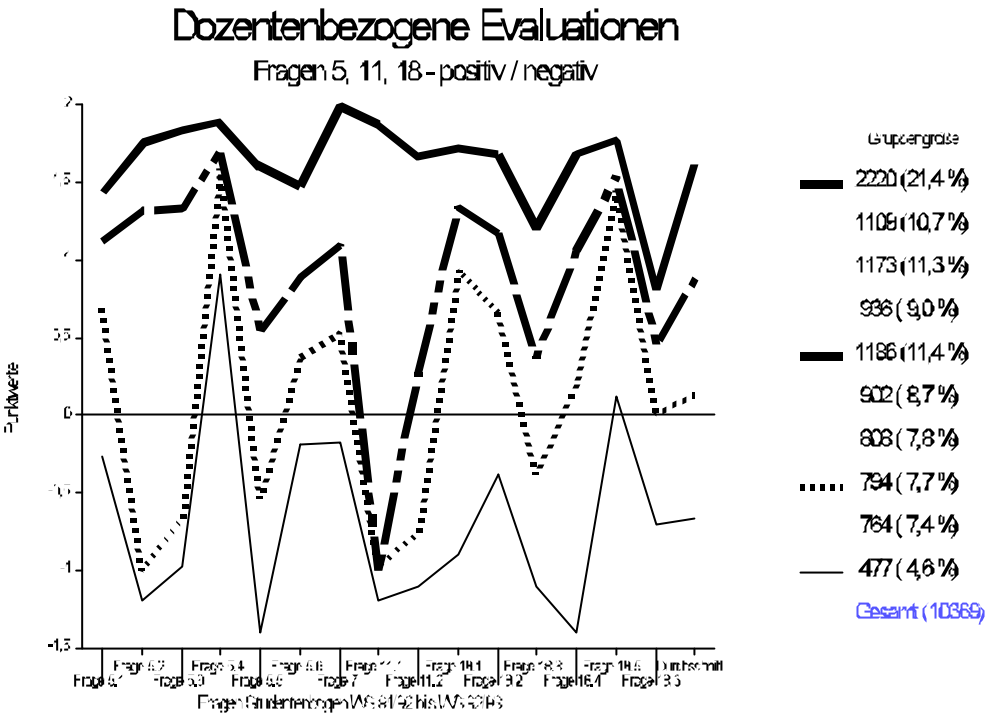

Figure 2 "konsistente" Urteilsprofile: von durchweg positiv bis (fast) durchweg negativ

- Durchweg positiv - selbst von den ansonsten überwiegend kritisch antwortenden Studentengruppen - wird die Themenzentrierung des Vortrags beurteilt (Frage 5.4: nicht zu eng am Thema klebend, aber auch nicht zu abschweifend).
- Ähnliches gilt für die Bereitschaft der Dozenten, auf Zwischenfragen aus dem Auditorium einzugehen (Frage 18.5; Ausnahmen: das mit Abstand negativste Profil 10, das jedoch nicht einmal 5 % der Hörer repräsentiert).

Ansonsten aber dominiert die Meinungs- und Urteilsvielfalt unter den Vorlesungsteilnehmern.

Auch eine Deutung der Evaluationsprofile als Antwort auf die Frage nach der "guten Lehre" ist nur zum Teil möglich. Zwar kann - was in solchen Fällen häufig geschieht - für jeden Befragten ein additiver Index "durchschnittliches Gesamturteil" aus den Einzelitems[12] gebildet und als Güte-Indikator interpretiert werden. Doch schon ein kurzer Blick auf den Verlauf der Evaluationsprofile demonstriert die mangelnde Gültigkeit dieses Konzepts "Gesamtbewertung als Summe der Einzelurteile". Nur vier der zehn Bewertungsprofile[13] weisen Kurvenverläufe auf, die eine "konsistente" Deutung des Gesamtindexwertes als Güte-Indikator erlauben (vgl. Abb. 2); d.h. ein höherer Gesamtwert bedeutet hier, daß auch alle Detail-Urteile positiver (zumindest nicht negativer) ausfallen. Diese vier "konsistenten" Profile repräsentieren lediglich 45 % der Befragten. Anders ausgedrückt: Für 55 % der Urteilenden hätte ein Durchschnittsindex über alle Items eine äußerst zweifelhafte Gültigkeit, da so unterschiedliche Aspekte wie freier Vortrag ("nicht ablesen"), Vortragstempo, Medieneinsatz, Schwierigkeitsgrad des Stoffs, Klarheit der Lernanforderungen, erkennbares Interesse des/der Lehrenden usw. als sich gegenseitig ausgleichende Aspekte der Lehrqualität gleichgewichtig in die Berechnung eingehen.

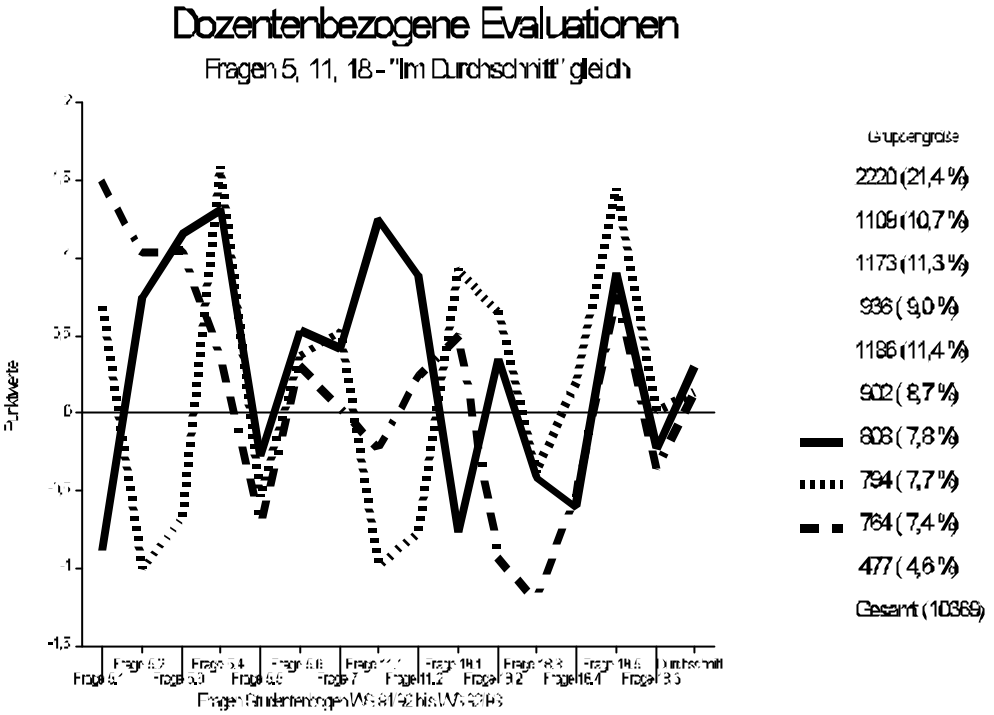

Figure 3 Beispiel: Urteilsprofile mit gleichem Durchschnittswert über alle Items

Wie problematisch ein solches Vorgehen ist, kann aus Abb. 3 abgelesen werden. In ihr ist die Gestalt von drei Evaluationsprofilen dargestellt, die hinsichtlich des Durchschnitts der Item-Punktwerte als gleich erscheinen (Gesamtindex jeweils ca. +0,2). Im Detail jedoch sind sich die drei so repräsentierten Gruppen von Urteilenden lediglich bei vier der 15 Items einig: Der sprachliche Vortrag ist nicht immer gut verständlich (Frage 5.5), obwohl meist nicht zu viele ungeläufige Begriffe und Fremdwörter verwendet werden (Frage 5.6); außerdem geht der/die Lehrende in richtigem Ausmaß auf Zwischenfragen der Hörer/innen ein (Frage 18.5), wenngleich die Bereitschaft zur Berücksichtigung studentischer Wünsche nur mittelmäßig beurteilt wird (Frage 18.6). Ansonsten schöpfen die Einschätzungen je nach Gruppe fast das ganze Spektrum von sehr positiv bis sehr negativ aus.

So bedeutet etwa das Profil 7 (durchgezogene Linie): überwiegend gelesen, nicht zu schnell, richtige Konzentration auf das Thema, nicht zuviel und nicht zu schwieriger Stoff, jedoch desinteressierter, den Studierenden gegenüber sich distanziert gebender Dozent; dagegen das Profil 8 (dünn gestrichelte Linie): überwiegend frei vorgetragen, aber zu schnell, zum Teil zuviel Redundanz, zuviel und zu schwieriger Stoff, jedoch persönlich interessierter, den Studierenden gegenüber aufgeschlossener, "ansprechbar" wirkender Dozent. Profil 9 (dick gestrichelte Linie) schließlich zeichnet sich durch Einschätzungen aus, die zum Teil zwischen den beiden vorhergehenden angesiedelt sind (etwa Menge und Schwierigkeitsgrad des Stoffs), teilweise solche Aspekte explizit kritisieren, die bei den anderen auf mittlere Werte stoßen (etwa mangelnde Vorlesungsstruktur, fehlende Klarheit der Lernanforderungen).

Es zeigt sich: Diese drei Beurteilungsmuster als "im Durchschnitt aller Einzelbewertungen etwa gleich" gelten zu lassen, das wäre nicht nur zu pauschal, sondern das wäre schlichtweg falsch! Im Detail betrachtet, erweisen sich die Urteile bei wesentlichen Einzelaspekten sogar als Gegensätze!

Die gleiche Konsequenz ist im übrigen für die Verwendung pauschalisierender Fragen zu ziehen, die vom Befragten Globalurteile abverlangen. Eine entsprechende Frage erbrachte Antworten, die mit dem oben geschilderten Durchschnittsindex hoch korrelieren und deren Durchschnitt pro

Bewertungsprofil weitgehend dem Indexdurchschnitt entspricht.

3.2 Evaluation von Lernprozeß und Lernerfolg: Bewertungsprofile II

In der Tendenz deutlich negativer, jedoch ähnlich heterogen - quantitativ sogar noch heterogener - fallen die Beurteilungsmuster aus, wenn man die Antworten auf diejenigen Fragen analysiert, die sich um den *Lernprozeß und die Einschätzung des eigenen Lernerfolgs* drehen.

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

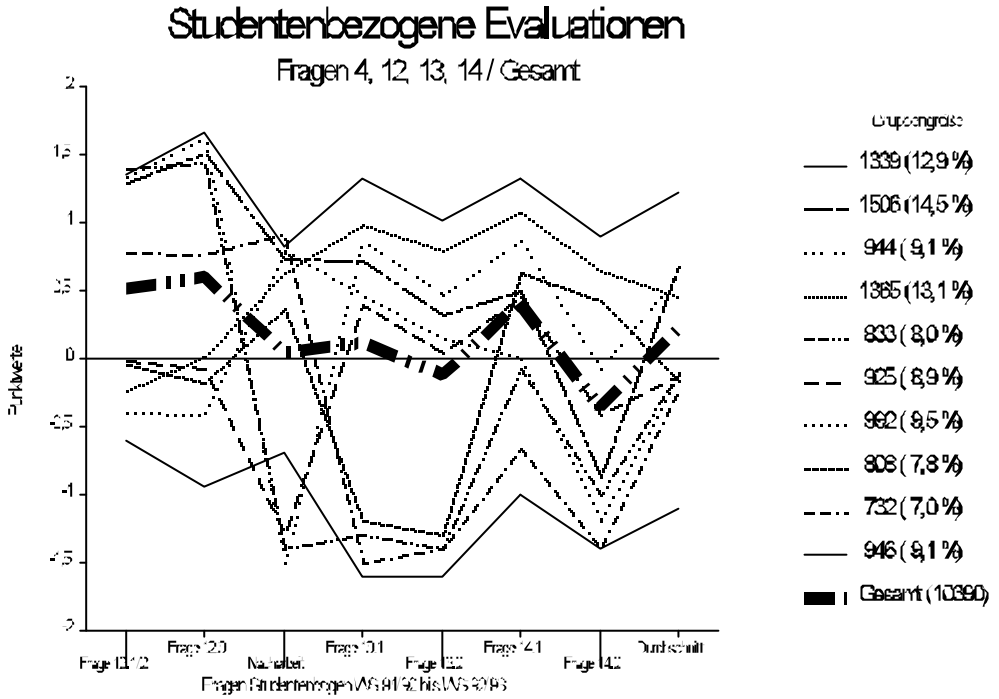

Figure 4 Gesamtübersicht

Auch hier wäre ein aus den Mittelwerten je Urteilsskala konstruiertes Urteilsprofil lediglich ein statistisches Konstrukt - ein Artefakt -, das unter den Vorlesungshörern in dieser Form kaum vorkommt, wie die Abb. 4 veranschaulicht.

Zwar sind hier (läßt man die eher deskriptiven Angaben über die eigene Arbeit außer Betracht) fünf Profile als "konsistente" Qualitätsbewertung der Lernsituation zu lesen (Profile 1, 2, 3, 7, 10, die zusammen 55 % der Beurteilungen repräsentieren).[14] Dafür zeichnen die verbleibenden fünf Profilverläufe umso heterogenere und in sich widersprüchlichere Bilder.

Frappierend ist insgesamt die Deutlichkeit, mit der sich die gern aufgestellte Gleichung "Gute Lehre = Lernerfolg" als eine zu stark vereinfachende monokausale Ursachenzuschreibung erweist. Wertet man die Aussagen "In der Vorlesung wurden Vorkenntnisse vorausgesetzt, die bei mir nicht vorhanden waren" (Frage 12.1) sowie "Deshalb [wegen vorausgesetzter, aber fehlender Vorkenntnisse] fiel es manchmal schwer, der Vorlesung zu folgen" (Frage 12.2) als Indikatoren für schlechte Lernvoraussetzungen (mangelnde Zielgruppenangemessenheit der Lehre), so zeigt sich:

Auch *bei ungünstigen Lernvoraussetzungen* kann der Stoff dennoch als interessant und anregend empfunden werden, und man kann nach dem eigenen Eindruck einiges gelernt haben. Bei gleichen Voraussetzungen ist aber auch der entgegengesetzte Verlauf möglich: kein Interesse geweckt, keine Anregung zur Weiterarbeit, wenig oder gar nichts gelernt. Oder: zwar nicht interessant und anregend, aber dennoch einiges gelernt; bzw. das Gegenteil davon: einigermaßen interessant und anregend, aber wenig gelernt. Ähnlich stellt sich die Situation *bei günstigen Lernvoraussetzungen* dar. Selbst wenn keine unangemessenen Vorkenntnisse vorausgesetzt wurden und keine Schwierigkeiten beim Verfolgen der Vorlesung auftraten, kann der Stoff als wenig interessant und anregend empfunden und kann der wahrgenommene Lernerfolg außerordentlich gering sein. Oder auch: der Stoff erscheint zwar einigermaßen interessant und anregend, dennoch herrscht der Eindruck vor, viel gelernt zu haben.

Einen Durchschnittsindex als Global-Indikator für so etwas wie "Qualität des Lernens" aus den einzelnen Items zu konstruieren, verbietet sich bei den hier aufgezeigten Konstellationen nun vollends.

3.3 Die Heterogenität der studentischen Urteile gilt für (fast) jede Vorlesung

Die bisherige Auswertung hat belegt: Durchschnittspunktwerte je Item sind untauglich zur Bewertung von Lehrveranstaltungen; und das nicht nur wegen der zu beobachtenden großen Beurteilungsvarianz je Item, sondern mehr noch wegen der nicht gegebenen Unabhängigkeit der Einschätzungen je Item: Es existieren jeweils beurteiler-typische Konstellationen ("Bewertungsprofile").

Für das Ziel einer hinreichend differenzierenden - also nicht unangemessen vereinfachenden - Evaluation muß allerdings dieses Faktum kein Mangel sein, sondern könnte gerade einen außerordentlichen Vorteil darstellen. Es könnte ja sein - und das wäre die eigentlich naheliegende Erwartung -, daß sich jeweils *vorlesungstypische* Konstellationen herausschälen, daß in der einen Vorlesung zwei oder drei "benachbarte" Bewertungsprofile dominieren, in der anderen Vorlesung zwei oder drei andere Profile. Die Heterogenität könnte sich also als Urteilsvarianz *zwischen* den Vorlesungen erweisen und somit gerade Ausdruck abgestufter Wertschätzung durch die Mehrzahl der Teilnehmer sein.

Aber auch diese Hoffnung erweist sich als irrig: Die Heterogenität gilt (mit ganz wenigen Ausnahmen) *innerhalb* jeder einzelnen der untersuchten Vorlesungen. Läßt man sehr kleine Vorlesungen (weniger als 25 Hörer) außer Betracht, so existiert keine einzige (mindestens mittelgroße) Veranstaltung, in der bei den dozentenbezogenen oder bei den lernprozeßbezogenen Urteilsaspekten weniger als fünf verschiedene Beurteilungsmuster vertreten wären. Und selbst bei den kleinen Vorlesungen (im allgemeinen Spezialvorlesungen im Hauptstudium mit homogenerer Hörerschaft) ist die (relativ) einheitliche Bewertung durch ihre Teilnehmer die seltene Ausnahme:

	dozentenbezogene Evaluationen (ClDoz)			lernprozeßbezogene Evaluationen (ClStud)		
	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

Die 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: Dasselbe Lehrangebot wird umso unterschiedlicher aufgenommen und umso widersprüchlicher bewertet, je vielfältiger die Interessenstruktur der Teilnehmer ist.

4. Welche lehrangebotsunabhängigen Aspekte beeinflussen die studentischen Einschätzungen?

Wenn es in ein und derselben Vorlesung so viele unterschiedliche, ja gegensätzliche Urteile über die Qualität der Lehre und noch mehr über die Qualität des Lernprozesses gibt, dann liegt die Frage nahe: Was sind die Gründe für diese Unterschiede? Als Indikator für ein zusammenfassendes Urteil der Vorlesungsteilnehmer sei die Frage gewählt: "Hat die Vorlesung bisher Ihre Erwartungen insgesamt erfüllt?" (s. oben, Tab. 1 und 2). Die Antworten hierauf weisen - wie schon angemerkt - die größte gemeinsame Korrelation mit einer Reihe von Detailbewertungen auf.

Die folgende Tabelle der Koeffizienten einer multiplen Regressionsanalyse zeigt, in welchem Ausmaß das Zufriedenheitsurteil von Sachverhalten abhängt, die die einzelne Lehrperson nicht beeinflussen kann; d.h. sie gibt Hinweise auf Bestimmungsgründe für die studentische Beurteilung, soweit sie in der Interessen- und Erwartungsstruktur der Teilnehmer, in deren Studierverhalten sowie in Einstellungen der "sozialen Umwelt" (Mitstudenten) liegen.[15]

Prädiktorvariable:	Regressionskoeff. (b)	
Motivlage der Vorlesungsteilnehmer: (vgl. oben, Tab. 2)		
negativ I (> 80 % Nur-Pflicht-Hörer)	-0,50	
negativ II (> 90 % Nur-Pflicht-Hörer)	-0,30	
positiv I (mehr als 70 % "intrinsisch" Motivierte)	+0,26	
positiv II (mehr als 75 % "intrinsisch" ohne Pflicht)	+0,36	1,42
Wahlmöglichkeit, pers. Interesse, Erwartungen:		
Vorlesungsbesuch aus persönl. Interesse	+0,18	
Wahlveranstaltung	+0,09	
Erwartung: Erörterung von Problemzusammenhängen	+0,13	
Pflichtveranstaltung	-0,14	
Prüfungsrelevanz/Klausur	-0,24	
mehr als nur ein Besuchsgrund	+0,18	
mehr als zwei Besuchsgründe	+0,29	1,25
Studierverhalten/Studienerfahrung:		
Vorlesungsversäumnis mind. 1x	-0,10	
" mind. 2x	-0,12	
" mehr als 3x	-0,20	
Vorarbeit des Stoffs	+0,19	
Nacharbeit des Stoffs	+0,19	
Erstsemester	-0,13	0,93
Hauptstudium	+0,03 (n.sign.)	
Summe:		3,60

Die unstandardisierten Regressionskoeffizienten (b) in der Tabelle sind wie folgt zu lesen:
Wenn alle übrigen Merkmale gleich sind (bzw. mathematisch präziser formuliert: wenn deren Einfluß statistisch "kontrolliert" wurde), dann unterscheiden sich die Zufriedenheitsurteile von Personen in einer Vorlesung mit überwiegend (> 80 %) "extrinsisch" motivierten Teilnehmern von denen in einer Vorlesung mit "gemischtem Publikum" im Durchschnitt um -0,50 Skalenpunkte, und zwar unabhängig von der eigenen Motivlage; bei mehr als 90 % "extrinsisch" Teilnehmenden kommt eine weitere Differenz von -0,30 Skalenpunkten hinzu. Im Durchschnitt um 0,26 bzw. 0,36 Punkte positiver fallen die individuellen Urteile in Vorlesungen aus, in denen die intrinsisch motivierten Hörer stark bzw. sehr stark überwiegen. Zusammengenommen macht der "Ansteckungseffekt" der studentischen Umwelt 1,42 Skalenpunkte aus; d.h.: Hörer in Vorlesungen mit über 75 % intrinsisch motivierten Teilnehmern urteilen im Durchschnitt um gut 1,4 Punkte positiver als Hörer in Vorlesungen mit über 90 % Nur-Pflicht-Teilnehmern.[16]

Einen ähnlich starken Einfluß auf die Evaluationen hat darüber hinaus die individuelle Interessens- und Motivlage des jeweiligen Urteilenden. Hatte der bzw. die Evaluierende Wahlmöglichkeiten zwischen verschiedenen Veranstaltungen? Ist er/sie aus eigenem Interesse (also intrinsisch motiviert) oder nur

pflichtweise in die Vorlesung gegangen? Wird die Erörterung von spezifischen Problemzusammenhängen oder nur eine möglichst gezielte Vorbereitung auf eine (lästige) Prüfung oder Klausur mit entsprechender Stoffbeschränkung gewünscht? Personen, die aus persönlichem Interesse in einer Wahlveranstaltung sind, unterscheiden sich von anderen, bei denen dies nicht der Fall ist und die nur eine schmalspurige Prüfungsvorbereitung wünschen, im Durchschnitt um weitere gut 1,25 Skalenpunkte.

An dritter Stelle folgen das faktische Studierverhalten und die Studienerfahrung der Hörer: regelmäßige Vor- und Nacharbeit, regelmäßige Teilnahme und die Tatsache, die Orientierungsprobleme des ersten Semesters überwunden zu haben, sind für nochmals mehr als 0,9 Punkte im Zufriedenheitsurteil verantwortlich.

Zusammen stehen diese wenigen in die Analyse einbezogenen Merkmale, die zudem noch wegen der Dichotomisierung der Variablen tendenziell das Ausmaß des Zusammenhangs *unter*schätzen, bereits für 3,6 Punkte der verwendeten 5-Punkte-Beurteilungsskala.
Verbalisiert:

- Wer als Dozent oder Dozentin vor freiwilligen Hörern mit Wahlmöglichkeiten zwischen Angebotsalternativen lehren darf,
- wessen Hörer persönliches Interesse mitbringen,
- wessen Hörer regelmäßig vor- und/oder nacharbeiten und die Veranstaltung regelmäßig besuchen
- und wessen Hörer "studienerfahren" sind,

erhält um 3,6 Punkte positivere Urteile als sein Kollege oder seine Kollegin mit entgegengesetzter Konstellation.

Wird die Analyse fakultätsweise durchgeführt und können die Veranstaltungen noch nach generell "beliebtem" oder "unbeliebtem" Stoff eingeordnet werden, ist der so "vorhersagbare" lehrunabhängige Punkteanteil sogar noch höher.

Um keine Mißdeutungen aufkommen zu lassen: Der aufgezeigte Zusammenhang zwischen dem Zustandekommen (individueller) studentischer Urteile über eine Lehrveranstaltung und lehrunabhängigen Merkmalen des Urteilenden sowie seiner studentischen Umwelt bezieht sich auf den Typ von Items, die als globale Zufriedenheitsfragen gekennzeichnet werden können. Deren geringer Informationswert als Indikator für Lehrqualität wurde bereits im vorhergehenden Kapitel nachgewiesen. Dennoch ist das Ergebnis nicht unerheblich. Der eine Grund ist die wiederkehrende Empfehlung mancher Kollegen unter Hinweis auf die "bewährte Praxis" in US-Hochschulen, möglichst kurze Fragebögen mit vor allem globalen Beurteilungs-Items einzusetzen. Der andere Grund ist die aktuelle Ranking-Diskussion, die zum Teil auch auf Lehrveranstaltungen und auf Lehrende bezogen wird.

4.1 Exkurs: Ranking von Vorlesungen per Teilnehmerbefragung?

Die Hochschule soll - so ist es ihr gesetzlich aufgetragen - "gewährleisten, daß ... die Formen der Lehre und des Studiums den methodischen und didaktischen Erkenntnissen entsprechen" (§ 6 Abs. 1 Ziff. 2 WissHG NW). Um dies aber gewährleisten zu können, muß - so argumentiert Webler (1991, 244) - die Hochschule sich Kenntnisse vom Zustand der Lehre verschaffen, eben durch Evaluation. Im Aktionsprogramm "Qualität der Lehre" wird u.a. als ein mögliches Instrument zur Förderung und Gewährleistung guter Lehrqualität die Einführung eines Anreizes in Form "besonderer Auszeichnungen für in der Lehre engagierte Wissenschaftler" zur Diskussion gestellt (MWF NW 1991, 35).

Von hier ist es nur noch ein kurzer Schritt zu der Empfehlung, auf der Basis methodisch kontrollierter empirischer Erhebung studentischer Teilnehmerurteile ein "Ranking" von Lehrveranstaltungen und damit von Lehrenden vorzunehmen (als vermeintlich objektivierende Form der Evaluation). Und in diesem Zusammenhang stößt man dann regelmäßig auf den bereits diskutierten Vorschlag, auf der

Basis einer geeigneten Beurteilungsskala Durchschnittswerte pro Veranstaltung zu berechnen und anhand dieser Werte eine Rangfolge der Veranstaltungen (und damit der Lehrenden) zu erstellen. Als "geeignet" gelten dabei üblicherweise globale Beurteilungs-Items.

Testweise wurde ein entsprechendes Ranking hier sowohl für eine Fakultät wie für den gesamten Datenbestand durchgeführt (beschränkt auf Veranstaltungen mit mehr als 15 Teilnehmern, um Mittelwerte zumindest statistisch sinnvoll berechnen zu können). Ergänzend wurden für jeden Befragten "bereinigte Netto-Beurteilungen" berechnet; "bereinigt" soll hier heißen: Unter Verwendung der oben geschilderten Prädiktor-Variablen wurden regressionsanalytisch "Schätzwerte" für das verwendete Beurteilungs-Item berechnet, wie sie sich unabhängig von der jeweiligen Lehrleistung aufgrund der Konstellation der Individualmerkmale und der sozialen Umgebung des Befragten bestimmen lassen. Brutto-Beurteilung (d.h. die Antwort im Fragebogen) abzüglich lehrunabhängigem Schätzwert ergibt dann als Rest (Regressions-Residuum) die "Netto-Beurteilung", d.h. den Skalenwert, der *nicht* den Individual- und Umgebungsmerkmalen zugerechnet werden kann. Berechnet man nun von diesen Residuen wiederum Durchschnittswerte pro Vorlesung und bildet auf ihrer Basis eine Rangreihe, dann erhält man eine Reihenfolge, wie sie sich in der hypothetischen Situation einstellen würde, daß jede Vorlesung eine identische Zusammensetzung der Hörerschaft hätte.

In der Tat existieren in den 22 berücksichtigten Vorlesungen der testweise herangezogenen Fakultät durchaus bemerkenswerte Unterschiede im (Brutto-)Urteilsdurchschnitt der Hörer: Die fünf ersten Plätze weisen Skalenwerte auf, die aussagen, daß "im Schnitt" die Erwartungen der Teilnehmer weitgehend bzw. voll erfüllt wurden. Danach folgt bis zum Platz 12 ein "Mittelfeld" mit Werten von 0,77 bis 0,97, d.h. in diesen Vorlesungen überwiegt jedenfalls die Zufriedenheit deutlich. Vom Rangplatz 18 abwärts ergibt sich ein negativer Durchschnitt: die Mehrheit hat die Skalenwerte -1 oder -2 gewählt (Erwartungen weitgehend nicht oder überhaupt nicht erfüllt). Damit hätte man also die fünf "Spitzenreiter" und die fünf "Schlußlichter" ermittelt. Die Dozenten, die die Rangplätze 1 bis 5 einnehmen, könnten als Vorbild gelten, von denen die fünf "Schlußlichter" lernen könnten, wie man "gute Lehre" gestaltet.

Nichts wäre allerdings falscher als diese Schlußfolgerung. Ein Vergleich dieser "Brutto-Rangordnung" mit der rechnerisch ermittelten (hypothetischen) "Netto-Rangordnung" zeigt: Beide weisen kaum noch Ähnlichkeit miteinander auf, teilweise fallen die Einordnungen geradezu entgegengesetzt aus. So findet sich die Veranstaltung mit dem Brutto-Rangplatz 2 auf dem Netto-Rangplatz 14 wieder; dagegen rückt die Nr. 21 (brutto) auf Platz 8 (netto) vor, die Nr. 6 verliert 10 Plätze, die Nr. 19 dagegen gewinnt 12 Plätze usw. Ähnlich ist das Ergebnis für die berücksichtigten 124 Vorlesungen der Gesamtdatenbasis. Hier finden sich zahlreiche Verschiebungen um +/- 60 und mehr Plätze; relativ stabile Einordnungen (Verschiebungen um weniger als 10 Plätze) sind dagegen selten.

Fazit: Ein Ranking von Lehrveranstaltungen anhand der Durchschnittswerte von per Befragung der Veranstaltungsteilnehmer ermittelten globalen Beurteilungsindikatoren ist unter methodischem Gesichtspunkt völlig untauglich, ist - deutlicher ausgedrückt - methodischer Unsinn. An eine solche Basis geknüpfte "incentives" für erfolgreiche Lehre würden gerade diejenigen zusätzlich belohnen, die es vergleichsweise leicht haben, und diejenigen bestrafen, die die schwereren Aufgaben übernehmen.

5. Zum Informationswert von Vorlesungskritiken

Die vorgetragene Kritik an der Aussagekraft studentischer Vorlesungsbeurteilungen *als Evaluation* bedeutet allerdings nicht, daß Veranstaltungs-Umfragen generell einen geringen Informationswert oder nur zweifelhafte Gültigkeit hätten. Sie liefern im Gegenteil - wie auch andere Daten der Umfrageforschung -, sofern auf methodisch angemessenem Niveau erhoben und hinreichend differenziert ausgewertet, wertvolle Informationen sowohl für die Hochschule als Institution wie für die betroffenen Lehrenden wie für die Studierenden. Sie informieren den Dozenten bzw. die Dozentin beispielsweise darüber,

- für wen sie überhaupt lehren (häufig genug bestehen, wie sich gezeigt hat, erstaunlich unzutreffende Vorstellungen darüber),

- mit welchen Interessen, Wünschen, Erwartungen oder auch Vorurteilen die Hörer in die Lehrveranstaltung kommen (dies gibt die Chance, die Lehre in geeignet erscheinender Weise darauf auszurichten),
- wie gut oder wie schlecht die Studierenden mit dem Lehrangebot zurechtkommen, ob und ggf. in welchem Ausmaß als bekannt vorausgesetzte Vorkenntnisse fehlen (dies öffnet die Möglichkeit, den Studierenden gezielt Lern- und Arbeitsempfehlungen zu geben oder ergänzende Übungen/Propädeutika anzubieten - oder auch ein eventuell überzogenes Anforderungsniveau zurückzuschrauben),
- wie gut oder wie schlecht sie als Lehrperson "ankommen", ob sie von den Studierenden als interessiert oder desinteressiert, als distanziert oder ansprechbar wahrgenommen werden, ob ihr Sprachstil, ihre Wortwahl zielgruppenangemessen ist, ob die Beispiele und Illustrationen verstanden werden (auch hier hat sich gezeigt, daß häufig genug erhebliche Diskrepanzen zwischen dem Selbstbild des/der Lehrenden und dem Fremdbild bestehen, das sich aus den studentischen Antworten herausschält), sowie schließlich - und nicht zuletzt -
- *wie* unterschiedlich die Beurteilungen und Wahrnehmungen der Veranstaltungsteilnehmer ausfallen (können), daß also Kritik von einzelnen nicht die Kritik *der* Studenten sein muß (ebensowenig wie von herangetragenem Lob auf generelle Zustimmung geschlossen werden darf).

Für die Hochschule als Institution und für die Studentenschaft liefern die Befragungsdaten (selbst wenn sie aus Datenschutzgründen nicht einzel-lehrveranstaltungsbezogen ausgewertet werden dürfen) Hinweise darauf, wie hoch die Zufriedenheit oder Unzufriedenheit mit der Lehre verbreitet ist, welche Schwachpunkte in der Lehre, aber auch im Studierverhalten existieren (z.B. Vor- und Nacharbeit, Regelmäßigkeit des Vorlesungsbesuchs, Motivationsmängel u.ä.), wo veranstaltungsübergreifend weitere Lern- und Arbeitshilfen bereitzustellen sind, welcher Lernstoff als beliebt und motivationsfördernd, welcher als unbeliebt und motivationshemmend erlebt wird, unter welchen Rahmenbedingungen (Räumlichkeiten, Ausstattung, Zeitbudget u.ä.) gelehrt und gelernt wird, und manches mehr. Alles dies sind wesentliche, die Leistungsbereitschaft von Lernenden wie von Lehrenden beeinflussende Faktoren, über die (gegenseitig) mehr stereotype Vorurteile als zutreffende Kenntnisse existieren.

Nicht zuletzt bietet eine verläßliche empirische Basis günstige Voraussetzungen für eine sachliche Diskussion zwischen allen Beteiligten (sowohl innerhalb der einzelnen Lehrveranstaltungen als auch auf anderen Ebenen) mit dem Ziel, auch unter den gegebenen und zumindest kurzfristig nicht zu behebenden Restriktionen das Bestmögliche zu erreichen. Ein (methodisch in keiner Weise zu rechtfertigender) Mißbrauch von per Befragung erhobener Lehrveranstaltungskritik als "Evaluation" der Lehrqualität wäre für dieses Ziel schädlich.

Zitierte Quellen:

DER SPIEGEL (1993): Willkommen im Labyrinth, Nr. 16, 80 ff.
Kromrey, H. (1987): Zur Verallgemeinerbarkeit empirischer Befunde bei nichtrepräsentativen Stichproben. In: Rundfunk und Fernsehen, H. 4, 478-499
Kromrey, H. (1988): Akzeptanz- und Begleitforschung. Methodische Ansätze, Möglichkeiten und Grenzen. In: Massacommunicatie, H. 3, 221-242
MWF - Ministerium für Wissenschaft und Forschung NW (Hg.) (1991): Aktionsprogramm Qualität der Lehre, Düsseldorf
Rost, J. (1990): LACORD. Latent Class Analysis for Ordinal Variables, Kiel: Institut für die Pädagogik der Naturwissenschaften
Schmidt, J. (1980): Evaluation I, Dozentenkurs, Essen: HDZ
Webler, W.-D. (1991): Kriterien für gute akademische Lehre. In: Das Hochschulwesen, H. 6, 243-249

Endnoten:

1. Die Datengrundlage für die folgenden Ausführungen sind Vorlesungsbefragungen im Rahmen des Projekts "Evaluation der Lehre an der Ruhr-Universität Bochum"

(durchgeführt i.A. des Rektorats in Zusammenarbeit mit der Universitätskommission für Lehre, Studium und Studienreform).

2. Etwa: Wie oft wurden technische Medien eingesetzt [regelmäßig, manchmal, gar nicht]? Oder: Wurden vorlesungsbegleitende Übungen angeboten?

3. Gemeint sind Aussagen wie im folgenden, schon etwas älteren Zitat von hochschuldidaktischer Seite, das aber genauso gut auch heute hätte formuliert werden können: "Urteile (Schätzungen) von Studenten über die Lehre sind verläßlicher als Urteile der Dozenten über die Leistung der Studenten und ebenso verläßlich wie Urteile von Kollegen über die Lehre, (...); sie sind - wenn man etwa 20-30 Studenten urteilen läßt - zuverlässig wie professionelle Testverfahren; sie sind von anderen Merkmalen der Studenten selbst und der Dozenten wenig beeinflußt." (Schmidt 1980, S. 51 f.) Alle in diesem Zitat aufgestellten Behauptungen erweisen sich als empirisch falsch.

4. Fragewortlaut: "Hat die Vorlesung bisher Ihre Erwartungen insgesamt erfüllt?" Beurteilungsskala von -2 (Erwartungen überhaupt nicht erfüllt) bis +2 (Erwartungen voll erfüllt).

5. Index der Motivation zur Teilnahme an der Vorlesung mit den Ausprägungen "extrinsisch" (= Teilnahme ausschließlich, weil es sich um eine Prüfungs- bzw. Klausurpflichtveranstaltung handelt), "intrinsisch ohne Pflicht" (= Teilnahme ohne Prüfungs- bzw. Klausurpflicht, z.B. aus persönlichem Interesse, Wahlveranstaltung, zur Auffrischung von Kenntnissen etc.), "intrinsisch + Pflicht" (= Teilnahme wg. Prüfungs- bzw. Klausurpflicht *und* aus anderen Gründen).

6. Befragungen in Vorlesungen der Fakultäten Wirtschaftswiss., Sozialwiss., Psychologie, Bauingenieurwesen, Physik, Medizin (jeweils "flächendeckend") sowie in einzelnen Vorlesungen anderer Fakultäten.

7. Index zur Beschreibung der Motivation der Hörerschaft, gebildet als Anteil der extrinsisch bzw. intrinsisch motivierten Teilnehmer in der jeweiligen Vorlesung; -2: mehr als 90 % Nur-Pflicht-Hörer, -1: mehr als 80 % Nur-Pflicht-Hörer; +2: mehr als 75 % "intrinsisch" Motivierte ohne Prüfungs-/Klausurpflicht, +1: mehr als 70 % "intrinsisch" Motivierte (mit oder ohne zusätzliche Prüfungspflicht), 0: übrige Veranstaltungen.

8. Darauf kann hier nicht im Detail eingegangen werden. Für ein Beispiel aus einem anderen Forschungskontext vgl. Kromrey 1987.

9. Dies wären die o.g. Ausnahmefälle von Veranstaltungen, in denen die Teilnehmerurteile "in großer Zahl" übereinstimmen.

10. Das eigentlich angemessenere Verfahren der Latent Class Analysis ist bei einem Umfang von über 10.000 Fällen nicht praktikabel. Testweise wurde jedoch eine LCA für ordinale Variablen (Programm LACORD von J. Rost) auf Stichproben der Gesamtdatenbasis angewendet. Ein Vergleich mit den Ergebnissen der deskriptiven Clusteranalyse (Quick Cluster mit Iterationen, SPSS für Windows) erbrachte weitgehende Übereinstimmungen.

11. Auf eine Detailbeschreibung der einzelnen Urteilsprofile wird hier verzichtet; Interessenten können ausführlichere Ergebnisberichte beim Autor anfordern.

12. In den Beschriftungen der Abbildungen 1 ff. als "Durchschnitt" bezeichnet.

13. Die 10-Cluster-Lösung erwies sich als die unter statistischen Kriterien optimale (hohe Homogenität in den Clustern, starke Heterogenität zwischen den Clustern).

14. Auf weitere Grafiken wird aus Platzgründen verzichtet.

15. Abhängige Variable: "Erwartungen erfüllt?"; Skala von -2 [überhaupt nicht erfüllt] bis +2 [voll erfüllt]; Regressionsmodell: lineare, additive Beziehungen, dichotomisierte, 0/1-codierte Variablen.

16. Der "Bruttowert" des Einflusses der studentischen Umwelt, wie er oben aus Tab. 2 berechnet werden kann, beträgt 1,71 Punkte (Durchschnittspunktwert in Vorlesungen mit über 75 % intrinsisch motivierten Teilnehmern: 1,42; in Vorlesungen mit mehr als 90 % extrinsisch Motivierten: -0,29).