Döring, N. (2005). **Für Evaluation und gegen Evaluitis. Warum und wie Lehrevaluation an deutschen Hochschulen verbessert werden sollte**. In B. Berendt, H.-P. Voss & J. Wildt (2005), *Neues Handbuch Hochschullehre* (Ergänzungslieferung Juli 2005). Berlin: Raabe. ISBN: 3-8183-0206-5.

Für Evaluation und gegen Evaluitis

Warum kann und wie sollte Lehrevaluation an deutschen Hochschulen verbessert werden

Nicola Döring

Zusammenfassung

An vielen Hochschulen werden mittlerweile die Lehrveranstaltungen routinemäßig mittels Fragebogenerhebungen unter Studierenden evaluiert. Die etablierte Praxis der Lehrveranstaltungsevaluation weist dabei zahlreiche inhaltliche, methodische und ethische Schwächen auf. Anhand des aktuellen Forschungsstandes ist zu konstatieren, dass die gängige Evaluation von Lehrveranstaltungen trotz relativ hohen Aufwandes und hoher Kosten in der Regel zu keiner Verbesserung der Lehre führt. Der vorliegende Beitrag erläutert praxisnah und empirisch begründet die Probleme der üblichen Lehrevaluation. Er gibt somit denjenigen Argumente an die Hand, die etablierte Lehrevaluationspraxis begründet kritisieren und verbessern wollen. Der Beitrag zeigt Wege auf zu einer Lehrevaluation, die sich nicht in sinnloser oder gar schädlicher „Evaluitis" erschöpft.

Gliederung Seite

1. Einleitung

Vier Anspruchsgruppen, vier Entwicklungsdekaden

Die Evaluation von Hochschullehre soll dazu dienen, die Lehrqualität zu messen und sie - sofern notwendig - zu verbessern. Eine qualitativ hochwertige Lehre ist im Interesse der Studierenden und der Lehrenden, sie ist im Interesse der breiten Öffentlichkeit und des Staates. Dementsprechend gibt es aus den genannten vier Anspruchsgruppen jeweils eigene Impulse zur Lehrevaluation. Diese Impulse lassen sich grob den letzten vier Dekaden der Hochschulentwicklung zuordnen:

1960er Jahre: Studentische Veranstaltungskritik

Mit der 1968er-Studentenbewegung begann an westdeutschen Hochschulen eine Tradition der studentischen Veranstaltungskritik, in der sich der Anspruch der Studierenden artikuliert, über Lehrinhalte sowie über Lehr- und Lernformen aktiv mitzubestimmen. Bis heute wird die Lehrevaluation in manchen Studiengängen ausschließlich oder maßgeblich von den studentischen Fachschaften durchgeführt.

1970er Jahre: Hochschuldidaktik

Das Engagement der Lehrenden für Lehrqualität wird exemplarisch markiert durch die Gründung des Arbeitskreises Hochschuldidaktik (AHD) im Jahr 1971. Durch hochschuldidaktische Forschung, Weiterbildung und oft auch durch autodidaktisches Lernen wirkt das akademische Personal an der Qualitätssicherung im Bereich Hochschullehre mit. Während in der ehemaligen DDR alle Hochschullehrerinnen und -lehrer hochschuldidaktisch geschult wurden, war und ist die Teilnahme an hochschuldidaktischer Weiterbildung in der Bundesrepublik in vielen Bundesländern auf Eigeninitiative beschränkt, was in der aktuellen Lehrqualitätsdebatte teilweise als eklatantes Defizit beklagt wird (z.B. Berendt 2002; Preißer 2003).

1980er Jahre: Hochschulrankings

Die Publikation des ersten Hochschulrankings im Jahr 1989 im Nachrichtenmagazin DER SPIEGEL unterstreicht die wachsende Sorge der breiten Öffentlichkeit um die Leistungsfähigkeit und Qualität der Hochschulen. Lehrqualität und Lehrevaluation sind - neben Forschungsleistungen und anderen Aspekten - typischerweise wichtige Komponenten von Hochschulrankings, die bis heute regelmäßig von unterschiedlichen Institutionen erstellt und publiziert werden.

1990er Jahre: Hochschulrahmengesetz

Seit der vierten Novelle des Hochschulrahmengesetzes (HRG) im Jahr 1998 wird den Hochschulen in § 6 von Seiten des Staates vorgeschrieben, ihre Leistungen regelmäßig zu evaluieren. Dazu gehört die Lehrevaluation und zwar ausdrücklich unter Beteiligung der Studierenden. Die Landeshochschulgesetze sowie die Grundordnungen der Hochschulen präzisieren jeweils die Anforderungen an und den Umgang mit Lehrevaluationen.

Monetäre Belohnungen

Lehrevaluation gewinnt in jüngster Zeit dadurch an Bedeutung und Brisanz, dass sie nicht nur mit Karrierechancen von einzelnen Wissenschaftlerinnen und Wissenschaftlern sowie mit dem öffentlichen Image von Studiengängen, Fachbereichen und Hochschulen verknüpft ist, sondern auch mit monetären Belohnungen bzw. Bestrafungen gekoppelt werden kann und teilweise bereits gekoppelt wird: Einzelne Hochschullehrerinnen und Hochschullehrer können im Rahmen der neuen W-Besoldung für besonders gute Lehrevaluationsergebnisse Zulagen beanspruchen und Fachbereiche, die besonders positiv evaluierte Lehre nachweisen, können im Zuge der leistungsbezogenen Mittelvergabe honoriert werden. Umgekehrt entziehen negative Evaluationsergebnisse den Betroffenen entsprechende Ressourcen.

Lehrevaluation soll der Messung und - sofern notwendig - der Verbesserung der Lehrqualität dienen. Hohe Lehrqualität ist im Interesse der Studierenden, der Lehrenden, der breiten Öffentlichkeit und des Staates. Evaluationsergebnisse werden zunehmend mit monetären Sanktionen verknüpft.

Da die Ziele der Lehrevaluation im allgemeinen Interesse und die mit den Evaluationsergebnissen verknüpften Sanktionen für Individuen und Institutionen teilweise von großer Tragweite sind, muss dafür Sorge getragen werden, dass Lehrevaluation wissenschaftlichen und ethischen Standards genügt. Auch muss nachweisbar sein, dass - nicht zuletzt angesichts der Kosten von Lehrevaluation - die angestrebte Qualitätssicherung tatsächlich erreicht wird.

Ziel des vorliegenden Beitrags ist es, den bisherigen Umgang mit Lehrevaluation an deutschen Hochschulen auf der Basis wissenschaftlicher Befunde und Kriterien zu analysieren und daraus begründete Kritikpunkte sowie Verbesserungsvorschläge für die Evaluationspraxis abzuleiten.

2. Was ist Lehrevaluation?

Lehrevaluation kann hinsichtlich des betrachteten Evaluationsobjektes sowohl auf der Makroebene als auch auf der Mikroebene stattfinden:

Lehrevaluation auf Makroebene

Auf der Makroebene adressiert Lehrevaluation die Qualität des Lehrangebotes eines gesamten Studiengangs oder Fachbereiches. Es geht beispielsweise um Fragen des Curriculums, um die Überschneidungsfreiheit von Lehrveranstaltungen, um die Studierbarkeit innerhalb der Regelstudienzeit usw.

Lehrevaluation auf Mikroebene

Auf der Mikroebene befasst sich Lehrevaluation mit der Qualität einzelner Lehrveranstaltungen bzw. Lehrkräfte. Diese Lehrveranstaltungsevaluation ist bislang in der Praxis sehr viel weiter verbreitet als Lehrevaluation auf Makroebene und steht deswegen im Zentrum des vorliegenden Beitrags. In der englischsprachigen Fachliteratur existieren eine Reihe von synonymen Bezeichnungen für Lehrveranstaltungsevaluation (Course Evaluation; Student Evaluation of Teaching SET; Student Ratings of Teaching SRT; Student Response to Instruction SRTI).

Professoren-Ratings

Außerhalb der offiziellen Evaluationsmaßnahmen an Universitäten haben sich inzwischen im Internet eine Reihe von Online-Plattformen etabliert, auf denen Studierende anonym Urteile über namentlich genannte Hochschullehrer/innen publizieren (z.B. für USA und Kanada auf www.ratemyprofessors.com). Solche Professoren-Ratings erfolgen auf Notenskalen sowie anhand von offenen Beschreibungen (z.B. „do not take this guy, terrible teacher, worse than davanzo ... absent 7 times which I liked but never explained the readings we had to do for our paper, midterm and final ... he gave me two D's ..."). Professoren-Ratings werden von Studierenden unter anderem als Hintergrundinformation genutzt, wenn es um die Auswahl von Kursen geht (Kindred/Mohammed 2005).

2.1 Evaluation

Qualitätsmerkmal

Evaluation ist definiert als Bewertung eines Gegenstandes hinsichtlich ausgewählter Qualitätsmerkmale. Ein Merkmal wird dann zum Qualitätsmerkmal, wenn begründete und konsensfähige Qualitätsstandards vorliegen. Qualitätsstandards sind Anforderungen an die Ausprägung eines Merkmals. Es kann sich bei Qualitätsstandards um Minimal-, Regel- und Maximalstandards handeln. Beispiel: Die durchschnittliche Studiendauer ist ein Merkmal eines Studiengangs. Es lässt sich im Rahmen von Lehrevaluationen auf Makroebene als Qualitätsmerkmal nutzen, da mit der Regelstudienzeit ein Qualitätsstandard (hier: Regelstandard) vorliegt. Ein Studiengang, in dem die durchschnittliche Studienzeit die Regelstudienzeit dauerhaft und nennenswert überschreitet, ist demnach hinsichtlich dieses Merkmals negativ zu evaluieren.

Problem: Definition von Qualitätsstandards für einzelne Merkmale der Lehre

Für eine Fülle von Merkmalen, mit denen sich Studiengänge oder auch einzelne Lehrveranstaltungen beschreiben lassen, sind bislang keine theoretisch und/oder empirisch begründeten und konsensfähigen Qualitätsstandards entwickelt worden. Ein Beispiel ist das Merkmal Teilnehmerschwund in Lehrveranstaltungen: Wenn eine Pflichtvorlesung, die einen Studierendenjahrgang der Stärke N=120 adressiert, de facto gegen Semesterende nur noch von

n=20 Studierenden besucht wird, ist dieser Schwund ein Indikator qualitativ schlechter Lehre? Wurde die Vorlesung „leergelesen"? Tatsächlich wird oft spekuliert, dass es sich bei Studierenden, die einer Pflichtveranstaltung fernbleiben (so genannte No-Shows) um unzufriedene, von der Veranstaltung enttäuschte Lernende handelt. Dies wäre jedoch empirisch zu belegen. Denkbar sind ergänzend oder alternativ auch Gründe für das Fernbleiben, die mit positiver Lehrqualität assoziiert sind: So könnten Studierende auch dann bewusst und rational zu No-Shows werden, wenn die Lehrveranstaltung so durchgeführt wird, dass durch umfangreiche und qualitativ hochwertige veranstaltungsbegleitende Lehr- und Lernmaterialien auch ein flexibles und selbstständiges Lernen jenseits des Hörsaales möglich ist.

Handout I 1.7-1 Definition von Qualitätsstandards für einzelne Merkmale der Lehre

2.2 Evaluationsforschung

Evaluationsforschung

Evaluation wird dann zur Evaluationsforschung, wenn sie auf der Basis sozialwissenschaftlicher Methoden erfolgt (Bortz/Döring 2002). Bei der Lehrveranstaltungsevaluation ist dies der Fall, denn hier wird auf die populärste sozialwissenschaftliche Methode überhaupt zurückgegriffen, nämlich auf die Fragebogenmethode.

Wenn knapp von „Lehrevaluation" die Rede ist, dann ist damit in der Praxis meist Lehrveranstaltungsevaluationsforschung gemeint, also die Bewertung einzelner Lehrveranstaltungen (Mikroebene) auf der Basis von Studierendenurteilen, die mittels sozialwissenschaftlicher Fragebogenmethodik erfasst werden.

Typen der Evaluation

Im Bereich der Evaluationsforschung lassen sich unterschiedliche Typen der Evaluation differenzieren:

Summative versus formative Evaluation

Summative Evaluation dient der abschließenden Qualitätsbewertung, während formative Evaluation auf begleitende Qualitätsentwicklung abzielt. Typischerweise ist Lehrveranstaltungsevaluation sowohl auf summative Evaluation (Wie gut oder schlecht ist eine Lehrveranstaltung insgesamt?) als auch auf formative Evaluation (Wo liegen im einzelnen konkrete Schwächen einer Lehrveranstaltungen, die behoben werden sollten?) ausgerichtet.

Interne versus externe Evaluation

Interne Evaluation wird innerhalb der betreffenden Institution abgewickelt, während bei externer Evaluation organisationsfremde Experten herangezogen werden. Lehrveranstaltungsevaluation ist typischerweise interne Evaluation, die von Angehörigen des Fachbereiches bzw. der Hochschule durchgeführt wird.

Kasuistische versus quasi-experimentelle versus experimentelle Evaluation

Kasuistische Evaluation betrachtet einzelne Evaluationsobjekte für sich genommen. Quasi-experimentelle Evaluation vergleicht vorgefundene Objekte bzw. Objektgruppen miteinander, während experimentelle Evaluation die zu vergleichenden Gruppen und Bedingungen aktiv herstellt, und dabei durch Zufallszuteilungen (Randomisierung) die Vergleichbarkeit maximiert. Lehrveranstaltungsevaluation erfolgt typischerweise als kasuistische oder quasi-experimentelle Evaluation. Eine experimentelle Manipulation der Lehrbedingungen ist in der Praxis dagegen in der Regel nicht möglich.

> Lehrveranstaltungsevaluation ist hinsichtlich ihrer Funktion und Einbettung in das Praxisfeld der Hochschullehre summativ und formativ angelegt. Sie ist als interne Evaluation zu kennzeichnen und folgt entweder einem kasuistischen oder quasi-experimentellen Design.

Aussagekraft einer Evaluation

Die Aussagekraft einer Evaluation hängt maßgeblich von der Auswahl der betrachteten Qualitätsmerkmale ab. Aus pragmatischen Gründen sollte auf möglichst wenige, dafür allerdings hochrelevante Merkmale zurückgegriffen werden, die empirisch erfassbar sind (zur Merkmalsauswahl in gängigen Evaluationsinstrumenten siehe Abschnitt 3.2). Weiterhin ist entscheidend, dass bei der Auswertung und Interpretation inhaltlich begründete und konsensfähige Qualitätsstandards angelegt werden (zu Qualitätsstandards bei der Auswertung und Interpretation von Evaluationsdaten siehe Abschnitt 3.4). Schließlich ist es eine wesentliche Leistung der Evaluation, die Ergebnisse hinsichtlich der Einzelmerkmale für eine globale Qualitätsbewertung adäquat zu bündeln (summative Funktion) und für ggf. notwendige Maßnahmen der Qualitätsentwicklung entsprechende Konsequenzen für die Praxis abzuleiten (formative Funktion).

2.3 Informanten und Evaluatoren

Evaluatoren und Informanten

Für die Planung, Durchführung, Auswertung und Interpretation von Evaluationsstudien sind üblicherweise Evaluatorinnen und Evaluatoren zuständig, die entsprechendes sozialwissenschaftliches Methoden-Know-how mitbringen sollten. Bei der Lehrveranstaltungsevaluation auf der Basis von studentischer Veranstaltungskritik, nehmen die Studierenden die Rolle von Informanten bzw. Datenlieferanten ein. Nur bei Lehrevaluationen, die von Fachschaften durchgeführt werden (siehe Einleitung), fungieren einzelne Studierende auch als Evaluatoren, die z.B. Evaluationsergebnisse interpretieren und publizieren. In der Regel wird bei Lehrveranstaltungsevaluationen die Evaluatoren-Rolle von unterschiedlichen Personen(gruppen) wahrgenommen: So können Evaluationsinstrumente beispielsweise von einzelnen Lehrenden oder von Gremien (z.B. Fakultätsrat) entwickelt werden, und Evaluationsergebnisse können unter anderem von den Lehrenden selbst, von der Controlling-Abteilung der Universität oder von Studiendekanen interpretiert werden. An wenigen Universitäten wird die Lehrevaluation zentral durch Evaluationsarbeitsgruppen oder Evaluationszentren organisiert.

Bei der Lehrveranstaltungsevaluation ist die Informanten-Rolle (Wer liefert Daten zu den untersuchten Lehrveranstaltungen?) von der Evaluatoren-Rolle (Wer organisiert die Evaluationsforschung?) zu differenzieren.

3. Wie wird Lehrveranstaltungsevaluation durchgeführt?

Die Praxis der Lehrveranstaltungsevaluation in Deutschland lässt sich grob charakterisieren hinsichtlich ihrer Verbreitung (3.1), der verwendeten Evaluationsinstrumente (3.2), der Datenerhebung und Stichproben (3.3) sowie der Datenauswertung und Interpretation (3.4).

3.1 Verbreitung der Lehrevaluation

Lehrveranstaltungsevaluation international

Wie verbreitet ist die Lehrveranstaltungsevaluation international? In einer repräsentativen Studie wurden Hochschullehrerinnen und Hochschullehrer aus 15 ausgewählten Ländern gefragt, ob an ihren Fachbereichen regelmäßig Lehrevaluation stattfindet. Während beispielsweise in den USA (97%), in England (94%), in Brasilien (93%), Australien (89%) und Russland (86%) die überwältigende Mehrzahl der Befragten bejahte, waren es in Deutschland nur 42% (Boyer/Altbach/Whitelaw 1994). Zwar mögen sich die Verhältnisse in den letzten Jahren noch verändert haben, es lässt sich jedoch konstatieren, dass Deutschland in der Lehrveranstaltungsevaluation keine Vorreiterrolle einnimmt, sondern sich eher am Ausland orientiert, oft an den USA, aber auch an den Niederlanden (für einen Praxisleitfaden zur Durchführung studentischer Lehrevaluation nach dem Niederländischen Modell siehe Richter 1994).

Betrachtet man die Situation in Deutschland genauer, so zeigt eine Studie von Schnell und Kopp (2000) für sozialwissenschaftliche Studiengänge grob eine Drittel-Aufteilung hinsichtlich der Lehrevaluationsfrequenz: In einem Drittel der sozialwissenschaftlichen Studiengänge werden jedes Semester die Lehrveranstaltungen evaluiert (33%), während in gut einem Drittel seltener (38%) und knapp einem Drittel noch nie evaluiert wurde (29%).

3.2 Evaluationsinstrumente

Ad-hoc Fragebögen

Weiterhin ergab die Befragung in allen 94 sozialwissenschaftlichen Studiengängen bzw. Fachbereichen in Deutschland (Rücklauf n=81), dass die Lehrveranstaltungsevaluation typischerweise mit Ad-hoc-Fragebögen abgewickelt wird (Schnell/Kopp 2000, S. 25f.). Ad-hoc-Fragebögen basieren weder auf einem elaborierten und explizierten theoretischen Modell der Lehrveranstaltungsqualität, noch sind sie in der Regel hinsichtlich ihrer testtheoretischen Gütemerkmale (vor allem Reliabilität und Validität) geprüft. Obwohl im deutschsprachigen Raum durchaus einige wissenschaftliche erprobte Instrumente zur Lehrveranstaltungsevaluation existieren (z.B. das Heidelberger Inventar zur Lehrveranstaltungsevaluation HILVE, siehe Rindermann 2001), werden in den Fachbereichen überwiegend selbst entwickelte Fragebögen verwendet.

Formative Fragen, summative Fragen und qualitative Urteile

Bei aller Unterschiedlichkeit ähneln sich diese Instrumente insofern, als sie in der Regel sowohl eine Reihe formativer Fragen enthalten, die sich auf Einzelaspekte der Lehrveranstaltung beziehen (z.B. Tempo, Lehrmaterialien, Erklärungen, Beispiele usw.) als auch summative Fragen umfassen, die auf eine Globalbewertung hinauslaufen (z.B. „Wie zufrieden waren Sie mit der Lehrveranstaltung insgesamt?“, zu beantworten auf einer Zufriedenheits- oder Schulnoten-Skala). Die Fragebögen erheben überwiegend quantitative Daten (Urteile auf Ratingskalen), erlauben den befragten Studierenden in der Regel aber auch qualitative Urteile in Form offener Kommentare (Lob, Kritik) sowie konkreter Verbesserungsvorschläge.

3.3 Datenerhebung und Stichproben

Paper-Pencil-Verfahren und Selbstselektions-stichproben

Die Datenerhebung mit Hilfe der Lehrevaluationsfragebögen erfolgt typischerweise im Paper-Pencil-Verfahren in den einzelnen Lehrveranstaltungen gegen Ende des Semesters. Befragt werden dementsprechend Selbstselektionsstichproben, das heißt, nur die physisch Anwesenden nehmen teil. Die Nichtberücksichtigung von No-Shows führt dabei zu systematisch verzerrten Stichproben. Mangels genauer empirischer Analysen zur Gruppe der No-Shows ist es spekulativ, pauschal von einer Verzerrung allein im Sinne besonders zufriedener Studierender auszugehen.

Internet-basierte Datenerhebung

In jüngster Zeit stellen immer mehr Universitäten auf Internet-basierte Datenerhebung um, da diese eine automatische Analyse sowie eine umfassende Archivierung der Evaluationsdaten ermöglicht (z.B. mittels EvaSys: www.evasys.de).

3.4 Auswertung und Interpretation

Mittelwertsbildung

Formative Fragen zu Einzelaspekten sowie die summativen Fragen zur Gesamtbewertung der Lehrveranstaltung werden typischerweise durch Mittelwertsbildung über alle befragten Studierenden ausgewertet. Dabei ist zu beachten, dass die Gesamtbewertung der Lehrveranstaltung sich nicht aus der Zusammenfassung der einzelnen formati-

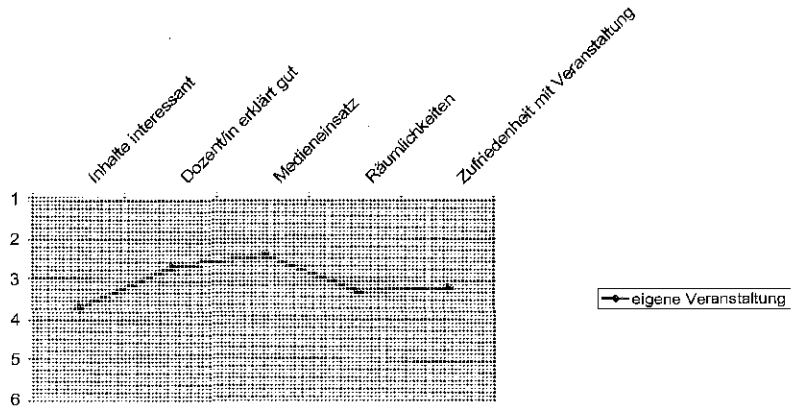

Abb. I 1.7-1 **Evaluationsergebnisse einer fiktiven Lehrveranstaltung**[1]

ven Fragen (d.h. Fragen nach Interessantheit der Inhalte, Medieneinsatz, Tempo, Beispielen, Fairness von Prüfungen etc.) ergibt, sondern als Ergebnis der direkten summativen Frage nach der Zufriedenheit mit der Lehrveranstaltung bzw. Lehrkraft. Als Antwortmodalitäten werden typischerweise fünf- bis siebenstufige Zufriedenheitsskalen oder Schulnotenskalen verwendet. Abbildung I 1.7-1 zeigt die Ergebnisse der Evaluation einer fiktiven Lehrveranstaltung, visualisiert über das Profil der Mittelwerte, die sich pro Merkmal jeweils durch Zusammenfassung der Einzelurteile der befragten Studierenden ergeben. In der Globalbewertung erreichte diese Lehrveranstaltung die Note 3,2. Dabei beurteilten die Studierenden die Interessantheit der Inhalte am schlechtesten (fast Note 4), während sie den Medieneinsatz mit der Note 2,4 am besten einstuften.

[1] Ergebnisse für vier formative und eine summative Variable. Jeweils Mittelwerte der Einzelurteile von Studierenden auf einer Schulnotenskala.

Qualitätsstandards bzw. Qualitätskriterien

Die Frage lautet nun: Wie ist die Qualität dieser konkreten Veranstaltung zu bewerten? Ist die Qualität ausreichend? Oder deuten die Daten auf so gravierende Defizite hin, dass eine Qualitätsverbesserung eingeleitet werden muss? Um diese Fragen beantworten zu können benötigen wir für die betrachteten Merkmale jeweils Qualitätsstandards bzw. Qualitätskriterien. Diese existieren jedoch nicht. A priori ist nicht festgelegt, ab welcher Durchschnittsnote einer Lehrveranstaltung hohe oder geringe Lehrqualität zugeschrieben werden soll.

Normorientiertes Testen

Anstelle eines kriteriumsorientierten Testens wird deswegen normorientiert vorgegangen: Das Notenprofil der betrachteten einzelnen Lehrveranstaltung wird beim normorientierten Testen mit dem Durchschnittsprofil aller Lehrveranstaltungen des selben Semesters im selben Studiengang bzw. in derselben Fakultät kontrastiert (siehe Abbildung I 1.7-2): Jenen Veranstaltungen, die bei der summativen Frage besser als die Durchschnittsveranstaltung abschneiden, wird dann ausreichende bzw. hohe Qualität, denen, die schlechter abschneiden geringe Qualität zugeschrieben. Im vorliegenden Beispiel lag die erzielte die durchschnittliche Veranstaltung eine Zufriedenheitsnote von 2,9, so dass die betrachtete Lehrveranstaltung mit der Zufriedenheitsnote 3,2 leicht unterdurchschnittlich ausfällt.

Ranking der Veranstaltungen

Exzellenz in der Lehre wird zusätzlich denjenigen Veranstaltungen zugebilligt, die sich unter den 10% der am besten bewerteten Veranstaltungen des jeweiligen Semesters befinden. Umgekehrt werden gravierende Defizite bei den Veranstaltungen bzw. Dozierenden diagnostiziert, die sich unter den 10% der am schlechtesten bewerteten Veranstaltungen befinden. Diese Art der Interpretation basiert auf einem Ranking der Veranstaltungen und damit auch der Lehrenden anhand der studentischen Lehrveranstaltungsbewertung.

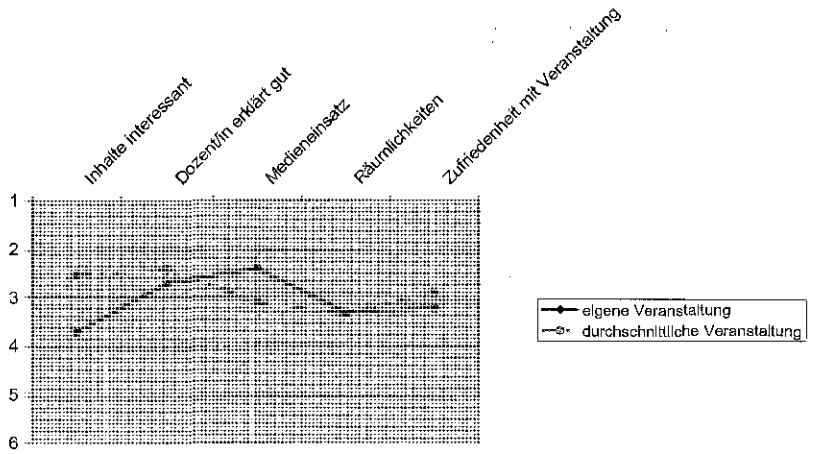

Abb. I 1.7-2 **Evaluationsergebnisse einer fiktiven Lehrveranstaltung**[2]

Qualitative Kommentare

Die qualitativen Kommentare werden in der Regel nicht systematisch (z.B. inhaltsanalytisch) ausgewertet, sondern nur kursorisch überflogen. Dabei besteht die Gefahr der selektiven Wahrnehmung. Zumal qualitative Kommentare ohnehin nur von einem Bruchteil der Studierenden überhaupt formuliert werden. Sowohl bei den quantitativen als auch bei den qualitativen Fragen besteht - vor allem bei großen und heterogenen Teilnehmergruppen - zudem das Problem der Urteilsdivergenz etwa in dem Sinne, dass Teilgruppen von Veranstaltungsteilnehmern konträre Veränderungen wünschen.

> Anhand der Evaluationsdaten wird einer Lehrveranstaltung gute Lehrqualität zugeschrieben, wenn sie im Vergleich zu anderen Lehrveranstaltungen überdurchschnittlich abschneidet. Exzellenz in der Lehre wird dem obersten Perzentil der Lehrveranstaltungen bescheinigt.

An der Wirtschaftsuniversität Wien müssen Lehrende, die in zwei aufeinander folgenden Semestern im untersten Perzentil gerankt waren, beim Studiendekan vorsprechen und werden verpflichtet, hochschuldidaktische Weiterbildungsmaßnahmen zu besuchen. Solche Negativsanktionen sind in Deutschland weitgehend unüblich. Man lässt es in der Regel dabei bewenden, die Lehrenden mit ihren - mehr oder minder positiven oder negativen - Ergebnissen zu konfrontieren und fordert sie auf, diese zu berücksichtigen sowie mit den Studierenden zu besprechen. Wie dies im Einzelnen geschehen soll, welche formativen Befunde in der welcher Weise zum Handeln anregen müssen, welche Arten von hochschuldidaktischen Veränderungen möglich und sinnvoll sind - z.B. auch unter Kosten-Nutzen-Erwägungen - all dies bleibt offen (siehe Abschnitt 5).

[2] Im Vergleich zur durchschnittlichen Bewertung aller Lehrveranstaltungen im selben Fachbereich im selben Semester (Schulnotenskala).

4. Welche Aussagekraft haben die gewonnenen Evaluationsdaten?

Den starken Forderungen nach Lehrevaluation stehen kritische Argumente gegenüber, die z.B. hinterfragen, ob die gängige Lehrevaluation überhaupt relevante Dimensionen der Lehrqualität erfasst (4.1), ob die Instrumente bzw. die Studierenden zuverlässige und gültige Urteile liefern (4.2), ob es zu systematischen Urteilsverzerrungen kommt (4.3) und ob die Datenauswertung mit Interpretationsproblemen (4.4) behaftet ist.

4.1 Erfasste Dimensionen der Lehrqualität

Die Qualität von Hochschullehre lässt sich allgemein in drei aufeinander aufbauende Teil-Qualitäten differenzieren (zu Modellen der Lehrqualität siehe Voss, 2004). Diese werden in der Praxis nur selektiv empirisch erfasst:

Strukturqualität

Strukturqualität (Voraussetzungen der Lehre). Diese Bedingungsfaktoren lassen sich bündeln in a) Umfeldfaktoren (z.B. Räumlichkeiten, Technikausstattung, Bibliothekszugang), b) Studierendenfaktoren (z.B. Studierfähigkeit, Vorwissen) und c) Lehrendenfaktoren (z.B. Lehrkompetenz einschließlich hochschuldidaktischer Expertise, Fachkompetenz, Motivation). Strukturqualität wird im Rahmen der Messung von Lehrveranstaltungsqualität allenfalls am Rande berücksichtigt.

Prozessqualität

Prozessqualität (Durchführung der Lehre). Wie viele und welche Aspekte der Durchführung der Lehre für die Qualität von Lehrveranstaltungen wichtig sind, wird in den Theorien der Lehrqualität, die nur teilweise auf Ergebnisse empirischer Studien über studentisches Lernen zurückgreifen, unterschiedlich beantwortet (vgl. Berendt 2000; Voss 2004). Einigkeit besteht bei den Evaluationsinstrumenten weitgehend darin, dass Merkmale des Lehrendenverhaltens für die Prozessqualität einschlägig sind (z.B. Enthusiasmus, Erklärungen, Tempo, Fairness, Freundlichkeit, Hilfsbereitschaft etc.). Das Lehrendenverhalten wird dabei in der Regel durch die Studierenden bewertet, manchmal auch durch die Lehrenden selbst. Das Studierendenverhalten (z.B. Vor- und Nachbereitung, aktive Beteiligung an Diskussionen) wird in gängigen Evaluationsfragebögen häufig randständiger einbezogen. Viele Fragebögen orientieren sich bei der Abfrage von Prozessvariablen ungenügend an der gängigen Lehr-Lern-Forschung (Preißer 2003).

Produktqualität

Produktqualität (Ergebnisse der Lehre). Gute Lehre sollte sich kurzfristig in objektivem Lernerfolg und langfristig in Berufserfolg sowie in einer positiven Lebensbewältigung widerspiegeln (vgl. Helmke 1996, S. 183). Doch diese zentralen Aspekte der Produktqualität werden in der Lehrveranstaltungsevaluation nicht direkt gemessen. Stattdessen werden für die summative Bewertung die subjektive Zufriedenheit der Studierenden mit der Lehrveranstaltung sowie teilweise der subjektive Lernerfolg erfasst. Da diese subjektiven Maße Verzerrungen unterworfen sind, bieten sie nur eine Annäherung an den objektiven Lernerfolg. Das Operieren mit subjektiven Maßen ist in der Praxis vor allem deswegen so populär, weil die Erfassung des objektiven Lernerfolgs sehr aufwändig ist: Sie erfordert z.B. Vorher- und Nachher-Messungen von Wissen, Einstellungen, Fertigkeiten. Forschungspraktisch noch aufwändiger ist die Erfassung von späterem Berufserfolg, der sich in Alumni-Studien allenfalls mit der Studiengangs-Qualität in Verbindung bringen lässt, kaum aber mit der Qualität einzelner Lehrveranstaltungen.

> Die Aussagekraft gängiger Lehrveranstaltungsevaluationen ist hinsichtlich der drei Dimensionen von Lehrqualität (Struktur, Prozess, Produkt) stark eingeschränkt, weil Aspekte der Prozessqualität im Zentrum stehen und die Produktqualität nur lückenhaft erfasst wird.

4.2 Zuverlässigkeit und Gültigkeit der Urteile

Urteilerübereinstimmung

Damit die Einzelurteile der Teilnehmerinnen und Teilnehmer einer Lehrveranstaltung sinnvoll zu Mittelwerten zusammengefasst werden können, muss sicher gestellt sein, dass die Urteilerübereinstimmung hinreichend groß ist. Zwar kann für jeden Mittelwert ergänzend angegeben werden, wie groß die Streuung der Urteile ist, für diese Streuungswerte existieren jedoch keine klaren Bewertungsnormen.

Reliabilität

Anders ist es bei der statistischen Urteilerübereinstimmung bzw. Reliabilität (Zuverlässigkeit bzw. Messgenauigkeit), für die Bewertungskriterien existieren. Bestimmt man beispielsweise die Split-Half-Reliabilität, so zeigen sich für die summativen Bewertungen von Lehrveranstaltungen Reliabilitätskoeffizienten von Rel ≤ 80 bei 10 bis 20 Urteilern und von Rel ≤ .90 bei 20 bis 40 Urteilern (Rindermann 2001, S. 131, Cashin 1988, S. 1), welche nach gängigen Standards als zufrieden stellend einzustufen sind. Es kann aber auch Konstellationen geben, in denen die Urteilerübereinstimmung sehr gering ist, so dass im Grunde separate Auswertungen für unterschiedliche Teilgruppen der Teilnehmer notwendig sind.

Sofern für eine Lehrveranstaltung Fragebogenergebnisse von mindestens 20 Studierenden vorliegen, resultieren in der Regel zuverlässige (reliable) Mittelwerte.

Validität

Selbst wenn Studierende übereinstimmend urteilen, stellt sich die noch wichtigere Frage, ob sie auch in einer Art und Weise urteilen, in der tatsächlich Lehrqualität abgebildet wird. Diese Frage nach der Gültigkeit der Urteile (Validität) lässt sich mit verschiedenen Validierungsmethoden überprüfen. Eine Methode ist die Kriteriumsvalidierung, bei der die studentischen Globalurteile verschiedener Veranstaltungen jeweils mit anderen Qualitätsindikatoren korreliert werden. Diese anderen Qualitätskriterien können z.B. Veranstaltungsbewertungen von Kollegen, geschulten Beobachtern oder Alumni sein. Tatsächlich zeigen Metaanalysen, in denen die Befunde zahlreicher Einzelstudien statistisch aggregiert wurden, dass der entsprechende Zusammenhang bei $r \approx .50$ liegt, was nach gängigen Kriterien für ausreichende Validität spricht (Rindermann 2001, S. 163ff.). Auch die Korrelation mit dem objektiven Lernerfolg fiel in einer klassischen Metaanalyse von Cohen (1981) mit $r = .46$ zufrieden stellend aus. Gleichzeitig ist zu beachten, dass bei Validitätskoeffizienten um $r = .50$ und dementsprechenden Determinationskoeffizienten von $r^2 = .25$ nur 25% der Varianz aufgeklärt werden und immerhin 75% der Unterschiedlichkeit der Studierendenurteile durch andere - und möglicherweise sachfremde - Faktoren beeinflusst sind.

Durchschnittliche Veranstaltungsbewertungen von Studierenden sind insofern als gültig (valide) anzuerkennen, als sie substantiell mit den Einschätzungen anderer Urteilergruppen sowie mit dem objektiven Lernerfolg korrelieren. Dies schließt jedoch systematische Urteilsverzerrungen nicht aus.

4.3 Urteilsverzerrungen

Bias-Variablen

Tatsächlich zeigt die umfangreiche internationale Forschung zu Lehrevaluationsdaten, dass studentische Lehrveranstaltungsbewertungen systematischen Verzerrungen unterliegen, beispielsweise im Hinblick auf folgende Bias-Variablen (Cashin 1988, Kromrey 1994, Rindermann 2001):

1. Teilnahmegrund:
   Pflichtveranstaltungen werden negativer bewertet als Wahlveranstaltungen.

2. Themenbeliebtheit:
   Veranstaltungen mit populären Themen werden positiver bewertet als Veranstaltungen mit unpopulären Themen.

3. Studierverhalten:
   Veranstaltungen, in denen Studierende als Kollektiv aktiver agieren, werden positiver bewertet als Veranstaltungen mit eher passivem Studierverhalten.

Keine oder allenfalls geringfügige Verzerrungen entstehen durch sozialstatistische Merkmale und Persönlichkeitsvariablen auf Seiten der Lehrenden und/oder Studierenden. Entgegen der verbreiteten Vermutung, dass Lehrveranstaltungen umso besser bewertet werden, je geringer die Anforderungen an die Studierenden sind, erzielen Lehrveranstaltungen mit einem als adäquat empfundenen (also auch nicht zu leichten) Anforderungsniveau die besten Bewertungen. Die schwache positive Korrelation zwischen Veranstaltungsbewertung und Benotung ist nicht eindeutig als Biasvariable zu interpretieren. Schließlich sollte sich theoriekonform eine gute Lehre sowohl in guten Evaluations- als auch in guten Lernergebnissen bzw. guten Noten widerspiegeln.

Vergleichbare Veranstaltungen

Damit Biasvariablen nicht zu verfälschten Schlussfolgerungen führen (z.B. gute Lehrkraft erhält in unpopulärer Pflichtvorlesung negative Bewertung oder umgekehrt), dürfen bei der normorientierten Auswertung nur vergleichbare Veranstaltungen aneinander gemessen werden (z.B. sollten Pflichtvorlesungen nur mit anderen Pflichtvorlesungen verglichen werden). Auch sind themen- und fachspezifische Bewertungsunterschiede einzubeziehen.

Peer-Review

Um den verschiedenen Verzerrungsfaktoren Rechnung zu tragen ist zudem ein direkter Rückschluss von der Bewertung einer Lehrveranstaltung auf die Lehrleistung der Lehrkraft nicht zulässig. Stattdessen darf eine Bewertung der Person nur erfolgen, wenn Evaluationsergebnisse für unterschiedliche Veranstaltungstypen und -themen vorliegen. In den USA, wo Lehrevaluationen schon länger eine wichtige Rolle für Karrierechancen spielen, wird studentisches Lehrveranstaltungsfeedback nicht nur von mehreren Veranstaltungen pro Lehrkraft berücksichtigt, sondern auch um weitere Datenquellen ergänzt, beispielsweise um ein Peer-Review zur fachwissenschaftlichen Aktualität und Qualität der Lehrinhalte und Lehrmaterialien (vgl. Stassen 2002).

> Studentische Veranstaltungsbewertungen sind durch eine Reihe von Faktoren verzerrt, die nicht direkt die Lehrqualität betreffen (z.B. Teilnahmemotivation, Themenpopularität). Um fehlerhafte Schlussfolgerungen zu vermeiden, dürfen bei der normorientierten Interpretation nur vergleichbare Veranstaltungen aneinander gemessen werden. Zudem müssen zur Bewertung der Lehrleistung einer Person die Ergebnisse mehrerer Veranstaltungen berücksichtigt sowie weitere Datenquellen (z.B. Peer-Review) herangezogen werden.

4.4 Interpretationsprobleme

Kosten-Nutzen-Relation

Die große Mehrzahl der Lehrveranstaltungen erzielt in der Evaluation Durchschnittsnoten zwischen 2 und 3. Dies bedeutet, dass messbare und statistisch signifikante Steigerungen nur bedingt überhaupt erreichbar sind. Zudem sind Verbesserungsinitiativen nur dann rational, wenn sie in vertretbarer Kosten-Nutzen-Relation stehen. Diese Einschränkung wird jedoch in der gängigen Evaluationsdiskussion überhaupt nicht beachtet. Wenn etwa ein Teil der Studierenden im Rahmen einer Evaluation fordert, dass Vorlesungsfolien in verschiedenen Formaten regelmäßig vor jeder Sitzung online und offline distribuiert werden, so fragt sich, ob der daraus resultierende Zeit- und Kostenaufwand in vernünftiger Relation zum tatsächlichen didaktischen Mehrwert dieser Maßnahme steht.

Negative Konsequenzen für den Lehr-Lern-Prozess

Die Bereitstellung von Folien im Vorfeld kann gemäß anekdotischen Erfahrungen sogar negative Konsequenzen für den Lehr-Lern-Prozess haben: So sind dramaturgische Spannungsbögen kaum mehr aufbaubar, wenn Studierende den inhaltlichen Verlauf einer Sitzung im Vorfeld oder parallel den Unterlagen entnehmen. Auch berichten eine Reihe von Studierenden in Evaluationsdiskussionen, dass die Bereitstellung von Folien ihre Motivation senke, überhaupt an Veranstaltungen teilzunehmen oder sich in den Sitzungen stark zu konzentrieren, weil sie den Stoff ja vermeintlich „schon in der Tasche haben". Auf Evaluationsbögen geäußerte Vorschläge müssen also immer wieder auf ihre Kosten-Nutzen-Bilanz und ihren didaktischen Mehrwert kritisch geprüft und dann zuweilen auch von der Lehrkraft verworfen werden.

Qualität eines Studiengangs

Mit Blick auf die globalen Ziele von Lehre und Studium (Kompetenzentwicklung, Vorbereitung auf die Tätigkeit in spezifischen Berufsfeldern etc.) ist zudem die Bedeutung von Qualitätsabstrichen in einzelnen Lehrveranstaltungen zu relativieren. Der Aufwand, der möglicherweise auf Mikroebene betrieben wird oder werden müsste, um den Unterhaltungswert von Dozentin X oder das Tafelbild von Dozent Y zu optimieren, steht womöglich in keinem Verhältnis zu der effektiven Bedeutung, die diese Details für die gesamte Qualität eines Studiengangs haben. Demgegenüber können - bislang in der Praxis eher vernachlässigte - Qualitätsdiskurse auf Makroebene eine unverhältnismäßig größere und nachhaltigere Wirkung entfalten, etwa wenn unter Beteiligung von Lehrenden, Studierenden, Alumni und zukünftigen Arbeitgebern vor allem die Inhalte der Lehre optimiert und aktualisiert werden (Kromrey 1994).

Normorientierte Interpretation

Hochgradig problematisch erscheint eine normorientierte Interpretation von Evaluationsdaten, in der stets die 10% am untersten Ende des Rankings als schlechte Lehrende kategorisiert werden. Dass ein Teil der Lehrenden unterdurchschnittlich und 10% im letzten Perzentil platziert sind, leitet sich nicht a priori aus Lehrdefiziten ab, sondern ist

statistische Gesetzmäßigkeit. Auch in einem hochschuldidaktisch optimal ausgebildeten Fachbereich mit sehr hoher Lehrqualität muss aus statistischen Gründen ein Teil der Lehrveranstaltungen unterdurchschnittlich abschneiden, ohne dass es dabei jedoch inhaltlich gerechtfertigt wäre, diesen schlechte Qualität zuzuschreiben.

Ein Fachbereich, der nachhaltig Sorge um die Lehrqualität trägt, hat keinen rationalen Anlass, jedes Semester alle Lehrveranstaltungen - auch die mit bewährten Konzepten und Dozenten - immer wieder aufs Neue zu evaluieren. Der bürokratisierten Überevaluation, der sich Dozierende kaum entziehen können, ohne dass Zweifel an ihrer Lehrkompetenz aufkommen, steht auf der anderen Seite ein eklatanter institutioneller Mangel an professionellen Maßnahmen der hochschuldidaktischen Interpretation und Umsetzung der Daten gegenüber.

Damit Lehrevaluation zur Verbesserung der Lehrqualität beitragen kann, müssen ihre Ergebnisse im Hinblick auf klare Handlungsvorgaben interpretierbar sein. Dies ist bislang nicht gegeben: Es fehlen Kosten-Nutzen-Erwägungen, Verknüpfungen zwischen Mikro- und Makroevaluation sowie Alternativen zur normorientierten Bewertung.

5. Welche Konsequenzen hat die Lehrveranstaltungsevaluation?

Selbstreflexion der Lehrenden und ...

Ergebnisse der Lehrveranstaltungsevaluation werden zunächst von den Lehrenden selbst - mehr oder minder intensiv - rezipiert und reflektiert. In einer Befragung von Lehrenden in den USA stellte sich heraus, dass immerhin 45% regelmäßig ihre Lehrveranstaltungen auf der Basis des studentischen Feedbacks modifizierten (Schmelkin/Spencer/Gellman 1997). Neben einer Selbstreflexion der Lehrenden wird in der Regel ein Dialog zwischen Lehrenden und Studierenden über die Evaluationsergebnisse empfohlen und durchgeführt.

... Dialog zwischen Lehrenden und Studierenden

Dementsprechend sollen die Evaluationsdaten bereits deutlich vor Semesterende erhoben werden, damit sie spätestens in der letzten Sitzung bereits ausgewertet vorliegen und somit gemeinsam besprochen werden können. Im Dialog möglicherweise identifizierte Schwächen und Verbesserungsmöglichkeiten können dann jedoch nur in Folgeveranstaltungen greifen.

Selbstreflexions-Hypothese und Dialog-Hypothese

Obwohl Selbstreflexion und Dialog als Nutzungsformen von Evaluationsdaten in der Fachliteratur einhellig als Königswege zur Sicherung der Lehrqualität sowie zur Verbesserung des Verhältnisses zwischen Lehrenden und Studierenden gewürdigt werden, muss man aus wissenschaftlicher Sicht fragen, ob diese Annahmen gerechtfertigt sind.

Gibt es überzeugende empirische Evidenzen für die Selbstreflexions-Hypothese und für die Dialog-Hypothese, die eine resultierende Qualitätssteigerung belegen? Überraschenderweise wurden beide Hypothesen bislang nur selten untersucht - und die vorliegenden Befunde sind negativ: Wenn Lehrende über die ihnen vorgelegten Evaluationsdaten nachdenken und mit Studierenden darüber reden, resultiert - auch über mehrere Semester hinweg - keine messbare Steigerung der Lehrqualität (Rindermann 2001; Schnell/Kopp 2000).

Psychologische und soziale Negativerfahrungen

Hinzu kommt, dass in diversen Evaluationsberichten deutlich wird, dass im Kontext des Evaluationsgeschehens sowohl auf Seiten der Lehrenden als auch der Lernenden psychologische und soziale Negativerfahrungen gemacht werden. So kann allein die Aufforderung an Studierende, ihre Wünsche zu artikulieren, zu einer Erhöhung von Anspruchsniveaus und dementsprechend zu Enttäuschungen führen - gerade wenn im Vorfeld unklar ist, ob und in welchem Maße welche Änderungswünsche überhaupt berücksichtigt werden können. Lehrende können durch Bewertungen, die sie als ungerecht empfinden sowie durch studentische Forderungen, die sie als überzogen einstufen, Frustrationen erleben, wodurch sich das Engagement in der Lehre nicht nur nicht erhöht, sondern möglicherweise sogar reduziert. Entsprechende Probleme, die sich auch negativ auf das Verhältnis zwischen Studierenden und Lehrenden niederschlagen können, sind in der Praxis unübersehbar, werden in der Literatur aber kaum thematisiert und sind bislang auch nicht Gegenstand empirischer Studien.

> Dass Lehrende über ihre Evaluationsergebnisse nachdenken (Selbstreflexion) und sie mit Studierenden besprechen (Dialog), wird als Königsweg für die Sicherung der Lehrqualität propagiert. Empirisch scheint dieser Ansatz im Hinblick auf Qualitätsverbesserung jedoch wirkungslos zu sein. Möglicherweise entstehen sogar auf psychologischer und sozialer Ebene nennenswerte Negativwirkungen (z.B. Anspruchseskalation, Enttäuschungen).

Beratungs-Ansatz

Damit Evaluationsergebnisse in der Praxis positive Konsequenzen entfalten können, fordert der Beratungs-Ansatz (Rindermann 2001, S. 250ff.) eine individuelle hochschuldidaktische Beratung bis hin zum Coaching für Lehrende, bei denen Verbesserungsbedarf besteht. Demgemäß ist spezifische hochschuldidaktische Expertise notwendig, um adäquate Veränderungsstrategien zu entwickeln sowie in die Praxis umzusetzen. Lehrende sollten hierbei nicht allein gelassen werden: Wer mit spezifischen negativen Evaluationsergebnissen konfrontiert wird, muss individuell zugeschnittene Lernchancen erhalten. In den USA ist dieser Ansatz vielfach etabliert und in seiner Wirksamkeit belegt (Marsh/Roche 1993). Die in der Bundesrepublik von hochschuldidaktischen Einrichtungen in fast allen Bundesländern (teilweise flächendeckend, z.B. Baden-Württemberg) durchgeführten hochschuldidaktischen Aus- und Weiterbildungsprogramme umfassen unter anderem auch Beratung und Coaching (vgl. Berendt 2005).

Wenn Lehrevaluation wirklich zu einer Verbesserung der Lehrqualität beitragen soll, müssen Lehrende in der gezielten Entwicklung ihrer Lehrkompetenz individuell unterstützt und auch individualisiert hochschuldidaktisch beraten werden.

Sanktions-Modell

Ein weiteres Wirkmodell der Lehrevaluation neben Selbstreflexion/Dialog und Beratung lässt sich als Sanktions-Modell kennzeichnen. Es geht davon aus, dass Bestrafungen bei negativen Evaluationsergebnissen und Belohnungen für positive Evaluationsbefunde die Lehrqualität regulieren. Während Bestrafungen, wie z.B. öffentliches „An-den-Pranger-Stellen" durch Publikation personalisierter Evaluationsdaten und Verleihung negativer Lehrpreise („schwarze Kreide") vermutlich eher zu Defensivreaktionen und Reaktanz führen, könnten positive Anreize durchaus das Lehrengagement steigern (vgl. Hackl/Sedlacek 2001). Dabei sollten Lehrleistungen dann jedoch genau wie Forschungsleistungen nicht normorientiert, sondern kriteriumsorientiert bemessen werden. Das heißt für Lehrveranstaltungen könnten in Abhängigkeit von der Erreichung vordefinierter fach- und themenspezifischer Ziele Leistungspunkte und daraus resultierend Mittel vergeben werden (z.B. Pflichtvorlesung ab Note 3; Seminar ab Note 2,5). Im Rahmen eines Anreiz-Modells ist dann auch eine regelmäßige Evaluation aller Lehrveranstaltungen sinnvoll.

Im Rahmen eines Sanktions-Modells sollten statt negativer Sanktionen eher positive Anreize für gute Lehrqualität geschaffen werden. Dabei sind jedoch angemessene Bewertungs- und Verteilungsmodelle bislang unklar. Auch die Wirksamkeit der Anreiz-Hypothese müsste empirisch noch geprüft werden.

6. Wie muss gute Lehrevaluation gestaltet werden?

Die in der Diagnose der Lehrevaluation aufgezeigten Probleme und Lösungsansätze lassen sich in vier Hauptpunkten zusammenfassen:

Qualitätsentwicklung im Zentrum

1. Lehrevaluation muss konsequent als Mittel zur Lehrqualitätsverbesserung konzipiert und eingesetzt werden und darf nicht zum Selbstzweck (z.B. routinemäßiges, aber konsequenzenloses Austeilen von Evaluationsbögen in jedem Semester in allen Lehrveranstaltungen) werden.

Lehrevaluation auf Mikroebene optimieren

2. Auf der Ebene der Lehrveranstaltungen gilt es, die Qualitätsmessung zu verfeinern (z.B. Berücksichtigung von Biasvariablen, Festlegung von begründeten Qualitätsstandards, Berücksichtigung weiterer Datenquellen wie z.B. Peer-Review). Auf dieser Basis ist es dann notwendig, im Sinne der Qualitätsentwicklung individuelle hochschuldidaktische Beratung und Betreuung zu implementieren.

Lehrevaluation auf Makroebene hinzufügen

3. Die Lehrveranstaltungsevaluation muss durch Lehrevaluation auf Makroebene ergänzt werden, um die Qualität des Studiums umfassend und nachhaltig sicher zu stellen. Für diese Makroevaluation liegen erprobte mehrstufige Konzepte vor, die von den Evaluationsverbünden (z.B. Nordverbund: www.uni-nordverbund.de; siehe auch Daniel/Mittag/Bornmann 2003) seit Jahren eingesetzt werden. Die Hochschulrektorenkonferenz (www.hrk.de) rückt im Rahmen ihres Projekts "Qualitätssicherung" in den letzten Jahren die Lehrevaluation auf Makroebene immer stärker in den Vordergrund. Perspektivisch sollte Studiengangsevaluation an Akkreditierung bzw. Re-Akkreditierung gekoppelt werden, wobei zu beachten ist, dass Akkreditierung mit Minimalstandards operiert, während Evaluation Regelstandards nutzt.

Meta-Evaluation betreiben

4. Lehrevaluation sollte als Lehrevaluationsforschung von professionellen Evaluatorinnen und Evaluatoren durchgeführt werden. Dabei muss sicher gestellt werden, dass die international verbindlichen wissenschaftlichen und ethischen Standards der Evaluationsforschung, wie sie von der Deutschen Gesellschaft für Evaluation adapiert wurden (2001), auch im Bereich der Lehrevaluation eingehalten werden. Notwendig ist also auch eine Bewertung der Evaluation im Sinne professioneller Meta-Evaluation.

Evaluation als Evaluationsforschung seriös zu betreiben und nachhaltig mit sinnvollen Interventionsmethoden zu verbinden, erfordert viel mehr Zeit, Geld und Fachexpertise als heute im Bereich Lehrevaluation investiert wird. Anstelle qualitativ hochwertiger Lehrevaluation trifft man vielerorts auf „Evaluitis“, das heißt auf Befragungsroutinen, die neuerdings durch Implementierung in Computersysteme zunehmend festgeschrieben werden. Evaluation wird auf diese Weise zum Selbstzweck. Theoretische, methodische und ethische Defizite werden gleichzeitig ignoriert. Die Kosten für diese Evaluitis sind hoch: Die Evaluationsgelder werden nutzlos ausgegeben. Zudem schaden unprofessionelle Evaluationen mit ihren falschen und teilweise unfairen Schlussfolgerungen nicht nur dem Evaluationsgedanken, sondern vermutlich auch der Lehrqualität und dem Verhältnis zwischen Studierenden, Lehrenden und Universitätsleitungen.

Die flächendeckende Festschreibung nutzloser Evaluationsroutinen bereitet gute Evaluation nicht vor, sondern verhindert sie. Denn Evaluitis suggeriert, das Evaluationsproblem sei gelöst, bevor es richtig in Angriff genommen wurde.

Literatur[3]

[1] Berendt, B. (2000): Was ist gute Hochschullehre. In: Zeitschrift für Pädagogik, 41, S. 247-260.

[2] Berendt, B. (2002): Academic Staff Development (ASD) als Bestandteil von Qualitätssicherung und -entwicklung. In: Berendt, B.; Voss, H.-P.; Wildt, J. (Hrsg.), Neues Handbuch Hochschullehre NHHL. Kapitel L 2.1. Berlin (Loseblattsammlung).

[3] Berendt, B. (2005): Academic Staff Development (ASD) im Kontext des Bologna-Prozesses: Stellenwert und Stand hochschuldidaktischer Aus- und Weiterbildung 2005 in der BRD. In: Berendt, B.; Voss, H.-P.; Wildt, J. (Hrsg.), Neues Handbuch Hochschullehre NHHL. Kapitel L 2.3. Berlin (Loseblattsammlung).

[4] Bortz, J.; Döring, N. (2002): Forschungsmethoden und Evaluation (3. Aufl.). Berlin.

[5] Boyer, E.; Altbach, P.; Whitelaw, M. (1994): The Academic Profession: An International Perspective. Princeton, New York.

[6] Cashin, W. (1988): Student Ratings of Teaching: A Summary of Research. IDEA Paper No 20, Kansas State University.

[7] Cohen, P. (1981): Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multi-Section Validity Studies. In: Review of Educational Research, 51 (3), S. 281-309.

[8] Daniel, H.-D.; Mittag, S.; Bornmann, L. (2003): Mehrstufige Evaluationsverfahren für Studium und Lehre. Empfehlungen zur Durchführung. In: Berendt, B.; Voss, H.-P.; Wildt, J. (Hrsg.), Neues Handbuch Hochschullehre. NHHL Kapitel I 2.2. Berlin (Loseblattsammlung).

[9] Deutsche Gesellschaft für Evaluation (2001): Standards für Evaluation. http://www.degeval.de/standards/

[10] Hackl, P.; Sedlacek, G. (2001): Evaluierung als Chance zur kontinuierlichen Verbesserung der Lehre: Das Beispiel Wirtschaftsuni Wien. In: Spiel, C. (Hrsg.), Evaluation universitärer Lehre - zwischen Qualitätsmanagement und Selbstzweck. Münster.

[11] Helmke, A. (1996): Studentische Evaluation der Lehre - Sackgassen und Perspektiven. In: Zeitschrift für Pädagogische Psychologie, 10 (3/4), S. 181-186.

[12] Kindred, J. / S. N. Mohammed (2005): "He Will Crush You Like an Academic Ninja!" Exploring Teacher Ratings on Ratemyprofessors.com. In: Journal of Computer-Mediated Communication, 10 (3), Article 9. http://jcmc.indiana.edu/vol10/issue3/kindred.html

[13] Kromrey, H. (1994): Wie erkennt man „gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen. In: Empirische Pädagogik, 8 (2), S. 153-168.

[3] Zugriffsdatum für alle elektronischen Dokumente: 6.5.2005.

[14] Preißer, R. (2003): Evaluation der Hochschullehre und Selbststeuerung des Lernens. In: Berendt, B.; Voss, H.-P.; Wildt, J. (Hrsg.), Neues Handbuch Hochschullehre. NHHL Kapitel I 2.3. Berlin (Loseblattsammlung und Ergänzungslieferungen).

[15] Marsh, H. / L. Roche (1993): The Use of Students' Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness. In: American Educational Research Journal, 30 (1), S. 217-251.

[16] Richter, R. (Hrsg.) (1994): Qualitätssorge in der Lehre. Leitfaden für die studentische Lehrevaluation. Neuwied.

[17] Rindermann, H. (2001): Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. Landau.

[18] Schmelkin, L. / K. Spencer / E. Gellman (1997): Faculty Perspectives on Course and Teacher Evaluation. In: Research in Higher Education, 38 (5), S. 575-592.

[19] Schnell, R.; Kopp, J. (2000): Theoretische und methodische Diskussionen der Lehrevaluationsforschung und deren praktische Bedeutung. Projektbericht des Projekts "Fakultätsinterne Evaluation der Lehre: Die Weiterentwicklung des bisherigen Evaluationskonzepts", Universität Konstanz. http://www.ub.uni-konstanz.de/kops/volltexte/2001/605/

[20] Stassen, M. (2002): Student Response to Instruction (SRTI) and Performance Appraisal. University of Massachusetts Amherst. http://www.umass.edu/oapa/SRTI/summative_evaluation_guide.pdf

[21] Voss, R. (2004): Lehrqualität und Lehrqualitätsmanagement an öffentlichen Hochschulen. Hamburg.