Natalija el Hage

Zur Validität studentischer Veranstaltungskritiken

Befunde empirischer Studien zu einem umstrittenen Verfahren

Hefte zur Bildungs- und Hochschulforschung (13)

Arbeitsgruppe Hochschulforschung, Sozialwissenschaftliche Fakultät,
Universität Konstanz, November 1995

Vorwort

Das Thema "Evaluation der Lehre" nimmt in der hochschulpolitischen Diskussion Deutschlands seit einigen Jahren einen hohen Stellenwert ein. Damit stellt sich zwangsläufig die Frage nach geeigneten Instrumenten und Indikatoren zur Messung der Effizienz und Qualität in den verschiedenen Hochschulbereichen. Diese befinden sich zum großen Teil allerdings noch in der Entwicklungs- oder Erprobungsphase. Bei Vorhaben der Lehrevaluation konzentrieren sich Interesse und Auseinandersetzung vorwiegend auf das Instrument der "studentischen Veranstaltungskritik". Die Beurteilung von Lehrenden durch Studierende ist jedoch umstritten.

Deshalb ist es nützlich und einer Versachlichung der Diskussion dienlich, sich anhand empirischer Studien darüber zu vergewissern, welche Zweifel gegenüber studentischen Veranstaltungskritiken begründet sind und welche nicht. Eine derartige Aufarbeitung kann nicht alle Fragen endgültig beantworten. Sie vermag aber, entsprechend dem Stand der empirischen Forschung, zu konstatieren, inwieweit studentische Urteile über Lehrveranstaltungen und Dozenten als gültige, zuverlässige und praktikable Grundlagen der Lehrevaluation herangezogen werden können.

Der vorliegende Bericht ist ein überarbeiteter Teil meiner Diplom-Arbeit gleichen Titels, die im Mai 1994 an der Universität Bonn am Institut für Psychologie vorgelegt wurde. Die Gelegenheit zur Überarbeitung erhielt ich im Rahmen der Arbeitsgruppe Hochschulforschung an der Universität Konstanz sowie durch Förderung des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie.

Eingangs des Berichtes skizziere ich die verschiedenen Ebenen der Evaluation an Hochschulen, die feststellbaren Konfliktlinien bei Vorhaben und Maßnahmen zur Lehrevaluation sowie - im Überblick - die Einwände gegen die Lehrevaluation, insbesondere gegen studentische Veranstaltungskritiken. Der Hauptteil der Arbeit gliedert sich nach den geläufigen und häufig vorgebrachten Vorbehalten gegenüber studentischen Veranstaltungsbeurteilungen; insgesamt wird auf zwölf solcher Argumente eingegangen. Dazu werden jeweils die einzelnen empirischen Untersuchungen knapp referiert und in ihrer Aussagefähigkeit kritisch gewürdigt, wobei einzelne Überschneidungen und manche methodischen Ausführungen unvermeidlich sind. Zu jedem Argument werden "Schlußfolgerungen" hinsichtlich ihrer Stichhaltigkeit gezogen. Zum Abschluß des Berichtes werden diese Folgerungen zusammengefaßt und bilanziert.

Die herangezogenen Untersuchungen stammen fast durchweg aus den USA, in denen studentische Veranstaltungskritik und Lehrevaluation in ihren verschiedenen Formen längst gebräuchlich sind. In letzter Zeit sind verschiedentlich aufschlußreiche deutsche Arbeiten zu diesem Thema erschienen (z.B. Kromrey 1993 und 1994; Daniel 1995; Rindermann 1995; Mohler 1995); diese werden in diesen Bericht nicht mehr eingearbeitet. Alles in allem bestätigen sie aber die amerikanischen Befunde zur Gültigkeit, Verläßlichkeit und Nützlichkeit studentischer Befragungen zur Lehre - vorausgesetzt, sie werden mit angemessenen Instrumenten durchgeführt, methodisch sinnvoll ausgewertet und in der Praxis der Hochschulen und Fachbereiche kompetent gehandhabt.

Nach der Durchsicht von fast 100 empirischen Studien zur Validität studentischer Veranstaltungskritik kann die Schlußfolgerung gezogen werden, daß der studentischen Urteilsfähigkeit durchaus zu vertrauen ist. Entsprechende Fragebögen lassen sich so konstruieren, daß sie Gütekriterien vergleichbarer Tests erreichen bzw. übertreffen. Studentische Stellungnahmen zu Lehrveranstaltungen und Lehrenden ermöglichen dann gültige und zutreffende Feststellungen zur Lehrqualität. Ein wichtiger Punkt dabei ist die multidimensionale Gestaltung von Fragebögen zur Lehrevaluation. Zu diesen Problemen der Erhebungsinstrumente wird ein eigener Bericht vorgelegt, in dem Instrumente studentischer Veranstaltungskritiken an deutschen Hochschulen präsentiert und verglichen werden (als Heft 14 dieser Reihe ""Instrumente studentischer Veranstaltungskritik - Gestaltungsprinzipien und Beispiele").

Sollen studentische Veranstaltungskritiken als Teil hochschulischer Evaluationsmaßnahmen eine Verbesserung der Lehre zur Folge haben, müssen Rahmenbedingungen geschaffen werden, die einer zielgerichteten und akzeptierten Verwendung der erhobenen Daten dienlich sind. Dreh- und Angelpunkt ist dabei die Stärkung der Wichtigkeit der Lehre an den Hochschulen. Bei der Einführung eines "sensiblen" Instruments wie der studentischen Veranstaltungskritik kommt deshalb der Einführungsstrategie eine besondere Bedeutung zu. Die Ansätze und Wege der Lehrevaluation in Deutschland, insbesondere in den einzelnen Bundesländern, sollen daher ebenfalls in einem eigenen, dritten Bericht dokumentiert und dargestellt werden (als Heft 15 dieser Reihe "Studienreform durch Lehrevaluation? Ansätze, Projekte und Verwendung").

Mit der Erarbeitung und Vorlage dieser drei Berichte wird beabsichtigt, einen Überblick zu wichtigen und strittigen Bereichen der "Lehrevaluation" zu liefern und damit eine Informationsgrundlage herzustellen. Dadurch soll der Austausch über Ansätze und Verfahren der Lehrevaluation und studentischer Lehrbeurteilung in einem breiteren Rahmen ermöglicht und versachlicht werden. Darüberhinaus besteht die Hoffnung, die Bemühungen um eine Verbesserung der Lehre vieler Dozenten und Fachbereiche anzuregen und zu unterstützen.

Natalija el Hage, Konstanz, November 1995

Inhalt

1 Lehrevaluation und Veranstaltungskritik im Meinungsstreit

1.1 Ebenen und Bereiche der Evaluation an Hochschulen

Um das Thema "Evaluation der Lehre" systematisch betrachten und diskutieren zu können, werden drei Ebenen unterschieden, die aufgrund der Organisation der Hochschulen in Deutschland zu beachten und zu behandeln sind: (I) die Hochschule in ihrer Gesamtheit; (II) die einzelnen Fachbereiche mit ihrem Kollegium; (III) die einzelnen Lehrveranstaltungen oder ihr Dozent bzw. ihre Dozentin.

Ebene I: Hochschule in ihrer Gesamtheit

Wird die Hochschule in ihrer Gesamtheit als Institution evaluiert, so sind die Leistungen in Forschung, Administration und Lehre getrennt zu erfassen. An dieser Stelle soll nicht weiter auf die Möglichkeiten der Bewertung von Forschungsleistungen eingegangen werden.

Die Administration einer Hochschule ist der Bereich, der am ehesten mit Mitteln und Methoden bewertet werden kann, die auch bei der Evaluation anderer Verwaltungen oder Unternehmungen (bzw. ihrer Abteilungen) angewandt werden. In Bezug auf Hochschulen würde es sich dabei um Fragen des effizienten Managements der Studierendenverwaltung, des Beschaffungswesens und der Mittelverwendung oder des Personaleinsatzes handeln.

Der Bewertung der Lehre auf Hochschulebene stehen eine ganze Reihe von Schwierigkeiten entgegen. Es erscheint angebracht, auf dieser Ebene die Qualität der Lehre als einen Teil der Studienbedingungen einer Hochschule insgesamt zu betrachten, wie Öffnungszeiten von Lehrvorrichtungen, Ausstattung mit studentischen Arbeitsplätzen, überfachliche Lehr- und Qualifikationsangebote (z.B. Studium generale, Sprachkurse, EDV-Anwendung).

Fraglich bleibt jedoch, inwieweit Bewertungen zur Lehrleistung, wie übrigens auch bei der Forschungsleistung, über eine gesamte Hochschule aussagekräftig und von praktischem Nutzen sein können. Die Untersuchung von Giesen und Jansen (1983) kommt zu dem Ergebnis, daß "... der Varianzanteil, der auf das spezifische Erleben der Fachgruppen zurückgeführt werden kann, stets größer ist als der universitätsspezifische Anteil" (S. 226).

Diese Ergebnisse geben einen wichtigen Hinweis für Evaluationsmaßnahmen zur Lehre: Auf der Hochschulebene sind nur ganz allgemeine - vor allem fachbereichsunspezifische - Studienbedingungen, d.h. Lehr- und Lernbedingungen, zu erfassen. Es erscheint wenig sinnvoll und zudem methodisch außerordentlich bedenklich, Durchschnittswerte über Fachbereiche hinaus für eine Hochschule insgesamt zu bilden, etwa in Form eines vergleichenden Rankings.

Ebene II: Fachbereiche als Organisationseinheit

Bei der Evaluierung eines Fachbereichs zeigt sich die administrativ-organisatorische Seite mit der Lehre eng verwoben. Der Lehrkörper eines Fachbereichs ist sowohl für allgemeine Faktoren der Studienbedingungen, wie z.B. Öffnungszeiten von Institutsbibliothek und Institutssektretariat, als auch für organisatorische Aspekte der Lehre, z.B. das Angebot genügender Pflichtveranstaltungen oder die Vermeidung zeitlicher Überschneidungen von Kursen, verantwortlich. Schwieriger ist zu sagen, inwieweit ein Fachbereich auch für die Lehrqualität der Veranstaltungen zuständig ist, da gegenwärtig vor allem die jeweiligen Lehrstuhlinhaber und -inhaberinnen für ihre Kurse zuständig sind.

Möglich und sinnvoll sind für den Fachbereich als Organisationseinheit somit Daten, die sich auf die Gesamtsituation und Auswirkungen der Lehre beziehen, wie z.B. Studienkonzepte und Schwerpunkte, Termineinhaltung, Prüfungsresultate und Semesterzahlen, Absolventen-/ Absolventinnenzahlen oder der Berufseinstieg und Berufsverlauf der Absolventen und Absolventinnen.

Ebene III: Lehrveranstaltungen und Lehrende

Für die Lehrveranstaltungen zeichnet vor allem der jeweilige Dozent/ die jeweilige Dozentin verantwortlich. Studentische Veranstaltungskritik bezieht sich gewöhnlich fast ausschließlich auf diese Ebene einzelner Lehrveranstaltungen und Lehrender. Es wäre jedoch angebracht, auch fachspezifische Studienbedingungen und den organisatorischen Aspekt der Lehre mit einem ähnlichen Instrument zu erfassen (d.h. zur Evaluation der Ebene des Fachbereichs die Erfahrungen und Urteile der Studierenden heranzuziehen und zu erfragen).

Lehrveranstaltungen und Lehrende durch Studierende beurteilen zu lassen, kann in ganz unterschiedlicher Weise erfolgen: z.B. als unmittelbare Rückmeldung auf eine einzelne Veranstaltung, als Beurteilung einer ganzen Veranstaltungsreihe, eines Semesters oder als Bewertung der didaktischen Qualitäten der Lehrenden insgesamt.

Inwieweit dabei schriftliche Fragebogen mit vorgegebenen Antwortmöglichkeiten einsetzbar sind, ist die entscheidende Frage. Das hängt freilich davon ab, ob die Studierenden in der Lage sind, zutreffend ihren Erfahrungen mitzuteilen und ihre Urteile abzugeben. Gerade diese Fähigkeit wird vielfach angezweifelt. Deshalb wird bestritten, daß eine Befragung in schriftlicher Form gültige und verläßliche Resultate über die Lehrleistung der Lehrenden liefern könne. Mit diesen Zweifeln und Vorbehalten setzt sich vorliegender Bericht auseinander.

Da die drei Ebenen -Hochschule, Fachbereich, Lehrveranstaltung- weitgehend unabhängig sind (es ist z.B. ein ausgezeichneter Kurs an einer Hochschule mit schlechten Studienbedingungen und chaotischer Organisation des Fachbereichs denkbar), ist die sorgfältige Unterscheidung der verschiedenen Ebenen unabdingbar. Auch die Westdeutsche Rektorenkonferenz (WRK) sprach sich bereits 1986 für "vergleichende Leistungsbeur-

teilungen zwischen gleichen Fächern", also dem "intradisziplinären Vergleich" aus. Sie warnte zurecht davor, die Hochschule in ihrer Ganzheit gleichzusetzen mit der Summe der Leistungen einzelner Fächer (vgl. WRK 1986). Dies gilt für die Forschung wie für die Lehre.

1.2 Konfliktlinien in der Evaluation von Lehrveranstaltungen und Dozenten

In der politischen und wissenschaftlichen Debatte um die Lehrevaluation der letzten Jahre sind verschiedene Konfliktlinien zu erkennen, auf die sich Befürwortung und Widerstand gegen mögliche Evalutions-Maßnahmen, wie zum Beispiel die Erstattung von Lehrberichten oder die Befragung der Studierenden, immer wieder beziehen. Die grundsätzlichen Konfliktlinien in diesem Bereich betreffen Fragen des Rechts, die Art und Verwendung von Lehrevaluationen sowie die studentische Kompetenz und Urteilsfähigkeit.

Lehrfreiheit und Datenschutz

Ein gewichtiger Konfliktpunkt ist der rechtliche Rahmen studentischer Veranstaltungskritiken und ihr möglicher öffentlicher Gebrauch. Häufig scheinen aber die Auseinandersetzungen um rechtliche Vorschriften nur die Verlagerung eines anderen Streitpunkts darzustellen: " ... anders als z.B. in den USA, wo Transparenz und Rechenschaftablegung auch in der Hochschullehre gesellschaftlich eingeklagt werden, kann sich bei uns Hochschullehre gegen gesellschaftliche Einsicht- und Einflußnahme weitgehend abschotten" (Heger, 1991, S. 52).

Noch deutlicher wird in einer Zwischenbilanz zur Lehrqualität des Ministeriums für Bildung und Wissenschaft ausgeführt: "Umgekehrt drängt sich der Verdacht auf, daß die Freiheit der Wissenschaft, die die Freiheit der Lehre umschließt, als Schutzargument mißbraucht wird, wenn sie zur systematischen und rechtlich abgesicherten Abschottung gegen jede Form der Kritik aufgefaßt wird. Dies ist um so erstaunlicher, als das andere große Standbein der Wissenschaft, die Forschung, öffentlich zugänglich verhandelt und auch kritisiert wird" (Bundesministerium für Bildung und Wissenschaft, 1993, S. 3).

Bei der Auseinandersetzung rechtlicher Art werden vor allem zwei Dimensionen thematisiert: zum einen Art. 5. Abs. 3 GG, der die Lehrfreiheit garantiert, und zum anderen Bestimmungen des Datenschutzes. Es wird diskutiert, ob es möglich ist "Dritten", z.B. Fachschaften oder Hochschulgruppen, zu verbieten, studentische Veranstaltungskritik durchzuführen, da die "Freiheit der Meinungsäußerung" jeden zur Veranstaltungskritik berechtige. Weiter ist zur Zeit umstritten, ob die Regelungen zur "Dienstaufsicht" von Hochschulleitung oder Wissensschaftsministerium sich derart ändern ließen, daß von dieser Seite Lehrbewertung organisiert und in Entscheidungen miteinbezogen werden könnten, und ob dies mit dem Art. 5 Abs. 3 GG vereinbar sei.

Grottian (vgl. 1992, S. 30 ff.) vertritt die Ansicht, daß die Lehre eine "wenigstens universitätsöffentliche Angelegenheit" sei. Die Bewertung der Lehre vergleicht er mit der der Forschung: "In der wissenschaftlichen Auseinandersetzung ertragen wir Kritik, bis ins Persönliche gehende Verrisse, mit einiger Gelassenheit; dagegen wird in der Lehre öffentliche Kritik zu einer Verletzung der persönlichen Intimsphäre hochstilisiert" (Grottian, 1992, S. 31). Er geht davon aus, daß studentische Veranstaltungskritik zur Dienstpflicht erklärt werden kann und beides mit dem Art. 5 Abs. 3 GG vereinbar ist. Gralki nennt Grottians Ansatz provokativ und kritisiert, daß im Vergleich zur wissenschaftlichen Kritik studentische Lehrbewertung anonym sei (vgl. Gralki, 1992).

Eine häufiger diskutierte Frage ist, ob von Studierenden oder der Verwaltung erhobene Bewertungsdaten zur Lehre veröffentlicht werden dürfen. Grottian vertritt die Auffassung: "Studierende haben als erste Adressaten der Lehre ein demokratisches Recht, die Befragungsergebnisse verantwortlich diskutieren zu können" (Grottian, 1992, S. 31). Erörtert wird, ob dies mit dem aus dem Artikel 2 Abs. 1GG abgeleiteten "informationellem Selbstbestimmungsrecht" vereinbar sei. Grottian dazu: "Zwar liegt in der Datenerhebung und namentlichen Veröffentlichung ein Eingriff in das Recht auf informationelle Selbstbestimmung, doch unterliegt das Recht auf informationelle Selbstbestimmung seinerseits Einschränkungen zugunsten der Allgemeinheit." (Grottian, 1992, S. 32) Informationelle Selbstbestimmung bleibt sicherlich im "unantastbaren Intimbereich" zu wahren, die Lehrtätigkeit eines Hochschullehrers ist jedoch allgemein zugänglich und prinzipiell eine "öffentliche Veranstaltung".

Detmer (1991) erläutert ein Urteil des Oberverwaltungsgerichts Berlin (OVG PV Bln 9.91 vom 18. Juli 1991), daß Datenerhebung und -veröffentlichungen mitbestimmungspflichtig im Sinne des Personalrechts und somit nur mit Einwilligung der Betroffen möglich sind. Der frühere hessische Datenschutzbeauftragte Spiros Simitis vertritt eine andere Sicht: Aufbau und Stil von Veranstaltungen seien keine "schützenswerten persönlichen Daten". Wird studentische Veranstaltungskritik allerdings von der Hochschulleitung und u.U. gegen den Willen der Lehrenden organisiert, würde dies in den Grundsatz der Lehrfreiheit eingreifen (vgl. SPIEGEL-SPEZIAL, 1993, S. 124 ff.). Somit würde weniger der Datenschutz als die "garantierte Lehrfreiheit" einer verpflichtenden und öffentlichen Veranstaltungskritik im Wege stehen.

Endgültig kann diese Rechtsunsicherheit vermutlich nur vom Bundesverfassungsgericht entschieden werden. Wie dies entschieden wird, ist schwer einzuschätzen, da verschiedene Ansichten auch zu der Rolle der Studierenden an den Hochschulen mit hineinspielen. So zitiert z.B. Krieger aus einem Verwaltungsgerichtsbeschluß (VG FK (Bln) - C - 6.91) des Verwaltungsgerichts Berlin vom 19. April 1991: "Auch wenn man dieser Rechtsprechung folgt, so setzt §85 Abs. 1 Nr. 13 Pers VG jedenfalls voraus, daß die technisierte Verhaltens- oder Leistungsüberwachung auf Daten beruht, die hierzu überhaupt geeignet sind. Von vornherein ungeeignet dürften neben Statusdaten (Name, Anschrift, Geburtstag usw.) Bewertungen durch Personen sein, die zu einer objektiven Beurteilung generell nicht in der Lage sind. Die Einschätzung von Lehrveranstaltungen durch Studenten, wie sie bei der hier streitigen Fragebogenaktion vorgesehen ist, dürfte zu dieser Kategorie ungeeigneter Daten gehören" (Krieger 1992, S. 59). Dieser Argu-

mentation schloß sich übrigens die zweite Instanz, das Berliner Oberverwaltungsgericht, nicht an. Es befand, daß es nur darauf ankäme, daß dieser subjektiven Leistungsbewertung Bedeutung für die Lehre beigemessen würde (vgl. Detmer, 1991).

Wie die rechtlichen Debatten ausgehen und von Gerichten entschieden werden, bleibt gegenwärtig weitgehend offen - zu kontrovers fallen die Stellungnahmen der Rechtsexperten aus, zu unterschiedlich urteilen verschiedene gerichtliche Instanzen. Jedoch ist es für den juristischen Streit keineswegs unwichtig, ob die "subjektive Leistungsbewertung" verschiedener Studierender eine "objektive Beurteilung" von Lehre und Lehrenden darstellen kann.

Art und Verwendung von Lehrevaluationen

Bei der Frage nach Art und Verwendung der Lehrevaluation gibt es vier große Konfliktfelder, die in Diskussionen immer wieder heftig thematisiert werden.

Dies ist erstens die grundsätzliche Frage nach dem Nutzen solcher Verfahren. Sehen manche Befürworter und Befürworterinnen in Maßnahmen der Lehrbewertung das Allheilmittel für die meisten Schwierigkeiten an den Hochschulen, so bezweifeln Gegner und Gegnerinnen deren prinzipielle Brauchbarkeit oder halten den Aufwand im Vergleich zum Nutzen für zu hoch. Als ein Teilbereich dieser Auseinandersetzung kann die Frage angesehen werden, ob es an den deutschen Hochschulen nicht gravierendere Probleme gäbe, als die Auseinandersetzung um die Lehrqualität. Damit wird auch die öffentliche Aufmerksamkeit kritisiert, die ihren Fokus in den letzten Jahren ganz besonders auf diesen Aspekt gelenkt hatte.

Der nächste Punkt der Auseinandersetzung in diesem Bereich, der gerade von übergeordneten Gremien wie den Landesrektorenkonferenzen oder der HRK häufiger aufgegriffen wird, ist die Problematik einer möglichen staatlichen Einflußnahme, und inwieweit diese die Autonomie der Hochschule bzw. die Freiheit der Lehrenden beeinträchtigen könnte. Dabei reichen die Ansichten von der Befürwortung einer Lehrevaluation, die zwar von den Hochschulen durchgeführt, von staatlicher Seite jedoch begleitet wird (wie dies z.B. in den Niederlanden der Fall ist), bis hin zu der immer noch existierenden Ansicht, daß Lehre ein Privatissimum des entsprechenden Professors ist und sich nicht einmal die jeweilige Hochschule einmischen darf.

Außer diesen beiden grundlegenden Abwägungsbereichen gibt es zudem ganz praktische Überlegungen und Ängste. Zum einen beziehen sie sich auf die mögliche Veröffentlichung von Evaluationsergebnissen. In den Hochschulsenaten ist ein Ringen und Feilschen zu beobachten. Werden Evaluationsdaten veröffentlicht? Sollen sich alle Lehrenden damit einverstanden erklären? Wie detailliert soll/ kann/ darf veröffentlicht werden? Nur kumulierte Ergebnisse? Oder mit Veranstaltungstitel? Gar mit dem eigenen Namen? In diesem Händel geht es nicht allein darum, professorale Privilegien zu verteidigen. Deutlich wird dabei zugleich die Unsicherheit, in einem Bereich bewertet zu werden, den man weder gelernt hat noch bislang dazu Erfahrungen sammeln konnte.

Zum anderen werden die möglichen Konsequenzen von Lehrbewertungen debattiert, wobei alle bisher genannten rechtlichen und methodischen Argumente einfließen. Die Streitpunkte reichen von der Frage, ob sich solche Bewertungen überhaupt auf Berufungen, Bleibeverhandlungen oder gar das Einkommen auswirken sollten, über die Frage einer darüber möglicherweise verstärkten staatlichen Steuerung (Mittelvergabe, Karriere, Gratifikationen) bis hin zu Fragen nach Art und Ausmaß möglicher Konsequenzen.

Kompetenz und Urteilsfähigkeit der Studierenden

Der Meinungsstreit, ob studentische Veranstaltungskritik als subjektive Meinungsbefragung (auch im Sinne einer Verbraucherbefragung), wie dies z.B. Anneliese Monika Grüger (1982) darstellt, oder als objektives Beurteilungsinstrument angesehen werden kann, bezieht sich zwar vordergründig auf methodische Aspekte der Art, des Einsatzes und der Auswertung solcher Instrumente. Dahinter verbergen sich aber prinzipielle Bedenken an der Berechtigung solcher Verfahren und an der Fähigkeit der Studierenden zu kompetenten Urteilen.

"Die Argumentationslinie beginnt beim grundsätzlichen Zweifel an der Kompetenz der Studierenden, Lehrleistung überhaupt bewerten zu können" (Preißer, 1991, S. 74). Der Frankfurter Psychologe Süllwold (1992) vertritt die Auffassung, daß Unterschiede in der Beurteilung verschiedener Hochschullehrer und -lehrerinnen einzig und allein auf unterschiedliches Bewertungsverhalten der Studierenden zurückzuführen ist, und nichts über deren tatsächliche Lehrqualität aussagen; er folgert daher: "Erhebungen über die Beurteilung von Hochschullehrern durch Studierende sind in fundamentaler Weise fehlerhaft, wenn die einzelnen Hochschullehrer durch verschiedene Studentengruppen beurteilt werden" (Süllwold, 1992, S. 34). Damit schließt Süllwold letztlich aus, daß objektive Instrumente, wie sie in vielen psychologischen Bereichen eingesetzt werden, für die studentische Veranstaltungskritik möglich sind.

1.3 Einwände gegen die Lehrevaluation durch Studierende

Rechtliche Probleme von Lehrfreiheit und Datenschutz, die mögliche Verwendung studentischer Lehrevaluation als Element der Studienreform oder Hochschulentwicklung werden in diesem Bericht nicht vertieft. Wie die Antworten dazu ausfallen, dürfte aber nicht zuletzt davon abhängen, inwieweit studentische Urteile zur Lehre ein gültiger und zuverlässiger Maßstab der Lehrqualität sein können. Die Einwände gegen die Validität studentischer Veranstaltungskritiken sind daher auf ihre Berechtigung hin empirisch zu prüfen.

Die in diesem Bericht dargestellten Ergebnisse zum Thema der Lehrevaluation durch Studierende sind Studien entnommen, die zum großen Teil aus den Vereinigten Staaten stammen. Daß gerade in den USA so viele Untersuchungen durchgeführt wurden, liegt an der langen Nutzungsdauer (bereits in den zwanziger Jahren wurde studentische Veranstaltungskritik eingeführt) und dem inzwischen weit verbreiteten Einsatz.

In Deutschland gibt es einige wenige Veröffentlichungen aus den siebziger Jahren (z.B. Keil/Piontkowski 1973; Ipsen/Portele 1976), die sich mit "Prozessen des Hochschulunterrichts" und der "Organisation von Lehre" befaßten und auf studentische Befragungen stützten, allerdings ohne als Arbeiten zur "Lehrevaluation" präsentiert und verstanden zu werden. In den letzten Jahren wurde zwar manches zum Thema "Evaluation der Lehre" veröffentlicht, mit Ausnahme weniger Studien konnten jedoch keine deutschen empirischen Untersuchungen zur Validität studentischer Veranstaltungskritik herangezogen werden. Die Instrumente neuerer Untersuchungen werden in einem eigenen Bericht behandelt, der sich mit Fragebogen und Tests studentischer Lehrevaluation befaßt (Heft 15).

Dies wirft die Frage der Übertragbarkeit auf: Inwieweit können Ergebnisse, die in großer Fülle in den USA gesammelt wurden, auf hiesige Verhältnisse angewandt werden? Bekanntlich unterscheidet sich das nordamerikanische Hochschulsystem in verschiedenen Punkten deutlich vom deutschen System. Dies wirkt sich auf die untersuchten Themen aus, von denen einige in Deutschland nicht die gleiche Rolle spielen dürften wie in den Staaten (z.B. Noten), während zu anderen (z.B. Auswirkung der "Vermassung" von Kursen auf die Lehrbewertung) Untersuchungen fehlen.

Allerdings scheinen die typischen Einwände gegen Lehrbewertungen durch Studierende von seiten der Dozierenden durchaus vergleichbar zu sein. So nennt der deutsche Wissenschaftler Süllwold (1992) als mögliche externe Einflußgrößen, welche die Lehrbewertung verzerren: Fachsemester, geschlechtsspezifische Zusammenhänge, voruniversitäre Bildung, Intelligenzniveau, Fleiß, Leistungsmotivation und ideologische Bedürfnisse. Dies deckt sich weitgehend mit den von Aleamoni (1987) aufgelisteten häufigsten Einwänden gegen studentische Veranstaltungskritik:

1. Studierende können die Lehre nicht konsistent beurteilen, da sie folgende Eigenschaften aufweisen: "Unreife, fehlende Erfahrung und Launenhaftigkeit".
2. Nur Kollegen/ Kolleginnen können die Lehre entsprechend kompetent bewerten.
3. Die meisten Bewertungsbögen sind nichts weiter als ein "Beliebtheitswettbewerb".
4. Studierende können die Lehre nicht bewerten, bis sie nicht genügend Distanz zum Kurs bzw. zur Lehranstalt besitzen.
5. Bewertungsbögen sind unreliabel und nicht valide.
6. Externe Variablen (z.B. Kursgröße) beeinflussen die Bewertung.
7. Noten (erwartete oder erhaltene) beeinflussen die Bewertung.

An dieser Aufstellung von Argumenten gegen studentische Veranstaltungskritik orientiert sich der vorliegende Bericht. Insgesamt wird auf zwölf Argumente eingegangen, die von unterschiedlicher Reichweite sind (vgl. Übersicht 1).

Grundsätzliche Einwände gegen die Gültigkeit und damit Verwendbarkeit studentischer Veranstaltungskritik heben in erster Linie die mangelnde Sachkenntnis und Unerfahrenheit der Studierenden hervor. Ihre Urteile würden deshalb nicht den inhaltlichen Wert einer Veranstaltung oder die tatsächliche Lehrleistung erfassen, sondern vielmehr deren Unterhaltungswert und die bloße "Beliebtheit" von Lehrenden.

Übersicht 1

Argumente gegen die Validität der Bewertung von Lehrleistungen durch Studierende

1 Mangelnde Sachkenntnis der Studierenden
Studierenden als Lernenden fehle die Sachkenntnis, um die Kompetenz der Lehrenden und die Lehrqualität überhaupt angemessen einschätzen zu können.

2 Unerfahrenheit und Unreife der Studierenden
Studierenden fehle es an Erfahrung und zum Teil am notwendigen Ernst. Diese Unreife führe zu beliebigen und beeinflußbaren Urteilen.

3 Der Unterhaltungswert, nicht der Inhaltswert steht für Studierende im Vordergrund
Studierende würden nur oder überwiegend einen spannenden und witzigen Vortragsstil beachten, dem wichtigen Inhaltswert der Lehre schenken sie kaum Aufmerksamkeit.

4 Die Beliebtheit, nicht die Lehrleistung wird von Studierenden beurteilt
Studierende würden vor allem emotional die Beliebtheit eines Lehrenden bewerten und nicht sachlich dessen Lehreffektivität und -qualität.

5 Nicht der Lehrende, sondern der Kurs wird bewertet
Lehrende würden in verschiedenen Kursen unterschiedlich bewertet, so daß ihre "Lehrqualität" vom Kurs abhinge, den sie abhalten. Deshalb sei eine vergleichende Lehrbewertung unzulässig.

6 Arten und Formen von Kursen verzerren die Lehrbeurteilung
Art, Größe, Anforderungen, Aufwand und Besucher verzerren die Bewertung der Veranstaltungsqualität.

7 Der Status der Lehrenden verzerrt die Lehrbewertung
Studierende lassen sich vom Status und Ruf eines/einer Lehrenden beeindrucken, so daß sie weniger die konkrete Lehrleistung beurteilen.

8 Persönliche Eigenschaften von Lehrenden verzerren die Bewertung
Studierende bewerten nicht die Leistungen in der Lehre, sondern statt dessen die Herzlichkeit, Freundlichkeit und auch den Humor eines Lehrenden.

9 Persönliche Eigenschaften der Studierenden verzerren die Bewertung
Studierende mit unterschiedlich ausgeprägten Eigenschaften (wie Motivation oder Intelligenz) kommen zu anderen Beurteilungen; insofern hängt die "Lehrqualität" weniger vom Lehrenden als von den Studierenden (und ihren Eigenschaften) ab.

10 Übereinstimmungen zwischen Lehrenden und Studierenden verzerren die Bewertung
Studierende, die in ihren Einstellungen und Haltungen mit denen der Lehrenden übereinstimmen, bewerten deren Veranstaltungen besser.

11 Das Geschlecht beeinflußt die Bewertung
Sowohl die Lehrenden, die bewertet werden, als auch die Studierenden, die ihre Urteile abgeben, richten sich nach dem Geschlecht des Gegenüber. Deshalb kommt es zu unterschiedlichen Beurteilungen.

12 Noten und Strenge beeinflussen die Bewertung
Bessere oder schlechtere Noten (erwartete oder erhaltene) ebenso wie der Eindruck der Milde oder Strenge des Lehrenden sind ausschlaggebend für die Lehrbewertung durch Studierende.

Andere Argumente beziehen sich darauf, daß die Bewertung eines Lehrenden vom Kurs, den er abhält, abhängig ist. Dabei werden sehr verschiedene Aspekte angesprochen: Kursart und Kursgröße, Kursaufwand und Kurslevel, Pflicht- oder Wahlkurs, Kurs für Hauptfach- oder Nebenfachstudierende. Damit werden vor allem Fragen der Vergleichbarkeit der Bewertung der Lehrenden thematisiert.

Schließlich werden Eigenschaften der Lehrenden wie der Studierenden angeführt, die in Verdacht stehen, die Bewertungen der Lehre zu verzerren. Darunter spielen Eigenschaften wie Strenge der Lehrenden, insbesondere in der Notengebung, oder die Leistungsmotivation der Studierenden eine Rolle. Aber auch das Geschlecht und der Status des Lehrenden (Professor, Assistent, Tutor) würden die Beurteilung des Lehrangebotes beeinflussen, da sich die Studierenden in der Bewertung stärker daran als an der zu erfassenden Lehrleistung und Lehrqualität orientierten.

Inwieweit empirische Untersuchungen derartige Vorbehalte gegenüber studentischer Veranstaltungskritik stützen, einschränken oder widerlegen, das soll im Folgenden geprüft werden. Es geht um die Qualität des "Produktes Lehrleistung" und deren Beurteilung durch Studierende.

2 Argumente und Forschungsergebnisse zur Validität studentischer Urteile

An Bewertungsinstrumente wird die berechtigte Forderung gestellt, daß sie objektiv, valide und reliabel zu sein haben. **Objektivität** gilt in den Sozialwissenschaften als gegeben, wenn Erfassung und Auswertung der Daten standardisiert sind. Sicherlich handelt es sich bei den Stellungnahmen von Studierenden zur Lehre oder zu Dozenten um persönlich-subjektive Aussagen und Wertungen. Diese lassen sich aber durchaus mit "objektiven" Instrumentarien erfassen. Insofern die studentischen Aussagen standardisiert erhoben und ausgewertet wurden, entsprechen sie der Forderung nach "Objektivität".

Als **Validität** wird die Gültigkeit von Instrumenten und Skalen in bezug auf einen bestimmten Sachverhalt bezeichnet. Mit anderen Worten: Die Frage nach der Validität bezieht sich darauf, ob Instrumente und deren Teile tatsächlich das messen, was sie zu messen vorgeben. Im Fall der studentischen Veranstaltungskritik zielt die Messung auf "die Lehrleistung der Dozenten" bzw. "die Lehrqualität einer Lehrveranstaltung aus studentischer Sicht". Dabei kann es aber zuweilen strittig sein, was Elemente der Lehrleistung sind und mit welchem Gewicht sie in einer Bilanz zu berücksichtigen wären.

Reliabilität meint die Meßgenauigkeit und Zuverlässigkeit der verwendeten Instrumente. Die Diskussion um Reliabilität wird bei der Veranstaltungskritik von Studierenden eher am Rande geführt, da viele Kritiker und Kritikerinnen von einer nicht gegebenen Validität ausgehen. Deshalb erübrige es sich, der Frage nachzugehen, ob das "Falsche" zudem noch ungenau gemessen wird. Läßt sich aber belegen, daß die Lehrleistung von Dozenten und die Qualität ihrer Veranstaltungen sich durchaus valide durch Studierende bestimmen läßt, dann wird die Frage nach der Meßgenauigkeit wichtig. Dies umso mehr, wenn mit der Erfassung Vergleiche verschiedener Dozenten oder Veranstaltungen beabsichtigt werden.

Reliabilität und Validität hängen zusammen. Denn erst, wenn ein Instrument den Sachverhalt überhaupt zutreffend erfaßt (d.h. valide ist), kann es genaue und zuverlässige Resultate liefern (d.h. reliabel sein). Und erst eine hohe Reliabilität ermöglicht eine gute Validität, d.h. nur ein genaues Instrument kann das gewünschte Kriterium zutreffend vermessen. Die große Auseinandersetzung findet jedoch weniger um die Reliabilität der Instrumente, um deren Meßgenauigkeit, statt, sondern ist weit umfassender. Es wird vor allem bezweifelt, ob es überhaupt möglich ist, die Lehrqualität und Lehrleistung mit Fragebogen der studentischen Veranstaltungskritik zu erfassen.

Die Kritik an der Validität der Erfassung studentischer Aussagen über die Qualität von Veranstaltungen und Lehrenden umfaßt eine Reihe von Argumenten und eine Vielzahl von angeführten Aspekten. Bei der Darstellung der empirischen Studien dazu ist es unvermeidlich, methodische Fachtermini zu verwenden, trotz aller Bemühungen, sie auf ein Mindestmaß zu beschränken. Zitate aus den amerikanischen Studien wurden der Lesbarkeit halber ins Deutsche übersetzt.

2.1 Erstes Argument: Mangelnde Sachkenntnis der Studierenden

"Mangelnde Sachkenntnis" ist eines der häufigsten Gegenargumente auf die Forderung nach studentischer Veranstaltungskritik. Studierende seien nicht in der Lage, die Lehre zu bewerten, weil sie keine Beurteilungsgrundlagen für die Unterscheidung zwischen guter und schlechter Lehre hätten. Sie könnten frühestens nach Beendigung ihres Studiums einschätzen, was sie gelernt haben und welchen Stellenwert dies hat. Daraus wird im Gegenzug abgeleitet, daß nur Kollegen/ Kolleginnen die Lehre sachgerecht beurteilen können.

Aus diesem Argument läßt sich ableiten, daß zwischen der Selbstbewertung der Lehrenden oder von Kollegen und Kolleginnen und der studentischen Veranstaltungskritik große Differenzen bestehen müßten. Dies wäre auch darauf zurückzuführen, daß zur Bewertung von Lehrveranstaltungen die Lehrenden und Studierenden jeweils ganz andersartige Kriterien heranziehen (vgl. dazu auch Abschnitt 3.2).

Bei dem Vorbehalt der mangelnden Sachkenntnis der Studierenden kann es bei der empirischen Prüfung nicht darum gehen, deren Vorhandensein oder Fehlen aufzuweisen. Vielmehr ist noch zu prüfen, ob und wie sie die Lehrbewertung möglicherweise beeinflußt. Dabei ist vor allem zu klären, ob tatsächlich eine Diskrepanz in der Lehrbewertung nachzuweisen ist.

2.1.1 Qualität der Lehre aus der Sicht junger Berufstätiger und Ehemaliger

Als Argument gegen die Bewertung der Lehre durch Studierende wird immer wieder deren fehlende Distanz zum Untersuchungsgegenstand genannt. Deshalb wird des weiteren behauptet, daß erst im Laufe ihrer späteren Berufstätigkeit Studierende einschätzen können, was sie gelernt haben.

Ziel der Untersuchung von Firth (1979) war es, gerade dieses Argument zu überprüfen. Dafür versuchte er, die Langzeitstabilität studentischer Evaluation zu ermitteln. Die Studie bestand aus insgesamt zwei Untersuchungsteilen. Zunächst wurden 216 Lehrende aus 83 Kursen mit einem Fragebogen von allen Studierenden des letzten Studienjahres der Wirtschafts-, Rechts- und Erziehungswissenschaften bewertet. Diese Fachbereiche wurden gewählt, da man dort den größten Zusammenhang zwischen Stoff und späteren Arbeitsanforderungen vermutete. Ein Jahr später wurden die dann bereits Graduierten angeschrieben und um eine Einschätzung der Lehreffektivität gebeten, die sie in Anbetracht ihrer jetzigen Berufserfahrung geben sollten.

Festgestellt werden konnte, daß die Bewertung zwischen dem letzten Jahr an der Universität und dem Jahr danach nur für 13 der 216 Lehrenden signifikant (5%-Niveau) unterschiedlich war. Die Spearman-rank-correlation betrug .903 ($p<0.05$), was eine hohe Konsistenz der Rangreihe belegt.

Mit Hilfe des Fragebogens "Student Assessment of Instruction" (SAI) bewerteten 1.374 Studierende (Rücklaufquote 85%) bei Overall und Marsh (1980) über drei Jahre hinweg Betriebswirtschaftskurse. Ein Jahr nach Abschluß ihres Programms (somit 1-3 Jahre

nach den besuchten Kursen) beantworteten noch einmal 65 Prozent der angeschriebenen Ehemaligen einen Fragebogen zur Lehrbewertung.

Ergebnisse: Die relative Übereinstimmung (d.h. die Übereinstimmung für jedes Item bei den 1.374 Studierenden), betrug r=.83 und die absolute Übereinstimmung (d.h. die Übereinstimmung des Durchschnittswerts bei jedem Item für die beiden Bewertungen), betrug 58 Prozent. Bei der Untersuchung von Untergruppen (ungraduate/graduate, qualitativ/quantitativ) ergaben Median-Korrelationen von Klassendurchschnitten r=.83-.86, Median-Korrelationen für die individuellen Antworten ergaben r=.53-.60.

Daher kommen die Autoren zu der Auffassung: "Die Ergebnisse dieser Studie beinhalten eine große Unterstützung für die vermutete Stabilität studentischer Bewertung ihrer Kurse und Lehrenden und lassen vermuten, daß diese Stabilität nicht systematisch mit Kursniveau oder -inhalt variiert" (Overall und Marsh 1980, S. 324).

Die Langzeitstabilität der Bewertungen von Studierenden und Ehemaligen (jungen Berufstätigen) und ihre Übereinstimmung konnte in beiden aufgeführten Studien nachgewiesen werden. Ob dies daran liegt, daß junge Berufstätige die verschiedenen Lehraspekte in gleichem Licht sehen, oder daran, daß sie sich an ihre früheren Bewertungen erinnern und sie nur wiederholen, ist dabei unerheblich, da die angeführte Kritik hinsichtlich der mangelnden Sachkenntnis der Studierenden davon nicht berührt wird.

2.1.2 Wie sich Lehrende eine Meinung zur Lehrqualität ihrer Kollegen und Kolleginnnen bilden

Mit dem Einwand der mangelnden Sachkenntnis der Studierenden wird zumeist, direkt oder indirekt, vorausgesetzt, daß nur Experten gleichen Ranges eine hinreichend qualifizierte Lehrbewertung vornehmen könnten. Es ist daher aufschlußreich, wie Lehrkollegen zu gegenseitigen Urteilen kommen und worauf diese gründen.

Eine Befragung von Aleamoni (1987) belegt, daß zwischen der Lehrbewertung von Kollegen/Kolleginnen und wissenschaftlicher Produktivität eine Korrelation von r=.07 besteht, wohingegen die Korrelation zwischen der Lehrqualitätseinschätzung von Lehrenden und Studierenden bei .70 liegt. In Gesprächen ermittelte der Autor, daß Lehrende ihre Informationen über die Lehrfähigkeit ihrer Kollegen und Kolleginnen nahezu ausschließlich von Studierenden erhielten. Dieser Punkt ist beachtenswert, da Aleamoni somit herausstellen konnte, daß Lehrende ihre Kollegen und Kolleginnen in der Lehre nicht nach den Kriterium "Ausmaß wissenschaftlicher Produktivität" bewerten.

Ziel der Untersuchung von McGaghie (1975) ist es, die Beziehung zwischen der Lehrbewertung von Studierenden und Lehrenden zu klären. Lehrende nennen folgende Bewertungsdimensionen: Forschungsaktivitäten, Teilnahme an der "Academic community", persönliche und wissenschaftliche Anerkennung, intellektuelle Größe, Beziehung zu Studierenden und Interesse an der Lehre. Es gäbe auch Hinweise, daß der akademische Rang die Bewertung von Kollegen beeinflußt.

Studierende würden dagegen die Lehre nach ganz anderen Kriterien bewerten: Kursorganisation, Präsentation, "Analytic-Synthetic Approach" und Beziehung Kurs - Dozent/ Dozentin. Trotz so unterschiedlicher Beurteilungskriterien waren sich die befragten Gruppen über den besten bzw. schlechtesten Lehrenden erstaunlich einig.

Die beiden Untersuchungen von McGaghie (1975) und Aleamoni (1987) beleuchten einen wenig untersuchten, dennoch außerordentlich wichtigen Aspekt der Lehrbewertung durch Kollegen. Dabei kommt es sehr häufig zu einer Bewertung der Lehre, ohne daß dem eine "peer-evaluation" zugrundeliegt, d.h. die Kollegen und Kolleginnen haben nicht eine oder mehrere Veranstaltungen besucht oder sich zumindest die Kursplanungen angesehen, sondern sie schließen aus verschiedenen anderen Indikatoren, die meist nicht definiert oder als Kriterien festgelegt sind, auf die Qualität der Lehre.

Dieser Aspekt der gegenseitigen Lehrbewertung durch Kollegen spielt in der wissenschaftlichen Forschung kaum eine Rolle. Das ist zu bedauern, da dieser Punkt in der hochschulpolitischen Diskussion häufig thematisiert wird. Es ist sogar zu vermuten, daß je weniger die Evaluation der Lehre in den Hochschulen etabliert und bereits Kriterien und Instrumente gefunden und akzeptiert sind, desto eher die Lehre auf eine Art und Weise in den Fachkollegien beurteilt wird, die kaum kontrollierbar und oft von weit weniger "Sachkenntnis" bestimmt ist, wie sie von den Studierenden verlangt wird.[1]

2.1.3 Meinung von Lehrenden und Studierenden zu allgemeinen Lehraspekten

Es liegen einige Studien vor, die zu ermitteln versuchen, ob Studierende und Lehrende gleiche oder verschiedene Kriterien zur Lehrbewertung benutzen bzw. diese gleich gewichten. Fraglich bleibt jedoch, welchen Einfluß es auf die Veranstaltungskritik selbst hat, wenn die Befragten unterschiedliche Kriterien der Bewertung zugrunde legen.

Mit ihrer Studie wollten Marquez, David und Dorfman (1979) mögliche Differenzen in der Gewichtung von Lehraspekten durch Lehrende und Studierende auffinden. Die Befragten, 40 Studierende und 40 Lehrende, wurden um die Bewertung der Lehreffektivität von Lehrenden gebeten, die ihnen in 100 Profilen vorgestellt worden waren. Nach Studierenden und Lehrenden unterschieden, ergaben sich nur geringe, aber signifikante Unterschiede. Lehrende betonen dabei geringfügig stärker Bereiche, die der übergeordneten Dimension "Inhalt" zugeordnet werden können, während den Studierenden die Bereiche der übergeordneten Dimension "Stil" etwas wichtiger sind.

Diese Unterschiede sind jedoch marginal. Dies kann z.T. an der großen Varianz zwischen den einzelnen Bewertern liegen. Eine Generalisierung der Ergebnisse ist bei dieser Studie aus verschiedenen Gründen schwer möglich.

In einer ähnlichen Studie von Baum und Brown (1980) hatten Studierende in sieben Kursen und 50 Lehrende die Gelegenheit, 100 Punkte über 10 Statements zu verteilen,

[1] Ein Beispiel für diese problematische Handhabung ist der Artikel von Jürgen Kriz "Wie gut sind unsere Universitäten?". (In: Stern, 1993, Bd. 16, S. 171-184), dessen Datenbasis rein auf die subjektive Einschätzung von Professorinnen und Professoren aus den entsprechenden Fachbereichen anderer Hochschulen beruht.

die allgemeine Aspekte der Lehre beinhalteten. Die Verteilung der Punkte zwischen Studierenden und Lehrenden unterscheidet sich im Gesamten auf dem 5%-Niveau, im einzelnen ist sie für 5 Statements ebenfalls signifikant unterschiedlich. Die Autoren schließen daraus, daß Lehrende und Studierende "fundamental unterschiedliche Kriterien" zur Lehrbewertung heranziehen würden.

Es wurde hier nicht, wie in den folgenden Studien, die Übereinstimmung zwischen der Beurteilung von Studierenden und Lehrenden in Bezug auf die Bewertung mit einem bestimmten Instrument ermittelt, sondern nur die Übereinstimmung in Bezug auf theoretische Lehraspekte und deren Wichtigkeit. Auf diese Weise lassen sich auch die Diskrepanzen zwischen dieser Studie und Untersuchungen mit relativ guter Übereinstimmung zwischen Studierenden und Lehrenden erklären.

2.1.4 Übereinstimmung von Lehrenden und Studierenden bei der Lehrevaluation

Ein Artikel, der eine gute Übereinstimmung zwischen Lehrenden und Studierenden bei der Einschätzung der Wichtigkeit von Items eines Instruments nachweist, ist der von Shatz und Best (1986), in dem Items identifiziert und verglichen werden, die zur Kursevaluation genutzt werden. Als Grundlage diente der "Students' Evaluations of Educational Quality" (SEEQ). 45 Lehrende und 106 Studierende wurden aufgefordert, die wichtigen Items des SEEQ zu benennen und daraus die wichtigsten zu identifizieren.

Beide Gruppen identifizierten durchschnittlich 22 Items als wichtig. Die benannten Items korrelierten mit .59, die wichtigsten sogar mit .75. Die Autoren kommen zum Schluß: "Obwohl Lehrende und Studierende die Kurse mit unterschiedlicher Verantwortlichkeit und Perspektiven betreten, scheinen sie bei der Frage übereinzustimmen, welche Aspekte der Kurse im Evaluierungsprozeß am wichtigsten sind" (Schatz und Best, 1986, S. 241).

Paula L. Stillman et al. (1983) ließen 85 Lehrende eines Jahres und 67 des nächsten Jahres mit einem 7 Item-Fragebogen und zwei offenen Fragen bewerten. Davon unterrichteten 41 Lehrenden den gleichen Kurs zum zweiten Mal. Außerdem wurde eine "peer evaluation" durchgeführt, bei der ein Lehrender/ eine Lehrende die Kurse besuchte und den gleichen Fragebogen beantwortete. Die Korrelationen zwischen studentischer Veranstaltungskritik und der "peer evaluation" rangierte zwischen .37 und .67 und war auf dem 0.001%-Niveau signifikant. Für die Lehrenden korrelierten die Ergebnisse zwischen .60-.90, die Reliabilität betrug .96 (Konsistenz).

In der Studie von Drews, Burroughs und Nokvich (1987) bewerteten Studierende und Lehrende in vier sehr unterschiedlichen Kursen 30 Prozent der Kursstunden mit einem kurzen 15-Item-Fragebogen jeweils am Ende einer Stunde. Die Korrelationen der Itembeantwortungen der Lehrenden mit den Studierenden waren bei sieben Items auf dem 1%-Niveau und bei 4 Items auf dem 5%-Niveau signifikant. Items zum Kursmaterial wurden sehr ähnlich eingeschätzt, gefolgt von Dozent-/Dozentinbewertung und Ge-

samtbewertung. Am eklatantesten unterschied sich die Einschätzung, ob die Unterrichtsstunde spannend oder einschläfernd war.

Schlußfolgerung

Es konnten keine Belege gefunden werden, die eine unzureichende Sachkenntnis der Studierenden vermuten lassen. Die Langzeitstabilität der Bewertungen ist beträchtlich. Studierende sehen die Lehrqualität nach Beendigung ihres Studiums nicht anders.

Im übrigen ist auffällig, daß Untersuchungen, die relativ allgemeine Kriterien der Lehre und ihre Wichtigkeit prüfen, eher größere Diskrepanzen in der Bewertung zwischen Lehrenden und Studierenden ermitteln. Untersuchungen, die konkrete Sachverhalte an bestimmten Aspekten oder in direkten Bewertungen der Lehre ermitteln, kommen dagegen zu hohen Übereinstimmungen zwischen Lehrenden und Studierenden.

Die unterschiedliche Einschätzung verschiedener Lehraspekte durch Studierende und Lehrende wirkt sich aber nicht auf die Beurteilung mit Hilfe entsprechender Fragebögen aus, wie die gute Übereinstimmung zwischen der Bewertung von Kollegen/ Kolleginnen und Studierenden zeigt. Eine Auswirkung "mangelnder Sachkenntnis" der Studierenden kann daher nicht angenommen werden.

Das Argument der "fehlenden Sachkenntnis" der Studierenden zur Beurteilung von Lehre setzt offensichtlich falsch an: Es unterstellt gleichsam, die Studierenden würden das "Wissen und die Kompetenz des Lehrenden" bewerten. Das trifft aber nicht zu, die Studierenden beurteilen vielmehr "die Vermittlung von Wissensbeständen" (nicht die Wissensbestände selbst)!

Daher ist gerade das Argument der mangelnden Sachkenntnis in doppeltem Sinne unzutreffend: Es trifft weder den Sachverhalt, um den es bei studentischer Veranstaltungskritik geht, noch finden sich in empirischen Studien stützende Belege.

2.2 Zweites Argument: Der Unterhaltungswert dominiert den Inhaltswert

Bedenken gegen die Lehrbewertung der Studierenden richten sich häufig darauf, daß Studierende nur oder überwiegend auf einen spannenden Vortragsstil achten und dabei dem Inhalt kaum Aufmerksamkeit schenken würden. Eine Reihe von Studien haben sich dieser Frage gewidmet: Bestimmt Ausdrucksstärke und Unterhaltungswert oder die Inhaltsdichte und der Lerngewinn das Urteil der Studierenden. Diese Arbeiten sind als Dr. Fox-Studium bekannt geworden (vgl. Übersicht 2).

Der Begriff "Dr.Fox-Effekt" stammt aus der Studie von Naftulin, Ware und Donnely (1973), die in sechs kurzen Videofilmen Inhaltsdichte (Anzahl genannter Fakten) und Expressivität (Unterhaltungswert) variierten. Der angebliche Dozent "Dr.Fox" wurde dabei von einem Schauspieler dargestellt. Mit diesem Effekt ist gemeint, daß eine Lehrveranstaltung von den Studierenden vor allem nach dem "unterhaltsamen" Vortrag bewertet wird, weniger nach der inhaltlichen Stoffdichte und dem Lernzugewinn.

Ware und Williams (1975) ließen 207 Studierende in sechs unterschiedlichen Kursen Videofilme (20 min) bewerten. Die Filme variierten bzgl. Inhaltsdichte und Vortragsart (expressiv, d.h. enthusiastisch, humorvoll, freundlich, ausdruckstark und charismatisch in der einen Version und wenig expressiv in der anderen). Danach füllten die Studierenden einen Fragebogen zur Lehre und einen Wissenstest aus. Sowohl der Inhalt als auch der Vortragsstil hatten im einzelnen Auswirkungen auf die Testergebnisse. Diese fielen für den expressiveren Vortragsstil insgesamt besser aus. Auch eine Interaktion zwischen Inhalt und Vortragsstil hatte keine Auswirkung auf die Ergebnisse des Wissentests.

Williams und Ware replizierten 1977 ihre Studie von 1975, um zu überprüfen, inwieweit Goffmans These (1959) zutrifft: "...expressives Verhalten kann eine Zuhörerschaft so stark oder stärker beeinflussen als der Inhalt, wenn es wenig Zeit oder Grund für die Zuhörerschaft gibt, die Präsentation zu beurteilen" (zitiert nach Ware und Williams, 1977, S. 455). Aus diesem Grund ließen sie nicht nur ein zwanzigminütiges Videoband bewerten, daß in der üblichen Art variierte, sondern zwei aufeinanderfolgende Bänder, die nacheinander bewertet wurden, um die Beobachtungsdauer zu verlängern. Goffmans These konnte zumindest für dieses Untersuchungsdesign nicht bestätigt werden, aufgrund der kurzen Präsentationszeit aber auch nicht widerlegt werden.

Meier und Feldhusen (1979) konnten in ihrer Studie viele der Befunde von Williams und Ware (1979) replizieren. Sie zeigten 161 Studierenden verschiedener Jahrgangsstufen jeweils einen der Videofilme von Ware und Williams (1975). Studierende bewerteten die ausdruckstärkere Darstellungsweise auf allen fünf in aufwendigen Vorstudien ermittelten Dimensionen (Instructor expressiveness, Instructor personality, Instructor explanations, Instructor preparation und Lecture content) höher.

Die Variation der Inhaltsdichte wirkte sich nur auf eine der fünf Dimensionen signifikant aus: die Bewertung auf der Subskala "Instructor explanations" war bei mittlerer Inhaltsdichte höher. Die Leistung der Studierenden war nach der ausdruckstärkeren Prä-

sentation etwas, jedoch nicht signifikant höher als bei der ausdrucksärmeren Präsentation.

Ziel der Studie von Marsh und Ware (1982) war ebenfalls die Überprüfung des "Dr. Fox"- Effekts mit Hilfe der Variation von Ausdrucksstärke und Inhaltsdichte. Jedoch versuchten die Autoren dabei sowohl die Multidimensionalität der Lehrbewertung zu erreichen als auch ein natürlicheres Setting zu wählen. Ausgehend von der These, daß Studierende mit einer bestimmten Zielmotivation in den Lehrveranstaltungen sitzen (z.B. mit Blick auf anschließende Prüfungen), wurde versucht, die extrinsische Motivation der Stoffaufnahme zu erhöhen.

Für diese Überprüfung wurden eine Reanalyse zweier älterer "Dr.Fox"-Studien durchgeführt. Eine war die geschilderte Studie von Ware und Williams (1975), die andere eine von den beiden Wissenschaftlern 1976 durchgeführte weitere Untersuchung. Diese wurde mit der Erweiterung wiederholt, daß die teilnehmenden Studierenden eine extrinsische Motivation bekamen: 1$ für die Teilnahme und 0,05$ für jede richtige Antwort eines anschließenden Tests. Über die "Belohnung" wurden 104 Studierende vor der Filmpräsentation und 108 Studierende nach der Filmpräsentation informiert. Bei der erneuten Analyse der Werte wurde der verwendete Fragebogen zur Bewertung der Lehre einer Faktorenanalyse unterzogen. Das Ergebnis ließ fünf gut interpretierbare Faktoren zu: Klarheit und Organisation, Gerichtetheit und Konzentration, Wissen und Kenntnisse, Begeisterung und Enthusiasmus, Lernanregung und Motivierung (in englisch: Instructor Clarity/Organisation, Instructor concern, Instructor Knowledge, Instructor Enthusiasm und Instructor Stimulated Learning). Diese wurden dann als abhängige Variablen angesehen, die den Einfluß experimenteller Manipulation demonstrieren sollten.

Diese Konstellation werten die Autoren wie folgt: "Diese Bedingungskonstellation ist besonders wichtig, da sie dem typischen universitären Kurs am nächsten kommt. Studierende wissen von Beginn des Kurses an, daß ihre Examensleistungen belohnt werden - Noten sind wahrscheinlich ein höherer Anreiz als das Angebot von 5 cents für jede richtige Antwort" (Marsh und Ware, 1982, S. 132).

Die Schlußfolgerung, die Marsh und Ware aus ihrer Untersuchung für die Lehre ziehen: "Augenscheinlich ist die Nutzung von Humor, Stimmmodulation und Enthusiasmus wichtig, wenn die extrinsische Motivation niedrig ist, aber weniger bedeutsam, wenn die Studierenden bereits motiviert sind" (Marsh und Ware, 1982, S. 133). Besonders interessant ist diese Schlußfolgerung, da Ware, der 1973 als Mitautor den Dr. Fox-Effekt mit"kreierte", damit seine früheren Ergebnisse deutlich relativiert.

Die Befunde machen deutlich, daß der Dr. Fox-Effekt vor allem in Situationen auftritt, die in der Hochschule selten sind, nämlich der Stoffpräsentation eines (wie in den älteren Dr.Fox-Studien) fachfremden Themas, dem das Auditorium in keinster Weise zweckgerichtet zuhört. Mit den überzeugenden Ergebnissen von Marsh und Ware (1982), die auf alle drei der vorher dargestellten Studien angewendet werden können, dürfte der fast ein Jahrzehnt durch die amerikanische Literatur zur studentischen Veran-

staltungskritik geisternde Effekt stichhaltig aufgeklärt und dadurch widerlegt worden sein.

Übersicht 2

"Dr. Fox"-Studien:
Inhaltsdichte und Unterhaltungswert
Befunde hinsichtlich der Effekte auf Lehrbewertung und Lernerfolgstest

1. Der erste Eindruck

Der deutlichste Effekt auf die Ergebnisse der Lehrbewertung durch Studierende geht von der Variation der Ausdruckstärke aus. Der Unterhaltungswert erscheint wichtiger als der Inhaltswert.
Allerdings wird die Lehrbewertung auch von den Variablen Inhaltsdichte und Belohnung beeinflußt. Zwar hatte die Änderung der Ausdrucksstärke einen ziemlich substantiellen Einfluß, dieser variiert jedoch in hohem Maße unter den fünf untersuchten Faktoren wie Klarheit, Wissen, Lernanregung, Begeisterung oder Konzentration.

2. Eine notwendige Differenzierung

Besonders ausgeprägt ist der Einfluß des Unterhaltungswertes (Ausdrucksstärke) auf die Dimension "Instructor enthusiasm" (Begeisterung), geringer bei allen anderen Faktoren.
Unter der Bedingung, daß die Belohnung bereits vorher angekündigt wird, hat die Ausdrucksstärke auf keine andere Dimension Einfluß mit Ausnahme des Faktors "Instructor Enthusiasm". Die Variation der Ausdrucksstärke beeinflußt demnach die Lehrbewertung nicht allgemein, sondern wirkt sich nur auf den damit verbundenen Faktor aus, was keine Verzerrung der Ergebnisse bedeuten würde.

3. Ein zweiter Blick

Der Einfluß der Inhaltsdichte auf eine günstige Lehrbewertung ist zwar geringer, variiert aber ebenfalls über die verschiedenen (fünf) Dimensionen. Diese Variable hat aber kaum Einfluß auf die Bewertung des Faktors "Instructor Enthusiasm" (Begeisterung). Der Einfluß der Inhaltsdichte ist jedoch bei niedriger Ausdrucksstärke größer.

4. Aufklärung und Revision des ersten Eindrucks

Die besten Ergebnisse im Wissensfest (Lernerfolg) werden unter der Bedingung erbracht: "Hohe Ausdruckstärke - Hohe Inhaltsdichte - Belohnungsankündigung vor Vortrag". Dabei hat der Zeitpunkt der Ankündigung der Belohnung einen hohen Einfluß auf die Testergebnisse. Liegt dieser Zeitpunkt nach der Präsentation des Vortrags, so hat die Ausdruckstärke einen stärkeren Einfluß auf die Ergebnisse als die Inhaltsdichte; wird die Belohnung jedoch vor dem Vortrag angekündigt, so ist der Einfluß der Inhaltsdichte hoch, der der Ausdrucksstärke jedoch vernachlässigbar.

Schlußfolgerung

In undifferenzierten Studien konnte der Eindruck entstehen, als würde der Unterhaltungswert (die Ausdrucksstärke) eines Vortrages von Studierenden mehr in ihre Urteile einbezogen als dessen Stoff- und Inhaltsdichte.

Weitere, differenzierende Studien haben diesen ersten Eindruck (als Dr. Fox-Effekt bekannt geworden) jedoch korrigiert, letztlich widerlegt. Dies aus zwei Gründen: Erstens trennen die Studierenden selbst sehr deutlich zwischen diesen beiden Aspekten (Dimensionen); zweitens achtet eine interessiert-motivierte Hörerschaft vermehrt auf den inhaltlichen Wert und Nutzen.

Die Befürchtung, daß Studierende vor allem den witzigen und unterhaltsamen Vortrag positiv bewerten, konnte nicht bestätigt werden. Studierende, die mit dem Ziel etwas zu lernen oder eine Prüfung zu bestehen, einen Kurs besuchen, bewerten diesen vor allem nach seinen Nutzen und nicht nach seinem Unterhaltungswert.

Im übrigen erscheint es durchaus angemessen, für die Qualität dieser Lehrleistung unter anderem auch die Ausdrucksstärke und Unterhaltsamkeit eines Vortrages einzubeziehen, da dadurch Aufmerksamkeit stabilisiert und die Lernmotivation erhöht wird. Dies begründet sich auch dadurch, daß der Lernerfolg bei den Studierenden ganz offensichtlich dann am größten ist, wenn beides zusammenkommt: eine hohe Ausdrucksstärke und eine hohe Inhaltsdichte.

In Instrumenten zur studentischen Veranstaltungskritik ist aber auf alle Fälle darauf zu achten, durch Fragen beide Aspekte einer Lehrveranstaltung getrennt zu erfassen, um mögliche Vermischungen zu vermeiden.

2.3 Drittes Argument: Beliebtheitsurteil statt Leistungsurteil

Wenn Lehrveranstaltungskritik der Vorwurf gemacht wird, nichts weiter als ein "Beliebtheitswettbewerb" zu sein, heißt das, daß ein starker "halo"- Effekt vermutet wird: Studierende würden überwiegend aufgrund der Beliebtheit eines Dozenten/einer Dozentin die Fragen beantworten. Marsh (1982) vermutet, daß dies vor allem dann der Fall sei, wenn es sich um kurze einfache Fragebögen handelt, die sehr global die Lehreffektivität zu ermitteln versuchen.

Die Studie von Barbara R. Sherman und Blackburn (1975) gehört zu den wenigen, die direkt mit dem Konstrukt "Beliebtheit" operieren. Sie wurde an einem College durchgeführt, wo üblicherweise die Studierenden die Mitglieder des Lehrkörpers mit einem "Institutional Research Instrument" bewerteten, das sowohl die Präsentation des Lehrstoffs als auch das Verhalten des Veranstaltenden gegenüber den Studierenden erfaßt. Weiter standen die Daten einer von Studierenden selbst entworfenen und durchgeführten Evaluation zur Verfügung, bei der 75 Studierende vor allem Fairneß, Nützlichkeit und Wichtigkeit einer Veranstaltung beurteilten. Als drittes Instrument wurde (in Anlehnung an Osgood) ein Semantisches Differential (Pequod) verwandt, mit dem 850 Studierende mit Hilfe von 30 bipolaren Begriffspaaren, die sich sowohl auf die Persönlichkeit als auch auf die Lehreffizienz bezogen, ihre Hochschullehrer und - lehrerinnen beurteilten.

An Ergebnissen ist aus dieser Studie festzuhalten:

1.) Die Korrelation zwischen den Werten aus dem "Institutional Research Instrument" und dem Polaritätenprofil bei 45 befragten Studierenden ist r = .77. Zwischen dem "Institutional Research Instrument" und den Eigeneinstufungen der Studierenden bei 63 befragten Studierenden beträgt die Korrelation r = .73 und zwischen dem Polaritätenprofil und dem "Pequod" bei 75 Studierenden ist r = .86. Alle Korrelationen sind auf dem 1%-Niveau signifikant.

2.) Auch bei der Multiplen Regressionsanalyse ergaben sich Regressionswerte zwischen R = .74-.88 für die Vorhersagbarkeit der Werte des "Pequod" und des "Institutional Research Instruments" aus den Ergebnissen des Polaritätenprofils, was nach Ansicht der Autorin und des Autors die Möglichkeit demonstriert: "... den Erfolg eines Lehrenden im Kurs auf Grundlage seiner Persönlichkeitseigenschaften vorherzusagen" (Sherman, Barbara R. und Blackburn, 1975, S. 130).

Barbara R. Sherman und Blackburn definieren leider an keiner Stelle in ihrem Artikel ihre Vorstellung des Begriffes "Persönlichkeitscharakteristika" im Vergleich zu den von ihnen verwendeten Konzept des Verhaltens eines/einer Lehrenden in Bezug auf "Kursfunktion und -aktivität". Somit bleibt die berechtigte Frage, ob mit den an Osgood angelehnten Instrument nur "Persönlichkeitscharakteristika" oder aber auch andere Parameter des "funktionalen Verhaltens" in der Lehrsituation erfaßt werden. Ist letzteres der Fall, so wäre nicht mehr genau zu bestimmen, welches Ausmaß der vorhersagbaren Varianz wirklich auf Persönlichkeitsfaktoren zurückzuführen wäre, sondern nur, daß die

drei verwandten Instumente gut übereinstimmen und auch wahrgenommene Persönlichkeitseigenschaften bei der Lehrbewertung eine Rolle spielen (vgl. Abschnitt 2.8).

Schlußfolgerungen

Die verschiedenen Studien machen deutlich, daß die Operationalisierung des Konstrukts "Beliebtheit" nicht einfach ist. Schwierigkeiten bereitet in solchen Fällen, Ursache und Wirkung exakt trennen und zuzuordnen. Kommt es z.B. zu einer hohen Beliebtheit, weil jemand ein guter Lehrer/ eine gute Lehrerin ist, oder erhält jemand gute Bewertungen, weil er/ sie beliebt ist?

Offensichtlich hängen Urteile zur Lehrqualität einer Veranstaltung und wahrgenommene Eigenschaften von Lehrenden miteinander zusammen; darauf werden wir später genauer zurückkommen (vgl. Argument 8).

Allerdings bleibt zu bedenken, daß ein solcher Zusammenhang nicht gegen die Validität der studentischen Veranstaltungskritik spricht, insofern für eine gute Lehrleistung bestimmte Eigenschaften vorauszusetzen sind bzw. in sie einfließen.

2.4 Viertes Argument: Unreife und Unerfahrenheit der Studierenden

Manche Lehrende bezweifeln, daß Studierende die Lehre beurteilen können, da sie zu unreif und unerfahren seien, manche unterstellen sogar, daß ihnen der notwendige Ernst fehle, weshalb ihre Urteile "beliebig" ausfallen und starke Schwankungen aufweisen würden. Im folgenden werden Studien zur Konsistenz der Beurteilung bei studentischer Veranstaltungskritik und zu Beurteilungsfehlern aufgrund von "Beeinflussung durch ältere Semester" oder "Beeinflussung durch expressiven Vortragstil" aufgeführt. Die Annahme, daß diese beiden Punkte auf die Lehrbewertung Einfluß haben, kann als vermuteter Mangel an Erfahrung (bzw. fehlende Beobachtungsschulung) gewertet werden.

2.4.1 Konsistenz oder Beliebigkeit

Konsistenz kann definiert werden als die Übereinstimmung von Lehrbewertungen für denselben Lehrenden bzw. dieselbe Lehrende in verschiedenen Kursen, sei es vor denselben Studierenden, zu verschiedenen Zeiten oder zu verschiedenen Themen.

Marsh (1982) vergleicht Daten aus 8.277 Kursen, die in 35 verschiedenen Fachbereichen über vier Jahre erhoben wurden. Das verwendete Instrument war der Fragebogen "Students' Evaluations of Educational Quality" (SEEQ), der 35 verschiedene Items und einige demographische Fragen enthält. Marsh fand Korrelationen von r=.70 für die Bewertung Lehrender, die den gleichen Kurs zu zwei verschiedenen Zeitpunkten abhielten. Dies drückt eine hohe Übereinstimmung und Konsistenz aus und gibt keine Hinweise auf die Beliebigkeit studentischer Urteile aufgrund ihrer "Unreife" oder Unerfahrenheit.

Seiler, Weybright und Stang (1977) nennen Korrelationen aus der studentischen Veranstungskritik:

- für Lehrende, die gleiche Kurse im gleichen Jahr jedoch zu verschiedenen Zeitpunkten gaben: r= .48;
- für Lehrende, die den gleichen Kurs in verschiedenen Jahren gaben: r= .38;
- für Lehrende, die unterschiedliche Kurse im gleichen Jahr unterrichteten: r= .30;
- und für Lehrende, die unterschiedliche Kurse in unterschiedlichen Jahren unterrichteten: r= .20.

Während eine noch hinreichende Konsistenz in dieser Untersuchung für gleiche Kurse zu verschiedenen Zeitpunkten oder Jahren gefunden wurde, fällt sie doch bei unterschiedlichen Kursen in unterschiedlichen Jahren stark ab.

2.4.2 Beeinflussung durch ältere Semester

Terry und McIntosh (1988) untersuchen den Einfluß, den Kurserwartungen von Psychologiestudierenden auf die Mitte- und Endsemesterbewertung haben. 61 Studierende beurteilten Kurs und Dozenten/ Dozentin insgesamt dreimal: vor Kursbeginn wurden die Erwartungen an den Kurs erfaßt. In einer Zusatzfrage gaben 81 Prozent der Studierenden an, daß sie etwas bzw. einiges über den Kurs gehört hätten, 85 Prozent davon gaben

an, daß ihre Informationen von anderen Studierenden stamme. Der Fragebogen wurde Mitte und Ende des Semesters wieder verteilt.

Insgesamt fanden die Autoren geringe, jedoch signifikante Korrelationen zwischen der ersten und dritten Beurteilung (r=.23) und zwischen der ersten und zweiten Beurteilung (r=.22). Die Korrelation zwischen der Mitte- und der Endsemesterbewertung dagegen betrug .61. Nach genauerer Durchsicht der Daten kamen die Autoren zu der Ansicht, daß vor allem zwei Erwartungsdimensionen die Bewertung beeinflußten: der Anforderungsgrad des Kurses und die Herzlichkeit des/ der Lehrenden. Sie vermuteten, daß dies die wichtigsten Dimensionen für Studienbeginner und -beginnerinnen seien.

Die Studie läßt die Frage offen, ob Vorinformationen älterer Semester valide sind und es so zum ermittelten Zusammenhang kommt, oder aber, ob Studierende in ihrer Bewertung von "falschen" Erwartungen beeinflußt werden. Da die Korrelationen zwischen Vor-Kursbeginn und späteren Erhebungen nicht sonderlich hoch sind, kann angenommen werden, daß sich die Studierenden während des Semesters eine eigene Meinung über die Fähigkeiten des/ der Lehrenden bilden.

Schlußfolgerungen

Das Argument, die Studierenden seien zu unreif die Lehre zu bewerten, konnte nicht belegt werden. Die umfangreiche Studie von Marsh (1982) läßt sogar ein ausgesprochen konsistentes Verhalten der Studierenden bei der Lehrbewertung vermuten.

Die Studierenden scheinen sich nur in sehr geringem Umfang von `Gerüchten' über die Lehrqualität der Lehrenden oder durch Informationen älterer Semester beeinflussen zu lassen. Diese Einflüsse erscheinen so gering, daß sie die Validität vergleichender Lehrbewertungen durch Studierende nicht bedeutsam vermindern.

Aufgrund der recht hohen Stabilität studentischer Urteile und ihrer geringen Beeinflußbarkeit kann daher ein Mangel an Erfahrung oder gar Unreife den Studierenden nicht unterstellt werden.

2.5 Fünftes Argument: Nicht die Lehrenden, sondern der Kurs bestimmt die Lehrqualität

Eine erstaunlicherweise nicht so häufig untersuchte Frage ist die nach dem Ausmaß des Kurseinflusses. Je nachdem, ob es sich um eine große dozentenzentrierte Veranstaltung oder ein kleines Seminar handelt, das stark von der Beteiligung der Studierenden abhängt, sind Unterschiede vorstellbar. Diese Unterschiede könnten sich auf so typische Fragen, wie die nach der Globalbewertung des Kurses oder auf die Einschätzung nach dem Ausmaß des Gelernten auswirken.

Romney (1976) versuchte, den Kurseffekt und den Effekt der/ des Lehrenden auf die Lehrbewertung mit Hilfe eines varianzanalytischen Verfahrens zu separieren. Ergebnis der varianzanalytischen Auswertung von 18 Kursen war, daß der Kurs die Lehrbewertung im gleichen Maße beeinflußt wie der Lehrende. Problematisch ist bei dieser Untersuchung der hohe Anteil an Kursen (15) aus dem ersten Studienjahr.

In verschiedenen anderen Studien wurde die unterschiedliche Art der Bewertung von Studierenden aus niedrigen und höheren Semestern untersucht. Es wäre nicht auszuschließen, daß höhere Semester es besser gelernt haben, den/ die Lehrende stärker vom Kurs zu trennen (zur methodischen Kritik an Romney vgl. Gillmore, 1977).

Kurseffekt und Lehrendeneffekt vesuchte auch Marsh (1982) mit Hilfe einer Pfadanalyse zu trennen. Die Daten aus 8.277 Kurse wurden zu 341 Sets zusammengestellt. Jedes Set beinhaltete folgendes:

- Zwei Bewertungen des gleichen Kurses, unterrichtet von der/dem gleichen Lehrenden zu zwei verschiedenen Zeitpunkten.
- Bewertung verschiedener Kurse bei dem/ der gleichen Lehrenden.
- Zwei Bewertungen des gleichen Kurses von zwei verschiedenen Lehrenden.
- Zwei Kurse unterrichtet von verschiedenen Lehrenden.

Die Korrelationen im Set ergaben als Ergebnisse:

- r=.70 für den gleichenden Instruktor/ die gleiche Instruktorin bei gleichen Kurs zu unterschiedlichen Zeitpunkten; da diese Übereinstimmung geringer ist, als die Reliabilität des verwandten Instruments (.90), kommt Marsh zu der Schlußfolgerung: "Dies läßt vermuten, daß ein substantieller Teil der Varianz (über 20%) bei der studentischen Bewertung reliabel ist, aber einmalig auf den bestimmten Kurs zurückzuführen ist" (ebd., S.51),
- r=.52 für den gleichen Instruktor/ die gleiche Instruktorin bei verschiedenen Kursen,
- durchschnittlich r=.14 bei verschiedenen Lehrenden, die den gleichen Kurs unterrichten, wobei eine große Variabilität zwischen den einzelnen Faktoren festgestellt wurde.

Mit verschiedenen Pfadanalysen versucht Marsh, den Kurs- bzw. den Lehrendeneffekt zu schätzen. Je nach Modell variierte die durchschnittliche Größe des Lehreffekts zwischen .50 und .57 und die des Kurseffekts zwischen .10 und .18.

Mit dieser differenzierten Analyse konnte Marsh belegen, daß Lehrende in verschiedenen Kursen zwar nicht völlig gleich "gut" bewertet werden, daß aber die Rolle des Kurses gegenüber dem Gewicht des Lehrenden für dessen Evaluation weit geringer, fast vernachlässigbar ist.

Schlußfolgerung

Die aufgeführten Untersuchungen zum Ausmaß des Kurseinflusses auf die Lehrbewertung müssen im Zusammenhang mit den Konsistenzstudien gesehen werden (vgl. Abschnitt 2.4). Insgesamt kann zwar davon ausgegangen werden, daß gewisse Varianzanteile existieren, die nur auf den entsprechenden Kurs zurückzuführen sind, doch der Anteil der konsistenten Beurteilung der Lehrenden liegt im Vergleich dazu deutlich höher und erscheint für eine vergleichende Lehrbewertung ausreichend.

Sicherlich ist es nach diesen Befunden nicht vertretbar, aufgrund der Evaluation eines Kurses allgemeine Aussagen über die Lehrleistung des/der Dozentin vornehmen zu können.

Ebensowenig sprechen aber unterschiedliche Bewertungen nach verschiedenen Kursen oder Zeiten für eine geringe Validität der messenden Instrumente. Das würde nur dann zutreffen, wenn unterstellt wird, daß Lehrende in allen ihren Veranstaltungen eine für die Studierenden gleiche Lehrleistung erbringen.

2.6 Sechstes Argument: Art und Form des Kurses beeinflussen die Bewertung

Neben der globalen Frage, ob der "Kurs" wichtiger als der "Lehrende" für die Bewertung sei - dies kann deutlich verneint werden - bleiben selbstverständlich eine ganze Reihe von Fragen offen, inwieweit Art, Form und Größe eines Kurses die Bewertung dessen Qualität beeinflußt. Sechs derartige "Kursmerkmale" werden nachfolgend betrachtet. Mögliche Unterschiede nach Art und Form der Kurse in der "Lehrbewertung" würden dafür sprechen, dies bei der vergleichenden Evaluation der Lehre von Dozenten zu berücksichtigen.

2.6.1 Hauptfach vs. Nebenfach

Die Frage, ob z.B. Studierende eines Magisterstudiengangs einen Unterschied in der Bewertung ihrer Kurse im Hauptfach und im Nebenfach machen, wurde meines Wissens in Deutschland noch nicht untersucht. Dies erscheint jedoch gerade für die geisteswissenschaftlichen Fächer als eine relevante Forschungsfrage.

Danielsen und White (1976) analysierten die Daten von über 5.300 Studierende, die 125 Lehrende bewertet hatten. Es konnten keine Unterschiede in der Bewertung von Haupt- bzw. Nebenfachstudierenden festgestellt werden.

Felicita F. Romeo und Weber (1985) ließen 773 Studierende aus 36 Kursen einen verkürzte "Student Course and Teacher Evaluation" (SCATE) ausfüllen. Sie ermittelten, daß Kurse im eigenen Hauptfach signifikant schlechter bewertet wurden.

Die Befunde zur Frage, ob Kurse im Haupt- oder Nebenfach unterschiedlich beurteilt werden, sind demnach nicht eindeutig.

2.6.2 Kurslevel bzw. Universitätsjahr der Studierenden

Immer wieder wird vermutet, daß niedrigere Semester ihre Lehrenden signifikant schlechter bewerten.

Danielsen und White (1976) und Patricia B. Elmore und Pohlmann (1978) konnten jedoch keine signifikanten Unterschiede in der Lehrbewertung unterschiedlicher Semester finden. Weder Blount et al. (1978) noch Felicita F. Romeo und Weber (1985) fanden Unterschiede in der Veranstaltungskritik von "graduate" und "ungraduate" Studierenden.

Richardson (1978) korrelierte die Antworten von 11.655 Studierende auf drei Globalitems "Lehrende", "Kurs" und "Examen" mit dem Kursniveau (Studienjahr). Alle drei Korrelationen waren auf dem 5%-Niveau signifikant, was aufgrund der großen Stichprobe schwer zu interpretieren ist, da bei großen Stichproben auch geringe Differenzen dann "signifikant" im statistischen Sinne werden.

In der Studie von Bruton und Crull (1982) hatte die Anzahl der Studienjahre einen geringen, aber doch signifikant negativen Einfluß auf die Kursbewertung. Allerdings waren die Kurse speziell für die Anfangssemester konzipiert.

Adriane Gaffuri et al. (1982) erhoben Daten in 85 Kursen. Es konnte nur eine Durchschnittskorrelation von .18 zwischen einem Item und dem Kurslevel gefunden werden.

Die meisten Studien erbringen keinen Zusammenhang zwischen Kurslevel bzw. Studienjahr und Lehrbewertung. Jedoch liegen einzelne Untersuchungen vor, bei denen zumeist geringe Unterschiede auftraten, wobei allerdings jeweils spezielle Bedingungen vorlagen.

2.6.3 Kursarten: Vorlesung, Übung, Seminar

Die externe Einflußgröße der Kursart - Vorlesung, Übung, Seminar - wurde in den Vereinigten Staaten nicht untersucht. Sie scheint jedoch in Deutschland eine eindeutige Auswirkung auf die Kursbewertung zu haben.

Preißer (1992) konnte an der TU Berlin ermitteln, daß Seminare sowohl im Grund- als auch im Hauptstudium besser bewertet wurden als Vorlesungen. Gleiches konnte auch Daniel (1994) sowohl für die Wirtschafts- als auch für die Geisteswissenschaften an der Universität Mannheim ermitteln.

2.6.4 Kursgröße

Eine Variable, die in den USA zwar häufig untersucht wurde, dort aber bei weiten nicht so relevant erscheint wie in Deutschland, ist die Kursgröße. Die Untersuchungen dazu beinhalten zwei Problembereiche. Einer davon ist eine mögliche Konfundierung zwischen Kursgröße und anderen Variablen. Große Kurse werden z.B. häufig in den ersten Semestern und in (ungeliebten) Pflichtkursen angeboten. Möglich errscheint auch, daß solche Kurse weniger gern unterrichtet werden oder an Lehrende mit niedrigerem akademischen Status und geringerer Lehrerfahrung abgegeben werden.

Der zweite Problembereich dieser Untersuchungen ist die Übertragbarkeit der Ergebnisse der amerikanischen Studien. Massenkurse wie sie in einigen Fächern in Deutschland üblich sind, sind in den Staaten nahezu unbekannt. Die Definition "Großer Kurs" fängt dort schon bei der Mitgliederzahl einer großen Schulklasse an.

Ein extremes Beispiel für diese Problematik ist die Arbeit von Felicita F. Romeo und Weber (1985), die keine sinifikanten Unterschiede für kleine vs. große Klassen fand. Dies könnte allerdings an ihrer Definition liegen, bei der "große Klassen" bereits bei mehr als 15 Studierenden beginnen.

Marcia K. Petchers und Chow (1988) ließen 856 Studierende des Fachbereichs Sozialarbeit (entspricht 81 Prozent der eingeschriebenen Studierenden) ihre Kurse mit dem "Student Evaluation of Course Instructors" (CECI) bewerten. Ein Einfluß der Kursgröße konnte nicht ermittelt werden.

Von einer negativen Korrelation mit der Klassengröße berichtet dagegen Scott (1977), d.h. in kleineren Kursen wurde die Lehre besser als in größeren beurteilt.

Patricia B. Elmore und Pohlmann (1978) untersuchten eine ganze Reihe von möglichen Einflußvariablen. Die Korrelationen waren insgesamt niedrig. Eine der noch am stärksten einwirkende Variable war die Kursgröße (negative Korrelation).

Adriane Gaffuri et al. (1982) korrelierten für 85 Kurse die Fragen ihres Fragebogens mit der Kursgröße. Die Durchschschnittskorrelation eines Items mit der Klassengröße betrug r=.-13.

Für unterschiedliche Klassengrößen ermittelten Cranton und Schmith (1986) Effektgrößen zwischen .10-.20. Bewertungen für kleine Kurse bzw. für Lehrende mit mehr Erfahrung fielen besser aus. Bei der Untersuchung des Effekts der Kursgröße nach einzelnen Semestern (aufgrund möglicher Konfundierungseffekte mit dem jeweiligen Semester), konnte nur für die Einschätzung des Gelernten ein Einfluß bestätigt werden. Am geringsten beeinflußte danach die Kursgröße die Bewertung der Effektivität des Lehrenden, dieser Effekt konnte aber nicht in allem Fachbereichen gefunden werden.

Eine Untersuchung, welche die Vermutung konfundierter Effekte unterstützt, haben Danielsen und White (1976) vorgestellt. Sie konnten einen signifikant positiven, allerdings nicht sehr großen Effekt für die Klassengröße ermitteln. Dies liegt jedoch u.U. an der Praxis der University of Georgia bzw. des College of Business Administration, an dem große Kurse nur von Lehrenden unterrichtet werden, die mit solchen Kursen gut zurecht kommen.

2.6.5 Kursschwierigkeit und Kursaufwand

Lehrende befürchten häufig, daß schwierige und/ oder aufwendigere Kurse schlechter bewertet werden. Leider liegen zu diesem Punkt nur wenige Ergebnisse vor.

In der Studie von Firth (1979) wurde mit Hilfe der studentischen Einschätzung der besten und schlechtesten Lehrenden nachgewiesen, daß das verwendete Evaluationsinstrument signifikante Unterschiede zwischen den beiden Extremgruppen (jeweils besten und schlechtesten 10% der Lehrenden) ermittelt. Dabei ermöglicht das Item "Schwierigkeit" und "Aufwand des Kurses" keine Unterscheidung zwischen den Gruppen. Das heißt, daß bei der Einschätzung eines Lehrenden die Kursschwierigkeit keine Rolle spielt. Centra (1977) konnte in einer ähnlichen Studie ebenfalls keinen signifikanten Einfluß ermitteln.

Coleman und McKeachie (1981) untersuchten die Auswirkung von Bewertungsergebnissen verschiedener Kurse auf die Kurswahl von Studierenden. Dafür wurden 582 zufällig ausgesuchten Studierenden Informationsmaterial über studentische Veranstaltungskritik gegeben. Darunter waren auch die Bewertungsergebnisse von vier verschiedenen Kursen der Politikwissenschaft aufgeführt. Drei Kurse hatten ungefähr das gleiche Ergebnis (Note 2-3), einer war insgesamt weit besser bewertet, aber als Kurs ausge-

wiesen, der mehr Arbeit erforderte als die anderen. Als Kontrollgruppen dienten 140 ebenfalls zufällig ausgesuchte, aber nicht informierte Studierende. Beim Vergleich zwischen den Kursen konnte festgestellt werden: In dem Kurs, der sehr positiv bewertet war, aber einen höheren Aufwand aufwies, hatten sich 40 Prozent mehr Studierende eingeschrieben, als es die Relation informierter Studierender vs. uninformierter Studierender voraussagen würde. Dieser Unterschied ist auf dem 1%-Niveau signifikant.

2.6.6 Pflicht- oder Wahlkurs

Brandenburg, Slinde und Batista (1977) untersuchten Korrelationen zwischen der Bewertung des "Illinois Course Evaluation Questionnaire" von 3.355 Studierenden und den Variablen: Klassengröße, Pflicht/ Wahlfach, erwartete Note, Kursniveau, Rang und Geschlecht des/der Lehrenden. Angesichts der Stichprobengröße verwundert es nicht, daß alle Effekte auf dem 1%-Niveau signifikant waren. Besonders hoch fielen sie jedoch bei der erwarteten Note und der Variable Pflicht- oder Wahlfach aus.

Patricia B. Elmore und Pohlmann (1978) konnten keine Unterschiede ermitteln für die Bewertung von Kursen, die unterschiedliche Anteile an Pflichtbesuchern hatten.

In der Untersuchung von Marcia K. Petchers und Chow (1988) wurde der Frage nachgegangen, inwieweit externe Rahmenbedingungen Einfluß auf studentische Lehrbewertungen haben. Einzig der Faktor "Wahl- oder Pflichtkurs" beeinflußte zwei der drei Subskalen signifikant.

Eine mögliche Erklärung dafür liefern DuCette und Jane Kenney (1982), die herausfanden, daß in einigen Kursen stärker als in anderen die Lehrbewertung mit den Noten korrelierte. Dies sei z.B. in Statistikkursen besonders ausgeprägt. Dazu vermuten DuCette und Jane Kenney, daß diese Art von Kursen Pflichtkurse sind, an denen die Studierenden kein großes inhaltliches Interesse haben und Noten daher um so wichtiger werden.

Eine andere Erklärungsmöglichkeit wäre, daß Plichtkurse häufig methodische o.ä. Kurse sind. Wie gut sie gemeistert werden, hängt dabei stärker als in anderen Kursen von Vorkenntnissen und Neigungen ab, die sich wiederum auf die Lernmotivation auswirken. Schätzen Studierende z.B. ihre erwartete Statistiknote gut ein, so kann man davon ausgehen, daß sie im entsprechenden Kurs gut mitgekommen sind. Gründe dafür sind außerdem ihre Vorkenntnisse und ihre Kursmotivation.

Hofman (1988) untersuchte, inwieweit verschiedene Studienmotivationsmerkmale studentische Veranstaltungskritik beeinflussen. 133 Studierende der Psychologie aus den Kursen: Statistik II, Handeln in komplexen Realitätsbereichen, Testtheorie II und Einführung in die Verhaltensanalyse wurden um die Bewertung ihrer Veranstaltungen gebeten. Beurteilt wurden sowohl die einzelnen Sitzungen, die Gesamtbeurteilung, der subjektive Lernfortschritt, das Interesse an verschiedenen psychologischen Themen (z.B. Sozialpsychologie, Methoden etc.), Leistungsmotivation, Arbeitsverhalten und Prüfungs- und Sprechängstlichkeit.

Das Ergebnis von Korrelationsuntersuchungen zeigte einen positiven Zusammenhang zwischen der Gesamtbewertung der beiden methodischen Veranstaltungen in mittlerer Höhe mit der Einstellung zum methodischen Vorgehen in der Psychologie und dem Interesse an kognitiver und mathematischen Psychologie. Ein negativer Zusammenhang ergab sich mit dem Interesse an psychoanalytischer und Sexualpsychologie. Korreliert wurden auch einzelne Dimensionen mit den Bewertungen. Zwischen 22 und 47 Prozent der Varianz der mittleren Beurteilungen einzelner Dimensionen ließen sich durch Kombinationen von zwei oder drei Motivationsmerkmalen vorhersagen.

Insgesamt schließt Hofmann: "Die Hypothese, daß Merkmale der Studienmotivation eine bedeutsame Determinante der Beurteilung von Veranstaltungen darstellen, wird somit insgesamt gut bestätigt" (Hofman, 1989, S. 125).

Kromrey (1994) geht sogar so weit, daß er die unterschiedlichen Lehrbewertungen vor allem auf eine unterschiedliche Motivation der Studierenden zurückführt. Daniel (1994) konnte jedoch ermitteln, daß weniger interessierte Studierende die Lehrenden zwar etwas schlechter bewerten, daß jedoch die Rangfolge der Bewertung mit denen interessierter Studierender identisch ist, sich nur parallel veschiebt.

Offenbar kann es sich in der Lehrbewertung niederschlagen, ob es sich bei einer Veranstaltung um einen Pflicht- oder Wahlkurs handelt. Dieser Faktor ist aber wohl hauptsächlich auf die unterschiedliche Motivation und das Interesse der Studierenden in diesen beiden Kursarten zurückzuführen (vgl. auch die behandelten Studien zum "Dr.-Fox-Effekt" sowie die Untersuchungen zum Argument 9: Eigenschaften der Studierenden).

Schlußfolgerung

Es kann nicht pauschal ausgeschlossen werden, daß externe Einflußgrößen Auswirkungen auf die studentische Lehrbewertung haben. Diese müssen jedoch im einzelnen betrachtet werden.

Insgesamt gibt es kaum einen Nachweis dafür, daß niedrigere Semester schlechtere Bewertungen abgeben. In Deutschland, wo die Zusammensetzung der Erstsemester eine andere ist als in den Vereinigten Staaten (späterer Studienbeginn, heterogenere Alterszusammensetzung, z.T. verschiedene Vorerfahrungen: Lehre, Zivildienst, Bundeswehr, Praktika etc.) müßten eigene Forschungprojekte mögliche Einflüsse, z.B. die einer abgeschlossenen Lehre, untersuchen.

Die Veranstaltungsart "Vorlesung" wird im Vergleich zu anderen Lehr-Lern-Formen relativ schlecht beurteilt. Dies müßte bei Kursvergleichen mitbedacht und berücksichtigt werden: Urteile zur Lehrqualität von Dozierenden sollten sich demnach bei Vergleichen nur auf die gleichen Veranstaltungsformen stützen.

Die in einigen Artikeln gefundenen Effekte zur Kursgröße sind alle eher niedrig. Dies kann jedoch auch an der verwendeten Definition eines großen Kurses, bzw. des Fehlens von "Massenkursen" in den USA liegen. Genauere Untersuchungen wären somit not-

wendig, um Aussagen für die Situation an deutsche Hochschulen machen zu können. Dabei müßte jedoch sichergestellt sein, daß die gefundenen Einflußgrößen tatsächlich auf die Kursgröße und nicht auf möglicherweise damit konfundierte Faktoren zurückzuführen sind.

Es wären ebenfalls weitere Studien an deutschen Hochschulen notwendig, um zu einer abschließenden Bewertung zu kommen, ob es eine Rolle spielt, daß ein Kurs in einem Haupt- oder Nebenfach angeboten wird.

Zusammengefaßt kann festgestellt werden, daß bei der Beurteilung von Lehrenden durch Studierende Kursschwierigkeit und -aufwand keine Auswirkungen auf die Lehrbewertung haben.

Man kann allerdings davon ausgehen, daß der Faktor Motivation die Lehrbewertung beeinflußt. Vermutlich lernen weniger motivierte Studierende weniger im entsprechenden Kurs und bewerten ihn daher schlechter. Dies wäre somit sogar ein Zusammenhang, der sich mit den Überlegungen zu "Lernzuwachs und Validität" decken würde. Lassen sich die Ergebnisse von Daniel bestätigen, so wäre auch eine Vergleichsmöglichkeit der Lehrenden gesichert, da sich Verfahren finden ließen, entsprechende Verzerrungen aufgrund unterschiedlicher Motivation zu eliminieren. Dazu wäre vorauszusetzen, sie bei der Lehrevaluation mit zu erheben.

Insgesamt müßten noch vermehrt Studien zu einigen der genannten Einflußgrößen durchgeführt werden, um zu einer abschließenden Beurteilung zu kommen. Dies wäre vor allem dann wichtig, wenn einzelne Kurse verschiedener Lehrender miteinander verglichen werden sollen. Auf alle Fälle ist es wichtig, bei der Erfassung und Analyse von Lehrbewertungen die Art und Form der Kurse und die Motivation der Studierenden zu berücksichtigen. Damit wäre bei der Interpretation der Daten zu vermeiden, der Lehrqualität von Hochschullehrenden (positiv oder negativ) zuzuschreiben, was im Grunde auf die Kursart und die Zusammensetzung der Studierenden im Kurs zurückzuführen ist.

2.7 Siebtes Argument: Der Status der Lehrenden verzerrt die Lehrbewertung

Gelegentlich wird vermutet, daß auch der Lehrstatus sich auf die Lehrbewertung auswirkt. Entsprechende Korrelationen ließen sich entweder auf die längere Lehrerfahrung von Lehrenden mit höherem Status zurückführen oder aber auf eine Urteilsverzerrung bei Studierenden, die sich durch den "Status" beeindrucken lassen und weniger die konkrete Lehrleistung beurteilen.

Patricia B. Elmore und Pohlmann (1978), Felicita F. Romeo und Weber (1985) und Marcia K. Petchers und Chow (1988) konnten jedoch keine Bewertungsunterschiede entsprechend dem Status der Lehrenden feststellen.

Die Autoren, Meredith und Schmitz (1986), gingen der Frage nach, ob signifikante Unterschiede in der Bewertung der Lehre von Dozenten/ Dozentinnen und (studentische) Tutoren/Tutorinnen existieren. Tutoren/ Tutorinnen erhielten signifikant bessere Noten für "Feedback" und schlechtere für "Beherrschung der Materie", "Enthusiasmus", "Zusammenfassung am Ende einer Diskussion" und "Seminarthemen waren interessant". Signifikante Unterschiede in Lernausmaß konnten nicht ermittelt werden. Bei diesen Ergebnissen stellt sich die Frage, inwieweit die Items "Enthusiasmus" und "Seminarthemen waren interessant" stärker wegen anderer Gründe (z.B. daß die von Tutoren/Tutorinnen unterrichteten Kurse oft Pflichtkurse des ersten Studienabschnitts waren) schlechter bewertet wurden.[2]

Centra (1980) bezieht sich u.a. auf eine eigene, frühere Studie und eine Untersuchung von Creech (1976), in der eine schlechtere Bewertung für Tutoren/Tutorinnen ermittelt wurde. In der Studie von Blount et al. (1978) konnten zwar nicht direkt bessere Ergebnisse für Lehrende mit höherem Status gefunden werden, jedoch ergaben paarweise untersuchte Werte (Newman-Keuls-Test), daß (auf dem 5%-Niveau signifikant) bessere Werte in einigen Dimensionen für Tutoren/Tutorinnen bzw. Assistenten und Assistentinnen als für Professoren und Professorinnen vergeben werden.

Schlußfolgerung

Insgesamt kann man nicht davon ausgehen, daß Studierende einen Dozenten/ eine Dozentin höher bewerten, nur weil dieser/ diese einen höheren akademischen Titel hat. Jedoch erscheint es möglich, daß z.B. Aspekte, die mit akademischen Titel häufig korrelieren, wie z.B. Lehrerfahrung oder Art des Kurses, von Einfluß für die Lehrbewertung sind.

[2] Insgesamt ist die Frage nach der Qualität der Tutoren und Tutorinnen auch im Zusammenhang mit der Diskussion um Ausweitung der Tutorenprogramme interessant. (Siehe: Aktionsprogramm "Qualität der Lehre", Abschlußbericht, Mai 1992, Hrg. Ministerium für Wissenschaft und Forschung, NRW).

2.8 Achtes Argument: Eigenschaften von Lehrenden verzerren die Bewertung

Wie die Diskussion um die Beliebtheit von Lehrenden (vgl. Argument 3) geht die Frage nach dem Einfluß von Persönlichkeitseigenschaften in eine ähnliche Richtung. Der Kernpunkt beider Diskussionen ist, streng gesehen, eine Validitätsfrage: Werden Lehrende aufgrund ihrer Leistungen in der Lehre oder aufgrund ihrer Beliebtheit, ihres Humors etc., d.h. unter der Einwirkung eines "Halo"-Effekts, eingeschätzt?

2.8.1 Herzlichkeit

Patricia B. Elmore und Pohlmann (1978) ermittelten, daß die "Selbstbeurteilung von Herzlichkeit" zwar nicht hoch, aber signifikant positiv mit der Lehrbewertung korreliert.

Die Herzlichkeit der Lehrenden wurde von Patricia B. Elmore und Karen A. LaPointe (1975) als wichtige Variable identifiziert, die die studentische Bewertung der Lehreffektivität beeinflußt.

Terry & McIntosh (1988) finden in ihrer Studie die Herzlichkeit des oder der Lehrenden als eine wichtige Erwartungsdimension der Studierenden, welche die spätere Bewertung beeinflußt.

Es ist zu vermuten, daß die Variable "Herzlichkeit" vor allem Einfluß auf Faktoren der Lehrbewertung hat, die die Beziehungsdimension der Lehre thematisieren und weniger auf andere.

2.8.2 Humor und Hilfsbereitschaft

In der Studie von Katherine van Giffen (1990) konnte ermittelt werden, daß der Einfluß von Humor insgesamt positiv mit der Lehreffektivität verbunden war ($r = .532$, $p < .01$), ebenso Freundlichkeit ($r = .639$) und Hilfsbereitschaft ($r = .791$), wobei die letzten beiden mit $r = .806$ miteinander korrelierten. Ähnlich wie beim Faktor der Herzlichkeit ist zu vermuten, daß Humor nur auf bestimmte Faktorender Lehrbewertung, vor allem der Beziehungsebene, Einfluß hat.

Schlußfolgerungen:

Bislang liegen nur wenige Studien zur Frage vor, ob persönliche Eigenschaften von Lehrenden die Bewertung ihrer Leistung in der Lehre beeinflussen. Sie sind zudem auf wenige Aspekte begrenzt. In der Regel handelt es sich um Faktoren der emotionalen Zuwendung.

Die bisher untersuchten Faktoren wie Herzlichkeit oder Humor haben einen Einfluß auf die Bewertung der Lehre. Allerdings bleibt zu fragen, ob dabei von "Verzerrung" gesprochen werden kann, da ein förderliches soziales Klima auch die inhaltliche Qualität zu steigern vermag.

Deutlich werden in den vorliegenden Studien vor allem zwei Punkte: (1) Es ist meist unmöglich, eine eindeutige Kausalbeziehung zu bestimmen. (2) Die instrumentelle Trennung zwischen allgemeinen Persönlichkeitseigenschaften und Eigenschaften, die sich positiv auf die Lehre auswirken, scheint in keiner der Studien gelungen. Um letzteres genauer zu erläutern: Es erscheint nachvollziehbar, daß jemand, der die Eigenschaft hat, besonders zuverlässig zu sein, pünktlich in den Unterricht kommt.

Ob allgemeine Eigenschaften von Lehrenden die Lehrbewertung beeinflussen und somit ein "Halo"-Effekt vorliegt, kann nur mit Instrumenten erfaßt werden, die trennscharfe Faktoren besitzen. In einer solchen Anordnung ist dann belegbar, ob z.B. die Eigenschaftszuschreibung "aufgeschlossen" nicht nur mit adäquaten Items (z.B. behandelt auch neuere Themengebiete), sondern auch mit inadäquaten (z.B. kommt pünktlich) signifikant korrelieren.

2.9 Neuntes Argument: Eigenschaften der Studierenden beeinflussen die Bewertung

Betrachtet man mögliche Auswirkungen von Eigenschaften der Lehrenden auf die Bewertung, so muß parallel dazu die Frage gestellt werden, ob unterschiedliche Studierendenpersönlichkeiten einen Einfluß auf die Ergebnisse studentischer Veranstaltungskritik haben. Anzunehmen ist, daß etwa leistungsmotivierte Studierende positiver urteilen können, aber auch die Zuschreibung von "Erfahrung" von Einfluß sein kann.

Um mögliche Eigenschaftszusammenhänge aufzudecken, führten Abrami et al. (1982) eine dreistufige Untersuchung durch.

1.) 388 Studierende der Psychologie wurde ein Video vorgeführt, auf dem Expressivität (niedrig/hoch) und Inhaltsdichte (niedrig/hoch) einer Lektion zum Thema Rollenverhalten manipuliert wurden. Die Studierenden beurteilten ihre eigene Persönlichkeit und ihre Einschätzung der Persönlichkeit des Dozenten auf Videoband mit Hilfe der "Adjective Check List" (ACL). Die Lehrbewertung wurde mit Hilfe der "Teacher Rating Form" (TRF/ 44 Items) ermittelt, das Ausmaß des Gelernten mit einem Multiple-choice-Test abgefragt.

2.) In einer zweiten Laboruntersuchung sahen 87 Studierende auf die gleiche Art manipulierte Videos, jedoch zu einem anderen Thema und beantworteten ebenfalls ACL, TRF und den Test.

3.) In einer Felduntersuchung bewerteten 108 Studierende 5 Dozenten (Freiwillige) mit Hilfe von ACL und TRF, füllten den TRF für sich aus und beantworteten Fragen eines Tests.

Ein genereller bzw. konsistenter Zusammenhang zwischen studentischer Veranstaltungskritik und Persönlichkeitseigenschaften der Studierenden konnte nicht gefunden werden. In den zwei Laboruntersuchungen konnte kein Zusammenhang ermittelt werden zwischen dem manipulierten Lehrverhalten und studentischen Eigenschaften. Nur in der Feldstudie traf ein bedingter Zusammenhang zwischen studentischen Eigenschaften und studentischer Veranstaltungskritik auf.

Insgesamt kommen die Autoren zu dem Schluß, daß studentische Lehrbeurteilung unabhängig von Persönlichkeitcharakteristika der Studierenden ist.

2.9.1 Studienerfolg und Ursachenzuschreibung

Gelegentlich wird vermutet, daß Studierende mit gutem Studienerfolg (Noten) zu eher besseren Kursbewertungen kommen als Studierende mit schlechten Erfolgsresultaten, da letztere den "Mißerfolg" eher dem Kurs zuschreiben dürften.

Snyder und Clair (1976) ermittelten an einer Stichprobe von 72 Studierenden, daß je besser die Note ausfiel, desto eher wurde internal attribuiert (eigene Leistung), je niedriger, desto eher kam es zu externalen Attributionen (äußere Faktoren).

Arkin und Maruyama (1979) ließen 116 Studierende einen Fragebogen ausfüllen, der auch Skalen zur externaler/ internaler Attribution enthielt. Es konnte ermittelt werden, daß erfolgreiche Studierende eher internal attribuieren. Dagegen attribuierten weniger erfolgreiche stärker external als der Durchschnitt. Erfolgreichere Studierende machen im Vergleich zu weniger erfolgreichen nicht nur für sich, sondern auch für Durchschnittsstudierende stärker internale Gründe für Erfolg/ Mißerfolg verantwortlich und sehen dies auch etwas stabiler.

Darüberhinaus konnte festgestellt werden, daß die Selbstattribution weniger erfolgreicher Studierender nicht mit der Lehrbewertung korreliert, eine internale Selbstattribution aber positiv. Je stärker erfolgreiche Studierenden internal attribuierten, desto besser bewerten sie die Lehrenden. Eine mögliche Erklärung geben die Autoren, in dem sie einen zugrundeliegenden dritten Faktor vermuten.

Bei 361 Studierenden aus Kursen der Erziehungspsychologie wurde von Ames und Lau (1979) ein Vergleich zwischen dem Locus of Control und den Lehrbewertungen durchgeführt. Die Daten wurden mit Multiplen Regressionalanalysen weiter analysiert. Ein konsistent positiver Zusammenhang zwischen internaler Attribution und guter Bewertung und externaler Attribution und weniger guter Bewertung konnte ermittelt werden.

Die Autoren kommen zu der Schlußfolgerung: "Die Ergebnisse zeigen eindeutig, daß attributionale Einschätzung über die Gründe studentischer Leistung in einem Kurs signifikant mit der Lehrbewertung verbunden sind ... Der Bewertungsprozeß ist, wie auch immer, ein Wahrnehmungsvorgang, und die vorliegende Studie zeigt, daß attributionalen Überzeugungen so maßgeblich bei der Lehrbewertung wie jeder objektive Unterschied zwischen Lehrenden sind" (Ames und Lau 1979, S. 35).

Diese Ergebnisse decken sich z.T. mit der Studie von Arkin und Maruyama (1979), sind jedoch weitreichender, da in der vorherigen Studie keine Korrelation zwischen externaler Attribution und schlechteren Lehrergebnissen gefunden wurde. Dies kann daran liegen, daß bei Arkin und Maruyama der Erfolg bzw. Mißerfolg in die Bewertung einbezogen wurde.

Ames und Lau kommen somit zu einer Kausalattribution, die auch für die Diskussion um den Einfluß von Noten eine Bedeutung hat, indem sie die bessere Lehrbewertung von Studierenden nicht auf das bessere Ergebnis, sondern auf internale Attribution zurückführen. Ein anderes Erklärungsmuster bietet die unterschiedliche Leistungsmotivation, die auch mit der unterschiedlichen Attributionsmustern zusammenhängen könnte.

2.9.2 Leistungsmotivation

Mit seiner Untersuchung wollte Kovac (1976) ermitteln, inwieweit verschiedene Persönlichkeitsmerkmale studentische Kursevaluation beeinflussen. Daten der Bewertung von 16 Dozenten/ Dozentinnen aus verschiedenen Fakultäten wurden gesammelt. 286 Studierende beantworteten folgende Fragebögen: "Mehrabian Achievement Scale", Rotters' "Locus of Control Scale", "Instructor Evaluation Form" und ein Semantisches Dif-

ferential mit Einschätzungen zu den Bereichen: Wert einer Veranstaltung, Effizienz eines Kurses und Interessensanregung.

Eine hohe Korrelation zwischen dem "Locus of Control" und dem "Leistungsanspruch" konnte festgestellt werden. Studierende mit einem hohen Leistungsanspruch bewerteten Inhalt und Struktur eines Kurses besser als Studierende mit einem geringeren Leistungsanspruch. Geht man davon aus, daß Studierende mit einem externalen Locus of Control einen geringeren Leistungsanspruch haben, so decken sich die Ergebnisse mit denen der vorherigen Studien.

215 Studierenden wurden von Strom et.al. (1982) Items zur Kursstruktur und Schwierigkeit vorgelegt. Der Fragebogen wurde mit den Ergebnissen eines anderen Fragebogen zu "Reizsuchverhalten", "Leistungsmotivation" und "Dogmatismus" korreliert. Die einzige auf dem 5%-Niveau signifikante Korrelation ergab, daß leistungsmotivierte Studierende weniger strukturierte und anspruchsvollere Kurse vorziehen.

Pamela J. Cooper et. al. (1982) ließen 240 Studierende einen Fragebogen beantworten. Die Fragebögen der verschiedenen Kurse wurden dann für die jeweiligen Lehrenden (75) zusammengefaßt. Im Vergleich zu den vorherigen Studien konnten keine Zusammenhänge zwischen der Lehrbewertung und dem Selbstkonzept bzw. der Leistungsmotivation der Studierenden festgestellt werden. Problematisch dabei ist eine sehr niedrige Rücklaufquote pro Kurs, die u.U. zu dem diskrepanten Ergebnis geführt haben könnte.

2.9.3 Lernstile und Vorlieben

Man kann versuchen, Studierende nicht nur entsprechend ihrer Leistungsmotivation oder ihres Locus of Control einzuteilen, sondern z.B. auch nach unterschiedlichen kognitiven Stilen bzw. Lernstilen.

Meredith (1981) untersucht inwieweit die von Kagan und Krathwohl (1967) postulierten "Lernerarten" die studentische Veranstaltungskritik beeinflussen. Sie gehen von zwei unterschiedlichen kognitiven Stilen aus: den "Focusern", die verstärkt Details und Fakten lernen, und den "Scannern", die sich eher auf das Begreifen von Konzepten und Prinzipien konzentrieren. Daten von 8.754 Studierenden aus 354 Kursen wurden mit Hilfe der "Faculty-Course Evaluation Scale" ermittelt. Fünf der Items wurden mit den beiden "Focus/Scanner"-Items korreliert.

Von den zehn Korrelationen waren fünf signifikant, was den Autor zu dem Schluß veranlaßte: "...das ist ein knapper Beweis, daß diese `kognitiven Stile' in einem engen Zusammenhang mit der Lehr-/Kursbewertung stehen" (Meredith, 1981, S. 620). Dieser Punkt müßte mit solchen Themen wie Leistungmotivation, Studienmotivation etc. gemeinsam untersucht werden.

In eine andere Richtung geht die Untersuchung von Drummond und McIntire (1977) in der 85 Studierende mit Hilfe des "Group Embedded Figures Test" (GEFT) in drei Gruppen eingeteilt wurden: "field dependent" (Studierende, die weniger strukturierte Kurse mit mehr Freiheitsgraden vorziehen), Mittelgruppe und "field independent" (Studieren-

de, die das traditionelle System befürworten). Zwischen den beiden Extremgruppen ergaben sich signifikante Unterschiede auf 14 der 28 Kursevaluationsitems.

Allerdings muß gesagt werden, daß diese Items nahezu ausschließliche Vorlieben und Interessen der Studierenden abfragen. Es stellt sich somit die Generalisierbarkeitsfrage dieser Daten. Durchaus möglich erscheint, daß Studierende unterschiedliche Gestaltungsvorlieben für einen Kurs haben, fraglich ist, inwieweit diese die Lehrbewertung beeinflussen.

Die Studie von Drummond und McIntire kann somit eher vernachlässigt werden. Die von Meredith angesprochenen Punkte müßten dagegen näher untersucht werden, um eine Verallgemeinerung zu ermöglichen.

Schlußfolgerung

Eine allgemeine Schlußfolgerung aus diesen Studien erscheint schwierig. Es beeindruckt zwar, daß es z.B. möglich ist, daß Studierende mit einer internalen Selbstattribution ihre Lehrenden besser bewerten als der Durchschnitt. Ob dies jedoch überhaupt in nennenswertem Grad geschieht, bleibt fraglich.

Die wenigen Studien zum Einfluß der Leistungsmotivation konnten keine prinzipiellen oder durchgängigen Unterschiede in der Lehrbewertung bedingt durch die Leistungsmotivation ermitteln.

Inwieweit unterschiedliche Lernstile und kognitive Stile der Studierenden für die Lehrbewertung eine Rolle spielen, läßt sich angesichts der Datenbasis und der vagen Befunde nicht abschließend beurteilen.

Spezifische Haltungen und Stile von Studierenden haben stets nur einen möglichen Einfluß auf einzelne Aspekte von Kursevaluationen, der zudem nicht sehr stark zu sein scheint.

Eine generelle Verzerrung im größeren Ausmaß aufgrund der unterschiedlichen Persönlichkeitseigenschaften der Studierenden kann allerdings nicht angenommen werden.

2.10 Zehntes Argument: Übereinstimmung von Lehrenden und Studierenden

Aufgrund der Vermutung, daß Studierende jene Lehrende mit einem ähnlichen Weltbild oder ähnlicher Persönlichkeitsstruktur bevorzugen, ist es sinnvoll zu überprüfen, inwieweit die Übereinstimmung von Lehrenden und Studierenden in Haltungen und Orientierungen die Bewertung der Lehre durch die Studierenden beeinflußt.

Untersucht wurde von Hofman und Liya Kremer (1980), inwieweit die Kongruenz von Lehrenden und Studierenden die studentische Veranstaltungskritik beeinflußt. Im ersten Teil der Untersuchung wurden 11 Lehrende von 455 Studierenden bewertet. Ermittelt wurde weiter die Einstellung zur höheren Bildung und Lehre, sowohl bei Studierenden wie auch bei Lehrenden. Ein Ähnlichkeitsindex (SIMIL) wurde für den Einstellungsfragebogen konzipiert, um das Ausmaß an Kongruenz zwischen Lehrenden und Studierenden ausfindig zu machen. Die Multiple Regression der Itemdurchschnittswerte und des SIMIL erklärten 17,7 Prozent der Kriteriumsvarianz. Autor und Autorin folgerten, daß die Gesamtbewertung sowohl von der studentischen Einstellung zu höherer Bildung als auch von der Kongruenz zwischen den Einstellungen von Studierenden und Lehrenden beeinflußt wird.

Im zweiten Teil der Untersuchung wurde das oben beschriebene Schema abgewandelt. 437 Studierende bewerteten 12 Lehrende. Ermittelt wurde auch die Einschätzung, die Lehrende von den Studierenden auf den vier Dimensionen der Einstellungsmessung hatten und umgekehrt. Erhoben wurde ein Ähnlichkeitsindex zwischen a) den Profilen der Lehrenden und der Studierenden, b) zwischen der studentischen Einstellung und der Einschätzung der Lehrenden, c) zwischen der Einstellung der Lehrenden und der studentischen Einstellung und d) zwischen den wahrgenommenen Einstellungen von Lehrenden und ihren tatsächlichen Einstellungen. Regression der Durchschnittsbewertungen des Itemsets und des Kongruenzsets ergaben 50 Prozent erklärter Kriteriumsvarianz

Der Kausalzusammenhang bleibt allerdings unklar: Entweder führt die Wahrnehmung guter Lehre dazu, daß Studierende davon ausgehen, daß die entsprechenden Lehrenden ihre bevorzugten Ansichten zur Lehre teilen; oder anders herum, daß Lehrende, von denen Studierende vermuten, sie würden ihre Ansichten teilen, bessere Bewertungen erhalten. Diese Studie ist in Zusammenhang mit der Diskussion über die mögliche Auswirkung von Meinungen über "gute Lehre" auf die tatsächliche Lehrbewertung mit einem bestimmten Instrument zu sehen (vgl. Abschnitt 2.1.3).

Thomas, Ribich und Freie (1982) gingen von der Hypothese aus, daß die Identifikation mit dem/der Lehrenden sich sowohl auf die Lernmotivation als auch auf die Lehrbewertung auswirkt. Sie benutzten als Grundlage für ihre Studie ein klassisches psychoanalytisches Identifikationsmodell. Sie kamen zu folgendem Ergebnis: Für fünf Faktoren (Kommunikation, Enthusiasmus, Organisation, Evaluation und Lehreffektivität) wurden signifikante Werte zwischen r= .20 und r= .43 erreicht. Eine Kausalbeziehung konnte aber nicht ermittelt werden. Demnach kann die Lehrbewertung entweder von einer

Identifikation mit dem Lehrenden beeinflußt werden, oder die Identifikation führt zu einer höheren Lernmotivation und beeinflußt darüber die Lehrbewertung positiv.

Morstain (1977) ließ in verschiedenen Kursen den "Student Orientation Survery" (SOS) und das "Student Instructional Report" (SIR) ausfüllen. Parallel dazu beantworteten die Mitglieder des Lehrkörpers den "Faculty Orientation Survey" (FOS). SOS und FOS sind Instrumente, die einen Überblick über die persönliche Orientierung von Studierenden (SOS) bzw. Lehrenden (FOS) ermöglichen. Gefragt wurde dabei nach Ansichten zu Art, Zweck und Prozessen der College-Ausbildung.

Ergebnisse: 7 der 24 möglichen Korrelationen der SOS- und der SIR-Werte korrelierten auf dem 5%-Niveau schwach positiv (r=.11-.14) miteinander. Einzig die Korrelation zwischen der Skala "Inquire" und dem angegebenen Kursnutzen für den Studierenden korrelierte auf dem 1%-Niveau signifikant (r=.28). Die Ergebnisse der Korrelation FOS und SIR waren weitaus höher. Von den 24 möglichen Kombinationen korrelierten vier schwach auf dem 5%-Niveau (r= -.11-.14) und sechs auf dem 1%-Niveau (r= -.70-.48).

Diese bedeutend höheren Korrelationen kommen nach Ansicht des Autors vor allem daher zustande, daß Klassenmittelwerte die studentischen Unterschiede bei der Beantwortung des SIR (Lehrbewertung) nivellierten. Die Gesamtauswertung der Multiplen Regressionsanalyse, in die alle signifikanten Skalen eingingen, zeigt, daß bei den vier untersuchten SIR-Skalen 10 bis 28 Prozent der aufgeklärten Varianz durch die "Orientierung" von Lehrenden und Studierenden erklärt werden kann.

Im letzten Teil seiner Untersuchung korrelierte Morstain inkongruente "Orientierungen" von Studierenden/Lehrenden. Dabei konnte kein signifikanter Wert ermittelt werden. Morstain spricht von einer "schwachen Tendenz", daß Kurse mit einer übereinstimmenden "Orientierung" von Lehrenden und Studierenden besser bewertet werden.

Dieser Befund kann jedoch auch folgendermaßen betrachtet werden: Ausgehend davon, daß Studierende ein Verhalten wahrnehmen und die Einstellungen von Dozenten/Dozentinnen beurteilen, ist es wenig verwunderlich, daß Einstellungsitems (SOS), die sich auf die Lehre beziehen, hoch mit der Lehrbewertung (SIR) korrelieren. Wenn Lehrende bestimmte Erwartungen erfüllen, wie sich Studierende die Lehre vorstellen, dann ist es durchaus ersichtlich, daß dies auf eine positivere Bewertung durchschlägt.

Schlußfolgerung

Nach den vorliegenden Ergebnissen kann man nicht davon ausgehen, daß die Übereinstimmungen zwischen Lehrenden und Studierenden größere Auswirkungen auf die Lehrbewertung haben. Die vorhandenen Befunde zur "schwachen Tendenz" einer besseren Lehrbewertung bei Übereinstimmung in Haltungen und Orientierungen lassen sich auf unterschiedliche Weise verstehen und erklären.

Außerdem haben die vorhandenen Studien keinen direkten Zusammenhang ermittelt. Zudem kann angenommen werden, daß die "Validität" der Instrumente dadurch nicht berührt wird.

2.11 Elftes Argument: Das Geschlecht beeinflußt die Bewertung

Der Geschlechtseinfluß auf die Bewertung von Veranstaltungen und Lehrenden muß unter zwei Gesichtspunkten betrachtet werden: Macht es Unterschiede, ob ein Kurs von einem Mann oder einer Frau unterrichtet wird? Und: Werden Lehrende von Studenten anders bewertet als von Studentinnen?

2.11.1 Unterschiede in der Bewertung von Dozenten und Dozentinnen

Ziel von Patricia B. Elmore und Pohlmann (1978) ist es, verschiedene Variablen, die in anderen Studien als beeinflussend für die Lehrbewertung ermittelt wurden, simultan zu erfassen. Daten wurden in der üblichen Jahreserhebung mit dem "Instructional Improvement Questionaire" an 174 Kursen gewonnen. Es konnten keine signifikant unterschiedlichen Bewertungen für Dozentinnen bzw. Dozenten ermittelt werden.

Die Autorinnen Thurston und Littlepage (1979) untersuchten die Hypothese, daß Professorinnen niedrigere Bewertungen erhalten als Professoren, daß sie weniger populär sind und ihre Kurse weniger häufig gewählt werden. Diese Hypothese stützt sich auf die Vermutung, daß Professorinnen mehr leisten müßten, um die gleichen Bewertungen zu erhalten.

Zwei Experimente wurden durchgeführt: Beim ersten Experiment wurde eine Scheinanmeldung für fiktive Kurse nicht existierender Professoren und Professorinnen durchgeführt. 47 Studenten und 52 Studentinnen sollten sich für Kurse anmelden, bei denen zum einen nur das Geschlecht der fiktiven Person und zum anderen sowohl der Kursinhalt als auch das Geschlecht variiert wurde. Ergebnis: Es konnte kein Geschlechtsunterschied bei der Wahl der Kurse, noch irgendeine Präferenz für einen der beiden Kurse gefunden werden. Auch das Geschlecht der Studierenden hatte keinen Einfluß auf die Wahl eines Kurse.

Im zweiten Experiment gewichteten 110 Studierende (66 Studentinnen und 44 Studenten) vier Aussagen, nachdem sie eine Persönlichkeitsbeschreibung einer fiktiven Professorin bzw., nur durch eine Änderung des Vornamen, eines fiktiven Professor gelesen hatten. Professorinnen bekamen bei keiner der vier Aussagen schlechtere Einschätzungen. Studenten und Studentinnen zeigten keine Unterschiede in ihren Bewertungen. Andere Wechselwirkungen konnten ebenfalls nicht ermittelt werden. Insgesamt sahen die Autorinnen eine mögliche Erklärung für ihre Ergebnisse, die diskrepant zu den Ergebnissen anderer Studien stehen, in der Genauigkeit und des Ausmaßes an "outside definition of excellence", das in den beiden Experimenten durch die Beschreibung der wissenschaftlichen Leistungen der Lehrenden erreicht wurde.

Nach Durchsicht verschiedener Studien sind Bray und Howard (1980) der Ansicht, daß die Beziehung zwischen den Persönlichkeitseigenschaften von Studierenden und Lehrenden und der Lehreffektivität komplexer Natur sind. Faktoren wie Herzlichkeit, Interesse und Geschlecht wurden bereits in vielen anderen Studien untersucht. Bray und

Howard gehen der Vermutung nach, daß dies Aspekte eines globaleren Konstrukts, der "sex role orientation" (SRO)" sind.

497 Studierende bewerteten 36 Lehrende eines sozialwissenschaftlichen College. Diese Lehrenden wurden aus einer Gruppe von einhundert Hochschullehrern und -lehrerinnen ausgewählt, die den "Bem Sex Role Inventory" (BSRI) beantwortet hatten. Aus dieser Gruppe wurden zwölf Lehrende, deren Werte in der Kategorie "feminin" lagen, und je zwölf weitere der Kategorien "maskulin" und "androgyn" ausgewählt. Die Studierenden beurteilten die Hochschullehrer und -lehrerinnen mit Hilfe eines Teils des "Instructional Development and Assessment Sysstem" (IDEA). Aus dieser Studie sind folgende Ergebnisse festzuhalten:

1. Androgyne Lehrende wurden signifikant höher bewertet als maskuline Lehrende, die Studierenden waren mit ihnen auch zufriedener.
2. Insgesamt wurden androgyne Lehrende nicht signifikant höher bewertet als feminine. Allerdings berichteten Studierende von androgynen Lehrenden über geringfügig mehr Lernfortschritte.
3. Entgegen der Vermutung der Autoren wurden keine Zusammenhänge zwischen dem BSRI der Studierenden und der Lehrenden festgestellt.
4. Studierende, die von Hochschullehrerinnen unterrichtet wurden, die sich selbst als maskulin einstuften, gaben einen signifikant höheren Lernfortschritt an, als Studierenden, die von Hochschullehrern unterrichtet wurden, die sich als maskulin einstuften.
5. Von den Hochschullehrerinnen bekamen die, die sich androgyn einstuften, die höchsten Zufriedenheitswerte.

Interaktion: Geschlecht - Humor

Katherine van Giffen (1990) beabsichtigte, unter naturalistischen Bedingungen, die eventuell unterschiedliche Nutzung von Humor durch Hochullehrer und Hochschullehrerinnen zu klären. 849 Studierende bewerteten zwölf Hochschullehrer und zwölf -lehrerinnen, die alle angaben, daß sie gelegentlich Humor in ihren Veranstaltungen verwenden würden.

Die Durchschnittswerte für Lehreffektivität und für die Nutzung von Humor können sowohl für Hochschullehrer und Hochschullehrerinnen als identisch betrachtet werden. Weder Kursniveau noch eine Wechselwirkung zwischen Kursniveau und Geschlecht konnten ermittelt werden.

Interaktion: Geschlecht - Herzlichkeit

22 Kurse, in gleicher Anzahl von Hochschullehrern und Hochschullehrerinnen unterrichtet, wurden von 838 Studierenden bewertet. Für die folgende Auswertung von Patricia B. Elmore und Karen A. LaPointe (1975) wurde der erste Teil des "Instructional Improvement Questionaire" verwandt. Von den zweiundzwanzig Kursleitern und Kursleiterinnen wurde die Beantwortung von zwei Fragen zur Selbsteinschätzung ihrer Herzlichkeit gesammelt. Für die Autorinnen waren von besonderem Interesse, Zusammen-

hänge zwischen den Faktoren Geschlecht des/ der Lehrenden, Geschlecht des/ der Studierenden und Herzlichkeit des/ der Lehrenden zu ermitteln.

Ergebnisse: Es wurden keine signifikanten Zusammenhänge zwischen dem Geschlecht (Lehrenden und Studierenden) und der Herzlichkeit des/der Lehrenden festgestellt. Insgesamt scheinen die Ergebnisse weniger von der Selbstauskunft über die Herzlichkeit der Lehrenden als vom Geschlecht der Studierenden beeinflußt.

2.11.2 Unterschiede in der Lehrbewertung durch Studenten vs. Studentinnen

Patricia B. Elmore und Pohlmann (1978) fanden keinen globalen Geschlechtseinfluß auf die Lehrbewertungen, die sie abgeben.

Ashton (1975) fand ebenfalls keine Korrelationen zwischen der Einschätzung von Persönlichkeitseigenschaften (erfragt mit dem "Survey of Interpersonal Values") und einem Fragebogen zur Lehrbewertung in einem Kurs von 84 Studierenden. Seperat nach Geschlecht untersucht, erwiesen sich jedoch zwei Beziehungsebenen bei Studentinnen als signifikant mit der Lehrbewertung korrelierend: Konformität (-.54) und Unabhängigkeit (.48). Der Autor schließt daraus, daß zwischenmenschliche Werte höher mit der Lehrbewertung von Studentinnen in Beziehung stehen, als daß sie die Bewertung von Studenten beeinflussen.

Dies würde sich auch mit den Ergebnissen von Daniel und Hornbostel (SPIEGEL-SPEZIAL, 1993, S. 157) decken. Einen Hinweis darauf, daß Studentinnen u.U. zu strengeren Bewertung neigen, hat Daniel gefunden. Er ermittelte in der zweiten SPIEGEL-Erhebung, daß die Beurteilungen von Studentinnen signifikant schlechter als die ihrer Kommilitonen waren. Ob dies jedoch auch konkret bei der Beurteilung eines/einer bestimmten Lehrenden der Fall wäre, ist fraglich, da in der zweiten SPIEGEL-Erhebung nicht einzelne Lehrende bewertet wurden, sondern eher die Studienbedingungen, die von den Studentinnen als ungünstiger empfunden wurden als von Studenten.

Die Autorinnen Patricia B. Elmore und Karen A. LaPointe (1975) konnten weiter beobachten: Studentinnen vergaben bessere Werte bei dem Item "einzelne Ziele des Kurses" als Studenten. Ein Item ("zeigte Interesse an den Studierenden") wurde von Studentinnen signifikant höher für ihre lehrenden Geschlechtsgenossinnen eingeschätzt als für Hochschullehrer. Auch bewerteten Studentinnen ihre Hochschullehrerinnen insgesamt besser, als das Studenten taten. Dagegen wurden Hochschullehrer mit diesem Item von Studenten fast immer signifikant höher bewertet als von Studentinnen. Dies deutet auf geschlechtsspezifische Übereinstimmung bestimmter Aspekte hin.

Dieses Ergebnis deckt sich mit den Untersuchungen von Bray und Howard (1980), die eine signifikante Beziehung zwischen Geschlecht der Lehrenden und Geschlecht der Studierenden in Bezug auf den berichteten Lernfortschritt ermittelten. Dabei gaben Studentinnen einen höheren Lernfortschritt in Veranstaltungen an, die von Frauen gehalten wurden, im Vergleich zu den von Männern gehaltenen. Auch gaben Studentinnen im Vergleich zu Studenten einen höheren Lernfortschritt bei Hochschullehrerinnen an.

Dies wäre eine Möglichkeiten, den oben genannten Befund zu erklären: Geht man von Validierungsstudien aus, so wäre es schlüssig, daß Studierende, die einen höheren Lernfortschritt in bestimmten Kursen beobachten, diese auch höher bewerten - dies wären in diesem Fall Studentinnen in Kursen, die von Dozentinnen gehalten wurden.

Schlußfolgerung

Nach den Studien zum Einfluß des Geschlechtes auf die Lehrevaluation ist zu vermuten, daß es keine direkt und prinzipiell unterschiedlichen Bewertungen zwischen Dozenten und Dozentinnen gibt (obwohl dieses Ergebnis kulturabhängig sein könnte), daß es aber möglicherweise zu Interaktionen zwischen Geschlecht und Persönlichkeitseigenschaften kommen kann.

Die untersuchten Interaktionen ließen jedoch keinen signifikanten Einfluß des Geschlechts der oder des Lehrenden auf die Ergebnisse der studentischen Veranstaltungskritik erkennen. Insgesamt ist somit zu vermuten, daß Dozenten und Dozentinnen keine signifikant unterschiedlichen Bewertungen allein aufgrund ihrer Geschlechtszugehörigkeit erhalten.

Die Untersuchungen über Studierende legen den Schluß nahe, daß Studentinnen nicht generell Hochschullehrer bzw. Hochschullehrerinnen anders bewerten als ihre männlichen Kollegen, sie aber einzelne Aspekte anders gewichten. Für genauere Aussagen müßten weitere Untersuchungen erfolgen. Da gerade Rollenbilder und -muster kulturabhängig sein können, wäre es notwendig, solche Untersuchungen zum Geschlechtseinfluß auf die Lehrbewertung auch in Deutschland, bzw. in vergleichbaren Ländern, durchzuführen, um verbindliche Aussagen treffen zu können.

2.12 Zwölftes Argument: Noten (erwartete oder erhaltene) beeinflussen die Bewertung.

Noten sind wohl die am exzessivsten untersuchte mögliche Einflußvariable der studentischen Veranstaltungskritik. Die Angst vor der möglichen "Bestechlichkeit" der Studierenden ist auch in der Umfrage von Marsh (1982b) zu finden, in der mehr als zwei Drittel der befragten Lehrenden angaben, daß sie eine Verzerrung der Lehrbewertungen durch die Art der Notenvergabe vermuten. Diese Ängste können genauer beschrieben werden: Zum einen wird befürchtet, daß strenger benotende Lehrende schlechter bewertet werden als mild bewertende Lehrende; zum anderen besteht die Angst, daß Studierende, die schlechter bewertet wurden, ihrerseits die Lehrenden schlechter bewerten.

2.12.1 Mild oder streng benotende Lehrende

Um feststellen zu können, ob mild benotende Lehrende bessere Werte in der Lehrevaluation erhalten als streng bewertende, ist es zwingend notwendig, Kursunterschiede zu vergleichen und nicht innerhalb von Kursen zu korrelieren. Auf diese Art und Weise werden auch mögliche Interaktionen (z.B. zwischen Studienmotivation, Note und Lehrbewertung) vermieden. Bedauerlicherweise ist die Art der Korrelation häufig nicht angegeben.

Ziel der Studie von Brown (1976) ist es, mit Hilfe einer großen Stichprobe und multivariater Techniken zu ermitteln, welchen Einfluß Noten auf die Bewertung der Lehrenden haben. Eine multiple Regressionsanalyse zeigte, daß die Durchschnittsnote eines Kurses der beste Prädiktor für die Bewertung des Lehrenden im Kurs war. Die Korrelation zwischen der Durchschnittsbewertung und der Durchschnittsnote betrug r =.35.

Auch in der Studie von Centra (1977) wird festgestellt, daß Korrelationen zwischen Leistungen und Gesamtlehrbewertung und in etwas geringerem Ausmaß zwischen Leistungen und Bewertung der Lehreffektivität sowie Qualität der empfohlenen Lektüre liegen. Am wenigsten korrelierte die Leistung mit der Kursschwierigkeit und der Arbeitsbelastung. Ein Problem liegt in der niedrigen Zahl der Studierenden in den Kursen, der Median lag bei 8 Studierenden.

In der Studie von Patricia B. Elmore und Pohlmann (1978) wirkt von den untersuchten Variablen noch am stärksten die erwartete Note im Kurs auf die Lehrbewertung ein (positive Beziehung). Adriane Gaffuri et al. (1982) erhoben Daten über drei Trimester in 85 Kurse mit einem Fragebogen: Die Durchschschnittskorrelation eines Items mit der Note betrug r=.35.

Die Studie von Chacko (1983) untersucht die Hypothese, daß Studierende, die eine schlechtere Bewertung in ihren "midterm examinations" erhielten, eher geneigt sind, die Effektivität der Lehrenden geringer zu bewerten, als Studierende, die keine ungünstigen Rückmeldungen erhalten haben. Achtundvierzig Studierende wurden in zwei vergleichbare Kurse eingeteilt, die von der gleichen Person unterrichtet wurden. Die Art der Bewertung der "midterm examination" wurde so variiert, daß die Bewertung des einen

Kurses (Section T) strenger als bei diesem Hochschullehrer sonst üblich ausfiel, während der andere Kurs (Sektion C) als eine Art Kontrollgruppe fungierte, da er den bei diesem Lehrenden üblichen Bewertungsstandard hatte. Die Studierenden beantworteten eine "Student Evaluations Form" eine Woche vor dem "midterm examination" und ein weiteres Mal eine Woche nach Bekanntgabe der Examensergebnisse. Dieser Evaluationsbogen beinhaltete ein weiteres Item, nämlich die Frage, ob das Examen zu hart, wie erwartet oder zu nachsichtig bewertet worden sei.

Ergebnisse: 1. Die Studierenden des ersten Kurses (Section T) nahmen die Examensbewertung als strenger wahr als die Studierenden der Kontrollgruppe (Section C). 2. In der zweiten Durchführung der studentischen Veranstaltungskritik konnten für "Section C" keine signifikanten Änderungen ermittelt werden. Für "Section T" dagegen bewerteten die Studierenden ihren Instruktor schlechter als vor Bekanntgabe der Noten. Sowohl die Itemsscores der "Inhaltsdimension" als auch die Scores der "Orientierung zu Studierenden" veränderte sich. 3. Die drei Dimensionen, die sich nicht änderten waren: "Kursorganisation", "Präsentation" und "Erreichbarkeit".

Problematisch ist bei dieser Studie zu bewerten, daß die Studierenden die Bewertungsstandards des Lehrenden vermutlich bereits vor der Untersuchung kannten und beurteilten, ob die vergebenen Noten aus dem Üblichen herausfielen. Daher kann nach dieser Untersuchung nicht geschlossen werden, ob sich die Studierenden ungerecht bewertet fühlten und daher ein anderes "Bild" vom Lehrenden bekamen, sowohl von seiner menschlichen wie auch seiner fachlichen Qualität, oder ob generell härter bewertende Lehrende schlechtere Ergebnisse erzielen.

2.12.2 Gut oder schlecht bewertete Studierende

Viele Lehrende befürchten, Studierende, die eine schlechte oder eine schlechter als erwartete Note erhalten, dann ihrerseits es den Lehrenden "mit gleicher Münze heimzahlen" - und ihre Lehre schlechter beurteilen. Um zu ermitteln, ob Studierende, die weniger gut bewertet wurden, ihrerseits den Dozenten/die Dozentin schlechter beurteilen, dürfen nicht die von den Kursteilnehmern und -teilnehmerinnen erwarteten, sondern es sind die nur tatsächlich erhaltenen Noten heranzuziehen. Für eine Korrelation: erwartete schlechte Note - schlechte Lehrbewertung gäbe es andere plausible Erklärungen, z.B.: Der Student hat wenig verstanden, also schließt er daraus, daß der Lehrende den Stoff schlecht vermitteln kann.

Yates und Ksarmos (1971) überprüfen ihre Überlegung, ob Studierende ihre Bewertung ändern, nachdem sie ihre Noten erfahren haben. 31 Studierende eines Kurses beantworteten zweimal innerhalb kürzester Zeit (ca. 1 Woche) den "Instructional Improvement Questionaire", wobei ihnen der Zweck der zweimaligen Bewertung nicht bekannt war. Insgesamt kamen die Wissenschaftler zu den Ergebnis, daß es zu einer sehr geringen, nicht signifikanten Verschlechterung der Bewertung kam. Ein Unterschied in der Art der Bewertung von gut vs. schlecht benoteten Studierenden konnte nicht gefunden werden.

Kritisch ist zu bedenken, daß die ähnliche Bewertung an der kurzen Zeit zwischen der ersten und zweiten Erhebung liegen könnte, was auch die Autoren als mögliche Begründung sehen. Es ist bedauerlich, daß bei dieser "quasiexperimentellen" Anordnung, nicht eine möglichst adäquate Täuschung der Studierenden versucht wurde. Eine solche Untersuchung an einer größeren Zahl von Studierenden wäre eine gute Möglichkeit, dem vermuteten Zusammenhang zwischen Noten und Bewertung nachzugehen.

Snyder und Clair (1976) untersuchten den Noteneinfluß auf Lehrbewertungen und versuchen dabei Fähigkeits- und Leistungseinflüsse auszuschließen. 72 Studierende wurden einem Intelligenztest (Wonderlic Personell Test) unterzogen und wahllos mit den Noten A, B und C gleich verteilt bewertet. Die Ergebnisse wurden den Studierenden mitgeteilt und ihnen wurde weiter erläutert, daß zwischen dem Wonderlic-Ergebnis und einem späteren Test üblicherweise eine hohe Korrelationen besteht. Danach sahen die Studierenden eine zehnminütige Lektion auf Video und füllten den Test aus. Zusätzlich sollten sie angeben, welche Bewertung sie im Test erwarten. Diese Tests wurden dann wiederum ohne Rücksicht auf reale Ergebnisse mit den gleich verteilten Noten A, B und C bewertet. Es wurde jedoch darauf geachtet, daß die Studierenden, die im Wonderlic ein A (bzw. B oder C) erhalten hatten, zu gleichen Teilen ein A, B oder C erhielten. Abschließend bewerteten die Studenten und Studentinnen den Lehrenden, dessen Lektion sie auf Video gesehen hatten, mit einer modifizierten Version einer "Teaching Assessment Blank" und eines weiteren Items zur Gesamtbeurteilung des Lehrenden.

Ergebnisse: 1. Je höher die erwartete Note, desto schlechter die Bewertung des Lehrenden. 2. Je höher die erhaltene Note, desto besser die Bewertung des Lehrenden. 3. Übertraf die erhaltene die erwartete Note, so gaben die Studierenden bessere Bewertungen ab. 4. Am schlechtesten fiel die Bewertung des Lehrenden dann aus, wenn die erhaltene Note die Erwartungen enttäuscht hatte.

Da diese Studie eher einem Experiment als einer Untersuchung im natürlichen Setting ähnelt, sind Verallgemeinerungen schwierig. Die Situation ist an verschiedenen Punkten so künstlich (Tape ist zu kurz, nur 10 min.; widerspricht gelernten Erwartungen in die eigene Intelligenz und dem gelernten Gefühl, wie man in einem Examen abgeschnitten hat), daß mögliche neue Einflüsse völlig ungeklärt sind. Es bleibt weiter untersuchenswert, warum erwartete gute Noten zu schlechterer Bewertung, erhaltenen gute Noten jedoch zu besserer Bewertung führten.

Brandenburg, Slinde und Batista (1977) fanden signifikante Korrelationen zwischen der Bewertung des "Illinois Course Evaluation Questionnaire" (EQ) und der erwarteten Noten. Aleamoni und Hexner (1980) konnten dagegen keine nennenswerte Korrelation zwischen den Ergebnissen des CEQ, der vor dem Abschlußexamen ausgefüllt wurde, und den späteren Noten ermitteln.

In einer Untersuchung von Lester (1982) mit 47 Studierenden wurden ebenfalls keine signifikante Beziehung zwischen der durchschnittlichen Gesamteinschätzung des Lehrenden/ bzw. des Kurses und den Noten festgestellt. Problematisch ist die geringe Anzahl an befragten Studierenden.

Eine mögliche Erklärung für die Korrelation zwischen externen Variablen und der Kursbewertung könnte in der Art des Items bzw. der Fragebogenkonstruktion liegen. Dies wurde bereits von Marsh an den verschiedensten Punkten erwähnt. Eine Bestätigung dafür liefern auch Aleamoni und Thomas (1980), die in ihrer Untersuchung signifikant höhere Korrelationen zwischen den drei generellen Fragen zu Lehrenden, Kurs und Inhalt feststellen konnten als im Vergleich zu einzelnen Subskalen.

Nicht auszuschließen sind spezifische Interaktionen: Gibt es Studierende, die sich stärker als andere von Noten beeinflussen lassen? Welche Persönlichkeitseigenschaften oder Denkstile sind dafür verantwortlich? Denkbar ist, daß nicht alle Studierenden sich im gleichen Maße von erhaltenen Noten beeinflussen lassen. Es liegen hier zwei Studien vor, die diesen Bereich untersuchen.

Blass (1980) ermittelte in einem Kurs der Einführung in die Psychologie eine signifikant positive Korrelation zwischen sechs der insgesamt neun Items des Bewertungsbogens. Mit einem von ihm selbst (1969) konstruierten Verfahren, teilte er daraufhin den Kurs in relativ objektiv urteilende und relativ subjektiv bewertende Studierende. Die Korrelation zwischen Note und Lehrbewertung erwies sich bei den subjektiv bewertenden Studierenden als insgesamt deutlicher als für die Gesamtgruppe, während in der relativ objektiv urteilenden Kurshälfte nur noch Korrelationen für zwei der neun Items ermittelt werden konnten.

Meredith (1982) überprüfte, inwieweit die von Eison (1981) postulierte bipolare Dimension "Learning Orientation/ Grade Orientation (LOGO)" mit Noten zusammenhängen. 9.400 Studierende aus 424 Kursen beantworteten eine 41-Item Skala am Ende des Semesters. Fünf Items korrelierten mit dem Gesamtwert der "Grade Orientation" signifikant auf dem 1%-Niveau: Erwartete Note (r=.27), Geistige Entwicklung (r=.34), Persönlichkeitsentwicklung (r=.35), Gesamtbewertung des Lehrenden (r=.48) und Gesamtkursbewertung (r=.49).

Insgesamt schloß der Autor, daß die persönliche Wichtigkeit, die die Noten für einen Studierenden haben, sowohl die Bewertung der Lehrenden als auch die Kurszufriedenheit beeinflußt. Wobei zu vermuten ist, daß die Wichtigkeit der Note und das Leistungsmotiv eng miteinander verknüpft sind (vgl. Abschnitt 2.9.2).

2.12.3 Erwartete Noten

Überraschend viele Studien erheben auch erwartete Noten bzw. stellen es Studierenden frei, erwartete oder erhaltene Noten anzugeben. Werden diese zu Kursdurchschnitten zusammengefaßt, so können diese Werte verwendet werden, um zu überprüfen, ob als "milde bewertend" eingeschätzte Lehrende ihrerseits besser bewertet werden. Korreliert man jedoch die Note einzelner Studierender mit ihrer Lehrbewertung, so ist dies nicht möglich.

Ducette und Jane Kenney (1982) verwendeten die Daten von ca. 5.900 Studierenden, die einen Fragebogen zur Evaluation der Lehre beantwortet hatten. Einige Dimensionen

der Lehrevaluation in der Studie von DuCette und Jane Kenney (1982) korrelieren positiv mit den erwarteten Noten, im extremsten Fall erklärten sie weniger als 20 Prozent der Varianz. Die stärkste Korrelation wurde konsistent in der Dimension "Kurseffektivität" gefunden. Im Vergleich dazu korrelierte die Dimension "Lehreffektivität" weitaus geringer mit den erwarteten Noten.

Diese Ergebnisse werteten DuCette und Jane Kenney als Beweis, daß ein "bias" als alleinige Erklärung für die Korrelation Note - Bewertung nicht genügt. Die Korrelation zwischen der zweiten Dimension und den Noten beurteilten sie als realistische Einschätzung der Studierenden. Die, die im Kurs mehr gelernt hätten, erwarteten bessere Noten. Die Korrelation zwischen der ersten Dimension und der studentischen Veranstaltungskritik unterstützt eher eine "bias"-Erklärung.

Diese Studie macht deutlich, daß ein "bias" entweder nicht vorliegt bzw. in dieser Versuchsanordnung kaum zu ermitteln ist. Erhebt man erwartete Noten, so muß überlegt werden, was eigentlich damit erfaßt wird. Welche Faktoren führen dazu, daß Studierende eine bestimmte Note erwarten? Es erscheint wahrscheinlich, daß mit der erwarteten Note im Grunde genommen eine ganze Reihe anderer Faktoren erfaßt werden. Diese sind u.a. vermutlich Vorkenntnisse, Motivation, der Erfahrungswert, wie man sonst in solchen Prüfungen abschneidet u.ä.. Die Ergebnisse von DuCett und Jane Kenney sind daher am sinnvollsten mit Studien zur Motivation der Studierenden im Zusammenhang zu betrachten.

Schlußfolgerung

Anhand der vorliegenden Studien über den Einfluß von Noten (erhaltene oder erwartete) auf die Lehrbewertung ist nur schwerlich zu folgern, daß strenger bewertende Lehrende häufiger schlechtere Ergebnisse erzielen, da zu vermuten ist, daß nicht alle Studien Werte ***zwischen*** *Kursen korreliert haben. Außerdem ist es möglich, daß die ermittelten Zusammenhänge auf unterliegende Faktoren, wie z.B. unterschiedliche Motivation, zurückzuführen sind.*

Aus den referierten Studien kann keine generelle Aussage abgeleitet werden. Es erscheint möglich, daß sich manche Studierende in einem begrenzten Ausmaß bei der Lehrbewertung von erhaltenen Noten beeinflussen lassen. Es besteht jedoch kein Nachweis, daß strenger bewertende Lehrende allein deshalb schlechtere Bewertungen erhalten.

Auch korrelieren erwartete oder erhaltene Noten in den Untersuchungen nur mit bestimmten Aspekten des Fragebogens zur Veranstaltungskritik. Um zu einen abschließenden Urteil zu kommen, scheinen zu dieser wichtigen und heiklen Problematik des Einflusses der Notenstrenge weitere Studien notwendig.

3 Validierung und Reliabilität studentischer Veranstaltungskritik

Wurden bisher allgemeine Einwände gegen die Validität studentischer Bewertungen betrachtet, so sind im folgenden Studien aufgeführt, die systematisch die Zuverlässigkeit und Genauigkeit studentischer Bewertungen mit vorher festgelegten und akzeptierten Kriterien vergleichen. Cohen (1981) nennt einige solcher Kriterien, an denen die Frage nach der Validität studentischer Bewertungen geknüpft werden können: "... (1) Bewertungen von Kollegen, (2) Bewertungen von Seiten der Verwaltung, (3) Selbstbewertungen, (4) Bewertung ehemaliger Studierender und (5) studentische Leistung" (Cohen, 1981, S. 283). Die von Cohen genannten Punkte 1, 3 und 4 wurde bereits weiter oben besprochen, zu Punkt 2 liegt mir keine Studie vor. Ergänzt werden soll an dieser Stelle Cohens Aufzählung durch weitere Validierungskriterien.

3.1 Lernzuwachs als Maß der Lehrqualität und -effektivität

Die Validierung an dem Ausmaß an gelerntem Wissen gilt als eines der besten, wenn nicht als das beste Kriterium, insofern sich die Lehrleistung der Dozierenden sich in einem möglichst hohen Lernzuwachs erweisen sollte.

Es erscheint nachvollziehbar, daß Lehrende, deren Aufgabe es ist, Wissen, Strukturen, Zusammenhänge etc. zu vermitteln, genau daran gemessen werden sollten, inwieweit sie ihrer Aufgabe gerecht wurden. Aufgrund der Unterschiede in der Kursorganisation zwischen den Vereinigten Staaten und Deutschland[3] stellt sich jedoch die Frage, inwieweit Lehrende hier überhaupt dafür verantwortlich gemacht werden können, was ihre Studenten und Studentinnen gelernt haben. Cohens Einwand zu dieser Art der Validierung: "Dies ist höchstens ein grober Index der Lehreffektivität, da eine Anzahl von Faktoren außerhalb der Kontrolle des Lehrenden - studentische Fähigkeit und Motivation z.B. - den Lernzuwachs beeinflussen" (Cohen, 1981, S. 281), trifft einen Teil der Problematik. Dieser wäre jedoch anhand von Vergleichen zwischen randomisierten Kursen noch kontrollierbar. D.h. die Fragestellung einer solchen Untersuchung müßte lauten: Haben die Studierenden eines Kurses beim Dozenten A weniger gelernt als im vergleichbaren Kurs von Dozentin B?

Die größere Schwierigkeit besteht in der Operationalisierung des Lernzuwachses, der auf einen bestimmten Kurs/Dozent/Dozentin zurückzuführen ist, und in der Vergleichbarkeit und Wertigkeit dieses Lernzuwachses. Hat zum Beispiel Dozent B, der seine Studierenden dazu gebracht hat, eine Unmenge von Fakten zu lernen, mehr erreicht, als Dozentin A, die den Studierenden vor allem Strukturzusammenhänge vermittelt hat, oder Dozent C, der in seiner Vorlesung das Interesse in sein Fachgebiet nachhaltig wecken konnte?

Cohen (1981) führte eine Meta-Analyse von 41 Studien durch. Eine der wichtigsten Voraussetzungen für die Auswahl der Studien war, daß die Daten Vergleiche zwischen

[3] In den Vereinigten Staaten sind Kurse häufig kleiner, stärker strukturiert und arbeitsaufwendiger als in Deutschland.

Klassen enthielten. Folgende Ergebnisse konnten vom Autor als Einflußfaktoren auf die Kursleistung (meist Klausurergebnisse oder Note) ermitteln werden: Große Effekte hatten Geschicklichkeit in Lehre und Kursstruktur, mittlere Effekte wurden bei der Beziehung Lehrende/Studierende, Rückmeldung und Bewertung gefunden. Die Kursschwierigkeit hatte keine Bedeutung. Insgesamt stellte Cohen fest: "Wir können relativ sicher sein, daß Dimensionen des Gesamtkurses und des Lehrenden ziemlich eng mit der Leistung des Studierenden zusammenhängen" (Cohen, 1981, S. 298).

Untersucht wurde dann, inwieweit ein Zusammenhang zwischen Charakteristika der einzelnen Studien und den gefundenen Effekten in 67 Kursen besteht. Drei Variablen korrelierten signifikant mit der Effektgröße: Kontrolle möglicher Einflußfaktoren auf die Kursbewertung, Zeitpunkt der Kursbewertung (vor oder nach Bekanntgabe der Noten) und die Erfahrung des/der Lehrenden. Diese drei Variablen machten gemeinsam 31 Prozent der Varianz in der Korrelation zwischen studentischer Veranstaltungskritik (verwandt wurde nur die Gesamtbewertung der Lehrenden) und der Leistung aus.

Die Korrelation zwischen Lehrbewertung und Noten betrug .85, wenn die Noten bekannt, und .38, wenn die Noten nicht bekannt waren. Dies kann als relativ eindeutiger Nachweis gelten, daß nicht nur vorheriges Interesse, sondern auch das direkte Wissen um die eigene Note die studentische Veranstaltungskritik beeinflußt. Der Autor zieht folgenden Schluß: "Die vorliegende Meta-Analyse beinhaltet eine große Unterstützung für die Validität studentischer Bewertung als Maß der Lehreffektivität" (ebd., S. 300).

Problematisch dabei erscheint nur, daß bei einer Definition des Lernzuwachses über Noten bzw. Ergebnisse von Abschlußklausuren es kaum möglich ist, zwischen Effekten aufgrund von Noten und Effekten aufgrund von Gelerntem zu trennen. Daher kann die Schlußfolgerung von Cohen auch interpretiert werden zugunsten des Vorwurfs, daß Lehrende, die weniger streng bewerten, auch selbst besser bewertet werden. Zwar analysierte der Autor nur Studien, die zwischen Kursen verglichen, dennoch kann gerade da ein derartiges "Bias" vermutet werden.

Um die Schwierigkeit der Schlußfolgerung des Lernzuwachses anhand von Noten zu umgehen und diesen präziser zu fassen, wurden von Prosser und Trigwell (1991) Daten von 122 Studierenden aus 11 Kursen in "Kommunikation in der Pflege" sowohl innerhalb des Kurses als auch zwischen den Kursen verglichen. Erhoben wurde:

a) ihre "akademische" Leistungsfähigkeit, mit Hilfe des Higer School Certificate Score,
b) quantitative Lernfortschritte,
c) Fragebogen zur Herangehensweise an das Lernen (tiefes Verständnis anstreben, Zusammenhänge erkennen, oberflächliches Verständnis aufbauen),
d) zwei Gesamtbewertungen zu Kurs und Dozent/ Dozentin und
e) qualitative Lernfortschritte (Aufsatz zum Kursziel), gemessen an fünf Kategorien von vorstrukturelles bis abstraktes Verständnis.

Bei der Auswahl der Datenklassen berufen sich die Autoren auf neuere Studien, die verstärkt die Qualität und weniger das Ausmaß des Gelernten in den Vordergrund stellen.

Ergebnisse: Der Vergleich zwischen den Kursen erbrachte, daß studentische Bewertung eines Kurses signifikant auf dem 1%-Niveau mit der Herangehensweise "tiefes Verständnis anstreben" und mit qualitativen Lernfortschritten korrelierte. Der Vergleich innerhalb der Kurse ergab positive Korrelationen zwischen der Gesamtkursbewertung und der Herangehensweise "tiefes Verständnis anstreben" und eine negative Korrelation zwischen der Gesamtbewertung des Kurses und der Note, beides auf dem 5%-Niveau signifikant. Die Autoren gehen deshalb davon aus, daß eine Validierung von Lehrbewertungen mit Hilfe von qualitativen Verfahren erfolgen müßte.

In der Arbeit von Katherine van Giffen (1990) korrelierte die Einschätzung von Lehreffektivität und die Selbsteinschätzung der Lernleistung durch die Studierenden positiv ($r = .789$).

Alle drei Studien machen die Schwierigkeit der präzisen Definition des Kriteriums "Lernzuwachs" deutlich. Es gibt Anhaltspunkte, daß Studierende, die mehr gelernt haben, auch ihre Lehrenden besser bewerten. Wie Prosser und Trigwell (1991) jedoch zeigen konnten, kann aber die genauere Klärung dieser Tendenzen nur mit Hilfe aufwendiger Untersuchungen gelingen, die verschiedene Aspekte des "Lernzuwachses" beleuchten würden.

3.2 Übereinstimmung zwischen Selbstbeurteilungen von Lehrenden und studentischer Veranstaltungskritik

Weitaus leichter als die Folgen von "Lernzuwachs" ist es, die Übereinstimmung zwischen der Selbstbeurteilung von Lehrenden und der Kursbewertung durch Studierende zu überprüfen. Diese Übereinstimmung hat eine größere praktische als wissenschaftliche Relevanz, da es zwar eine gute Methode ist, Lehrende davon zu überzeugen, daß studentische Veranstaltungskritik nützliche Ergebnisse liefert, jedoch die Frage nicht geklärt ist, inwieweit Selbstbeurteilungen ihrerseits valide sind (z.B. in Zusammenhang mit dem Kriterium Lernzuwachs).

Eine der wohl größten Validitätsuntersuchungen in diesem Bereich ist von Marsh (1982b) durchgeführt worden. Er nutzte dafür die Multitrait-Multimethode-Methode (MTMM) und untersuchte die konvergente und diskriminante Validität zwischen den Selbstbeurteilungen und den Lehrbewertungen. 181 Lehrende mit 329 Kursen füllten Bögen zur Lehrevaluation aus, wobei ihnen explizit mitgeteilt wurde, daß ihre Bewertungen sich durchaus von denen der Studierenden unterscheiden könnten. Der gleiche Fragebogen wurde auch von den Studierenden bewertet.

Die Ergebnisse der Korrelationsuntersuchungen ergaben gute Übereinstimmungen. Sie lagen bei den meisten Faktoren bei $r= .45$, was u.a. als Beweis für konvergente Validität diente. Dagegen lag die Korrelation verschiedener Faktoren bei $r= .02$ (Nachweis für di-

vergente Validität). Mit einer durchgeführten ANOVA konnte jedoch auch ein gewisser Methoden- bzw. Halo-Effekt ermittelt werden, der vor allem an den studentischen Werten zu bemerken war.

Nach dieser umfänglichen Studie läßt sich folgern: Erstens unterscheiden Lehrende und Studierende in nahezu übereinstimmender Weise die verschiedenen Elemente der Lehrqualität (Dimensionen oder Faktoren). Zweitens weisen ihre Urteile eine hohe Übereinstimmung auf.

3.3 Übereinstimmung zwischen Studierenden und Beobachtern/Beobachterinnen

Ähnlich der Übereinstimmung zwischen Selbstbewertung der Lehrenden und der Bewertung durch Studierende ist die Ermittlung zwischen der Lehrbewertung von Studierenden und anderen Beobachtern/Beobachterinnen angelegt. Sie hat im Vergleich zum vorherigen Kriterium den Vorteil, daß Werte von Beobachtern/Beobachterinnen mit Hilfe von präzise festgelegten Kriterien Beobachterschulung etc. leichter zu validieren sind.

McKeachie und Lin (1978) führten die einzige dazu vorliegende Studie durch. 20 Lehrende mit Kursen von durchschnittlich 25 Studierenden wurden sechs mal in vierzehn Semesterwochen von trainierten Beobachtern in Beobachtungkategorien eingeteilt. Zwei Kategorien wurden für die Studie von McKeachie und Lin (1978) weiterverwendet: Wärme und Akzeptanz. Studierende füllten eine Lehrevaluation "Student Perception of Teaching and Learning" aus.

Die drei Items des Faktors "Rapport" (erlaubend, freundlich, fordert zu Kritik auf) wurde mit der Beobachtereinschätzung von "Wärme und Akzeptanz" korreliert. Eine signifikante Korrelation ergab sich zwischen dem Fragebogenitem "freundlich" und der beobachteten "Akzeptanz". Auch korrelierte der Gesamtfaktor "Rapport" signifikant mit der Akzeptanz. Die Autoren kommen zu dem Schluß: "Diese Studie leistet eine gewisse empirische Unterstützung der Annahme, daß studentische Bewertung auf dem Verhalten der Lehrenden beruht" (McKeachie und Lin, 1978, S. 46).

Eine solche Schlußfolgerung erscheint jedoch zu weitreichend. Die Kategorien der geschulten Beobachter und Beobachterinnen ließen sich alle auf einer "Beziehungsebene" anordnen, Kategorien einer didaktisch-sachlichen Ebene fehlten völlig. Dennoch läßt sich nach dieser Untersuchung feststellen, daß studentische Bewertung, wenn sie mit relativ konkreten Items ermittelt wird (wie dies in der vorliegenden Studie der Fall war), zumindest auf einer "Beziehungsdimension" auf dem Verhalten von Lehrenden beruht.

3.4 Verbesserung der Lehrbewertung nach Veränderung kritischer Aspekte

Eine besonders elegante Art der Validierung ist die der Messung von Veränderung der Lehrbewertungen, nachdem kritisch bewertete Aspekte der Lehre verbessert wurden. Auch zu diesem wichtigen Punkt liegt allerdings nur eine Studie vor.

Zwei Studienjahre eines vierwöchigen Medizinkurses wurden von Lynch, Tamburrino und Nagel (1989) mit einem Fragebogen evaluiert. Drei Problembereiche wurden identifiziert und Verbesserungen durchgeführt.

Die Bewertungen der Kurse in den folgenden drei Semestern wurden erhoben. Signifikante Änderungen der Ergebnisse wurden für die modifizierten Bereiche ermittelt. Alle anderen Ergebnisse unterschieden sich nicht signifikant von den ursprünglichen. Da sich die Bewertung nur bei den modifizierten Bereichen änderte, kann dies als ein Zeichen externer Validität der studentischen Veranstaltungskritik betrachtet werden.

3.5 Weiterempfehlung bzw. Besuch eines weiteren Kurses

Eine weitere Möglichkeit, Validität zu erfassen - wenn als eines der Lehrziele "Interesse am entsprechenden Fach zu wecken" gilt - ist die Ermittlung, ob Studierende den Kurs weiterempfehlen und weitere Kurse in der entsprechenden Disziplin zu besuchen planen. (Letzteres ist etwas problematisch, wenn man das Ausmaß an Veranstaltungen bedenkt, die verpflichtend zu belegen sind.)

In der Untersuchung von Bruton und Crull (1982) stuften 1.314 Studierende ihren Kurs in eine der fünf Kategorien von weit unter- bis weit überdurchschnittlich ein; sie gaben zudem eine Gesamtbewertung für den Lehrenden ab. Mit Hilfe einer Pfadanalyse wurden Zusammenhänge analysiert. Obwohl der Kurs explizit unabhängig von den Lehrenden zu bewerten war, war der Dozent/die Dozentin doch der wichtigste Faktor der Kursbewertung.

Ergebnis war, daß Studierende, die kleinere Veranstaltungen besucht haben, eher bereit sind, einen weiteren Kurs derselben Disziplin zu besuchen. Dies weist darauf hin, daß die "Kursgröße" u.U. mit der Gesamtbewertung "Dozent/Dozentin" interagiert. Um genauere Aussagen zu ermöglichen, müßten Kursvergleiche so durchgeführt werden, daß der Varianzanteil bei einer Weiterempfehlung, der auf den/die Lehrende zurückzuführen ist, ermittelt werden kann.

3.6 Reliabilität studentischer Veranstaltungskritik

Eine Definition der Reliabilität für studentische Veranstaltungskritik nennen Overall und Marsh (1980): "Reliabilität, wie sie sich in den meisten früheren Studien darstellt, ist gefaßt als die relative Übereinstimmung zwischen den Beurteilungen verschiedener Studierender innerhalb derselben Klasse - unter der Annahme, daß jede Varianz des ein-

zelnen Studierenden zufällig ist und als Fehlervarianz betrachtet werden kann" (Overall und Marsh, 1980, S. 324).

Bledsoe (1978) ermittelte eine Reliabilität von .93 für Klassen und .96 für Items in 42 Statistikkursen und damit einen hohen Grad an Übereinstimmung und Genauigkeit.

Centra (1980) berechnet die Reliabilitäten für die Bewertungsergebnisse von 30 Lehrenden, variiert dabei die Anzahl der Studierenden und kommt zu folgenden Ergebnissen: "Bei 10 Bewertern war der Reliabilitätskoefizient für die meisten Items um .70 und um .78 für die Gesamtbewertung der Lehrenden. Die geschätzte Reliabilität für 15 Studierende war bei .80 für die Items und für 20 Bewerter waren die Reliabilität bei .90. Diese durchschnittlichen Reliabilitäten waren nahe an .90. Diese Durchschnittsreliabilitäten sind ähnlich wie die in anderen Studien berichteten" (Centra, 1980, S.27).

Viele Instrumente, die in anderen Arbeiten untersucht wurden, kamen zu vergleichbaren Ergebnissen. Es ist zu vermuten, daß bei einer entsprechend sorgfältigen Erstellung solcher Instrumente die Reliabiltät kein größeres Problem darstellen dürfte. Es erscheint somit möglich Fragebögen zu gestalten, die das definierte Kriterium (Lehrqualität aus studentischer Sicht) in ausreichend präzisem Maße erfassen.

4 Zusammenfassung und Folgerungen

Die gesichteten Studien zur studentischen Veranstaltungskritik - insgesamt etwa 100 aus amerikanischen Hochschulen - lassen sich verschiedenen Einwänden und Befürchtungen hinsichtlich der Validität von Bewertungen der Lehre durch Studierende zuordnen. Auf zwölf derartige Argumente wurde eingegangen, um zu klären, inwieweit sie berechtigt sind. Dabei handelt es sich einerseits um grundsätzlichere Vorbehalte, welche den Studierenden eine Urteilskompetenz überhaupt absprechen (z.B. Unerfahrenheit, mangelnde Sachkenntnis, falsche Kriterien), andererseits um Hinweise auf Faktoren, welche die studentischen Bewertungen verzerren oder beeinflussen könnten (wie Kursschwierigkeit, Motivation oder Notenstrenge).

Darüberhinaus wurde auf Studien eingegangen, welche die studentische Veranstaltungskritik über den Vergleich mit anderen Kriterien (z.B. Lernzuwachs, Verbesserung der Lehrbewertung, Weiterempfehlung) oder mit anderen Beurteilern (Lehrende bzw. Fremdbeobachter/innen) hinsichtlich ihrer Validität untersucht haben.

Die gesichteten Studien und ihre Ergebnisse lassen die Schlußfolgerung zu, daß es möglich ist, entsprechend gestaltete Fragebögen zur Bewertung der Lehrenden einzusetzen. Sie besitzen eine Validität, die es zuläßt, sie auch für den Vergleich von Lehrleistungen verschiedener Lehrender zu verwenden. Diese Bilanz sei nachfolgend für die verschiedenen Argumente und Faktoren zusammenfassend begründet und ergänzt.

Insgesamt kann gesagt werden, daß Studierende konsistente Urteile in der Lehrbewertung treffen können, die auch über Monate (oder Jahre) stabil bleiben. Von einer "mangelnden Distanz zum Bewertungsobjekt" (Aleamoni, 1987) kann daher nicht ausgegangen werden. Der Varianzanteil einer Bewertung, der auf den entsprechenden Kurs zurückzuführen ist, ist weitaus geringer, als der Varianzanteil, der auf den entsprechenden Lehrenden zurückzuführen ist. Zwar gewichten Studierende verschiedene Aspekte der Lehre anders als Lehrende, jedoch wirkt sich das nicht auf die Bewertung mit Hilfe eines entsprechenden Fragebogens aus. Eine gute Übereinstimmung der Bewertung von Lehrenden durch Kollegen/ Kolleginnen und Studierende konnte ermittelt werden.

Der Verdacht, daß studentische Veranstaltungskritik eher einem bloßen Beliebtheitswettbewerb gleichkommt, konnte nicht bestätigt werden. Es scheint möglich in diesem Bereich Instrumente zu entwickeln, die reliabel und valide sind, und die die Gefahr möglicher "Halo"-Effekte minimieren oder zu kontrollieren erlauben.

Von der Einwirkung möglicher Einflußvariablen konnten keine direkte Ergebnisverzerrung aufgrund des Geschlechts der Lehrenden ermittelt werden. Es gibt jedoch Hinweise, daß Studentinnen einige Lehraspekte anders beurteilen als Studenten. Ebenfalls keine Auswirkungen haben Kursaufwand und -schwierigkeit. Die Angst der "Bestechlichkeit" von Studierenden erscheint somit unbegründet. Auch erhalten "milder" benotende Lehrende insgesamt keine besseren Bewertungen als "strenger" benotende. Keinen direkten Einfluß scheinen auch Persönlichkeitseigenschaften sowohl von Lehrenden als

auch von Studierenden auf die Kursbewertung zu haben. Ein mittelbarer Einfluß über andere Faktoren ist jedoch möglich.

In den vorhandenen Studien wurden keine Einflüsse von Semesterzahl und Kursgröße gefunden. Diese Faktoren müßten jedoch in Deutschland unter den hiesigen Bedingungen noch einmal genauer untersucht werden. Die wenigen deutschen Studien lassen vermuten, daß Vorlesungen insgesamt schlechter bewertet werden als Seminare.

Einen eindeutigen Einfluß hat der Faktor "Motivation" auf die Lehrbewertung. Dieser scheint jedoch kalkulierbar zu sein. Möglich ist auch, daß sich erhaltene Noten bei bestimmten Studierenden auf ihre spätere Lehrbewertung auswirken.

Zum Einfluß externer Faktoren auf die studentischen Bewertungen liegen wohl die meisten Studien vor: Es gibt wenig, von dem nicht vermutet wurde, daß es einen Einfluß auf die Lehrbewertung haben könnte. Bei dieser Art von Studien werden häufig entweder nur einzelne Aspekte untersucht oder aber es werden in großen Untersuchungen eine beträchtlichen Anzahl externer Variablen mit der Lehrbewertung korreliert.

Eine Ausnahme von dieser Herangehensweise ist die Studie von Scott (1977). Scott erhob Daten von 3.625 Studierende aus 195 Kursen (138 Lehrende), um anhand einer Liste häufiger Klagen von Lehrenden erst einmal zu ermitteln, welche Rahmenbedingungen u.U. zu einer niedrigeren Lehrbewertung führen könnten. Die genannten Schwierigkeiten waren: der Kurs wurde erstmalig unterrichtet, der Stoff sei zu schwierig, als daß man ihn adäquat im Kurs präsentieren könnte, die Klasse sei zu groß, der Kurs wurde nicht im eigenen Fachschwerpunkt unterrichtet, der Kurs war neu konzipiert, es wurden methodische Neuerungen eingeführt und eine inhaltliche Überarbeitung wurde versucht.

Bei 83 Kursen wurde von Seiten der Lehrenden mindestens ein "mildernder Umstand" aufgeführt. Korrelationsstudien ergaben, daß nur eine hohe Klasssengröße einen signifikant (negativen) Einfluß auf die Durchschnittsbewertung des Kurses hatte. Der Autor sieht keinen statistischen Beweis für externe Einflußgrößen.

An dieser Studie erscheint die Herangehensweise interessanter als das Ergebnis. Denn auf diese Art und Weise können Lehrbewertungen eine höhere soziale Validität erhalten und mögliche Einflüsse auf eine niedrigere Lehrbewertung aus der Sicht der Betroffenen gesammelt werden.

Es gibt die These, die von Marsh in verschiedenen Studien betont wird, daß sich der Einfluß externer Variablen nur auf bestimmte Teile eines Fragebogens beschränkt, nämlich auf diejenigen, die inhaltlich mit dieser Variablen zusammenhängen. Dieser Punkt soll an dieser Stelle noch einmal aufgegriffen werden.

Patricia B. Elmore und Pohlmann (1978) ermittelten, daß Items, die am besten aufgrund von Eigenschaften der Lehrenden und Studierenden vorhergesagt werden konnten, alle auf einer "Beziehungsdimension" lagen. Daß andere Dimensionen der Lehrbewertung weniger oder gar nicht betroffen waren, werteten die Autorin und der Autor als Hinweis auf die diskriminante Validität.

Um konkretere Aussagen zu dem Einfluß der "externen Variablen" zu treffen, sollten künftige Analysen nicht nur global ausgelegt sein. Notwendig wäre, daß mit Fragebogen, die bereits als reliabel und valide gelten und multidimensional gestaltet sind, detailliert versucht wird zu ermitteln, welche Variablen welche Faktoren beeinflussen. Dabei ist es wichtig, nicht nur mit Vergleichen innerhalb der Kurse, sondern auch mit Kursdurchschnitten zu arbeiten, da damit viele "response biases" herausfallen. Zudem sind es Durchschnitte, Verteilungen und Profilvergleiche, die üblicherweise als Grundlage der Lehrbewertung verwandt werden.

Allerdings beinhaltet der Methodenstreit, ob eine vergleichende Lehrbewertung auf der Grundlage studentischer Urteile möglich ist oder nicht, auch die Frage nach dem politischen Willen, solche Instrumente einsetzen zu wollen. Die Frage nach der methodischen Validität kann auch als "Nebenkriegsschauplatz" von Abwehr und Verweigerung betrachtet werden, vor allem wenn die Anforderungen so hoch geschraubt werden, wie sie bei der Beurteilung von Studierenden durch Lehrende diesen nicht abverlangt werden. So vergleicht Weiss (1991) die Diskussion um die Lehrbewertung mit der Bewertung an Schulen: "In Anbetracht der großen Bedeutung für das Schicksal der Beurteilten müssen an die Leistungsbeurteilungen in den Schulen strenge Maßstäbe angelegt werden. Sie müßten objektiv, zuverlässig und gültig sein ... Diese 'Gütekriterien' werden in den Schulen nur unzureichend erfüllt ..." (S. 36).

Ähnlich sieht es auch Aleamoni: "Was wäre, wenn der Spieß umgedreht werden würde und die Bewertung der Studierenden in Zweifel gezogen würde? ... Wie viele Nachweise könnten wir vorzeigen, um die Studierenden zu überzeugen, daß die Beurteilungen ihrer Leistungen ausschließlich auf objektiven Kriterien beruhen? ... Wie viele Nachweise könnten wir vorzeigen, um die Studierenden zu überzeugen, daß Bewertungen ihrer Leistungen im Kurs sowohl valide als auch reliabel sind?" (Aleamoni 1987, S. 29-30).

Eine angemessene Erfassung von Lehrleistungen durch studentische Veranstaltungskritik ist mit mehrdimensionalen Fragebogen, welche die relevanten Sachverhalte ansprechen, möglich. In den Vereinigten Staaten und anderen Ländern, in den letzten Jahren vermehrt auch in Deutschland, sind derartige Instrumente zur Evaluation von Lehrveranstaltungen und für Lehrberichte entwickelt und geprüft worden. Sie genügen Ansprüchen an die Validität und Reliabilität von Messungen im sozialwissenschaftlichen Bereich, vermögen sie zum Teil sogar zu übertreffen. Deshalb ist ihre Verwendung im Rahmen der vergleichenden Lehrevaluation und Lehrberichterstattung berechtigt. Einwände können sich jedenfalls nicht auf die unzureichende Validität studentischer Urteile und Veranstaltungskritiken berufen. Denn keines der Argumente dagegen erwies sich nach eingehender Prüfung empirischer Studien dazu letztlich als stichhaltig.

Literatur

Abrami, P. C.; Perry, R. P. und Leventhal, L. (1982). The Relationship Between Student Personality Characteristics, Teacher Ratings, and Students Achievement. In: Journal of Educational Psychology, Bd. 74, S. 111-125

Aleamoni, L. M. (1987). Typical Faculty Concerns About Student Evaluation of Teaching. New Direction for Teaching and Learning, Bd. 31, S. 25-31

Aleamoni, L.M. und Hexner, Pamela (1980). A Review of the Research on Student Evaluation and a Report on the Effect of Different Sets of Instructions on Student Course and Instructor Evaluation.. Instructional Science, Bd. 9, S. 67-84

Ames, R. und Lau, S. (1979). An Attributional Approach to the Validity of Student Ratings of Instruction. Contemporary Educational Psychology, Bd. 4, S. 26-39

Arkin, R. M. und Maruyama, G. M. (1979). Attribution, Affect, and College Exam Performance. Journal of Educational Psychology, Bd. 71 (1), S. 85-93

Ashton, R. H. (1975). Correlates of Ratings of Teaching Effectiveness Gordon´s Survey of Interpersonal Values. Psychological Reports, Bd. 36, S.890

Beatty, M. J. und Zahn, C. J. (1990). Are Student Ratings of Communication Instructors Due to `Easy´ Grading Practices?: An Analysis of Teacher Credibility and Student-Reported Performance Levels. Communication Education, Bd. 39, S. 275-282

Blass, T. (1980). What Do Positive Correlations Between Student Grades and Teacher Evaluation Mean? Teaching of Psychology, Bd. 7 (3), S. 186-187

Bledsoe, J. C. (1978). Insigt into one´s Own Teaching: Stability of Students´ Evaluations Across Classes. Psychological Reports, Bd. 42, S. 1071-1074

Blount, H. P., Stallings, W. M. und Gupta, V. G. (1978). The Effects of Different Instructions on Student Ratings of Universitiy Courses and Teachers. Journal of Educational Research, Bd.71, S.149-152

Brandenburg, D. C., Slinde, J. A. und Batista, E. E. (1977). Student Ratings of Instruction: Validity and Normative Interpretation. Research in Higher Education, Bd. 7, S. 67-78

Bray, J. H. und Howard, G. S. (1980). Interaction of Teacher and Student Sex and Sex Role Orientations and Student Evaluations of College Instruction. Conteporary Educational Psychology, Bd. 5, S. 241-248

Brown, D. L. (1976). Faculty Ratings and Student Grades: A University-wide Multiple Regression Analysis. Journal of Educational Psychology, Bd. 68 (5), S. 573-576

Bruton, B. T. und Crull, Sue R. (1982). Causes and Consequences of Student Evaluation of Instruction. Research in Higher Education, Bd. 17 (3), S. 195-206

Bundesministerium für Bildung und Wissenschaft (Hrsg.) (1993). Qualität und Wettbewerb in der akademischen Lehre - Zwischenbilanz zum Modellprogramm des Bundesministeriums für Bildung und Wissenschaft. Reihe: Bildund - Wissenschaft - Aktuell, Bd. 5, Bonn

Centra, J. A. (1977). Student Ratings of Instruction and Their Relationship to Student Learning. American Educational Research Journal, Bd. 14, S. 17-24

Centra, J. A. (1980). Determining Faculty Effectiveness. San Francisco: Jossey-Bass Publishers

Chacko, T. I. (1983). Student Ratings of Instruction: A Function of Grading Standards. Educational Research Quarterly, Bd. 8 (2), S. 19-25

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. Review of Educational Research, Bd. 51 (3), S. 281-309

Coleman, J. und McKeachie, W.J. (1981). Effects of Instructors/ Course Evaluations on Student Course Selection. Journal of Educational, Bd. 73 (2), S.224-226

Cooper, Pamela J.; Steward, Lea P. und Gudykunst, W. B. (1982). Relationship with Instructor and other Variables Influencing Student Evaluations of Instruction. Communication Quarterly, Bd. 30 (4), S.308-314

Cranton, Patricia A. und Schmith, R. A. (1986). A New Look at the Effect of Course Characteristics on Student Ratings of Instruction. American Educational Research Journal, Bd. 23 (1), S.117-128

Daniel, H.-D. (1994). Hörerbefragung an der Universität Mannheim: Konzeption, Erhebung, Auswertung. Empirische Pädagogik, 1994, 8 (2), S. 109-130

Daniel, H. D.,Thoma, Michaela& Bandilla, W. (1995) Das Modellprojekt „Evaluation der Lehre an der Universität Mannheim. Teil 1: Planung und Durchführung von Befragungen in Lehrveranstaltungen. In: Mohler, P.(Hrsg.) Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung. Münster: Waxmann

Daniel, H.D. und Hornbostel (1993). Provozierende Fragen. SPIEGEL-SPEZIAL, 1993, S.157

Danielsen, A. L. und White, R. A. (1976). Some Evidence on the Variables Associated with Student Evalution of Teachers. The Journal of Economic Education, Bd. 7, S.117-119

DER SPIEGEL, (1994). Famose Ergebnisse. Bd.15, S.55-57

Drews, D. R., Burroughs, J. W. und Nokvich, DeeAnn (1987). Teacher Self-Ratings as a Validity Criterion for Student Evaluations. Teaching of Psychology, Bd. 14(1), S. 23-25

Drummond, R. J. und McIntire, W. G. (1977). The Role of Cognitive Style in Student Evaluation of Instruction. College Student Journal, Bd. 11, S. 220-223

DuCette, J. und Kenney, Jane (1982). Do Grading Standards Affect Student Evaluations of Teaching? Some New Evidence on an Old Question. Journal of Educational Psychology, Bd. 3, S. 308-314

Elmore, Patricia B. und LaPointe, Karen A. (1975). Effect of Teacher Sex, Student Sex, and Teacher Warmth on the Evaluation of College Instructors. Journal of Educational Psychology, Bd. 67 (3); S. 368-374

Elmore, Patricia B. und Pohlmann, J. T. (1978). Effect of Teacher, Student, and Class Characteristics on the Evaluation of College Instructors. Journal of Educational Psychology, Bd.70, S.187-192

Firth, M. (1979). Impact of Work Experience on the Validity of Student Evaluaton of Teaching Effectiveness. Journal of Educational Psychology, Bd. 71, S. 726-730

Gaffuri, Adriane, Wrench, D., Karr, C. und Kopp, R. (1982). Exploring Some Pitfalls in Student Evaluation of Teaching. Teaching of Psychology, Bd. 9 (4), S.229-230

Giesen,H. und Jansen, R. (1983). Universität aus der Sicht ihrer Studenten. Zeitschrift für Entwicklungspsychologische und Pädagogische Psychologie, Bd.15, S. 222-223

Giffen, Katherine van (1990). Influence of professor gender and perceived use of humor on course evaluations. Humor, Bd. 3(1), S.65-73

Gillmore, G. M. (1977). How Large is the Course Effect? Research in Higher Education, Bd. 7, S. 187-189

Gralski, H. O. (1992). Provokative Evaluation - der falsche Weg zur Verbesserung der Lehre. In: Grühn, D. und Gattwinkel, H. (Hrsg.) Evaluation von Lehrveranstaltungen. S. 35-41, Berlin. Zentrale Universitäts-Druckerei

Grottian, P. (1992). Der geschlossene Vortrag - ein Lehrstück für die Pervetierung der Lehrbewertung. In: Grühn, D. und Gattwinkel, H. (Hrsg.) Evaluation von Lehrveranstaltungen. S. 27-33, Berlin. Zentrale Universitäts-Druckerei

Heger, M. (1991). Lehrveranstaltungskritik als angewandte Hochschuldidaktik. Zeitschrift für Hochschuldidaktik, Bd. 15, S.43-69

Hofman, J. E. und Kremen, Liya (1980). Attitudes Toward Higher Education and Course Evaluation. Journal of Educational Psychology, Bd. 72 (5), S. 610-617

Hofman, J. M. (1988).Studienmotivation und Veranstaltungsbeurteilung. Psychologie, Erziehung, Unterricht, Bd. 35, S.119-126

Ipsen, D. & Portele, G. (1976). Organisation von Forschung und Lehre an westdeutschen Hochschulen. Serie: Hochschulplanung. München: Verlag Dokumentation

Keil, W. & Piontkowski, U. (1973).Strukturen und Prozesse im Hochschulunterricht. Weinheim, Basel: Beltz Verlag

Kovac, Robert (1976). Personality Correlates of Faculty and Course Evaluations. Research in Higher Education, Bd. 5, S. 335-344

Krieger, W. (1992). Evaluation von Hochschulunterricht. In: Grühn, D. und Gattwinkel, H. (Hrsg.) Evaluation von Lehrveranstaltungen. S. 45-59, Berlin. Zentrale Universitäts-Druckerei

Kriz, J. (1993). Wie gut sind unsere Universitäten?. Stern, 1993, Bd.16, S.171-184,

Kromrey, H., (1994), Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: Studien zur Hochschullehre. Münster: Waxman-Verlag

Lester, D. (1982). Students´ Evaluation of Teaching and Course Performance. Psychological Reports, Bd. 50, S.1126

Linzer Schwartz, Lita (1980). Criteria for Effective University Teaching. Improving College and University Teaching, Bd. 28 (3), S. 120-123

Lynch, D. J.; Tamburrino, M. und Nagel, R. (1989). Students´ Reactions as Guide to Course Revision. Psychological Reports, Bd. 65, S. 989-990

Marsh, H. W. (1982). The Use of Path Analysis to Estimate Teacher and Course Effects in Student Ratings of Instructional Effectiveness. Applied Psychological Measurement, Bd.6 (1), S.47-59

Marsh, H. W. (1982b). Validity of Students´ Evaluation of College Teaching. A Multitrait-Multimethod Analysis. Journal of Educational Psychology, Bd. 74 (2), S. 264-279

Marsh, W. und Ware Jr., J. E. (1982). Effects of Expressiveness, Content Coverage, and Incentive on Multidimensional Student Rating Scales: New Interpretation of the Dr. Fox Effect. Journal of Educational Psychology, Bd. 74(1), S. 126-134

McGaghie, W. C. (1975). Student and Faculty Ratings of Instruction: Another Look. Journal of Medical Education, Bd. 50, S. 387-389

McKeachie, W.J. und Lin, Y.G. (1978). A Note on Validity of Student Ratings of Teaching. Educational Research Quarterly, Bd. 4 (3), S. 45-47

Meier, R. S. und Feldhusen, J. F. (1979). Another Look at Dr. Fox: Effect of Stated Purpose for Evaluation, Lecturer Expressiveness, and Density of Lecture Content on Student Ratings. Journal of Educational Psychology, Bd. 71 (3), S. 339-345

Meredith, G. M. (1981). Focus-Scan Learning Strategy Correlates of Students´ Appraisal of Instruction. Perceptual and Motor Skills, Bd. 53, S. 620

Meredith, G. M. (1981b). Preferred Lenght of Scales for Students´ Evaluation of Instruction. Perceptual and Motor Skills, Bd. 53, S.490

Meredith, G. M. (1982). Grade-related Attitude Correlation of Instructor/ Course Satisfaction among College Students. Pschological Reports, Bd. 50, S. 1142

Meredith, G. M. und Schmitz, E. D. (1986). Student-taught and Faculty-taught Seminars in Undergraduate Education: Another Look. Perceptual and Motor Skills, Bd. 62, S. 593-594

Ministerium für Wissenschaft und Forschung NRW (Hrsg.), (Mai 1992). Aktionsprogramm Qualität der Lehre, Abschlußbericht, zweite erweiterte Auflage, Düsseldorf

Mishra, S. P. (1979). The Use of Instructors´ Self-Selected Items in Evaluationg Teaching Effectiveness. Journal of Psychology, Bd. 102. S. 173-177

Mohler, P. (1995). Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung. Münster: Waxmann

Morstain, B. R. (1977). Relationship of Student and Instructor Educational Orentations wirh Course Ratings. Journal of Educational Psychology, Bd. 69 (4), S. 388-398

Naftulin, D.H., Ware, J.E. und Donelly, F.A. (1973). The Dr. Fox Lecture: A Pardigm of Educational Seduction. Journal of Medical Psychology, Bd. 48, S. 630-635

Overall, J. U. und Marsh, H. W. (1980). Students´ Evaluations of Instruction: A Longitudinal Study of Their Stability. Journal of Educational Psychology, Bd. 72 (3), S. 321-325

Petchers, Marcia K. und Chow, J. C. (1988). Sources of Variation in Students´ Evaluation of Instruction in a Graduate Social Work Program. Journal of Social Work Education, Bd. 1, S. 35-42

Preißer, R. (1991). Bewertung der Qualität der Lehre an der Technischen Universität Berlin. In: Webler, W. D. und Otto, H. U. (Hrsg.) Der Ort der Lehre in der Hochschule.Lehrleistungen, Prestige und Hochschulwettbewerb. Weinheim: Deutscher Studienverlag

Preißer, R. (1992). Verwirklichungsbedingungen der Evaluation der Lehre und der Verbesserung der Lehre - Konsequenzen aus den bisherigen Erfahrungen mit Lehrveranstaltungskritik. In: Grühn, D. und Gattwinkel, H. (Hrsg.) Evaluation von Lehrveranstaltungen. S. 197-217, Berlin. Zentrale Universitäts-Druckerei

Prosser, M. und Trigwell, K. (1991). Student Evaluations of Teaching and Courses: Student Learning Approaches and Outcomes as Critera of Validity. Contemporary Educational Psychology, Bd. 16, S. 293-301

Richardson, M. D. (1978). A Study of Class Level Course Evaluation Trends. College Student Journal, Bd. 12, S. 132-134

Romeo, Felicia F. und Weber, W. A. (1985). An Examination of Variables which Influence Student Ratings of University Faculty. College Student Journal, Bd. 19, S. 133-140

Romney, D. (1976). Course Effect vs. Teacher Effect on Student´ Ratings of Teaching Competence Research. Higher Education, Bd. 5, S. 345-350

Schott, E. (1973). Zur empirischen und theoretischen Grundlegung eines Bewertungsinstrumentes für Vorlesungen. Serie: Blickpunkt Hochschuldidaktik, Bd. 28

Scott, C. S. (1977). Student Ratings and Instructor-Definded Extenuating Circumstances. Journal of Educational Psychology, Bd. 69 (6), S. 744-747

Seiler, L. H.; Weybrigt, Loren und Stang, D. J. (1977). How Useful Are Published Evaluations Ratings to Student Selecting Courses and Instructors?. Teaching of Psychology, Bd. 4, S. 174-177

Shatz, M. und Best, J. B. (1986). Selection of Items for Course Evaluation by Faculty and Students. Psychological Reports, Bd.58, S. 239-242

Sherman, Barbara R. und Blackburn, R. T. (1975). Personal Characteristics and Teaching Effectivness of Colege Faculty. Journal of Educational Psychology, Bd. 67 (1), S. 124-131

Snyder, C.R. und Clair, M. (1976).Effects of Expected and Obtained Grades on Teacher Evaluation and Attribution of Performance. Journal of Educational Psychology, Bd. 68 (1), S. 75-82

SPIEGEL-SPEZIAL (1993). Bd.3, Hamburg

Stillman, Paula. L., Gillers, M.; Heins, M., Nicholson, G. und Sabers, D. L. (1983). Effect of Immediate Student Evaluations On a Multi-Instructor Course. Journal of Medical Education, Bd. 58, S .176-178

Strom, B., Hocevar, D., Zimmer, J. und Michael, W. B. (1982). The Course Structure Inventory: Discriminant and Construct Validy. Educational and Psychological Measurement, Bd. 42, S. 1125-1133

Süllwold, F. (1992). Welche Realität wird bei der Beurteilung von Hochschullehrern durch Studierende erfahren?. Mitteilungen des Hochschulverbandes, Bd. 1, S. 34-36

Terry, R. L. und Mcintosh, D. E. (1988). Do Students´ Expencancies Affect Their Course Evaluations?. Educational and Psychological Measurement, Bd. 48, S. 787-798

Thomas, D., Ribich, F. und Freie, J. (1982). The Relationship Between Psychological Identification with Instructors and Student Ratings of College Courses. Instructional Science, Bd. 11, S. 139-154

Thurston Barnett, Linda und Littlepage, Glenn (1979). Course preferences and evaluations of male and female professors by male and female students. Bulletin of the Psychonomic Society, Bd. 13(1), S. 44-46

Ware Jr., J. E. und Williams, R. G. (1975). The Dr. Fox Effect: A Study of Lecturer Effectinveness and Rating of Instruction. Journal of Medical Education, Bd. 50, S. 149-156

Weiss, R. (1991). Ziele und Probleme einer Lehrveranstaltungskritik. Zeitschrift für Hochschuldidaktik, Bd. 15, S. 35-42

Westdeutsche Rektorenkonferenz (1986). Zur Beurteilung und Entwicklung der Ansätze zur Leistungsbewertung und -messung von Hochschulen. Stellungnahme des 149. Plenum der Westdeutschen Rektorenkonferenz, Bonn

Williams, R. G. und Ware, J. E. Jr. (1977). An Extended visit with Dr. Fox: Validity of Student Satisfaction with Instruction Ratings after Repeated Exposures to a Lecturer. American Educational Research Journal, Bd. 14 (4), S.449-457

Yates, J.W. und Karmos, J.S. (1971). Final Grade Predictions and Instructional Evaluation at Different Levels of Course Achievement. Journal of Instructional Psychology, Bd. 4 (3), S. 38-40

Lehrevaluation und Studienqualität

Beiträge zur studentischen Veranstaltungskritik und zur Lehrberichterstattung an Hochschulen.

Vorgelegt von: Natalija el Hage, Arbeitsgruppe Hochschulforschung, Universität Konstanz.

Hefte zur Bildungs- und Hochschulforschung

Heft 13: Zur Validität studentischer Veranstaltungskritiken- Befunde empirischer Studien zu einem umstrittenen Verfahren.

Studentische Veranstaltungskritik ist ein in Deutschland umstrittenes Verfahren. Mit diesem Heft soll zur Versachlichung der Diskussion beigetragen werden. Als erstes werden daher die Konfliktlinien bei Vorhaben zur Lehrevaluation aufgeführt. Daran schließen sich häufig vorgebrachte Vorbehalte gegen die studentische Veranstaltungskritik an. Diesen insgesamt 12 immer wieder genannten Bedenken werden Studien zum entsprechenden Thema zugeordnet. Alles in allem werden ca. 100 Studien, vor allem aus den Vereinigten Staaten, kurz beschrieben und ihre Befunde dargestellt.

Insgesamt kann bei entsprechend konstruierten Instrumenten davon ausgegangen werden, daß studentische Veranstaltungskritik methodischen Kriterien vergleichbarer Verfahren genügt.

Heft 14: Instrumente studentischer Veranstaltungskritik- Gestaltungsprinzipien und Beispiele.

In Ländern, in denen studentische Lehrbewertung schon seit längerem verwendet wird (z.B. U.S.A.) sind auch Untersuchungen zu den Konstruktionskriterien zu entsprechenden Fragebögen bekannt. Um adäquaten methodischen Gütekriterien zu genügen, müssen Instrumente studentischer Veranstaltungskritik entsprechend sorgfältig konstruiert werden. Neben der Wichtigkeit der Multidimensionalität der Verfahren ist es u.a. notwendig, daß verwendete Items von beobachtbaren Kriterien ausgehen. Abgesehen von Bedingungen, die sich auf die Güte der Fragebögen auswirken, werden auch wichtige Anwendungsqualitäten (optimale Länge eines Instruments, Anonymität, etc.) diskutiert.

In Deutschland wird bereits eine unüberschaubare von Fragebögen der unterschiedlichsten Güte verwendet. Vorgestellt und kurz beurteilt werden in diesem Heft diejenigen, die nicht nur einen praktischen sondern auch einen wissenschaftlichen Zweck verfolgen.

Heft 15: Studienreform durch Lehrevaluation? Ansätze, Projekte und Verwendungen der Lehrbewertung

In Deutschland ist in den letzen Jahren bereits eine große Anzahl von Maßnahmen im Bereich der Lehrevaluation durchgeführt worden. Kaum eine wissenschaftliche Institution, die sich nicht an der Diskussion beteiligen oder sogar eigene Projekte durchführen würde. Eine Übersicht über die wichtigsten Vorhaben, sowohl vom Bundesbildungsministerium als auch von den Wissenschaftsministern der Länder, von verschiedenen Wissenschaftsinstitutionen, wie auch von den Studierenden wird aufgeführt.

Bereits heute ist deutlich, daß eine reine Problemaufdeckung keine größeren Verbesserungen in der Lehre bringen wird. Damit studentische Veranstaltungskritik und andere Datensammlungen der Lehrqualität eine Wirkung haben können, müssen entsprechende Rahmenbedingungen geschaffen werden und die gewonnen Daten systematisch verwendet werden. Für diesem Zweck werden entsprechende Ansätze und Möglichkeiten der Datenverwendung aufgezeigt.