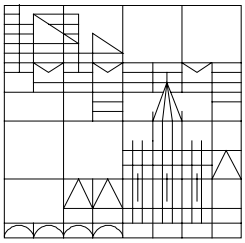
Universität Konstanz

Fachbereich für Politik- und Verwaltungswissenschaft

Prof. Dr. Rainer Schnell
PD Dr. Johannes Kopp



Theoretische und methodische Diskussionen der Lehrevaluationsforschung und deren praktische Bedeutung.

Forschungsbericht des durch das Ministerium für Wissenschaft, Forschung und Kunst aus den Haushaltsmitteln zur Verkürzung der Studienzeiten und zur Stärkung der Lehre (Kapitel 1423, Titelgruppe 71) geförderten Forschungsprojektes „Fakultätsinterne Evaluation der Lehre – die Weiterentwicklung des bisherigen Evaluationskonzepts“

Dezember 2000

Gliederung

Anhänge:

- **A: Ein Überblick über die zur Evaluation der Lehre in sozialwissenschaftlichen Studiengängen eingesetzten Instrumente**
- **B: Einige weitere in der Literatur zu findende Instrumente zur Evaluation der Lehre**
- **C: Dokumentation des Befragungsmaterials (Fragebogen, Anschreiben)**
- **D: Konstanzer Lehrevaluation Entwicklungsprofile**
- **E: Ein erweitertes Instrument zur Lehrevaluation im Fachbereich für Politik- und Verwaltungswissenschaft**

1. Vorbemerkung

Die moderne Universität steht wie wohl nur selten zuvor in ihrer Geschichte im Mittelpunkt einer öffentlichen Diskussion, die sich vor allem um die Frage fokussiert, welche Aufgaben dieser Institution zukommen und wie diese Aufgaben am besten zu erfüllen sind. Kritik an den Strukturen der Universität – und den dadurch bedingten vermeintlichen oder realen Defiziten der in den Hochschulen vermittelten Kenntnisse – findet sich dabei mutmaßlich fast genau so lange wie die Universität selbst, jedoch nur selten zuvor in der neueren Geschichte scheinen die Forderungen nach einem Strukturwandel und einem Umbau der Universität so schwerwiegend zu sein, wie heute zum Teil beobachtbar. Die zunehmende Diskussion um die Leistungsfähigkeit der Universität ist dabei in den letzten Jahren sicher nicht zufällig deutlich angewachsen. Vielmehr gehen diese Diskussionen mit einigen bedeutsamen gesellschaftlichen Veränderungen einher und ohne einen Blick auf diese Debatten kann auch die hochschulpolitische Diskussion nicht richtig eingeschätzt werden.

So läßt sich in den letzten Jahrzehnten ein unvergleichbarer Ausbau der Bildungsbeteiligung und eine entsprechende Steigerung der Zahl der Studierenden beobachten (vgl. zu den Anfängen dieser Diskussion Dahrendorf 1965 und zur aktuellen Entwicklung Arbeitsgruppe Bildungsbericht 1994). In Anbetracht realer oder vermeintlicher Bildungsrückstände vervierfachte sich die Zahl der Studierenden zwischen 1960 und 1980 in der alten Bundesrepublik. Schon 1980 waren mehr als eine Million Studierende an den verschiedenen Hochschulen immatrikuliert. In der Zwischenzeit hat sich diese Zahl – jetzt natürlich auch einschließlich der Neuen Bundesländer – erneut fast verdoppelt. Im Wintersemester 1998/1999 studierten mehr als 1.8 Millionen Menschen an den Hochschulen der Bundesrepublik (Zifonun 1999). Entsprechende Prognosen gehen davon aus, dass sich diese Zahl trotz aktueller Schwankungen allein aufgrund der entsprechenden demographischen Entwicklungen bis zum Ende des nun beginnenden Jahrzehntes noch weiter steigern wird (vgl. hierzu Statistisches Bundesamt 1997: 63). Diese zunehmende Nutzung der Hochschulen – bei einem sich nur wenig verändernden Budget – geht mit einer zunehmenden Anerkennung der Bedeutung von Bildung als einer der wichtigsten Entwicklungsbedingungen moderner Gesellschaften einher. Der Wandel der Gesellschaft hin zu einer dienstleistungsorientierten Wissensgesellschaft ist eng mit der Zunahme von Humankapital verbunden. An dieser Stelle kann nicht auf die vielfältigen Diskussionen über die weitere gesellschaftliche Entwicklung eingegangen werden, es erscheint jedoch unbestreitbar, dass Bildung bei diesen Entwicklungsprozessen eine nicht zu überschätzende Rolle einnehmen wird (vgl. etwa Krais 1983).

Gerade in Anbetracht zunehmender Modernisierung und der damit einher gehenden gesellschaftlichen Differenzierung und Globalisierung hat dann das Bildungssystem auch für gesamtgesellschaftliche Prozesse eine große Bedeutung. Diese Entwicklung kann nun für die

Organisation dieses gesellschaftlichen Teilbereiches nicht folgenlos bleiben und dementsprechend hat sich die Universität im Laufe der letzten Jahrzehnte deutlich verändert (vgl. House 1993: 56ff): Verstärkt durch die ökonomische Situation halten auch deshalb an den Universitäten Effizienzkriterien Einzug.[1] Auch wenn diese Gesichtspunkte teilweise als systemfremde Codes in einem an dem binären Schematismus von Wahrheit und Falschheit orientierten Teilsystem betrachtet werden, so läßt sich eine zunehmende Orientierung an derartigen Aspekten wohl nicht mehr aufhalten. Dabei werden nun jedoch nicht rein ökonomische Gesichtspunkte berücksichtigt– dies erscheint auch schwer möglich. Vielmehr schließt sich auch die Hochschule an die Diskussion am Qualitätssicherung beziehungsweise Qualitätsmanagements an, die in der Zwischenzeit fast alle Bereiche und hierbei vor allem auch die öffentliche Verwaltung erfasst hat (vgl. einleitend und für weitere Hinweise Schenker-Wicki 1996: 37ff).

Auch wenn es nicht möglich ist, eine einzige Ursache für diese zunehmende Beachtung von Qualitäts- und Evaluationskriterien in der hochschulpolitischen Diskussion auszumachen, so kann als ein wichtiger Auslöser dieser Diskussionen das im Dezember 1989 vom Spiegel publizierte Hochschulranking genannt werden (Bülow-Schramm 1994: S16).[2] Mit einem derartigen Vorgehen wird auch für die Bundesrepublik versucht, an die entsprechenden Entwicklungen in den Vereinigten Staaten anzuschließen (vgl. hierzu Herbst 1991; Kellermann 1992; Fallon 1998).

Diese Bemühungen entstanden nun keineswegs zufällig: Sicherlich durch die skizzierten allgemeinen gesellschaftlichen Veränderungen bedingt oder zumindest verstärkt, lassen nun durchaus verschiedene Probleme der klassischen Hochschule beobachten. Hierbei wurde und wird vor allem die Qualität der Lehre und Ausbildung diskutiert. Gerade auch in der öffentlichen und veröffentlichten Meinung kreiste die Diskussion um überfüllte Seminare und Vorlesungen, das mangelnde Engagement der Lehrenden, die unzureichende Ausbildung der Studierenden und dies vor allem in Hinsicht auf eine spätere Erwerbstätigkeit, das im internationalen Vergleich hohe Alter der Studienabsolventen, die in einigen Fächer die Mehrheit bildende Zahl der Studienabbrecher und die durch all diese Faktoren bedingte Fehlallokation öf-

---

[1] Schwermer (1999: 58) sieht die Forderung nach Evaluation „aus der Knappheit der finanziellen Ressourcen geboren und mit Rationalisierungsdruck verbunden".

[2] Die Veröffentlichung dieser Rangliste hat eine große und teilweise heftige Diskussion ausgelöst. Zumindest unter methodischen Gesichtpunkten kann man heute festhalten, dass die hier durchgeführte Untersuchung den Qualitätsanforderungen an eine Evaluation nicht genügte (vgl. für eine detaillierte Kritik etwa Lamnek 1900 sowie Gräf 1991; siehe des weiteren auch Scheuch 1990 und Balke/Stiensmeier-Pelser/Welzel 1991). Lamnek (1990: 99f) kommt zu folgendem Fazit: die „Studie ist unter methodischen Gesichtspunkten äußerst kritisch zu beurteilen: 1. Sie suggeriert Professionalität, obgleich sie dilettantisch gemacht ist. 2. Sie gibt die Widerspiegelung objektiver Verhältnisse an der Universität vor, obgleich sie (nicht notwendigerweise zutreffende) Perzeptionen einer universitären Teilgruppe wiedergibt. 3. Die Studie ist nicht repräsentativ, weshalb die Befunde in keiner Weise hätten generalisiert werden dürfen. 4. Auch das Erhebungsinstrument läßt zu wünschen übrig, weshalb bei der Interpretation ausgesprochene Zurückhaltung hätte praktiziert werden sollen" (vgl. für die Reaktionen zu einer ähnlichen Studie in Österreich den Überblick von Mayer 1992).

fentlicher Mittel denn um die Ergebnisse der wissenschaftlichen Forschung. Verschärft wurde die Situation der Hochschule zudem noch durch die immer knapper werdenden öffentlichen Mittel. Aus diesen Gründen war es fast eine zwangsläufige Entwicklung, dass auch im Bereich der Hochschule verstärkt Evaluationsbemühungen und eine Diskussion um die Qualität von Lehre und Forschung zu finden sind – auch wenn dies in einigen zeitgeistlichen Analysen als Beleg einer zunehmenden Kolonialisierung der Lebenswelt oder eine Einschränkung der wissenschaftlichen Freiheit interpretiert wird (Musgnug 1992).

Das Forschungsprojekt, dessen Ergebnisse in diesem Bericht vorgestellt werden sollen, setzt sich nun schwerpunktmäßig nicht mit derart allgemeinen Fragestellungen auseinander. Ausgangspunkt des Interesses ist vielmehr eine recht konkretes empirisches Problem: Im Rahmen der nun eben seit rund einem Jahrzehnt stattfindenden Evaluationen und Bewertungen der Hochschule spielen interne Evaluationsinstrumente und hierbei vor allem der Einbezug der Studierenden bei der Einschätzung der jeweiligen Lehrveranstaltungen eine große Rolle. Fast genauso lange wie diese internen Evaluationsmaßnahmen findet sich aber auch eine Diskussion darüber, ob die mit Hilfe derartiger Befragungen erzielbaren Ergebnisse eigentlich auch eine gewisse inhaltliche Gültigkeit beanspruchen dürfen und somit überhaupt einen Wert bei der Einschätzung der Qualität der Lehre haben können oder nicht. Genau dieser Frage soll im folgenden weiter nachgegangen werden.

Dabei erscheint es jedoch sinnvoll, die in diesem Bereich zu findenden Diskussionen nicht unabhängig vorzustellen, sondern in den allgemeineren Rahmen der Evaluationsforschung einzubetten. Aus diesem Grunde soll im folgenden zunächst ein wirklich kurzer Überblick über die verschiedenen Ansätze und Diskussionen dieser Forschungstradition (Kapitel 2) und vor allem der verschiedenen Möglichkeiten der Hochschulevaluation (Kapitel 3) gegeben werden. In beiden Bereichen ist die Fülle der entsprechenden Literatur fast nicht mehr überschaubar und allein deshalb soll nicht einmal versucht werden, den Anspruch auf Vollständigkeit zu erheben. Vielmehr sollen nur grob die wichtigsten Entwicklungslinien aufgezeichnet werden, um die Diskussion über den Status von Studierendenbefragungen im Rahmen der Hochschulevaluation richtig beurteilen zu können.[3]

Anschließend an diese allgemeine Diskussion kann dann konkreter auf die Befragung von Studierenden als Evaluationsmittel eingegangen werden (Kapitel 4). Hierzu soll zuerst auf die Zielsetzung und die Verbreitung interner Evaluationen mit Hilfe studentischer Befragungen, aber auch auf die konkrete Gestaltung und Umsetzung eingegangen werden. Um diese Frage

[3] In der Zwischenzeit finden sich eine Reihe guter Übersichtsarbeiten in die Evaluationsforschung, die die erste Orientierung erleichtern, vgl. hierzu etwa Rossi und Freeman (1993), Bortz und Döring (1995), Wottawa und Thierau (1998) oder Clarke und Dawson (1999).

zu beantworten, kann auf eine im Rahmen des Projektes „Fakultätsinterne Evaluation der Lehre“ durchgeführte Befragung derjenigen bundesrepublikanischen Hochschulen zurückgriffen werden, die einen Studium in einem sozialwissenschaftlichen Studiengang anbieten.[4] Hierbei soll vor allem ein Überblick über die zum Einsatz kommenden Befragungsinstrumente gegeben werden.[5] Zum Abschluß dieses vierten Kapitels wird anhand der an der Fakultät für Verwaltungswissenschaft durchgeführten Lehrevaluationen der letzten Jahre untersucht, inwieweit der bei Studierendenbefragungen – meist implizit – unterstellte Wirkungsmechanismus eine gewisse Gültigkeit für sich beanspruchen kann.

In Kapitel 5 sollen danach die verschiedenen in der deutschen, vor allem aber in der internationalen Literatur zu findenden Diskussionen über die Möglichkeiten von Studierendenbefragungen und deren methodischen Problemen vorgestellt werden. Vor dem Hintergrund dieser verschiedenen Diskussionen soll hier kurz im Abschnitt 6 auf das an der Fakultät für Verwaltungswissenschaft der Universität Konstanz zum Einsatz kommende Evaluationsinstrument eingegangen werden.

Im Anschluß an diese Diskussionen kann dann näher auf verschiedene Möglichkeiten einer Validierung derartiger studentischer Einschätzungen zur Lehrqualität eingegangen werden (Kapitel 7). Hierzu sollen einerseits die verschiedenen denkbaren Vorgehensweisen vorgestellt und deren Probleme diskutiert werden. Andererseits ist es in diesem Abschnitt auch möglich, hinsichtlich einem der möglichen Validierungsansätze – der Einbeziehung der Absolventen eines Studienganges – kurz auf eigene empirische Untersuchungen zurückzugreifen. Abschließend gilt es, zusammenfassend die Folgerungen dieser Diskussionen für die Einschätzung der Qualität der Lehre und damit für eine umfassende Evaluation der Hochschule einzuschätzen (Kapitel 8).

---

[4] Wie unten genauer ausgeführt wird, werden hierunter im folgenden Studienabschlüsse in Soziologie, Politikwissenschaften und Verwaltungswissenschaften verstanden. Eine Berücksichtigung der verschiedenen Evaluationen in anderen, teilweise sicherlich den Sozialwissenschaften zurechenbaren Teildisziplinen wie etwa der Sozialpsychologie oder der Ökonomie (vgl. hierzu Frey 1990) erfolgte nicht, da dies den Rahmen der Möglichkeiten einer selbst durchgeführten Untersuchung übersteigen würde.

[5] Im Anhang A dieses Berichtes findet sich eine Dokumentation der verschiedenen hier zu findenden Evaluationsinstrumente. Neben den im Rahmen der eigenen Befragung gesammelten Fragebögen sind im Anhang B auch einige wenige in der entsprechenden Literatur veröffentlichten Instrumente dokumentiert. Schon die Verschiedenartigkeit der in diesen beiden Anhängen zusammengestellten Befragungsinstrumente – vor allem hinsichtlich der Detailliertheit der Fragen sowie der Länge des eingesetzten Instrumentes – macht deutlich, dass noch lange nicht von einem einheitlichen Ansatz der Veranstaltungsevaluation ausgegangen werden kann.

2. Zur Evaluationsforschung: Einige kurze Anmerkungen zu einem Forschungsfeld und seinem Selbstverständnis

Ganz allgemein kann man unter Evaluation die Bewertung einer Maßnahme oder eines Objektes und vor allem die Qualitäts- und Erfolgskontrolle verstehen (vgl. Heid 2000: 101). Obwohl die Evaluationsforschung häufig als ein Teil der Sozialwissenschaften betrachtet wird, finden sich derartige Untersuchungen in den verschiedensten Bereichen der Wissenschaft – und natürlich auch darüber hinaus. Gerade in Rahmen der Medizin, der technischen Disziplinen oder der Organisationsforschung gehört es zur wissenschaftlichen Tradition, die Wirksamkeit einzelner Maßnahmen in konkreten und praktischen Handlungsbezügen, also außerhalb der theoretischen und empirisch-experimentellen Grundlagenforschung, einschätzen zu können. Die Nützlichkeit eines Medikamentes oder einer neuen Operationsmethode, die Chancen und Folgen der großtechnische Umsetzung einer bestimmten Innovation oder die praktischen Folgen einer Umstrukturierung bestimmter Organisationsbereiche zu bestimmen, ist hier wesentlicher Bestandteil des Forschungsprozesses (vgl. hier etwa Austin 1996 sowie die verschiedenen Beiträge in Becher/Kuhlmann 1995; für eine Abgrenzung der verschiedenen Prüfverfahren wie Revision, Controlling und Evaluation vgl. Schenker-Wicki 1996: 14ff; für eine kurze Geschichte der Evaluationsforschung Stamm 1998: 23ff sowie zu einigen wichtigen internen Diskussionen Sechrest/Figueredo 1993). In der Zwischenzeit finden sich jedoch auch in den verschiedenen Teilbereichen der Sozialwissenschaft immer mehr Versuche, die in diesem Bereich entwickelten Kenntnisse und Theorien auch praktisch einzusetzen – und daher auch einzuschätzen und dementsprechend als Folge im Rahmen dieser Bemühungen auch immer mehr Arbeiten, die sich mit der Evaluation dieser Programme und Aktionen beschäftigen (vgl. für Hinweise auf einige historische Studien Freeman/Solomon 1981; Hellstern/Wollmann 1984; Bortz/Döring 1995: 95f sowie Rossi/Freeman 1993: 9ff; vgl. auch Kromrey 1995a).

Die allgemeine Evaluationsforschung stellt also nicht nur ein Teilgebiet der Sozialwissenschaften dar. Aus diesem Grunde kann und wird die Evaluation teilweise recht allgemein definiert: „Evaluation is defined as a form of disciplined inquiry (...) that applies scientific procedures to the collection and analysis of information about the content, structure and outcome of programmes, projects and planned interventions“ (Clarke/Dawson 1999: 1). Evaluation dient dabei generell der rationalen Entscheidungsfindung und letztlich der Effizienzsteigerung.[6] Trotz dieser allgemeinen Definition und der letztlich auch breiten Anwendungsfelder

[6] In den Vereinigten Staaten findet sich hierzu das „United States General Accounting Office“, das seine Aufgabe ganz allgemein wie folgt beschreibt: „The General Accounting Office is the investigative arm of Congress. GAO's mission is to help the Congress oversee federal programs and operations to assure accountability to the American people. GAO's evaluators, auditors, lawyers, economists, public policy analysts, information technology specialists, and other multi-disciplinary professionals seek to enhance the economy, efficiency, effectiveness, and credibility of the federal government both in fact and in the eyes of the American

hat es sich eingebürgert, unter Evaluation meist die systematische Untersuchung sozialer Interventionen zu verstehen.[7] Hierbei lassen sich eine Vielzahl von empirischen Anwendungsfeldern aufzählen, die von groß angelegten Fragestellungen wie etwa nach institutionellen Lösungsmöglichkeiten zur Verbesserung und Effizienzsteigerung im Gesundheitswesen bis hin zu eher lokalen Fragen wie der Preisgestaltung bei der städtischen Parkraumbewirtschaftung reichen.

In den Vereinigten Staaten finden sich eher sozialwissenschaftliche Evaluationsstudien spätestens seit den dreißiger Jahren des 20. Jahrhunderts und stellen seitdem einen integralen Bestandteil der Sozialpolitik dar (Bortz/Döring 1995: 95). Nicht zufälligerweise weisen viele sozialwissenschaftliche Studien in den Vereinigten Staaten von Amerika einen großen Praxisbezug auf und bemühen sich die konkreten Einflußmöglichkeiten und Konsequenzen bestimmter Maßnahmen zu ergründen. So finden sich etwa im Bereich der Familienpolitik und Familienentwicklung gerade in Hinsicht auf die ethnische Pluralität der Gesellschaft und die doch recht großen ökonomischen Unterschiede zwischen den einzelnen Schichten und Ethnien immer wieder Beiträge, die versuchen, die Möglichkeiten und Grenzen staatlicher Interventionen zu untersuchen (vgl. etwa Jencks 1992). Große sozialpolitische Programme in den Vereinigten Staaten – die Idee der Great Society zur Bekämpfung der Armut oder der Versuch, mit Hilfe der sogenannten "affirmative action" die Benachteiligung ethnischer Minderheiten auszugleichen – finden sich in dieser Entwicklungslinie und fast immer findet sich ein Versuch, den Erfolg dieser Maßnahmen auch durch begleitende Evaluationsmaßnahmen einschätzen zu können[8]. Eine entsprechende Evaluationstradition in der Bundesrepublik ist relativ kurz, hat jedoch gerade in der Einschätzung entsprechender bildungs- und ausbildungsbezogene Maßnahmen eine starke Verwurzelung. Ein heute fast schon klassisches Beispiel in diesem Zusammenhang ist die Einführung und die daran anschließende Bewertung der Gesamtschule (vgl. Fend 1982).

Die Evaluationsforschung wird dabei in der Zwischenzeit häufig als ein eigenständiges wissenschaftliches Feld angesehen, das eine Vielzahl von praktischen Anwendungsgebieten auf-

---

people. GAO accomplishes its mission through a variety of activities including financial audits, program reviews, investigations, legal support, and policy/program analyses. GAO is dedicated to good government through its commitment to the values of accountability, integrity, and reliability“ (http://www.gao.gov). Diese staatliche Stelle dient also ganz allgemein der Evaluation der verschiedensten staatlichen Aktivitäten und Programme und verfolgt dabei vor allem Effizienzgesichtspunkte. Ob in der Bundesrepublik der Bundesrechnungshof eine vergleichbare Institution ist, erscheint eine offene Frage (vgl. Dorn 1984).

[7] Dies wird etwa in der Antwort auf die Frage „What is Evaluation Research?“ deutlich, die sich zu Beginn eines der sicherlich bekanntesten Lehrbuches in diesem Bereich von Peter H. Rossi und Howard E. Freeman findet: „Evaluation research is the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of social intervention programs“ (Rossi/Freeman 1995: 5).

[8] Vgl. allgemein Hellstern/Wollmann (1984: 27ff); Rossi/Freeman (1993: 9ff); hinsichtlich des Erfolgs der "affirmative action" siehe Bowen/Bok (1998).

weist. Sie stellt dabei einen Teil der anwendungsorientierten und vor allem empirischen Forschung dar. Zielsetzung dieses Abschnittes kann es aufgrund der Vielzahl verschiedener Bereiche, Themen und Diskussionen der Evaluationsforschung gar nicht sein, eine Darstellung der umfangreichen Forschungstätigkeiten in diesem Feld zu versuchen. Vielmehr sollen hier einige wenige der grundlegenden Erörterungen der Evaluationsforschung skizziert werden, die die Einordnung der weiteren Darstellung und speziell die Diskussionen innerhalb der aktuellen Hochschulevaluation besser verständlich machen.

Generell lassen sich drei verschiedene Klassen von Evaluationsstudien unterscheiden: „(1) analysis related to the conceptualization and design of interventions; (2) monitoring of program implementation; and (3) assessment of program effectiveness and efficiency“ (Rossi/Freeman 1993: 34). Es ist leicht einsichtig, dass gerade im Rahmen der entsprechenden hochschulpolitischen Diskussion auch alle Arten von Evaluation zum Einsatz kommen können, auch wenn jeweils die genauen Einsatzgebiete sicher sehr unterschiedlich sind. Die hier im Mittelpunkt stehenden Bewertungen von einzelnen Lehrveranstaltungen im Rahmen der Einschätzung der Qualität der Lehre können dabei bestenfalls in die dritte Gruppe von Evaluationsstudien eingestuft werden. Allen Evaluationsstudien sind jedoch gewisse Kriterien gemeinsam, denen sie nach allgemeiner Einschätzung genügen müssen (vgl. auch Brandstädter 1990; Schwermer 1999: 59ff):

- Evaluationsstudien sollen die Nützlichkeit eines bestimmten Programmes einschätzen können. Hierzu ist es notwendig, sowohl den Grad der Zielverwirklichung als auch die dafür eingesetzten Mittel zu kennen. Nur durch die Kenntnis beider Faktoren ist es möglich, die Effizienz eines Programmes einzuschätzen.
- Um den Grad der Zielerreichung einschätzen zu können, ist es eine unabdingbare Voraussetzung, dass die entsprechenden Ziele klar bestimmt sind. Auch wenn dieser Punkt vielleicht trivial erscheint, so stehen viele Evaluationsstudien genau an diesem Punkt vor einem großem Problem. Häufig ist nicht genau klar, welche Dimensionen eines meist recht vielfältigen Bereiches wie eingeschätzt werden sollen (vgl. hinsichtlich der Evaluationen von Hochschulen hierzu Altrichter/Schratz 1992: 15f).

Im Rahmen dieser Evaluationsforschung finden sich darüber hinaus nun eine Reihe wichtiger Diskussionstraditionen, die für das Verständnis der weiteren Ausführungen bedeutsam sind. Dies betrifft vor allem eine Reihe recht grundlegender methodischer und methodologischer Fragestellungen. Die Vielzahl von Anwendungsfeldern und die Unterschiedlichkeit der entsprechenden Zielsetzungen bedingt wohl die doch recht große Heterogenität bei diesen Gesichtspunkten. So lassen sich hier sowohl hinsichtlich der zeitlichen Verortung einer Evaluation – soll sie begleitend und damit formativ oder nach der Implementation einer entsprechenden Maßnahme und damit summativ erfolgen? – der diskursiven oder administrativen Form,

der internen oder externen formalen Organisation der Evaluation als auch der eher qualitativen oder quantitativen Durchführung der entsprechenden Untersuchung gänzlich unterschiedliche Positionen finden (vgl. Rossi/Freeman 1993: 135ff sowie hinsichtlich der Evaluation der Lehre Bülow-Schramm 1994). Auch über die Nützlichkeit von Evaluationen generell finden sich entsprechend gegensätzliche Auffassungen (vgl. schon Alkin/Daillak/White 1979: 14ff). Generell kann man jedoch festhalten, dass die Evaluationsforschung vor allem „der Analyse der Aus- und Nebenwirklungen von Interventionen und Programmen besteht“ (Brandstädter 1990: 223), sie jedoch häufig auch nicht explizit genannte oder gar latente Funktionen aufweist. „Evaluationsforschung dient nicht nur der Verbesserung von Lern- und Entwicklungsbedingungen, der Verbesserung von Allokationsentscheidungen in der Verteilung öffentlicher Mittel oder der Konsensbildung in strittigen (...) Fragen, sondern wird nicht selten auch veranstaltet zur Durchsetzung politischer Interessen, zur publikumswirksamen Selbstprofilierung oder zur Abkühlung von Konflikten durch dilatorische Behandlung“ (Brandstädter 1990: 224).

Im folgenden soll nun jedoch nicht näher auf diese teilweise recht allgemein diskutierten Fragen, die vielfach auch Fragen ethischer und wertbedingter Urteile betreffen (vgl. Chen 1990: 87ff; House 1993: 163) eingegangen werden, da sich gerade in dem hier im Mittelpunkt stehenden Problemfeld viele der in der allgemeinen Evaluationsforschung diskutierten Fragen so gut wie nicht stellen.

3. Zur Evaluation von Hochschulen: Ansätze, Diskussionen, Ergebnisse

Die Evaluation von Bildungseinrichtungen und Bildungsmaßnahmen stellt eine der ältesten Anwendungen der Evaluationsforschung dar (Cashin 1986). Dies betrifft historisch vor allem auch Maßnahmen im Rahmen der grundlegenden Ausbildung, wie etwa das Head-Start-Programm in den Vereinigten Staaten (Cooley/Lohnes 1976; Hellstern/Wollman 1984) oder der Sesamstraße (vgl. Cook et al. 1975) und in der Bundesrepublik die entsprechenden Arbeiten der Schulforschung (zusammenfassend Fend 1998), umfasst aber in den letzten Jahren aus den oben skizzierten Gründen der zunehmenden Bedeutung von Humankapital auch immer mehr die Hochschule (vgl. allgemein Anderson 1998 sowie Rau 1996 für einen ersten Überblick über die umfangreiche Literatur). Bereits in der Einleitung dieses Berichtes werden einige weitere allgemeine gesellschaftlichen Entwicklungen vorgestellt, die diese zunehmende Beschäftigung mit der Hochschule als Bildungsinstitution und ihrer Evaluation verständlich machen können.

Im folgenden sollen die wichtigsten Ansatzpunkte, Diskussionen, vor allem aber auch Ergebnisse der Forschungsarbeiten zur Evaluation von Hochschulen vorgestellt werden. Nur im Rahmen dieser allgemeineren Darstellung kann dann auch die in den weiteren Abschnitten diskutierten Befragungen von Studierenden entsprechend verortet werden.

3.1 Grundlegende Hypothesen und Konsequenzen der Hochschulevaluation

In Hinsicht auf die Evaluation von Hochschulen kann auf eine ganze Reihe von Arbeiten zurückgegriffen werden, die einen Überblick über dieses Feld und die damit verbundenen Probleme anbieten (vgl. unter anderem Altrichter/Schratz 1992; Holtkamp/Schnitzer 1992; Berendt/Stary 1993; Gerlich 1993; Bülow-Schramm 1994; Sturm 1994; Willems/Gijselaers/de Bie 1994; Bülow-Schramm/Carstensen 1995; Carstensen/Reissert 1995; Rauch 1995; Cashin 1996; Rau 1996; Webler 1996; Schenker-Wicki 1996: 92ff; Lohnert/Rolfes 1997; Stamm 1998; Hochschulrektorenkonferenz 1998).

Wie schon der Umfang dieser nur partiellen Auflistung einschlägiger Literatur zeigt, wird das Thema der Qualitätsmessung, -sicherung, -management oder -steigerung der universitären Ausbildung und der gesamten Institution der Universität seit einiger Zeit heftig diskutiert (vgl. auch Wissenschaftsrat 1997). Ausgangspunkt dieser Diskussion sind dabei die oben im Abschnitt 1 skizzierten Mängel (lange Studienzeiten, hohe Abbruchsquoten, angeblich unzureichende Qualität der Wissensvermittlung) (Webler 1996: 121). Carstensen (1997: 5) führt die verstärkten Hochschulevaluationen entsprechend auf die Hypothese zurück, „(...) daß ein curricular gut strukturierte, transparente, mit hinreichenden Studienberatungsangeboten versehene, gut organisierte, von den Lehrenden engagiert getragene, ressourciell gut ausgestattete,

disziplinär begründete und profilierte Studienstruktur die Voraussetzung dafür ist, dass ein effizientes Studium im Sinne von hoher Erfolgsquote, geringem Schwund und kurzen Studienzeiten bis zum ersten Abschluß gewährleistet werden kann".

Die empirisch diagnostizierte Situation der universitären Ausbildung und die daraus meist gefolgerten Mängel in der universitären Lehre und ihre realen oder vermeintlichen Konsequenzen für die Qualität der Wissensvermittlung führten zu einem ganzen Reigen von Maßnahmen, die – natürlich in unterschiedlichem Ausmaße – an den einzelnen Universitäten und Hochschulen eingesetzt wurden.

Zu diesen Maßnahmen gehört

- die Aufwertung der didaktisch-pädagogischen Fähigkeiten bei der Habilitation und Berufungen,
- die Steigerung der Verantwortung der Fachbereiche für die Lehrleistung und das Lehrangebot (beispielsweise durch die Einführung von regelmäßigen Lehrberichten und einer besseren Abstimmung zwischen einzelnen Studiengängen und Fakultäten),
- die Verbesserung des Prüfungswesens und der Studienberatung,
- Mentoren- und Tutorenprogramme,
- die Einführung von Lehrpreisen,
- und die mehr oder weniger dauerhafte Institutionalisierung von Lehrevaluationen.[9]

Im folgenden sollen diese Maßnahmen kurz im Zusammenhang mit der allgemeinen Diskussion über den Stand der Qualitätsdiskussion universitärer Bildung und Forschung vorgestellt und ein allgemeines Ablaufmodells zur Evaluation der Lehre an Hochschulen skizziert werden.[10]

---

[9] Vgl. für diese Auflistung Webler (1996: 124). Webler (196: 125) zeichnet ein recht pessimistisches Bild, wenn es um die Akzeptanz dieser Maßnahmen geht: „Von der wirklichen Dringlichkeit einer Verbesserung der Qualität von Lehre und Studium sind – von Einzelpersonen abgesehen – die deutschen Hochschulen noch in den wenigsten Fällen überzeugt". Die im Hochschulrahmengesetz vorgegebenen Zielvorgaben und die zu ihrer Durchsetzung geeigneten Maßnahmen würden „nicht beachtet, nicht nachgehalten oder (...) nicht nachgefragt" (Webler 1996: 125).

[10] Diese Skizze stützt sich dabei unter anderem auf ein von der zentralen Evaluationsagentur der niedersächsischen Hochschule herausgegebenen Handbuches zur Evaluation von Lehre und Studium an Hochschulen (Lohnert /Rolfes 1997). Der Grad der Institutionalisierung entsprechender Evaluationsmaßnahmen unterscheidet sich entsprechend der föderativen Struktur des Bildungswesens trotz recht eindeutiger Vorgaben in den entsprechenden Gesetzen (vgl. Webler 1996) immer noch deutlich (Hochschulrektorenkonferenz 1998: 7). Vor allem das Land Niedersachsen hat – teilweise in Kooperation mit verschiedenen anderen norddeutschen Universitäten und zum Teil in Kooperation etwa mit niederländischen Hochschulen – hier entsprechende Vorarbeiten geleistet (vgl. Palandt 1997; Seidel 1997). Der von der Hochschulrektorenkonferenz (1998) herausgegebene Bericht liefert einen kurzen Überblick über die verschiedenen Bemühungen und die daran beteiligten Institutionen und Einrichtungen.

3.2 Schritte einer Hochschulevaluation

Eine Hochschulevaluation soll dabei der Erfassung von Stärken und Schwächen von Studium und Lehre ganzer Fachgebiete und Institute dienen (Lohnert/Rolfes 1997: 10). Dies beginnt mit der Einschätzung der Stellung eines Fachbereiches im Gesamtsystem einer Universität, der Beschäftigung mit dem Gesamtkanon eines Fachbereiches und seinen einzelnen Fächern und führt über Fragen des einzelnen Fachgebietes – etwa nach der Aktualität von Studienplänen und anderen Regelungen – bis hin schließlich zur Ebene der einzelnen Lehrveranstaltung. Insgesamt handelt es sich bei einer Evaluation um eine sehr umfangreiche und vor allem auch recht zeitintensive Aufgabe. So ist dabei mindestens davon auszugehen, dass zwischen der Initiierung einer Selbstevaluation und dem Ende dieses Verfahrens zwischen einem und zwei Jahren vergehen können (vgl. etwa für einen entsprechenden Zeitplan und die für notwendig angesehene Logistik der einzelnen Arbeitsschritte Lohnert/Rolfes 1997: 19ff).

3.2.1 Verwendung prozessproduzierter Daten

Der erste Schritt einer Hochschulevaluation besteht aus der Analyse der jeweiligen durch die Hochschulstatistik verfügbaren Bestands- und Verlaufsdaten, also der quantitativen Grunddaten des jeweiligen Faches oder Fachbereiches. Mit Hilfe derartiger Informationen lassen sich bereits wichtige Basisinformation über den Studienverlauf – wie etwa Zahl der Studienanfänger, Zulassungsbeschränkungen, Studiendauer oder spezifische kritischen Situationen hinsichtlich des Studienabbruches, aber auch die Dauer des Studium bis zu einer entsprechenden Zwischen- oder Hauptprüfung und hierbei vor allem der Studierenden, die diesen Studienabschluß in der Regelstudienzeit absolviert haben, die Altersstruktur der Studierenden und die Zahl der Studienabschlüsse – gut analysieren. Es muß allerdings darauf hingewiesen werden, dass in deutschen Universitäten ein großer Teil dieser Daten bisher entweder erst gar nicht durch die Hochschulverwaltung erfasst wird oder diese Daten aufgrund tatsächlicher oder vermeintlicher datenschutzrechtlicher Bedenken faktisch nicht zur Verfügung steht. Hierzu gehören Daten zur Analyse der entsprechenden Prüfungsstatistiken, der Organisation und des Aufbaus der einzelnen Studiengänge, der Dienstleistungen für andere Fächer, aber auch die aus anderen Fächern entgegengenommenen Leistungen, aber auch der materiellen Ausrüstung der einzelnen Fachbereiche und der Universität – beginnend mit der Erstellung des vor allem in der amerikanischen Diskussion wichtigen "student-teacher-ratio", über den Bestand und die Ausstattung der entsprechenden Bibliotheken bis hin zur Zahl entsprechender Labor- oder PC-Arbeitsplätze und der gesamten räumlichen Ausstattung (vgl. Hochschulrektorenkonferenz 1998: 12; einen Überblick über die hier denkbaren Indikatoren findet sich bei Schenker-Wicki 1996: 114ff). Hier ist auch die Berücksichtigung weiterer Leistungsindikatoren vorstellbar. Als diese Maßnahmen dienen dazu, den institutionellen Rahmen und die strukturell vorgegebenen Bedingungen zu erfassen, die einen ausgesprochen wichtigen Beitrag zur Qualität der

universitären Ausbildung beitragen. Im Rahmen dieses ersten Schrittes kann auch der Inhalt und die Struktur des Lehrangebotes, wie etwa der zeitliche und curriculare Studienablauf, der Prüfungsaufwand, der Studienplan sowie die Studienordnung gesichtet und bewertet werden.

3.2.2 Befragungen

Der zweiter Schritt einer Hochschulevaluation besteht in der Befragung der einzelnen Gruppen zur Qualität der Lehre. Hierunter fallen neben den Studierenden – hierauf wird ausführlich im Abschnitt 4 einzugehen sein[11] – einerseits die Absolventen eines Studienganges[12] und andererseits die Befragung der Lehrenden. Die verschiedenen hier gesammelten Daten werden dann durch weitere Interviews und Gruppendiskussionen ergänzt. Ziel ist es hierbei, für ausgewählte Problembereiche weitere Informationen zu gewinnen – wie etwa bei Studienabbrechern oder Studienfachwechseln.

3.2.3 Externe Begutachtungen

Diese verschiedenen internen Evaluationsmaßnahmen können durch externe Begutachtungen ergänzt werden. Im Mittelpunkt steht dabei meist die Begehung des Fachgebietes durch eine externe Gutachtergruppe (Bülow-Schramm/Carstensen 1995). Im Rahmen der bereits erwähnten niedersächsischen Evaluationsagentur ist dabei die Zusammensetzung dieser sogenannten "peer-group" detailliert vorgeschrieben (Lohnert/Rolfes 1997: 79). Dieser hier dann zu erstellende Bericht dient der kritischen Würdigung der internen Evaluation, soll aber auch weiterführende Aufgaben lösen (Hochschulrektorenkonferenz 1998: 11), indem etwa zusätzliche Probleme angesprochen oder weiterführende Lösungsmöglichkeiten diskutiert werden. Am Ende dieses Evaluationsprozesses – als sogenanntes "follow-up" – sollten die Ergebnisse der internen sowie der externen Bewertung zu einzelnen Maßnahmen konkretisiert werden, die eine (weitere) Verbesserung der Lehre zum Ziel haben.[13] Die Evaluation sollte dabei sowohl die generelle Ausbildungssituation des Faches beziehungsweise des Fachbereiches, seine

---

[11] Oben wurde darauf hingewiesen, dass ohne eine deutliche Zielvorgabe nur sehr schwer eine Evaluation möglich ist. Was nun jedoch genau unter einer guten akademischen Lehre zu verstehen ist, wird meist nicht offen diskutiert. Aufgrund einschlägiger Ergebnisse hochschuldidaktischer Forschungen und entsprechender theoretischer Überlegungen der Lern- und Motivationspsychologie, der Kleingruppentheorie sowie der Kommunikationsforschung lassen jedoch sehr wohl einschlägige Kriterien formulieren (vgl. hierzu Webler 1991, aber auch Gelfert 1992 sowie Webler 1992; vgl. auch Willems/Gijselaers/de Bie 1994: 7ff).

[12] Vgl. generell Lohnert/Rolfes (1997: 44) sowie Teichler (1992); zu den Ergebnissen der Befragung der Absolventen des Konstanzer Studienganges der Verwaltungswissenschaft Kreuter/Kopp (2000).

[13] Selbstverständlich ist es vorstellbar, neben ganzen Fachbereichen oder Studiengängen auch nur einzelne neue Lehrformen oder Unterrichtsmedien zu untersuchen (vgl. hierzu etwa Keil-Slaweik 1999 oder Baumgartner 1999).

interne Struktur, die Veranstaltungs- und Prüfungsinhalte als auch das jeweilige methodische und didaktische Niveau der einzelnen Lehrveranstaltungen umfassen.

3.2.4 Umsetzung der Ergebnisse

Die entsprechenden Ergebnisse werden dann idealiter in eine konkrete Zielvereinbarung zwischen dem Fachbereich und der Universitätsleitung überführt, die vor allem auch detaillierte Umsetzungen der einzelnen Schritte umfassen sollte und dabei auch einen festgelegten Zeitrahmen vorgibt. Abschließend sollten auch die entsprechenden Evaluationsschritte selbst wiederum kritisch überprüft und sozusagen selbst noch einmal evaluiert werden. Insgesamt ist durch die hier geschilderte Mischung verschiedener Evaluationsschritte – Befragung verschiedener Gruppen mit Hilfe verschiedener Methoden, interne und externe Evaluationsschritte – versucht, möglichst umfassend und zielorientiert auf das Niveau der universitären Ausbildung einzuwirken. Verständlicherweise kann eine derartige Evaluation nicht dauernd oder auch nur in relativ kurzen Abständen erfolgen. Hierfür ist der damit verbundene zeitliche und materielle Aufwand und das für eine erfolgreiche Evaluation notwendige Engagement zu hoch. Dementsprechend liegt etwa die Empfehlung der Hochschulrektorenkonferenz für die externe Evaluation bei einem Zyklus von 10 Jahren, die zentrale Evaluationsstelle des Landes Niedersachsens versucht, die einzelnen Fächer alle fünf Jahre zu bewerten.

3.2.5 Hochschulevaluation und Selbststeuerung der Hochschulen

Trotz der mit einer Evaluation verbundenen Schwierigkeiten und dem in einigen Bundesländern noch recht geringen Grad der Institutionalisierung wird die Evaluation als eines der wichtigsten neueren Instrumente der Selbststeuerung in einer autonomen Hochschule angesehen (Künzel 1997). Zumindest die Hochschulrektorenkonferenz (1998: 13) ist recht optimistisch und vermutet: Es ist „wahrscheinlich, dass Evaluation sich in den kommenden Jahren in den deutschen Hochschulen als allgemein akzeptiertes Verfahren durchsetzen wird“. Durch diese hier skizzierte Mischung aus interner und externer Evaluation gelingt es, zwei der drei denkbaren Kontrollinstanzen zu vereinen und den meist aus verschiedenen Gründen unerwünschten Einfluß staatlicher Stellen gering zu halten.[14] Es ist an dieser Stelle jedoch noch einmal zu betonen, dass es jedoch überhaupt einmal einen gewissen Handlungsspielraum der Universität geben muß, um Ergebnisse einer Evaluation entsprechend umzusetzen. So formu-

[14] „There are three models of educational accountability: (a) state or public control, which entail elected representatives, appointed officials, and heads managing schools; (b) professional control, which entails control by teachers, professors, and professional administrators; and (c) consumer control, which can operate either through direct participation of the public or through market mechanisms derived from the private sector“ (House 1993: 59; vgl. auch Altrichter/Schratz 1992: 9ff).

lieren bereits 1992 Altrichter und Schratz in einem Bericht über die Lage der universitären Evaluation in Österreich: „Eine Erhöhung der Außenkontrolle durch staatlich orientierte Rechenschaftslegung ohne eine grundlegende Erhöhung der Autonomie und internen Manövrierfähigkeit der Universitäten (die zumindest Personal- und Finanzhoheit quantitativ und qualitativ umfassen müßte) macht nicht sehr viel Sinn“ (Altrichter/Schratz 1992: 13f).

Bei einem derartigen umfassenden, sowohl interne wie externe Maßnahmen einschließenden Evaluationsansatzes kann man dabei auf verschiedene internationale Erfahrungen zurückgreifen (vgl. etwa Altrichter/Schratz 1992; Holtkamp/Schnitzer 1992; House 1993 sowie für eine beispielhafte Zusammenfassung hinsichtlich der niederländischen Forschungsevaluation Kieser 1998). Teilweise werden Evaluationen explizit international vergleichend angelegt (Hochschulrektorenkonferenz 1995). Dabei zeigt sich, dass hier teilweise auf eine lange Evaluationstradition in den Hochschulen und damit auf eine große auch praktische Erfahrung zurückgegriffen werden kann (vgl. etwa Webler 1995).[15] Es sei an dieser Stelle jedoch darauf hingewiesen, dass es durchaus eine offene und nur empirisch zu beantwortende Frage ist, ob die Universität als Organisation durch derartige Evaluationsprozesse wirklich veränderbar ist. Ein Teil der hier geschilderten Maßnahmen setzt auf die Verhandlung der strittigen Fragen zwischen den Beteiligten. Teilweise wird die Lernfähigkeit von Organisationen durch Verhandlungen sehr kritisch beurteilt (vgl. Huber 1999).

3.2.6 Die Rolle von Studierendenbefragungen in der Hochschulevaluation

Diese Darstellung des Vorgehens bei Hochschulevaluationen war erforderlich, um in dem gesamten Diskussionszusammenhang die Rolle von Studierendenbefragungen besser einschätzen zu können. Immer wieder wird betont, dass es zwar nicht darum gehe, einzelne Lehrende zu kritisieren, jedoch seien die einzelnen Lehrveranstaltungen genau der Ort, an dem universitäre Lehre nun einmal stattfindet und Studierendenbefragungen deshalb ein fast unerläßlicher Schritt bei der Bewertung der Lehrsituation in einem Fachbereich oder einen einzelnen Fach. In der Praxis führt dies ab und an jedoch durchaus dazu, dass zwar mehr oder weniger regelmäßig entsprechende Untersuchungen durchgeführt werden und sogar entsprechende Ranglisten zumindest fach- oder fakultätsintern veröffentlicht werden, weitere Evaluationsmaßnahmen jedoch nicht unternommen werden. Die kurze Skizze in diesem Abschnitt sollte deutlich gemacht haben, dass Studierendenbefragungen zwar eine wichtige Rolle bei Evaluationsbemühungen einnehmen, dass ohne weitere unterstützende Maßnahmen jedoch nur schwer von einer wirklichen Evaluation gesprochen werden kann und dass darüber hinaus al-

[15] Stamm (1998: 26ff) liefert einen kurzen Überblick über die Qualitätsevaluationen in den Hochschulen in Großbritannien, Frankreich, den Niederlanden, Deutschland und der Schweiz.

lein aufgrund derartiger Befragungen wohl auch nicht mit einer Verbesserung der Lehrsituation und damit zu einer Beseitigung der diagnostizierten Mängel beizutragen ist.

4. Befragung von Studierenden als Evaluationsinstrument? Erfahrungen im Bereich sozialwissenschaftlicher Studiengänge

In fast allen der oben aufgeführten Anleitungen und Übersichtsarbeiten zur Evaluation der universitären Lehre kommt der Befragung von Studierenden eine sehr große Bedeutung zu. In ihrem Überblick über Evaluationsverfahren der Lehre beschreibt Bülow-Schramm (1994: S9) die studentische Lehrbeurteilung als wichtigste erprobte Form der Evaluation. Das klassische Evaluationsinstrument ist dabei die Erhebung der studentischen Meinungen mit Hilfe eines Fragebogens. Trotz der Heterogenität der hier zum Einsatz kommenden Instrumente werden in fast allen Befragungen folgende Kernbereiche thematisiert:

- Das Lernverhalten der Studierenden;
- das Lehrverhalten der Dozierenden;
- das Unterrichtsmaterial;
- der Betreuungsaufwand durch die Dozierenden:
- der Lernerfolg der Studierenden;
- der Bezug zu anderen Veranstaltungen sowie schließlich
- die äußeren Bedingungen der jeweiligen Veranstaltung (Bülow-Schramm 1994: S9).[16]

In dem nun folgenden Abschnitt dieses Berichtes sollen diese Evaluationsbemühungen mit Hilfe studentischer Befragungen näher untersucht werden. Dabei soll zuerst die Zielsetzung derartiger Befragungen vorgestellt werden (Abschnitt 4.1), um daran anschließend aufgrund einer eigenen Erhebung die Verbreitung von Studierendenbefragungen in den sozialwissenschaftlichen Studiengängen in der Bundesrepublik zu untersuchen (Abschnitt 4.2).[17] In einem dritten Absatz sollen die verschiedenen, im Rahmen der eigenen Untersuchung erfassten oder in der entsprechenden Literatur abgedruckten Erhebungsinstrumente vorgestellt und miteinander verglichen werden (Abschnitt 4.3). Zum Abschluß dieses Kapitels soll dann anhand der in der Fakultät für Verwaltungswissenschaft der Universität Konstanz regelmäßig erhobenen Studierendenbefragungen die empirische Gültigkeit des bei diesen Evaluationsinstrumenten meist implizit unterstellten Wirkungsmodells ansatzweise untersucht werden (Abschnitt 4.4).

---

[16] Eine ähnliche Einteilung findet sich auch bei der Analyse amerikanischer Evaluationsstudien, vgl. hierzu Braskamp/Ory (1994) sowie Cashin (1995).

[17] Bereits oben wurde kurz die Definition sozialwissenschaftlicher Studiengänge problematisiert. Man kann sich an dieser Stelle jedoch auch fragen, warum nicht umfassender die jeweiligen Evaluationsinstrumente erhoben wurden. Zur Begründung muß hier auf die Knappheit der zur Verfügung stehenden Mittel und den begrenzten Zeithorizont des hier zusammenfassend dargestellten Projektes verwiesen werden. Darüber hinaus ist anzumerken, dass für die wichtigste Zielsetzung des Projektes – der Frage nach der methodischen Bewertung der verwendeten Bewertungsinstrumente – auch mit Hilfe einer eingeschränkten Auswahl eine sinnvolle Antwort möglich erscheint.

4.1 Die Zielsetzung studentischer Lehrbefragungen

In dem bei Studierendenbefragungen impliziten einfachen Modell werden die Studierenden als die Verbraucher oder Konsumenten universitärer Lehre betrachtet, deren Urteil über die angebotenen Produkte eine wichtige Rolle bei der Bewertung deren Qualität darstellen sollte. Ein derartiges Verständnis universitärer Ausbildung wird sicherlich nicht überall geteilt – im Abschnitt 5 werden die wichtigsten Argumente in diesem Zusammenhang diskutiert. Unbestritten ist jedoch, dass es ohne die Einbeziehung studentischer Urteile wohl kaum zu einer angemessenen Bewertung einzelner Veranstaltungen oder der gesamten Lehrsituation kommen kann.

Meist erfolgen diese Evaluationen am Ende eines Seminars oder einer Vorlesung und weisen damit einen summativen Charakter aus (Bülow-Schramm 1994: S9). Teilweise sollen die entsprechenden Befragungen aber entweder zu Beginn einer Veranstaltung – hier werden dann allerdings aus verständlichen Gründen eher soziodemographische Hintergrundsvariablen oder allgemeine Einstellungen und Erwartungen erfragt[18] – oder etwa zur Mitte einer entsprechenden Veranstaltung durchgeführt werden.

Zielsetzung all dieser Bemühungen ist es dabei, „den Lehrprozeß – die Vermittlung von Kenntnissen durch Professoren und die Aneignung von Wissen und Ausbildung durch die Studierenden – intern in Fachbereichen und Lehrveranstaltungen zu verbessern. Lehrveranstaltungen und Lehrkräfte variieren stark in ihrer Lehrqualität: Veranstaltungen gleichen Inhalts können bei einem Dozenten interessant und anregend sein, können eine effektive Prüfungsvorbereitung darstellen und wichtige Kenntnisse vermitteln, bei einem anderen Dozenten können sie jedoch ermüdend und intellektuell fruchtlos, unklar aufgebaut und ineffektiv sein. Im gleichen Sinne gibt es Veranstaltungs- und Lehrformen, die Studenten zu Mitarbeit und stofflicher Auseinandersetzung anregen, während andere zu passivem Zu- (oder Weg-) hören verleiten. Hier, an individuellen Lehrdefiziten und suboptimalen Lehrmethoden können Lehrveranstaltungsevaluationen ansetzen und können durch Informationen des Dozenten und veranstaltungsinterne Besprechung zwischen Lehrenden und Lernenden Optimierungsprozesse angestoßen werden“ (Rindermann 1996a: 12).

Wie unterschiedlich auch die genaue Umsetzung diesen Vorhabens in den einzelnen Fachbereichen und Universitäten ausfallen mag, der grundlegende Mechanismus, der im gerade angeführten Zitat dargestellt wird, ist überall gleich: Studierende sollen aus ihrer Sicht eine inhaltliche Rückkopplung über die Lehrveranstaltung an die Lehrenden geben, um so den Prozeß

[18] Vergleiche hierzu etwa die im Anhang A dokumentierte Befragung des Instituts für Soziologie der Freien Universität Berlin (Anhang A 3).

der Wissensvermittlung zu optimieren. Bei der Durchführung von Lehrevaluationen ist es aber häufig unklar, welche Dimensionen und welche Objekte (einzelne Lehrveranstaltungen, Studiengänge, Dozenten) überhaupt bewertet werden sollen. Auch aus diesem Grund wurde im Rahmen dieses Projekts versucht, durch eine empirische Untersuchung an allen sozialwissenschaftlichen Fakultäten in der Bundesrepublik, einen Überblick über die Praxis der Berücksichtigung studentischer Beurteilungen im Rahmen sozialwissenschaftlicher Studiengänge zu gewinnen.

4.2 Die Verbreitung studentischer Lehrbefragungen in sozialwissenschaftlichen Studiengängen an bundesdeutschen Hochschulen

Um die Frage nach der Verbreitung derartiger studentischer Lehrbefragungen im Rahmen sozialwissenschaftlicher Studiengänge an den Hochschulen der Bundesrepublik Deutschland zu beantworten, wurde im Rahmen des hier vorgestellten Projektes eine Befragung der entsprechenden Institute, Fachbereiche oder Fakultäten durchgeführt.

Eine Zielsetzung dieses Teilprojektes war es, einen Überblick über die bislang in sozialwissenschaftlichen Studiengängen eingesetzten Evaluationsinstrumente zu erhalten. Hierzu war sinnvoll, neben den verschiedentlich veröffentlichten Befragungsinstrumenten auch die anderen zur Verwendung kommenden, bislang jedoch nicht veröffentlichten Instrumente zu betrachten. Aus diesem Grunde wurde eine schriftliche Befragung durchgeführt.

4.2.1 Definition der Grundgesamtheit

Es sollten Daten über diejenigen Fachbereiche, Institute oder Fakultäten erhoben werden, an denen ein Studium der Sozialwissenschaften beziehungsweise eines ihrer Teilfächer möglich ist. Entsprechend wurden die Adressen aller universitären Einrichtungen ermittelt, die die Fächer Sozialwissenschaften, Soziologie, Politikwissenschaften sowie schließlich eben Verwaltungswissenschaft als Studienfach anbieten. Als Grundlage der entsprechenden Adressensammlung diente der bundesweite Internet-Studienführer, der durch die zentrale Studienberatung der Universität Münster organisiert wird. Insgesamt 94 verschiedene Adressaten ergaben sich als Ergebnis dieser Recherche.

Nicht immer konnte dabei eindeutig eine für die Durchführung der sozialwissenschaftlichen Studiengänge federführende Institution ausgemacht werden. Teilweise wurden aus diesem Grunde mehrere Fachgebiete umfassende Fachbereiche oder Fakultäten angeschrieben. Die entsprechenden Antworten beziehen sich aus diesem Grunde in dem ein oder anderen Fall also auf umfassendere Einheiten als die einzelnen Studiengänge der Sozialwissenschaften. In

den meisten Fällen war es jedoch möglich, die für die Fächer Soziologie und Politikwissenschaften verantwortlichen Einheiten direkt anzuschreiben.

Entsprechend muß beachtet werden, dass die Befragungseinheiten sehr heterogene Organisationen mit unterschiedlichsten Größen der vertretenen Fächer, sowohl in Hinsicht auf die Zahl der Studierenden als auch in Hinsicht auf die Zahl der Lehrenden, darstellen.

4.2.2 Durchführung der Befragung

Für die Befragung wurde ein nur eine Seite umfassender Fragebogen entwickelt, der zusammen mit einem Anschreiben an die Befragungseinheiten versandt wurde (Anschreiben und Fragebogen sind in Anhang C dokumentiert). Die Adressaten wurden gebeten, neben dem Fragebogen die in ihrer Institution eingesetzten Befragungsinstrumente sowie eventuelle vorhandene weitere Dokumente zur internen Lehrevaluation zurück zusenden.

Die Versendung der entsprechenden Anfragen erfolgte Anfang November 1999, ein Erinnerungsschreiben wurde nach drei Wochen im Dezember 1999 verschickt.

Von den angeschriebenen 94 Fachbereiche, Fakultäten und Institute antworteten insgesamt 81 auf unsere Anfrage. Die Antwortquote von 86,2 Prozent liegt deutlich über dem sonst bei schriftlichen Befragungen zu erwartenden Maße und kann daher wohl als ein Indiz für das Interesse der Adressaten für derartige Themen interpretiert werden.

4.2.3 Ergebnisse der Befragung

Von den an dieser Befragung teilnehmenden 81 Fachbereichen, Instituten oder Fakultäten mit sozialwissenschaftlichen Studiengängen haben mehr als zwei Drittel bereits mindestens einmal eine Lehrevaluation durchgeführt.[19] In Tabelle 4.1 sind die entsprechenden Antworten im Überblick dargestellt.

[19] Zwei Institute reagierten zwar auf die Anfrage, beantworteten jedoch nicht den beigefügten Fragebogen. In einem der beiden Fälle geschah dies mit dem Hinweis auf die sehr unterschiedliche Handhabung der Evaluation innerhalb der Fakultät. Freundlicherweise wurde jedoch das hier dann teilweise zum Einsatz kommende Befragungsinstrument beigelegt, so dass es hier Berücksichtigung finden kann. Für die in diesem Abschnitt des Textes vorgenommenen Analysen vermindert sich jedoch die Stichprobe auf 79 Einheiten.

Tabelle 4.1: Erfahrungen mit Lehrevaluation

	Anzahl	Anteil in Prozent
Lehrevaluation in jedem Semester	26	32,9
Lehrevaluation schon durchgeführt, aber nicht jedes Semester	30	38,0
Noch keine Lehrevaluation durchgeführt	23	29,1

Interessant ist nun jedoch auch, in welchem Umfang die entsprechenden Evaluationen denn nun durchgeführt werden. Hierzu wurde sowohl untersucht, ob alle, oder nur bestimmte Personengruppen innerhalb des Faches als auch, ob alle oder nur bestimmte Veranstaltungstypen evaluiert werden. Zuerst wird in Tabelle 4.2 die unterschiedliche Evaluation der einzelnen in der Lehre engagierten Gruppen betrachtet.

Tabelle 4.2: Lehrevaluation bei einzelnen Gruppen

	Ja (in Prozent)
Lehrevaluation bei Professoren	96,3
Lehrevaluation bei Mitarbeitern (Mittelbau)	96,3
Lehrevaluation bei Tutoren	77,7

In insgesamt drei Viertel aller Evaluation betreibenden Institutionen werden alle bei der Lehre Beteiligten – also Professoren, der Mittelbau als auch Tutoren – bewertet. Wenn einzelne Teilgruppen von der Evaluation ausgenommen werden, so sind dies vor allem die Tutoren. Diese werden jedoch immer noch in mehr als 77 Prozent aller Universitäten evaluiert. Auch hinsichtlich der Bewertung einzelner Lehrveranstaltungstypen findet sich ein ähnliches Bild. In fast 70 Prozent aller hier befragten Institutionen werden bei Evaluationen alle Arten von Lehrveranstaltungen bewertet, allerdings finden sich auch hier typische Unterschiede. In Tabelle 4.3 sind die entsprechenden Ergebnisse zusammengefasst.

Tabelle 4.3: Lehrevaluation bei einzelnen Veranstaltungstypen

	Ja (in Prozent)
Lehrevaluation bei Vorlesungen	94,4
Lehrevaluation bei Seminaren/Übungen	92,6
Lehrevaluation bei Tutorien	75,9

Wie bereits aus den in Tabelle 4.2 vorgestellten Ergebnissen hinsichtlich der einzelnen universitären Gruppen erwartbar, so werden vor allem die Vorlesungen und – zu einem nur unbedeutsam geringeren Anteil – Seminare und Übungen bewertet.

Wenn man die bislang vorgestellten Ergebnisse zusammenfassen will, so sind Lehrevaluationen im Rahmen sozialwissenschaftlicher Studiengänge bereits auf breiter Front eingeführt, jedoch nur in knapp einem Drittel aller an der Befragung teilnehmenden Institutionen finden diese Evaluationen in jedem Semester statt. Ein weiteres Drittel führt sie unregelmäßig durch und immerhin nahezu ein weiteres Drittel hat bislang überhaupt keine Erfahrungen mit der Evaluation von Lehrveranstaltungen gesammelt. Wenn evaluiert wird, so werden hiervon meist alle Veranstaltungen und alle lehrenden Gruppen betroffen. Eine kleine Ausnahme bilden dabei Tutorien und die entsprechenden Tutoren.

Zielsetzung des hier vorgestellten Projektes ist es nun jedoch nicht in erster Linie einen Überblick über die Evaluationspraxis in den Sozialwissenschaften zu geben, sondern vielmehr die hier zum Einsatz kommenden Befragungsinstrumente näher und kritisch zu betrachten. Aus diesem Grunde sollen im folgenden hier gesammelten Instrumente genauer analysiert werden. Wie bereits erwähnt, findet sich im Anhang A eine Dokumentation dieser Instrumente.[20]

4.2.3.1 Überblick über die eingesetzten Befragungsinstrumente

Von den angeschriebenen Instituten, Fachbereichen oder Fakultäten wurden insgesamt 37 Fragebögen zurückgesandt. Zwei dieser Institutionen verwiesen auf die Dokumentation ihrer Evaluationsbemühungen im Internet, aber immerhin 18 der Institutionen, die angaben, bereits einmal eine Evaluation durchgeführt zu haben, haben trotz der deutliche Bitte keinen entsprechenden Bogen beigelegt.[21]

Schon ein kurzer Blick auf die verschiedenen Evaluationsinstrumente zeigt die große Unterschiedlichkeit im Umfang der einzelnen Instrumente. Während einige Befragungsinstrumente bei einer sorgfältigen Bearbeitung durch die Studierenden sicherlich einen hohen Zeitaufwand mit sich bringen, umfassen andere Evaluationsbögen nur wenige Statements. Weiterhin handelt es sich nicht bei allen zugesandten Materialien um fertige Befragungsinstrumente; zum Teil (vgl. etwa A 6) handelt es sich um Itemsammlungen, bei denen aus einer Vielzahl von Bewertungsdimensionen ein eigenes Instrument zusammengestellt werden kann.

Zusammenfassend lassen sich bei den Evaluationsbögen zwei Idealtypen unterscheiden:

---

[20] Die zur Verfügung stehenden Befragungsinstrumente werden jeweils getrennt dokumentiert und von uns mit einer eigenen Ordnungsnummer versehen. Zu Beginn des Anhangs A findet sich eine Liste der aufgeführten Instrumente. Von einigen Institutionen wurden mehrere Versionen, etwa getrennt nach Vorlesungen und Seminaren, aber auch aufgrund individueller Unterschiede bei einzelnen Dozierenden, zugesandt. Im Anhang A sind alle eingeschickten Befragungsinstrumente dokumentiert.

[21] Hierbei besteht übrigens kein statistisch bedeutsamer Zusammenhang zwischen der Regelmäßigkeit der Lehrevaluation und der Bereitschaft, einen Fragebogen zurückzusenden.

- Auf der einen Seite steht dabei die kurze und im Rahmen üblicher Veranstaltungen wohl recht problemlos durchzuführende Befragung, deren Instrument meist auf einer einzigen Seite zusammengefasst ist (vgl. beispielsweise im Anhang die Instrumente A 1 oder A 26).
- Auf der anderen Seite befinden sich Instrumente, die eigentlich nicht der Evaluation einzelner Veranstaltungen, sondern einer erweiterten Analyse des Studierendenverhaltens und der allgemeinen Studienbedingungen und dem Studienablauf dienen sollen und die aus diesem Grunde auch wesentlich komplexer und vor allem auch umfangreicher gestaltet sind (vgl. etwa A 2 oder A 14).

Trotz dieser Heterogenität des zugrundeliegenden Materials lassen sich einige wichtige Dimensionen benennen, die fast durchgehend zum Standardrepertoire der an deutschen Hochschulen – zumindest in dem breit definierten Feld der Sozialwissenschaften – Verwendung findenden Evaluationsinstrumente gehören.

4.2.3.2 Dimensionen der Lehrevaluation

Insgesamt lassen sich in den verwendeten Instrumenten 11 verschiedene Dimensionen unterscheiden, auf die nun näher eingegangen werden soll.[22] Im folgenden sollen diese Dimensionen kurz einzeln besprochen und vorgestellt werden.[23] In diesem Rahmen werden dann gegebenenfalls die jeweiligen methodischen Probleme mit einzelnen Fragen und Operationalisierungen thematisiert.

4.2.3.2.1 Dimension 1: Um welche (Art von) Veranstaltung handelt es sich?

Fast durchgängig wird nach dem genauen Titel der Veranstaltung oder nach einem anderen entsprechend eindeutigen Identifikator wie etwa den an manchen Universitäten üblichen Veranstaltungsnummern gefragt. Wenn – wie es aus der Logik dieses Verfahrens eigentlich sinn-

---

[22] Vgl. hier auch die bereits oben vorgestellte Unterteilung bei Bülow-Schramm (1994: S9) sowie Kreuzer (1999:76ff).

[23] Selbstverständlich ist die Zuordnung des ein oder anderen Items sicher auch in einer anderen Art und Weise möglich als es im folgenden geschieht. Die hier zu findende Auflistung und Einordnung soll nur einen Überblick über die Vielgestaltigkeit von Lehrevaluationsinstrumenten geben und kann nicht eine entsprechende multivariate Analyse ersetzen, die notwendig wäre, um die einzelnen Dimensionen empirisch nachprüfbar zu unterscheiden (vgl. hierzu etwa Rindermann 1996a). Hierzu müssten aber natürlich die Rohdaten der Lehrevaluationen vorliegen. Es scheint bisher keinen Versuch gegeben zu haben, solche Ergebnisse der Lehrevaluationen verschiedener Universitäten systematisch auf der Basis von Rohdaten zu vergleichen. Abgesehen von der Datenschutzproblematik sind die dabei auftretenden methodischen Probleme sind nicht völlig trivial, vgl. hierzu allgemein Van de Vijver/Leung (1997).

voll und notwendig ist (vgl. hierzu noch Abschnitt 4.4) – eine individuelle Rückmeldung der Evaluationsergebnisse an die einzelnen Lehrenden geplant ist, erscheint dies auch unabdingbar. Nur in einigen Ausnahmen wird eine entsprechende Frage nicht gestellt (vgl. etwa A 2), hier handelt es sich dann aber auch nicht um Evaluationsinstrumente im herkömmlichen und oben diskutierten Sinne. Vielmehr wird hier etwa untersucht, wie die Studienmotivation und der Studienhintergrund der einzelnen Befragten gestaltet ist. Ein konkreter Bezug zu bestimmten Lehrveranstaltungen wird nicht hergestellt. Stattdessen finden sich etwa Fragen über die Bedeutung des Studiums im Vergleich zu anderen Lebensbereichen. Selbstverständlich werden jedoch auch in diesen Befragungen entsprechende Identifikatoren erhoben, um die spätere Zuordnung der zu verschiedenen Zeitpunkten erhobenen Informationen möglich zu machen. Selbst wenn nicht geplant ist, die jeweiligen Ergebnisse individuell zurückzumelden, so sollte doch für die weitere Analyse mindestens der Veranstaltungstyp – Vorlesung, Proseminar, Hauptseminar, Übung, Vorkurs, Grundkurs oder was sonst an Veranstaltungstypen möglich ist – erhoben werden. Nur so erscheint es möglich, weiterführende Analysen, etwa über den Grund der studentischen Zufriedenheit mit einzelnen Studienabschnitten, zu untersuchen.

4.2.3.2.2 Dimension 2: Einige demographische Angaben zu den Studierenden

In nahezu allen Bewertungsinstrumenten finden sich einige Fragen nach grundlegenden demographischen Angaben der Studierenden. So wird meist nach dem Alter und der Geschlechtszugehörigkeit, vor allem aber auch nach dem jeweiligen Studiengang und den Studiensemestern beziehungsweise Fachsemestern gefragt. Teilweise (vgl. etwa A 9) wird auch untersucht, wie hoch die individuelle zeitliche Belastung der Studierenden durch Lehrveranstaltungen ist. Für eine genauere Untersuchung der Zufriedenheit mit der universitären Lehre reichen diese wenigen Informationen aber selbstverständlich meist nicht aus. Deshalb werden hier häufig wesentlich detailliertere Abfragen durchgeführt (vgl. unten). Anhand der wesentlich weiter verbreiteten Erhebung basisdemographischer Eigenschaften läßt sich jedoch zumindest feststellen, ob etwa Studierende verschiedener Fachrichtungen oder mit unterschiedlichen Erfahrungen mit universitären Lehrveranstaltungen die Veranstaltung unterschiedlich einschätzen. Bemerkenswerterweise findet sich in keinem Lehrevaluationsbogen die Möglichkeit, die verschiedenen Fragebögen eines Studierenden einander zuordnen zu können. Obwohl dies aus der Sicht der Studierenden zunächst wünschenswert erscheinen mag (Vermeidung vermeintlicher Sanktionierung), ist dies aus methodischer Hinsicht bedauerlich. Es sind daher keinerlei testtheoretische Untersuchungen der Instrumente in verschiedenen Kontexten möglich (vgl. hierzu beispielsweise Rost 1996). Sollten Studierende ihre Bewertungskriterien nach Veranstaltungstypen, mit zunehmender Studiendauer oder zunehmender Erfahrung im Umgang mit Evaluationsinstrumenten verändern, so wirkt sich dies auf die Ergebnisse aus, kann

aber nicht festgestellt werden. Weiterhin ist zu erwarten, dass eine individuelle Zuordnenbarkeit der Instrumente zu einer Erhöhung der Ernsthaftigkeit (oder des „Commitments“) der Beantwortung führen würde. Die Vergabe der Identifikationsschlüssel könnte dabei durchaus anonym erfolgen, so dass keine Zuordnung der Evaluationsbögen z.B. zu Matrikelnummern möglich ist.

4.2.3.2.3 Dimension 3: Zur Konzeption und Struktur der Lehrveranstaltung

Wesentlich interessanter für die Evaluation von Lehrveranstaltungen sind Fragen, die sich mit der Konzeption und der Struktur der jeweiligen Übung oder Vorlesung beschäftigen. Auch hier sind jedoch sehr große Unterschiede im Umfang und damit der Tiefe der jeweiligen Fragen festzustellen.

In dem schon erwähnten kürzesten Bogen (vgl. A 26) wird nur danach gefragt, ob das Thema angemessen behandelt wird, die Veranstaltung übersichtlich strukturiert und die Arbeitsformen für die Veranstaltung geeignet sind. Allerdings sind mit den genannten drei Fragen einige wesentliche Dimensionen angesprochen. In fast allen Erhebungen finden sich Fragen nach der Strukturiertheit der Veranstaltung, der Klarheit des Aufbaus und der Gliederung. Häufig wird auch erhoben, ob innerhalb der Lehrveranstaltung der Bezug zu der Gesamtgliederung und den anderen Sitzungen deutlich geworden ist. Weiterhin wird meist auch die Frage gestellt, inwieweit ein Bezug zu anderen Lehrveranstaltungen beziehungsweise der Gesamtstruktur des Studienaufbaus hergestellt und deutlich gemacht wurde.

Da die Auswahl der entsprechenden Lehrveranstaltung – soweit es sich nicht um eine Pflichtveranstaltung handelt – stark durch die Ankündigung in den entsprechenden Vorlesungsverzeichnisses bestimmt wird, wird häufig danach gefragt, inwieweit der Inhalt der Veranstaltung mit den entsprechenden Ankündigungen übereinstimmte oder deutliche Abweichungen aufwies.

Ebenso finden sich auch mehrmals Fragen nach der Wissenschaftlichkeit der jeweiligen Präsentation durch die Dozierenden. Hier wird beispielsweise nicht nur nach der Angemessenheit der thematischen Behandlung (A 26) und dem Einbezug neuerer Forschungsergebnisse (A 5) gefragt, sondern auch, ob der Dozent fachliche Fehler mache (A6), oder über genügend Fachwissen verfüge (A 7).[24] Es erscheint uns mehr als problematisch, derartige Einschätzungen in

[24] Wenn man wirklich an einer Einschätzung der Kompetenz der Lehrenden durch die Studierenden interessiert ist, so kann man höchstens nach einer subjektiven Einschätzung fragen, wie dies etwa durch die Frage „Die Dozentin/der Dozent scheint mit den Lehrinhalten voll vertraut“ (vgl. A 10) geschieht. Alle anderen Fragen zielen wohl eher auf die allgemeine Darstellungsfähigkeit der Dozierenden denn auf ihre fachliche Kompetenz.

Lehrevaluationen vornehmen zu lassen. Studierende können derartige Fragen wohl kaum sinnvoll beantworten.

Wesentlich angemessener erscheint es, nach dem subjektiven Lernerfolg (A 10) zu fragen. Auch die Abstimmung der jeweiligen Lehrveranstaltung mit eventuell stattfindenden Tutorien oder Übungen fällt in diesen Bereich und ist sicherlich durch Studierende einschätzbar.

Relativ häufig wird nach dem wahrgenommenen Schwierigkeitsgrad der Veranstaltung und der Angemessenheit des stofflichen Umfangs der Lehrveranstaltung gefragt (vgl. etwa A 11).

In einigen Instrumenten wird danach gefragt, ob die Veranstaltung die Erwartung der Studierenden erfüllt wurden (vgl. A 13). Wenn – wie in diesem Beispiel - jedoch diese Erwartungen selbst nicht erhoben wurden, bleibt es offen, ob eine positive Antwort auf diese Frage als Indikator für eine gute oder eine schlechte Lehrveranstaltung anzusehen ist.[25]

Prinzipiell erwägenswert erscheinen Fragen zur wahrgenommenen Qualität der gehaltenen Referate anderer Seminarteilnehmer, auch wenn die Attribuierung der Ursache für die wahrgenommene Performanz fraglich bleiben muß. Allerdings könnte dieses Item einen möglichen Prädiktor für eine eventuell vorgenommene Gesamtbenotung der Veranstaltung darstellen.

Häufig wird nach dem Einsatz bestimmter Medien und – natürlich vor allem bei Vorlesungen – nach der Verfügbarkeit eines Skripts gefragt. Auch hierbei ist nicht unbedingt klar, wie die jeweiligen Antworten einzuordnen sind. Wenn etwa die Zielsetzung einer Veranstaltung die selbständige Erarbeitung der entsprechenden Literatur darstellt, kann ein Skript eher kontraproduktiv sein und die Erreichung dieser Zielsetzung negativ beeinflussen (Mußgnug 1992).

4.2.3.2.4 Dimension 4: Zur Vorgehens- und Verhaltensweise der Dozierenden

Neben den Fragen zur Bewertung der Veranstaltung an sich findet sich in nahezu allen Evaluationen auch vielfältige Fragen über den jeweiligen Dozenten und dessen Vorgehens- und Verhaltensweisen. Auch hier ist eine große Vielfalt – sowohl inhaltlich als hinsichtlich des Umfangs – zu beobachten und natürlich finden sich für die unterschiedlichen Veranstaltungs-

[25] Wie auch an anderen Stellen in diesem Bericht sollen diese Einwände nicht als Kritik an den jeweils die entsprechende Lehrevaluation durchführenden Institutionen oder gar Personen verstanden werden. So war etwa die hier angeführte Befragung zum Studium der Sozialwissenschaften an der Universität Bochum so angelegt, dass mit Hilfe eines einzigen Evaluationsbogens alle Veranstaltungen im Rahmen einer Interviewsituation bewertet werden sollten. Aufgrund dieser Aufgabenstellung war es gar nicht möglich, ausführlich die Erwartungen an die einzelnen Lehrveranstaltungen zu erheben (vgl. das entsprechende Instrument im Anhang A 13 und die Umsetzung der Fragen auf der letzten Seite). Häufig werden an Lehrbewertungen große Ansprüche gestellt, ohne die entsprechenden Maßnahmen mit genügend Mitteln auszustatten.

formen spezifische Fragen. Vor allem ist dabei zwischen naturgemäß sehr dozentenorientierten Vorlesungen und häufig stark von den Studierenden mitgeprägten Seminaren oder Übungen zu unterscheiden.

Unabhängig von der Veranstaltungsform wird jedoch fast immer das Engagement der Dozierenden erfragt, sei dies nun, indem man nach dem Interesse des Dozenten oder der Dozentin am Thema der Veranstaltung (A 1) oder direkt danach fragt, ob der Dozent seine Veranstaltung mit Engagement bestritt (A 24) beziehungsweise seine Lehre nur als lästige Pflicht ansehe (A 36).

Vor allem bei der Bewertung des Vortrags finden sich eine fast unerschöpfliche Variation der Subkategorien – lebendig, verständlich, systematisch, abschweifend, anregend, praxisbezogen, um nur eine kleine Auswahl zu nennen. Aber auch das generelle Tempo der Veranstaltung und recht allgemeine Einschätzungen – beispielsweise die Frage, ob die Lehre von den Dozierenden wichtig genommen wird, ob die Lehrveranstaltung gut vorbereitet war oder nach der Motivierungsfähigkeit der Dozierenden (vgl. A 1) – finden sich hier.

Sehr häufig wird auch versucht, die allgemeine wissenschaftliche und pädagogische Fähigkeit einzuschätzen, indem danach gefragt wird, ob der Dozent in der Lage sei, auch komplexe Sachverhalte verständlich darzustellen. Hinsichtlich des Verhaltens in eher seminaristischen Veranstaltungen steht vor allem die Betreuung der Studierenden bei Hausarbeiten und Referaten sowie die Anregungen und Leitung der inhaltlichen Diskussion im Mittelpunkt. So wird hier gefragt, ob es eine ausreichende Unterstützung bei den schriftlichen Arbeiten sowie genügend Gelegenheit zur Nachbesprechung gab (A 16) oder ob generell die Referatsthemen gut organisiert waren, es genügend Hinweise zur Vorbereitung der Referate und eine hilfreiche Kritik gab (A 5) und ob die Referate jeweils sinnvoll ergänzt worden sind (A 6). Zu diesem Bereich zählt unter anderem auch die Ansprechbarkeit des jeweils Dozierenden (vgl. beispielsweise A 28).

4.2.3.2.5 Dimension 5: Fragen zum eigenen Engagement der Studierenden

Einen fünften großen Frageblock stellen Statements zum eigenen Engagement der Studierenden selbst dar. Hier wird vor allem nach der regelmäßigen Anwesenheit in der Veranstaltung und gegebenenfalls nach den Gründen des Fehlens gefragt. Neben dieser Kontrolle wird häufig der eigene Arbeitsaufwand der Studierenden für die Veranstaltung erfragt. Dabei wird sowohl das faktische Verhalten (Durcharbeiten eine Skripts oder der eigenen Aufzeichnungen, Lektüre von grundlegender oder weiterführender Literatur sowie die Mitarbeit) als auch der zeitliche Aufwand erhoben.

Die konkrete Operationalisierung ist dabei durchaus kritisierbar. Fast immer wird nach der durchschnittlich verwendeten Zeit gefragt. Eine derartige Abfrage erfordert verschiedene Operationen und erscheint insgesamt weniger sinnig als etwa die konkrete Frage nach der Vorbereitungszeit in der letzten Woche.

Häufig wird auch nach dem eigenen Interesse an der entsprechenden Lehrveranstaltung gefragt. Darüber hinaus finden sich Fragen, die die generelle Studienmotivation erfassen sollen (vgl. A 9). Auch die Validität dieser Fragen scheint jedoch durchaus anzweifelbar.

Schließlich finden sich in einigen Instrumenten Fragen nach der Einschätzung des Engagements der Mitstudierenden, beispielsweise ob diese pünktlich und regelmäßig in den einzelnen Sitzungen erschienen, ob sie aufmerksam und interessiert waren und wie die allgemeine Einstellung zu der entsprechenden Lehrveranstaltung eingeschätzt wird. Obwohl sich mit dem vorliegenden Datenmaterial keine Untersuchungen zur Reliabilität und Validität solcher Angaben durchführen lassen, erscheinen beide Kriterien für diese Indikatoren äußerst fraglich.

4.2.3.2.6 Dimension 6: Art der Interaktionsbeziehung in den Sitzungen

Als eine weitere Dimension kann die Evaluation der Interaktionsbeziehungen in den Sitzungen einer Lehrveranstaltung angesehen werden. Diese Einschätzung hängt selbstverständlich sowohl von den Eigenschaften und Verhaltensweisen der Lehrenden sowie der Studierenden ab.

In diese Dimension fallen generelle Fragen nach dem Klima der Lehrveranstaltung (A 3) als auch detailliertere Abfragen anhand von Antonymskalen wie z.B. kooperativ vs. konkurrierend, interessant vs. langweilig, diszipliniert vs. chaotisch oder anregend vs. nicht anregend (vgl. A 1). Des weiteren finden sich hier Fragen zum Ausmaß der Beteiligung der Mitstudierenden.

Schließlich wird in einigen Evaluationsinstrumenten die soziale Kompetenz des Dozenten eingeschätzt (A 6). Hier finden sich etwa Items wie „der Dozent/die Dozentin behandelt die Studierenden arrogant von oben herab“ oder „gibt den Studierenden ausreichend Gelegenheit, selbst zu Wort zu kommen“.

4.2.3.2.7 Dimension 7: Warum wurde die zu evaluierende Veranstaltung besucht?

Vor allem bei der – in Abschnitt 5 ausführlicher dargestellten – Diskussion um die Erklärung eventuell unterschiedlicher Evaluationen einzelner Lehrveranstaltungen wird der Motivation der Studierenden häufig eine große Rolle zugeschrieben. Sicherlich auch aus diesem Grunde finden sich in vielen der hier untersuchten Evaluationsinstrumente Fragen danach, warum die zu evaluierende Veranstaltung besucht wurde.

Neben der Möglichkeit zur offenen Antwort finden sich hier eine ganze Reihe denkbarer Ursachen. Der Katalog der vorgegebenen Antwortmöglichkeiten umfasst als mögliche Ursachen für den Besuch der Veranstaltung unter anderem die folgenden Items:

- Es handelt sich um eine Pflichtveranstaltung
- Es soll ein Leistungsnachweis in dieser Veranstaltung erzielt werden
- Aufgrund des Praxisbezugs
- Die Lehrveranstaltung ist allgemein wichtig
- Die Lehrveranstaltung eignet sich als Prüfungsvorbereitung
- Persönliches Interesse am Thema
- Besondere Befähigung der Lehrperson
- Die Veranstaltung liegt zeitlich oder räumlich günstig
- Die Veranstaltung wird von der Fachschaft oder anderen Studenten empfohlen
- Die Veranstaltung erfolgt aufgrund einer Empfehlung der Studienberatung
- Die Veranstaltung wird wegen der Überfüllung anderer Veranstaltungen besucht
- Freunde und Bekannte besuchen diese Veranstaltung

Auch diese recht ausführliche Auflistung läßt sich sicherlich noch problemlos verlängern. Für eine ernsthafte Analyse der Studienmotivation müssten Daten zu Antworten auf eine vollständige Aufzählung mit Hilfe multivariater Analyseverfahren untersucht werden. Erstaunlicherweise liegen solche Untersuchungen bislang nicht vor.

4.2.3.2.8 Dimension 8: Unter welchen Rahmenbedingungen findet die Veranstaltung statt?

Gerade in der öffentlichen und veröffentlichten Diskussion um den Zustand der universitären Ausbildung und Lehre wird auch immer auf die äußeren Rahmenbedingungen des Studiums verwiesen. Überfüllte Seminare, eine mangelhafte technische Ausstattung vieler Hörsäle und Seminarräume oder eine mit immer knapper werdenden Mitteln haushaltende Bibliothek werden in dieser Diskussion immer wieder angesprochen und nicht zufälligerweise plazieren sich

die neuen und damit in dieser Hinsicht sicher besseren Universitäten in entsprechenden Rating- und Ranglisten relativ gut. Um den Einfluß derartiger Faktoren bei der Bewertung von Lehrveranstaltungen sinnvoll berücksichtigen zu können, werden diese äußeren Umstände in einer Vielzahl von Evaluationsinstrumenten miterhoben.

So wird hier nach dem Platzangebot im Raum, der Raumgröße, der Akustik und den Sichtbedingungen im Veranstaltungsraum, aber auch der Lärmbelästigung von außen sowie der technischen und apparativen Ausstattung des Raumes gefragt.

In einigen Instrumenten wird danach gefragt, ob die Veranstaltung überfüllt war. Daneben wird in einigen Instrumenten der Zugang zu entsprechender Fachliteratur, die Öffnungszeiten und die Ausstattung der Bibliothek, die Verfügbarkeit entsprechender Arbeitsplätze und in letzter Zeit vermehrt auch der Zugang zu PC-Arbeitsplätzen erhoben.

4.2.3.2.9 Dimension 9: Ein Gesamturteil über die Veranstaltung

Neben dieser Fülle von Detailinformationen und Einzelabfragen findet sich aber auch fast durchgängig in allen Evaluationsinstrumenten eine zusammenfassende Beurteilung der Lehrveranstaltung.

Meist wird hier um die Antwort auf eine einzelne Frage gebeten, die allerdings wiederum unterschiedliche inhaltliche Dimensionen ansprechen kann. So wird hier beispielsweise auch danach gefragt, ob insgesamt der Lerneffekt hoch war oder eher nicht (A 1). Häufiger finden sich jedoch allgemeine Statements, wie „Insgesamt bin ich mit der Veranstaltung zufrieden“ (A 3, ähnlich auch in etlichen anderen Evaluationsinstrumenten wie etwa A 31), „alles in allem habe ich in der Vorlesung viel gelernt“, „die Vorlesung hat angeregt, mich weiter mit der Thematik zu beschäftigen“ (A 35), „wie zufrieden bist Du mit dieser Veranstaltung insgesamt“ (A 33) oder schließlich eine einfache Gesamtbeurteilung der Lernveranstaltung (A 35).

Diese allgemeinen Urteile werden nun ab und an für einen Vergleich verschiedener Lehrveranstaltungen und als zusammenfassende Bewertung herangezogen. Es sollte aber aus der bisher vorgestellten Diskussion deutlich geworden sein, dass es sich bei der Evaluation von Lehrveranstaltungen um kein einfaches und einheitliches Konstrukt handelt, das mit Hilfe eines derart einfachen Indikators – etwa auch mit Hilfe einer einfachen Notenskala – abzubilden ist. So verlockend es auch sein mag, sich auf diese Gesamteinschätzung zu beziehen, so sollte doch immer die Vielschichtigkeit des Evaluationsvorgangs berücksichtigt werden.

4.2.3.2.10 Dimension 10: Die Möglichkeit zu einer nicht vorstrukturierten Bewertung

Ebenso wie die einfache, vorstrukturierte zusammenfassende Bewertung der gesamten Lehrveranstaltungen, so finden sich auch in nahezu allen hier betrachteten Instrumenten die Möglichkeit, bestimmte besonders positive oder negative Punkte offen zu formulieren. Wenn die Veranstaltungsbewertung jedoch zu einer fest institutionalisierten Vorgehensweise werden soll, so ist zu bedenken, dass derartige offene Abfragen so gut wie nie hinreichend ausgewertet werden. Sie haben eher eine ritualisierte Funktion.

4.2.3.2.11 Dimension 11: Ausführlichere Angaben zum soziodemographischen Hintergrund sowie die Erhebung der Studienmotivation und den Vorstellungen zum Studium

Einige der eingeschickten Erhebungsinstrumente dienen nicht in erster Linie zur Beurteilung einzelner Lehrveranstaltungen, sondern versuchen allgemeinere Fragestellungen im Zusammenhang mit dem Studienverhalten zu beantworten (vgl. insbesondere A 2, A 8 oder A 14).

In einem solchen Zusammenhang ist es dann auch möglich, wesentlich ausführlicher beispielsweise den sozialen Hintergrund, die Studienmotivation und generell die Lebenssituation von Studierenden zu erheben. Dies erlaubt dann natürlich z.B. fundiertere Aussagen zur Erklärung von Abbruchquoten oder der Studiendauer.

In solchen Instrumenten finden sich dann Fragen zu eventuellen Vorerfahrungen an der Universität, der genauen Studienpläne, der Bedeutung des Studiums im Vergleich zu anderen Lebensbereichen, konkreten Vorstellungen über die Berufswahl, der genauen Studienmotivation, schulischen Vorbedingungen und Leistungen, allgemeinen Angaben über Lernen und Studieren, der bisherigen beruflichen Erfahrung, zur Finanzierung des Studiums oder zur sozialen Herkunft.

Zwar wäre es interessant, derartige Angaben auch für die Erklärung eventuell unterschiedlicher Evaluationseinschätzungen heranzuziehen, für den standardisierten Gebrauch im universitären Alltag sind derartige Erhebungsinstrumente jedoch allein aufgrund des Umfangs kaum einzusetzen.

4.2.3.3 Zusammenfassung

Diese Darstellung der verschiedenen, in sozialwissenschaftlichen Studiengängen zum Einsatz kommenden Evaluationsinstrumente soll abschließend noch einmal quantitativ betrachtet werden. Welche der einzelnen oben skizzierten Dimensionen wie häufig in den Instrumenten

verwendet wird, findet sich in der Tabelle 4.4. Aufgeführt ist jeweils der Prozentsatz der Befragungsinstrumente, die Fragen aus der entsprechenden Dimension enthalten. Ausgangspunkt sind dabei die 38 Instrumente, die auch im Anhang A dokumentiert sind. Es ist hier aber nochmals darauf hinzuweisen, dass sich hinter einer positiven Nennung durchaus sehr unterschiedlich umfangreiche und auch sehr unterschiedlich sinnvolle Fragen und Operationalisierungen finden lassen.

Tabelle 4.4: Berücksichtigung der einzelnen Dimensionen in den vorhandenen Instrumenten

Dimension	Anteil in Prozent
Welche Veranstaltung	71,1
Demographische Kurzangaben	65,8
Konzeption und Struktur der Lehrveranstaltung	84,2
Vorgehensweise der Dozierenden	84,2
Engagement der Studierenden	60,5
Art der Interaktionsbeziehungen in den Sitzungen	31,6
Motivation zum Besuch der Lehrveranstaltung	21,1
Rahmenbedingungen	23,7
Gesamturteil	28,9
Möglichkeit zur offenen Bewertung	63,2
Ausführlichere demographische Angaben	28,9

Es zeigen sich hier durchaus interessante Unterschiede: Fast immer findet sich eine Einschätzung zur Konzeption und Struktur der Lehrveranstaltung oder zur Vorgehensweise der Dozierenden.

Seltener – und vor allem eher bei Untersuchungen zur allgemeinen Studienmotivation als bei konkreten Veranstaltungsbewertungen – finden sich ausführlichere Fragen zum sozialen Hintergrund und anderen demographischen Variablen.

Ein gründlicherer Blick auf die im Anhang dokumentierten Instrumente macht zudem deutlich, dass viele der hier zu findenden Fragebögen recht konkret auf die Situation in den einzelnen Fachbereichen, Instituten oder Fakultäten bezogen ist.

Die allermeisten dieser Befragungsinstrumente sind nicht entsprechend den methodischen Standards getestet worden.[26] Nicht zufälligerweise wird etwa in einer aus den Niederlanden

[26] Vgl. hierzu einführend Schnell/Hill/Esser (1999: 121ff). Eine Ausnahme bildet übrigens der von der Fakultät für Verwaltungswissenschaft in Konstanz verwendete Fragebogen, der systematischen Pretests unterworfen wurde.

übernommenen Anleitung zur studentischen Evaluation der Lehre darauf hingewiesen, dass bei der Konstruktion eines Fragebogens selbstverständlich die in der empirischen Sozialforschung üblichen Standards Verwendung finden sollten – und aus gutem Grund werden diese Regeln dort dann auch noch einmal kurz wiederholt (Willems/Gijselaers/de Bie 1994: 60).

Es wird häufiger anscheinend in den Hintergrund gedrängt, dass auch die Erhebung von Evaluationen bei Studierenden einen Meßprozeß darstellen soll, der eben den hier geltenden Regeln zu unterwerfen ist. Letztlich sollten also auch die hier Verwendung findenden Instrumente vorab auf ihre Eigenschaften geprüft werden. Nur selten scheinen diese Forderungen jedoch erfüllt zu sein. Eine Berücksichtigung dieser Regeln würde zumindest helfen, die gröbsten Fehler zu vermeiden.

Eine wesentliche und bedeutsame Ausnahme stellt hierbei jedoch das sogenannte Heidelberger Inventar zur Lehrveranstaltungs-Evaluation dar.[27] Dieses Instrument durchlief verschiedene Test- und Erprobungsphasen und wird hier zum Vergleich im Anhang B 1 dokumentiert.[28] Natürlich finden sich hier viele Parallelen zu den oben besprochenen, im Rahmen dieses Projekts gesammelten Instrumenten. Aus diesem Grunde läßt sich durchaus vermuten, dass in der Regel auch mit den üblichen Instrumenten nicht reine Artefakte produziert werden. Trotzdem muß man auch bei Lehrevaluationen den Standards der empirischen Sozialforschung Geltung verschaffen. Dies setzt einen nicht unerheblichen finanziellen und personellen Aufwand bei der Instrumentenkonstruktion voraus.

4.3 Zum Wirkungsmechanismus von Studierendenbefragungen

Es sollte deutlich geworden sein, dass es sich bei der Befragung Studierender sicherlich um das am häufigsten zum Einsatz kommende Evaluationsverfahren handelt. Erstaunlicherweise wird dabei die konkrete Zielsetzung jedoch nur recht selten thematisiert. Wie sollen Studierendenbefragungen eigentlich wirken, um – kurz- oder langfristig – dazu beizutragen, die Qualität universitärer Lehre zu verbessern?

Auf die implizit unterstellten theoretischen Wirkmechanismen wird in der Literatur und den Berichten über durchgeführte Lehrevaluationen nur sehr selten eingegangen. Bereits oben wurde eine der wenigen Ausnahmen zitiert: Rindermann (1996a: 12) schreibt, dass Lehrevaluationen versuchen, durch Informationsprozesse interne Optimierungsprozesse anzustoßen.

---

[27] Vgl. hierzu ausführlich Rindermann (1996a) und den Anhang B dieses Berichts.

[28] Im Anhang B finden sich auch noch einige weitere, in der Literatur zu findende Evaluationsinstrumente, die die Übersicht der insgesamt eingesetzten Erhebungsmöglichkeiten ergänzen sollen. Selbstverständlich soll damit aber kein Anspruch auf Vollständigkeit verbunden sein.

Nun ist es in den meisten Fällen so, dass die entsprechenden Bewertungsmaßnahmen am Ende einer Lehrveranstaltung durchgeführt werden – und somit nur meist als summative Evaluation zu verstehen sind. Einen formativen Charakter erhalten sie jedoch dann, wenn man mehrere Lehrveranstaltungen im zeitlichen Ablauf betrachtet. Es ist wohl gedacht, dass durch die Kritik in der einen Lehrveranstaltung die zeitlich darauf folgenden Seminare oder Vorlesungen qualitativ besser werden, da die Dozierenden durch diese Kritik auf bestimmte Mängel ihrer Lehre hingewiesen worden sind.

Es ist hier nicht der Ort, diese sehr positive und harmonische, vielleicht aber etwas naive Vorstellung über die Ursachen und Bestimmungsgrößen der Qualität universitärer Lehre näher zu diskutieren. Die Vorstellung, dass allein die Information über die geringe Qualität der Lehrveranstaltung genügen könnte, in folgenden Seminaren und Vorlesungen besser zu werden, entspricht kaum dem Stand der Forschung im Bereich der Handlungstheorien.

Anstelle einer letztlich doch fruchtlosen theoretischen Diskussion des angenommenen Wirkungsmechanismus soll im folgenden kurz auf empirischen Wege diese Erklärung näher untersucht werden. Hierbei kann auf verschiedene Lehrevaluationen zurückgegriffen werden, die an der Fakultät für Verwaltungswissenschaft der Universität Konstanz durchgeführt worden sind. Anhand dieser Daten kann getestet werden, ob es wirklich im Verlaufe längerer Evaluationsprozesse zu einer Verbesserung der Lehrqualität gekommen ist.[29]

4.3.1 Beschreibung der Datenbasis

An der Fakultät für Verwaltungswissenschaft der Universität Konstanz hat die Evaluation von Lehrveranstaltungen bereits eine relativ lange Tradition. In der Zwischenzeit liegen seit für insgesamt sechs Semester, beginnend mit dem Sommersemester 1997, entsprechende Daten vor. Ein Blick auf das entsprechende Erhebungsinstrument (vgl. im Anhang A 38) zeigt, dass der hier zum Einsatz kommende Fragebogen den üblichen Instrumenten gleicht (vgl. Kapitel 4.2).

Auch in diesem Instrument wird zwischen der Bewertung des Dozierendenverhaltens und der Einschätzung der Veranstaltung unterschieden. Ebenso finden sich Fragen nach dem Engagement der Studierenden für die Lehrveranstaltung sowie die Möglichkeit zur offenen Kritik und Stellungnahme. Wie in vielen Universitäten, so wird auch dieser Bogen gegen Ende der Veranstaltung eingesetzt und beruht auf der freiwilligen Bereitschaft sowohl der Studierenden wie auch der Veranstaltungsleitung. Für die im folgenden vorzustellende Analyse wurde auf die

[29] Die bisherigen Auswertungen der entsprechenden Lehrevaluationen beziehen sich immer nur auf die jeweils aktuelle Erhebung (vgl. Fakultät für Verwaltungswissenschaft 1999a; 1999b).

Erhebungen aus den Wintersemester 1997/98 sowie 1998/99 und den Sommersemester 1998 und 1999, also insgesamt aus vier unterschiedlichen, aufeinanderfolgenden Zeitpunkten, zurückgegriffen.[30] Tabelle 4.5 gibt einen Überblick über die Zahl der hier beteiligten Studierenden, der bewerteten Dozenten und der bewerteten Veranstaltungen.

Tabelle 4.5: Beteiligung an den Konstanzer Lehrevaluationen

	Studierende	Dozenten	Veranstaltungen
Wintersemester 1997/98	1.257	33	62
Sommersemester 1998	1.093	28	51
Wintersemester 1997/98	1.513	43	57
Sommersemester 1999	1.058	42	40

In den weiteren Analysen werden nur diejenigen Dozenten berücksichtigt, für die aus mindestens drei der vier hier untersuchten Semester Bewertungen vorliegen. Bei insgesamt 16 Lehrenden war dies der Fall.[31]

Sicherlich wäre es wünschenswert, nur Lehrveranstaltungen mit identischem Lehrinhalt zu vergleichen, da ja vor allem hier Lernprozesse für die Lehrgestaltung erwartbar sind. Ein derartiges Vorgehen würde jedoch die Fallzahl fast auf null reduzieren. Aus diesem Grunde werden hier nur themenübergreifende Bewertungen weiter berücksichtigt.

Dazu gehört die Frage nach dem Ausmaß der Vorbereitung des Dozierenden sowie die Frage nach einer zusammenfassenden Bewertung des Dozenten oder der Dozentin. Bei beiden Abfragen sollten die Studierenden ihre Dozenten auf einer Notenskala von „sehr gut“ bis „mangelhaft“ bewerten. Für die Analyse muß (fälschlich) angenommen werden, dass diese Eigenschaften unabhängig vom Veranstaltungsthema sind und deshalb ein Vergleich über die verschiedenen Semester möglich ist.

[30] Die Angaben zum Wintersemester 1999/2000 lagen zum Zeitpunkt der Analyse noch nicht vor, die entsprechenden Angaben aus dem Sommersemester 1997 konnten leider mehr nicht den einzelnen Veranstaltungen zugeordnet werden.

[31] Die im Vergleich dazu hohe Zahl der insgesamt evaluierten Personen (vgl. Tabelle 4.5) beruht darauf, dass hier auch Tutoren und Tutorinnen enthalten sind, die nie über mehrere Semester bewertet wurden. Ein Blick auf die hier weiter analysierten Personen zeigt, dass hier fast alle Universitätsprofessoren erfasst wurden.

4.3.2 Ergebnisse

Die Tabelle 4.6 zeigt die entsprechend der Veranstaltungsgröße gewichteten Mittelwerte der beiden Bewertungsdimensionen sowie zusätzlich die schlechteste durchschnittliche Bewertung einer Veranstaltung in jedem Semester.

Tabelle 4.6: Ergebnisse der Evaluation

		Vorbereitung		Zusammenfassende Bewertung	
	n	Mittelwert	Schlechtester Wert	Mittelwert	Schlechtester Wert
WS 1997/98	732	2,09	2,92	2,09	2,52
SoS 1998	745	2,04	2,59	2,06	2,57
WS 1997/98	726	2,08	2,61	2,25	2,77
SoS 1999	650	2,03	2,80	1,96	2,53

Als erstes Ergebnis läßt sich hier festhalten, dass hinsichtlich der beiden hier betrachteten Bewertungsmaßstäbe – der Vorbereitung der Dozierenden sowie ihre allgemeine Einschätzung – die Konstanzer Studierenden eine sehr positive Einschätzung vornehmen. Wie auch an anderen Universitäten, so zeigt sich auch hier, dass zumindest aus der Sicht der Betroffenen die Lehrveranstaltungen und vor allem ihre Leiter und Leiterinnen ein gutes und engagiertes Bild abgeben.

Dieses Ergebnis zeigt sich dabei jedoch nicht erst am Ende des hier beobachteten Evaluationsprozesses. Die Einschätzungen waren auch schon zu Beginn recht positiv. Allein aus diesem Grunde kann der meist unterstellte positive Effekt von Lehrevaluationen hier fast gar nicht eintreten. Dementsprechend sieht man auch bei einem weiteren Blick auf die Tabelle, dass sich kein Entwicklungstrend bei den vier hier analysierten Evaluationen feststellen läßt. Die Veränderungen zwischen den einzelnen Jahren sind vielmehr – wenn man die entsprechenden statistischen Standardfehler berücksichtigt – eher zufällige Schwankungen.

Nun soll nicht vorschnell aufgrund dieser Ergebnisse ein positiver Effekt von Evaluationsmaßnahmen ausgeschlossen werden. Ein erster berechtigter Einwand gegen die gerade vorgestellten Analysen besteht darin, dass man nicht die aggregierten, sondern die individuellen Übergänge betrachten muß.

Eine detaillierte Analyse der individuellen Veränderungen im Laufe der vier Evaluationen zeigt, dass sich kaum eine der mehr als 70 möglichen Veränderungen statistisch als bedeutsam

erweist.[32] Fast alle Veränderungen müssen als statistisch zufällig klassifiziert werden. Bei den wenigen signifikanten Ausnahmen läßt sich kein systematischer Trend feststellen. Diese Beobachtung gilt dabei sowohl für die konkrete Einschätzung als auch für die allgemeine Bewertung.

Ein zweites Gegenargument könnte nun lauten, dass vor allem die zuvor als relativ schlecht eingestuften Lehrveranstaltungen von einer Evaluation positiv beeinflußt werden.

Auch dieses Argument ist empirisch wohl nicht haltbar. Dies wird deutlich, wenn man sich die individuellen Entwicklungsprofile betrachtet, die im Anhang D dokumentiert sind.[33] Zwar ist auf den ersten Blick eine gewisse Veränderung festzuhalten, diese Unterschiede sind jedoch ebenfalls statistisch nicht bedeutsam und eher als Schwankungen um einen gemeinsamen Mittelwert denn als inhaltliche Effekte der Verbesserung der Lehre aufgrund der erfolgten Bewertung einzuschätzen. Insgesamt lassen sich von den 16 Dozenten mit ausreichender Datenbasis bei 7 keine, bei 5 eine positive Veränderung und bei 4 eine negative Veränderung feststellen. Es muß aber bedacht werden, dass die Teilnehmerzahlen von Semester zu Semester schwanken – und sich hier zwischen 3 und 160 bewegen – und das sich auch die Art der Veranstaltung – Vorlesung oder Seminar, Pflichtveranstaltung oder nicht – ändern kann. So lassen sich zwar signifikante Veränderungen feststellen, aber kaum eine signifikante und konsistente Tendenz.

4.3.3 Zusammenfassung

Die Ergebnisse der hier nur ausschnittsweise berichteten Analysen der Konstanzer Lehrevaluationen sprechen kaum für den zumeist unterstellten Lerneffekt der Lehrenden aufgrund der studentischen Bewertung. Dieses Resultat steht in Übereinstimmung mit den Ergebnissen anderer Studien. Rindermann (1996b: 139) berichtet von einer Studie, bei der neben der reinen Rückmeldung der Bewertungsergebnisse auch noch die Möglichkeit der internen Besprechung bestand: „Feedback oder Feedback und Besprechung führten zu keiner nachweisbaren Verbesserung der Lehre“ (Rindermann 1996b: 139). Sohr (1993: 170) sieht denn auch in der totalen Konsequenzenlosigkeit das Hauptproblem der bisherigen Evaluationen. Auch für Untersuchungen etwa an amerikanischen Universitäten lassen sich ähnliche Ergebnisse berichten (vgl. hierzu die Angaben in Rindermann 1996: 139f).

---

[32] Auf eine detaillierte Darstellung muß hier aus Platzgründen verzichtet werden.

[33] In den entsprechenden Abbildungen sind neben dem Mittelwert auch die Konfidenzintervalle der einzelnen Werte abgebildet.

Bei der Bewertung der nur geringen Konsequenzen der Konstanzer Lehrevaluation muß aber bedacht werden, dass generell die Qualität der Lehre in den einzelnen Veranstaltungen von den Studierenden als hoch eingeschätzt wird. Selbst die schlechteste Einzelbewertung lautet noch besser als „befriedigend“ und insgesamt liegt der Mittelwert bei „gut“. Angesichts dieses Ergebnisses stellt sich ohnehin die Frage nach dem generellen Sinn der Lehrevaluation.

5. Eignen sich Studierendenbefragungen zur Lehrevaluation? Zum Stand der methodischen Diskussion

Neben der Dokumentation der hier zum Einsatz kommenden Befragungsinstrumente, war es ein weiteres Hauptziel des hier vorgestellten Projektes, einen Literaturüberblick über methodische Diskussionen über die Evaluation von Lehrveranstaltungen mit Hilfe studentischer Lehrevaluationen zu erstellen. Die zentrale Frage hierbei ist die Frage nach der methodischen Güte der Beurteilungen. Sollten die Reliabilität und/oder Validität der Studierendenbefragungen fraglich sein, dann können die Ergebnisse nicht die Grundlage eines erfolgreichen Handelns mit dem Ziel der Verbesserung der Lehre sein.

Genau zu diesem Punkt findet sich in den letzten Jahren eine sehr ausführliche Diskussion in der deutschsprachigen Literatur. In diesem Abschnitt sollen die in dieser Diskussion vorgetragenen Argumente vorgestellt und gegeneinander abgewogen werden. Gerade bei der Diskussion um die Möglichkeiten studentischer Lehrevaluation erscheint eine Darstellung der entsprechenden Debatten jedoch mehr als unvollständig, wenn das Augenmerk nur auf die deutschsprachige Diskussion gerichtet bleibt.[34] Gerade in den Vereinigten Staaten haben Lehrevaluationen eine wesentlich längere Tradition – und übrigens meistens auch eine wesentlich größere Bedeutung – und es ist deshalb eigentlich wenig erstaunlich, dass sich hier eine lange Forschungstradition über die Güte entsprechender Einschätzungen findet. Aus diesem Grunde soll diese, in der bundesrepublikanischen Debatte erstaunlicherweise zum Teil häufig einfach ignorierte, Forschungslinie in einem zweiten Schritt vorgestellt und ihre wichtigsten Ergebnisse skizziert werden.

5.1 ‚Können Studierende Lehrveranstaltungen evaluieren?' – Skizze einer aktuellen Diskussion

Auf diese einfache, in der Überschrift des Kapitels wiedergegebenen Frage (vgl. Gold 1996) läßt sich der Gegenstand einer Diskussion zusammenfassen, die die deutschsprachige Diskussion über studentische Lehrevaluation zumindest in Teilen wesentlich bestimmt hat.

Ausgangspunkt dieser Debatte war dabei eine Reihe von Beiträgen, in denen aufgrund methodischer Fragestellungen die Nützlichkeit studentischer Evaluationen kritisiert wurde (vgl. Kromrey 1993a; 1993b; 1994a; 1994b; 1995a; 1995b; vgl. auch Süllwold 1992).[35]

---

[34] Vgl. zu dieser Forderung sowohl Kromrey (1995b) als auch Rindermann (1996b).

[35] Neben dieser methodischen Kritik findet sich auch häufig auch eine eher auf diffusen Gründen beruhende Ablehnung von Lehrevaluationen, die darin etwa die wissenschaftliche Freiheit beschränkt sehen (Mußgnug 1992), eine aufkommende Institutionalisierung der „Schnüffelei" (Giessen 1995: 44) ohne datenschutzrechtliche Grundlage befürchten oder glauben, dass Lehrevaluationen wenig Sinn machen, da „die Lehrkompetenz

Ein erster Kritikpunkt betrifft dabei die ungerechtfertigte und falsche Verwendung des Evaluationsbegriffs für studentische Befragungen: „Bei den in Frage stehenden Erhebungen handelt es sich nicht um ‚Evaluationen' von Lehre im Sinne eines methodisch kontrollierten sozialwissenschaftlichen Untersuchungsansatzes, sondern lediglich um die Erhebung bewerteter (also ‚evaluativer') Aussagen von ‚Betroffenen' (hier: Studierenden), also um Lehrveranstaltungs-Umfragen" (Kromrey 1993a: 43; vgl. auch Kromrey 1995b).

Doch auch abgesehen von dieser eher formalen Kritik, erscheinen die entsprechenden Ergebnisse wenig brauchbar, denn diese Aussagen seien keine Evaluationen, sondern nur „höchst subjektive und individuell unterschiedlich zustande kommende (...) ‚Akzeptanz-Aussagen'" (Kromrey 1993: 44). Die Verwendung von Veranstaltungsbefragungen zur Lehrevaluation sei dabei ein Gebiet voller methodischer Fallstricke, die meist nicht gesehen und berücksichtigt werden (Kromrey 1994a).

Die Kritik richtet sich dabei vor allem gegen die Verwendung dieser Instrumente zur allgemeinen Messung einer Lehrqualität und dem daraus abgeleiteten Versuch, Vergleiche zwischen einzelnen Lehrveranstaltungen oder zwischen verschiedenen Lehrenden zu ziehen oder schließlich ein Ranking verschiedener Lehrenden, Veranstaltungen oder Universitäten zu errichten. Lehrveranstaltungskritik als Rückmeldung an den jeweiligen Dozenten oder als Akzeptanzerhebung steht weniger im Brennpunkt der Kritik (Kromrey 1994a: 107). Evaluiert werde dabei aber weniger die Lehre als die Studierenden selbst (Kromrey 1994a: 108ff). Besonders kritisch wird aber die Verwendung von Durchschnittswerten als Indikator der Lehrqualität betrachtet. Hinter einem derartigen Vorgehen stehe die Begründung, dass studentische Wertungen aussagekräftig sind, wenn sie in großer Zahl übereinstimmen. Wäre diese Bedingung erfüllt, „wäre sicher auch die weitere (ergänzende) Unterstellung zu rechtfertigen, dass die verbleibenden individuellen Abweichungen keine ins Gewicht fallende Verzerrung verursachen, sondern sich bei hinreichend großer Zahl von Befragten ausgleichen" (Kromrey 1994a: 112). Es muß jedoch von einer recht hohen Heterogenität der Bewertung ausgegangen werden. Empirisch finden sich sehr unterschiedliche Bewertungsprofile, so dass wohl nur selten eine größere Gruppe von Studierenden die jeweiligen Durchschnittsprofile vertritt. Es dominiert die Meinungs- und Urteilsvielfalt unter den Studierenden. Diese Vielfalt findet sich natürlich vor allem, wenn neben globalen Einschätzungen auch noch sehr detaillierte Einzeldimensionen bewertet werden sollen (vgl. hierzu Kapitel 4). Die große Heterogenität schließe deshalb die Verwendung einfacher Mittelwerte aus, denn: „Bekanntlich ist ein Mittelwert nur dann sinnvoll interpretierbar, wenn sich die einzelnen Beobachtungswerte mehr oder weniger

---

von Dozenten (...) im Urteil von Studierenden ein derart unspezifischer Urteilsgegenstand [sei], dass er am ehesten aus kontextuellen Merkmalen der Lehrveranstaltung herleitbar ist" (Scholz 1995: 500).

stark in einem ‚typischen' Wertebereich konzentrieren" (Kromrey 1994b: 157). Trotz dieser Kritik glaubt Kromrey, dass Vorlesungsbeurteilungen wichtige Informationen, etwa über die Zusammensetzung der jeweiligen Zuhörerschaft, über deren Interessen und Wünsche und über die Akzeptanz der jeweiligen Veranstaltung liefern können. In einer Antwort auf diese Vorwürfe verweist Rindermann (1996b: 134ff) auf die internationale Forschung, die sehr wohl einfache Mittelwerte verwendet. Natürlich ist es jederzeit möglich, durch die Hinzunahme weiterer Informationen, etwa über die jeweilige Streuung der Urteile, die Ergebnisse besser zu fundieren.

Wichtiger als diese Diskussion erscheint für die hier im Mittelpunkt stehende Frage jedoch die Diskussion über die Urteilskompetenz der Studierenden, die letztlich die Validität der jeweiligen Bewertung beeinflußt. Die Kritik reicht dabei von dem Hinweis auf die Beeinflussung durch das Thema und die jeweiligen Rahmenbedingungen über die Einstufung der entsprechenden Erhebungen als reine Meinungsforschung bis hin zur Behauptung, dass Studierende prinzipiell nicht in der Lage seien, die Qualität universitärer Lehre festzustellen.[36]

Um diese Kritik zu überprüfen, hat Rindermann versucht, die Validität durch „Korrelationen mit Fremdbeurteilungen und mit Leistungsmaßen, durch Untersuchung der Generalisierbarkeit dozentenbezogener Lehrevaluationen und durch Zusammenhänge und Biasvariablen" zu bestimmen (Rindermann 1996b: 135). In seinen Untersuchungen korrelierten dabei alle externen Messungen mit einem Wert größer als 0.40 mit den studentischen Urteilen – ein Ergebnis, dass Rindermann als mittlere Größe der Validität bezeichnet (Rindermann 1996b: 136).

In einem zweiten Schritt weist Rindermann darauf hin, dass auch Studien über den Zusammenhang zwischen Lehreinschätzung und dem Lernerfolg anhand von Leistungstests ähnliche Korrelationen beobachten.

Schließlich wurden einzelne Biasvariablen untersucht, die als individuelle studentische Merkmale die Urteile verzerren könnten (Rindermann 1996b: 138). Die einzelnen Studien zusammenfassend schreibt Rindermann (1996b: 138), dass Beurteilungen Studierender kein unrealistisches Bild der Lehre zeichnen und sensibel für Lernerfolg seien. Zudem reflektierten sie mehr das Lehrverhalten des Lehrenden als Rahmenbedingungen wie Veranstaltungsthema oder Veranstaltungstyp. Deshalb könne begründet von studentischen Lehrevaluationen als einem Maß universitärer Lehrqualität (und nicht bloßer Messung von Akzeptanz oder ‚Selbstbezichtigung') gesprochen werden. „Zusammenfassend möchten wir (...) die Interpretation

[36] Für weitere Hinweise vgl. Rindermann (1996b: 135).

anbieten, daß bei Beachtung einiger methodischer Sorgfaltsregeln von hinreichender Validität studentischer Evaluationen auszugehen ist“ (Rindermann/Amelang 1994b: 23).[37]

Unter diese methodischen Sorgfaltsregeln fallen alle Maßnahmen, die zur Verringerung von Antworttendenzen beitragen: Feger (1992: 14) führt hier unter anderem die Anonymität der Befragung, Maßnahmen zur Verringerung des Gruppendrucks, der Einsatz über verschiedene Lehrveranstaltungen hinweg sowie die „glaubwürdige Zusage des Dozenten, dass die Kritik so weit wie möglich durch Änderungen der Lehrveranstaltung berücksichtigt wird“, auf.

Uns erscheinen hier zwei Bemerkungen notwendig. Zum einen sind bivariate Korrelationen von 0,4 kaum ein Hinweis auf die Substituierbarkeit zweier Messungen: 16 Prozent gemeinsame Varianz belegt zwar eine gemeinsame Kovariation, aber keineswegs, dass es sich um austauschbare Indikatoren handelt. Sollte das von Kromrey genannte Argument (Mischverteilungen in der Population der Studierenden) korrekt sein, dann sind auch diese (ohnehin für angeblich austauschbare Indikatoren geringen) Korrelationen Artefakte. Dies ließe sich aber nur durch sorgfältige weitere empirische Untersuchungen klären. Zum anderen zeigt jede Lektüre der ausgefüllten Lehrevaluationsbögen, dass es zumindest Subgruppen unter den Studierenden gibt, die die Lehrevaluation als Messversuch nicht ernstnehmen. Auch für diese Subgruppe gilt das genannte Mischverteilungsargument.

Neben der methodologischen Kritik an Lehrevaluationen findet sich überraschenderweise immer noch eine eher ideologiekritische Tradition. So gibt Ritter (1993) eine generelle Kritik an standardisierten Evaluationsverfahren: Hier wird von vielen enttäuschenden Erfahrungen mit Evaluation berichtet, deren Ursache die „unzureichende methodenkritische, erkenntnistheoretische Reflektion, die zu einem teilweise sogar kontraproduktiven Einsatz der Evaluation führte“ (Ritter 1993: 180), war. Wichtig wäre es, die Zusammenhänge zwischen Erkenntnissubjekt, Erkenntnisobjekt, Erkenntnismethode, Erkenntnisziel und dem Zweck, der Verwendung der Evaluationsergebnisse zu klären. Ritter (1993: 185) sieht sogar die Gefahr, durch Evaluation besonders kreative und inhaltlich originelle Lehrende auszufiltern und nur noch das Mittelmaß zu fördern. Zudem wird von einem impliziten Rückgang der inhaltlichen Anforderungen ausgegangen, um dadurch bessere Evaluationen zu erzielen. „Keinesfalls darf die Beliebtheit bei den Studenten für fachlich durchschnittliche Hochschullehrer zum Auswahlkriterium werden. Geht es um die Verbesserung der Lehre durch die Auswahl geeigneter Hochschullehrer, so ist die institutionalisierte Vorlesungskritik mittels Fragebogen nur sehr

[37] Empirischer Ausgangspunkt ist das sogenannte Heidelberger Inventar zur Evaluation von Lehrveranstaltungen (Rindermann/Amelang 1994a; vgl. auch Rindermann 1996a). Dieses, allerdings recht umfangreiche Befragungsinstrument ist in den entsprechenden Beiträgen ausführlich dokumentiert.

begrenzt aussagefähig“ (Ritter 1993: 185). Ritter stützt die hier skizzierten Äußerungen jedoch nicht durch empirische Belege.

Wie bei vielen Diskussionen in diesem Feld könnte man auch bei der Frage nach der Validität der entsprechenden studentischen Urteile eine moderierende Mittelposition einnehmen, wie sie etwa generell Brandstätter formuliert, wenn er anmerkt: „Ein dogmatischer experminenteller Rigorismus ist zumal bei der Evaluation komplexer Reform- und Interventionsprojekte (...) ebenso unangemessen wie ein einseitiger qualitativer Impressionismus“ (Brandtstätter 1990: 219). Gerade im Bereich der hochschulinternen Evaluation mit Hilfe von Studierendenbefragungen muß man aber keine neutrale mittlere Position einnehmen, sondern kann sich auf methodisch fundierte Urteile der amerikanischen Forschung stützen.

5.2 ‚Students‘ Evaluations of University Teaching’ – Ein Überblick über die amerikanische Forschung

Andreas Gold nimmt in seinem Beitrag zur aktuellen bundesrepublikanischen Diskussion über die Validität von studentischen Lehrevaluationen das im folgenden zu explizierende Argument vorweg, wenn er zum Stand der Diskussion schreibt: „Völlig neuartige Argumente (...) vermag ich der aktuellen Debatte bislang nicht zu entnehmen“ (Gold 1996: 147). Dabei bezieht er sich vor allem auf einen Vergleich mit der schon eine recht lange Tradition aufweisenden amerikanischen Forschung.[38] Im folgenden sollen deshalb die wichtigsten Argumente und Ergebnisse dieser Untersuchungen vorgestellt werden.[39]

Eine der deutschen Literatur entsprechende Fundamentalkritik an Lehrevaluationsbemühungen findet sich jedoch dabei kaum: die Einbeziehung studentischer Urteile kann in den Vereinigten Staaten auf eine sehr lange und fast unhinterfragte Tradition zurückblicken. Anstelle der Frage nach der grundlegenden Möglichkeit studentischer Evaluationen findet sich jedoch eine Fülle konkreter Forschungsarbeiten zu der methodischen Qualität entsprechender Arbeiten.

Im folgenden soll vor allem auf die Frage nach der Validität eingegangen werden, die Cashin (1995: 2) wie folgt formuliert: „In educational measurement, the basic question concerning

---

[38] Dabei ist die Zahl der entsprechenden Arbeiten in der Zwischenzeit fast nicht mehr überschaubar; Cashin (1995: 1) berichtet etwa von mehr als 1.500 Arbeiten, die sich mit dem Problem studentischer Lehrevaluationen befassen.

[39] Hierbei besteht die Möglichkeit sich auf mehrere, teilweise auch ältere zusammenfassende Arbeiten aus den Vereinigten Staaten zu stützen, die insgesamt ein sehr kohärentes Bild über die Forschungslage und die wichtigsten Ergebnisse vermitteln. Vergleiche dazu im folgenden Abrami/d’Apollonia/Cohen 1990; Cashin 1988; 1990; 1995; Cohen 1981; Freeman 1994; Marsh 1984; 1987 sowie Marsh/Roche 1993).

validity is: does the test measure what it is supposed to measure? For student ratings this translates into: to what extent do student rating items measure some aspect of teaching effectiveness“. Insgesamt kann man dabei zwischen drei unterschiedlichen Forschungsfragen unterscheiden (Marsh 1982: 82ff; Marsh 1984: 719; Cashin 1988; 1995):[40]

- Der Lernerfolg der Studierenden
- Die Selbsteinschätzung der Lehrenden
- Einschätzung der andere Gruppen (Absolventen, peer-Evaluation).

Die wichtigsten Ergebnisse der relevanten Untersuchungen sollen im folgenden kurz dargestellt werden. Als weiterer und damit vierter Punkt ist dabei auf die Suche nach Faktoren einzugehen, die das Urteil verzerren können. Ein wesentlicher Teil der Forschung behandelt derartige Faktoren.

5.2.1 Lernerfolg

Wenn das Ziel von Lehrevaluation die Messung der Lehreffektivität ist, so scheint es mehr als naheliegend, als Gütekriterium der entsprechenden Messungen die Lernerfahrungen der Studierenden heranzuziehen (Cohen 1981). Seminare, Übungen und Vorlesungen sollten dann gute Bewertungen durch die Studierenden erfahren, wenn der Lernerfolg in den entsprechenden Veranstaltungen hoch ist. Hierbei gilt jedoch: „It is difficult to validate students‘ evaluations against student learning measured by objective examination, because examination scores in different courses normally cannot be compared“ (Marsh 1984: 720).

Um dieses Problem zu umgehen, werden sogenannte ‚multiple-section courses‘ (Cashin 1995: 3) untersucht, in denen unterschiedliche Lehrer bei einem identischen Lehrplan und einer gemeinsamen Abschlußprüfung verschiedene Teile eines Kurses unterrichten. Die Korrelationen der entsprechenden Abschlußnote mit einzelnen Aspekten der studentischen Bewertung liegt dabei zwischen 0,22 für eine Bewertung der Interaktionsfähigkeit und 0,57 für die Vorbereitung des Dozenten (vgl. Cashin 1995: 3 sowie für die Originalstudien Cohen 1981 und Feldman 1989).

Wie immer die Höhe dieser Korrelation nun auch eingeschätzt werden soll, so wird hier deutlich, dass es einen beachtlichen Zusammenhang zwischen der Bewertung einer Veranstaltung durch die Studierenden und ihrem Lernerfolg (unterstellt, der Lernerfolg spiegelt sich in einer

[40] Das Problem entsprechender Untersuchungen fasst Marsh (1984: 719) wie folgt zusammen: „Student ratings, which constitute one measure of teaching effectiveness, are difficult to validate because there is no single criterion of effective teaching“.

Examensnote wider) gibt. Deutlich wird aber auch, dass ein Großteil der Variation in der entsprechenden Note nicht durch die Unterschiede in der Lehrqualität erklärt werden kann.

5.2.2 Selbsteinschätzung der Lehrenden

Ein zweiter Weg, die studentischen Einschätzungen zu überprüfen, besteht in einem Vergleich dieser Einschätzungen mit entsprechenden Selbstbewertungen der Dozenten. „Hence, the validity of student ratings will continue to be questioned until criteria are utilized that are both applicable across a wide range of courses and widely accepted as a indicator of teaching effectiveness. Instructor's self-evaluations of their own teaching effectiveness are a criterion that satisfies both of these requirements" (Marsh 1984: 723).

In den entsprechenden Studien finden sich Korrelationen zwischen 0,31 und 0,62 zwischen den Einschätzungen der Kursteilnehmer und der Lehrenden hinsichtlich der Effektivität des Unterrichts (vgl. Marsh 1984). Die Einschätzung dieser Ergebnisse ist dabei jedoch sehr unterschiedlich: Während Marsh (1984: 723) etwa eine deutliche Übereinstimmung in den Urteilen und daher ein Indiz für eine hohe Validität der entsprechenden studentischen Bewertungen sieht, betont Rindermann (1996a: 81ff) die kritischen Aspekte: Lehrende haben eine andere Perspektive und wenig Erfahrung in der Einschätzung von Lehrveranstaltungen und zudem – und dies ist sicher der wichtigste Punkt – besteht die Gefahr sozialer Erwünschtheit. Das Fazit ist dementsprechend wenig ermutigend: „Die Selbstbeurteilung der Veranstaltungen durch Dozenten kann deshalb nur mit Einschränkungen als ein Kriterium unter anderen zur Beurteilung der Validität von studentischen Evaluationen herangezogen werden" (Rindermann 1996a: 82).[41]

5.2.3 Einschätzung anderer Gruppen

Eine dritte Herangehensweise stellt der Vergleich der studentischen Bewertungen mit den Urteilen anderer Dritter dar. Hierbei können sowohl die Urteile der entsprechenden Hochschulverwaltung, der Kollegen und Kolleginnen, der ehemaligen Studierenden beziehungsweise der Absolventen oder schließlich von speziell geschulten Personen als Vergleichsgröße herangezogen werden (Cashin 1995: 3f). Die einzelnen Verfahren unterscheiden sich dabei natürlich ganz deutlich in dem mit ihnen verbundenen Aufwand.[42]

---

[41] An der entsprechenden Stelle (Rindermann 1996a: 85) findet sich eine Auflistung der entsprechenden Korrelationen zwischen diesen Selbsteinschätzungen und den studentischen Urteilen für die einzelnen bewerteten Items. Trotz der teilweise hohen Korrelation ist eine gewisse Vorsicht angebracht, da sich auch bei den studentischen Urteilen kaum negative Bewertungen finden.

[42] Vgl. hierzu Kreuter/Kopp (2000) sowie für das peer-rating-Verfahren Marsh (1984: 725f).

Zumindest die Ergebnisse mit Hilfe des peer-Ratings stimmen wenig zuversichtlich: „In summary, peer ratings based on classroom visitation do not appear to be substantially correlated with student ratings or with any other indicator of effective teaching“ (Marsh 1984: 725). Cashin (1995: 3) berichtet zwar durchaus beachtliche Zusammenhänge, zweifelt den Nutzen entsprechender Einschätzungen jedoch aufgrund von Reliabilitätsproblemen stark an.

Unabhängig davon, wie diese verschiedenen Ergebnisse einzuschätzen sind, wird in diesem Zusammenhang ein anderer Punkt deutlich: Solange kein grundlegender Konsens über die Zielsetzungen universitärer Lehrveranstaltungen besteht, scheint ein Vergleich verschiedener, teilweise recht allgemeiner Urteile über diese Lehrveranstaltung wenig hilfreich. Es ist eine offene Frage, ob die beobachtbaren Einschätzungsunterschiede vielleicht einfach auf einer unterschiedlichen Konzeption guter Lehrveranstaltungen beruhen (Cashin 1995).

Zumindest in diesem Punkt können Absolventenbefragungen prinzipiell durchaus sehr hilfreich sein: Hier lassen sich aufgrund der hier meist untersuchten Fragestellungen sehr gut einige wichtige und meist unbestrittene Kriterien – wie etwa der rasche Übergang in eine adäquate berufliche Position – formulieren und der Einfluß bestimmter Lehrveranstaltungen auf diese Variablen untersuchen.[43]

5.2.4 Bias-Variablen

Wie bereits erwähnt besteht der bei weitem größte Teil entsprechender Untersuchungen jedoch in der Suche nach sogenannten Bias-Variablen (Marsh 1984: 730ff).[44] Hierunter werden Faktoren verstanden, die das Urteil der Studierenden über die Qualität der Lehrveranstaltung beeinflussen. Genauer genommen sollen natürlich nur diejenigen Faktoren statistisch kontrolliert werden, die nicht mit der – wie auch immer genau definierbaren – Güte der Lehrveranstaltung in Verbindung stehen. Gerade in den Vereinigten Staaten, in denen in Folge dieser Evaluationen durchaus konsequenzenreiche Entscheidungen – etwa im Personalbereich – gefällt werden, besteht ein größeres Interesse an diesen Untersuchungen: „Instructors should not

---

[43] Unten (vgl. Punkt 7) wird zu sehen sein, dass diese recht optimistische Einschätzung in der Realität durch eine Vielzahl praktischer Probleme deutlich relativiert werden muß. Allein aus befragungstechnischen Gründen erscheint etwa eine Erhebung des genaueren Studienverlaufes so gut wie unmöglich (vgl. für die praktischen Probleme derartiger Befragungen die entsprechenden Abschnitte und weiteren Literaturhinweise in Kreuter/Kopp 2000).

[44] Wenn man die Vielzahl entsprechender Studien betrachtet, so muß man jedoch auch festhalten, dass selbst bei innerhalb der Sozialwissenschaften hohen Anteilen an erklärter Varianz der nicht durch die kontrollierten Variablen erklärte Anteil bei weitem größer ist. So bleiben etwa bei Marsh (1980) rund vier Fünftel der Varianz unaufgeklärt.

be faulted if their less effective teaching large classes of unmotivated students than their colleagues who were teaching small classes of motivated students“ (Cashin 1995: 4).

Wenn man die Ergebnisse der hier zu findenden Vielzahl empirischer Studien zusammenfasst, so finden sich drei große Variablenblöcke, die keiner und drei weitere Blöcke, die einer Kontrolle bei der Einschätzung studentischer Bewertungen von Lehrveranstaltungen bedürften.[45] Hier soll mit den Faktoren begonnen werden, die sich in den verschiedensten Untersuchungen als relativ unkritisch erwiesen haben:

- Zuerst sind hier einige Merkmale des jeweils Lehrenden zu nennen: Alter und Lehrerfahrung, aber auch das Geschlecht, die ethnische Zugehörigkeit und einige – durch verschiedene unabhängige Tests erhobene – Persönlichkeitsmerkmale wie etwa Selbstvertrauen, aber auch der Erfolg in der Forschung korrelieren nur relativ gering mit der Einschätzung der Studierenden (vgl. Basow/Howe 1987; Freeman 1994).
- Auch die entsprechenden Merkmale der Studierenden – Alter, Geschlecht, Semesterzahl, Persönlichkeitsmerkmale – üben keinen Einfluß auf die Bewertung der einzelnen Lehrveranstaltung aus.
- Erstaunlicherweise berichtet etwa Cashin (1995: 5) auch bei Variablen wie der Klassengröße von einem recht geringen Zusammenhang mit den Einschätzungen der Lehrqualität (vgl. aber dagegen die Ergebnisse in Kapitel 6). Sowohl die Tageszeit der Veranstaltung, wie auch die Zeit im Semester spielen ebenso keine bedeutsame Rolle.

Als Kriterium wurde hier überall der Zusammenhang mit der Einschätzung der Lehrqualität herangezogen. Bei einigen Variablenblöcken zeigt sich jedoch ein deutlicher Zusammenhang und demzufolge sollten diese Variablen genauer betrachtet werden, wobei hier dann inhaltlich zu unterscheiden ist, ob es sich um eine Störgröße oder einen inhaltlich zu interpretierenden Effekt handelt:

- Zuerst sollen auch hier wieder Merkmale des Lehrenden untersucht werden. Hier wird unter anderem die Ausdruckskraft, die Expressivität des Lehrenden untersucht. Nun stellt diese Variable sicherlich eine wichtige Größe bei der Bestimmung guter Lehre dar, aus diesem Grunde muß sie also nicht statistisch kontrolliert, sondern eher inhaltlich als mögliche Interventionsgröße betrachtet werden.
- Als zweiter Punkt ist hier auf Faktoren einzugehen, die die Studierenden betreffen. Zuerst ist hierbei die Motivation der Studierenden zu nennen. Bei Wahlveranstaltungen oder einem bereits vorab bestehenden inhaltlichen Interesse finden sich deutliche positivere Ein-

[45] Vgl. im folgenden unter anderem Marsh (1984; 1987); Cashin (1988; 1995); Freeman (1994) sowie Abrami/d’Appolonia/Cohen (1990).

schätzungen der Lehrleistungen (vgl. auch Esser 1995). Ebenso findet sich ein positiver, wenn auch relativ geringer Zusammenhang zwischen der erwarteten Bewertung und der Evaluation (Cashin 1995: 5). Während der zweite genannte Zusammenhang durchaus inhaltliche Gründe haben kann – Studierende, die in einer Veranstaltung mehr lernen, erzielen bessere Ergebnisse – sollte die vorab gemessene Motivation der Studierenden kontrolliert werden, um eine unverzerrte Einschätzung der Lehrqualität zu erhalten.

- Als dritter Block ist auf eine Reihe eher formaler oder institutioneller Settings hinzuweisen, die zu einer Beeinflussung der Bewertung führen können und deshalb kontrolliert werden sollten. Hierunter fallen auch recht selbstverständliche Dinge wie etwa die Möglichkeit zur anonymen Ausfüllung der entsprechenden Fragebögen (vgl. weiter Cashin 1995).

Insgesamt zeigen die amerikanischen Untersuchungen, dass die datentechnische Qualität studentischer Befragungen als recht positiv einzuschätzen ist. Trotz dieser Tatsache ist es auch in der Literatur weiterhin eine offene Frage, ob studentische Urteile in allgemeine Evaluationsbemühungen integriert werden können. Mit dieser Frage soll sich der folgende Abschnitt auseinandersetzen.

5.3 ‚Was tun?' Lohnt sich der Einsatz studentischer Befragungen zur Bewertung universitärer Lehrveranstaltungen?

Sollen studentische Lehrbewertungen nun also als Evaluationsinstrument zum Einsatz kommen? Die entsprechenden Untersuchungen und Evaluationsratgeber (vgl. die unter Punkt 3 des vorliegenden Berichtes genannte Literatur) sehen dies zu einem großen Teil vor – und auch die gerade vorgestellten Ergebnisse sprechen aus methodischer Sicht auch nicht gegen eine derartige Berücksichtigung. Es gibt jedoch eine ganze Reihe von Punkten, die zu berücksichtigen sind, wenn studentische Urteile zum Einsatz kommen sollen.

Zuallererst einmal ist festzuhalten, dass auch studentische Evaluationsbefragungen den Standards empirischer Sozialforschung entsprechen sollten. Schlecht oder mangelhaft durchgeführte Befragungen sprechen nicht gegen das gesamte Verfahren[46]. Ein Plädoyer gegen schlechte Sozialforschung (Kromrey 1996) muß kein Plädoyer gegen studentische Evaluation von Lehrveranstaltungen sein. Unübersichtliche oder mehrdimensionale Skalen und ohne Berücksichtigung der bisherigen Forschung meist ad-hoc-zusammengestellte Befragungsinstrumente sind ein Zeichen schlechter Sozialforschung. Ob diese Fehler allerdings in Evaluationsstudien häufiger oder auch seltener begangen werden als in anderen Bereichen der Sozialforschung ist jedoch eine müßige Fragestellung.

[46] vgl. für ein Beispiel die in Endruweit (1992) kritisierte Studie.

Zwei Probleme sollen jedoch hier näher besprochen werden: die Auswahl der zu Befragenden und die hierdurch eventuell entstehenden Probleme einerseits und die theoretische Fundierung der entsprechenden Studien und die damit verbundenen Fragestellungen.

Allein aufgrund der nicht geringen zeitlichen und für manche Fachbereiche auch finanziellen Belastungen durch Lehrevaluationen werden die meisten Untersuchungen als einmalige Befragung gegen Ende des Semesters durchgeführt. Selbst bei Pflichtveranstaltungen ist hier aber mit einer positiven Selektion und deshalb mit entsprechenden Verzerrungen zu rechnen (vgl. etwa Daniel 1998). Es ist davon auszugehen, dass die besonders unzufriedenen Studierenden bereits gar nicht mehr in die Evaluationssituation geraten. Ohne eine vollständige Erhebung oder eine genaue Modellierung des Ausfallprozesses kann hierüber jedoch nur spekuliert werden. Vielleicht liegen gerade hier die Gründe für die Diskrepanz zwischen der öffentlichen Wahrnehmung hinsichtlich der Qualität universitärer Lehre und denn dann doch immer wieder erstaunlich positiven Resultaten der einzelnen Evaluationen.

Diese Problematik wird durch zwei weitere Verzerrungsmöglichkeiten verschärft. Einerseits besteht kaum die Möglichkeit, alle Veranstaltungen in die Evaluation einzubeziehen. So wird etwa von einer der ersten Evaluationen an der Freien Universität Berlin berichtet, dass rund 80 Prozent aller lehrenden Professoren ihr Einverständnis zur Durchführung der entsprechenden Befragung erteilt haben (Grühn 1992). Es ist dabei kaum anzunehmen, dass es sich um den in der Evaluation berücksichtigten Veranstaltungen um eine zufällige Auswahl handelt. Andererseits ist selbst in den dann Berücksichtigung findenden Lehrveranstaltungen die Ausschöpfung recht eingeschränkt. Auch bei einer auf den ersten Blick recht hohen Rücklaufquote von etwa 70 Prozent (Grühn 1992) ist es unklar, welche Ausfallmechanismen hier wirksam sind. In Anbetracht der ab und an resignativen Einstellung auf Seiten der Studierenden können gerade die kritischeren Studierenden auf eine Einschätzung verzichten. Auch hier läßt sich über die genauen Prozesse ohne eine Non-Response-Studie nur spekulieren.

Zu bedenken ist jedoch, dass selbst bei den gerade berichteten recht hohen Quoten insgesamt nur rund 56 Prozent der Studierenden befragt werden. Ein weiteres Ergebnis der im Rahmen dieses Berichts durchgeführten Erhebung zur Evaluationspraxis in den sozialwissenschaftlichen Studiengängen war zudem, dass immerhin in mehr als einem Viertel aller Fälle die Teilnahmebereitschaft der Studierenden und gar in fast einem Drittel der Fälle die Teilnahmebereitschaft der Lehrenden als mittelmäßig oder schlechter eingestuft wurde. Eine einfache Interpretation der in Lehrevaluationen erzielten Mittelwerte als generelle Zufriedenheit mit der Lehrsituation ist also mit gutem Grund anzweifelbar.

Diese und ähnliche Fragen lassen sich mit Hilfe entsprechender Untersuchungen durchaus beantworten. Man kann auch gerade die sogenannten No-Shows (Esser 1995) beziehungsweise die eine Veranstaltung nicht weiter besuchenden Studierenden befragen oder generell mindestens zwei Befragungszeitpunkte vorsehen. Solche Untersuchungen sind aber mit einem sehr hohen zeitlichen und finanziellen Aufwand verbundenen und deshalb so gut wie nie zu finden.

Schließlich verbleibt eine weitere grundlegendere Problematik: Zwar scheinen studentische Urteile über einzelne Dimensionen der Lehrveranstaltungen zwar prinzipiell brauchbare Resultate zu erbringen, solange aber keine Einigkeit darüber besteht, was eigentlich die Qualität einer Lehrveranstaltung ausmacht, ist der weitere Einsatz studentischer Evaluationsmaßnahmen immer wieder kritisierbar. So schreiben etwa Willems, Gijselaers und de Bie (1994: 59ff) in einer Einführung in die Lehrevaluation durch Studierende, dass sich letztlich zwei Ansichten von Qualität unterscheiden lassen: die Qualität aus Sicht des „Verbrauchers“ und die Qualität aus Sicht des „Produzenten“. „Welche Ansicht für die Qualität schließlich bestimmend ist oder welche die beste ist, sei dahingestellt (...). Es ist aber wichtig zu erkennen, dass verschiedene Parteien die Ansichten der jeweils anderen Partei über Qualität nicht immer teilen“ (Willems/Gijselaers/de Bie 194: 59f).

Gerade die Tatsache, dass ein und dieselbe Lehrveranstaltung von Studierenden sehr unterschiedlich bewertet werden kann, macht deutlich, dass es einer theoretischen Modellierung der entsprechenden Bedingungsvariablen studentischer Einstufungen, aber auch der Qualität der Lehre bedarf. So formulieren Martin Schweer und Bernhard Rosemann (1995) ein psychologisch-handlungstheoretisches Modell, das die studentischen Urteile über Lehre erklären soll. Auffallend sind die sehr heterogenen Antworten, wenn man die Studierenden selbst nach Dimensionen guter Lehre befragt. Diese Ergebnisse sollen durch ein Modell der transaktionalen-Person-Situation-Verschränkung erklärt werden. „Dessen ungeachtet wird klar, dass die Evaluation einer Lehrsituation zu einem erheblichen Teil durch im strengen Sinne ‘lehrfremde’ Faktoren determiniert wird, also durch Faktoren, die mit der tatsächlichen Qualität der Lehre wenig zu tun hat“ (Schweer/Rosemann 1995: 195).[47] Ohne eine eindeutige Theorie des Lehrens und des Lernens fällt es schwer, sinnvolle und aussagekräftige Evaluationsinstrumente zu konstruieren (vgl. schon Marsh 1984: 715ff).

Dieser Vorwurf einer fehlenden theoretischen Orientierung steht letztlich auch hinter der oben skizzierten bundesdeutschen Debatte über die Möglichkeiten der studentischen Lehrevaluation. So fordert ja etwa Kromrey (1996: 158f), dass als Vorbedingung eines Messprozesses die zu erreichenden Ziele präzise angegeben werden sollen sowie die entsprechenden Kenntnisse

[47] Die Autoren ziehen deshalb ein recht pessimistisches Fazit und stellen fest, dass „das meßtheoretische Problem der Validität von Lehrevaluationen bislang nicht gelöst ist“ (Schweer/Rosemann 1995: 195).

über die angemessene Vorgehensweise vorliegen muß (vgl. auch Helmke 1996 oder Krempkow 1998). Es läßt sich dabei vermuten, dass studentische Urteile eben nicht nur von den gezeigten Verhaltensvariationen der Dozenten, sondern eben auch von impliziten Annahmen und Theorien der Studierenden beeinflußt werden (Astleitner 1991; Astleitner/Krumm 1991). Als bedeutsamster Mangel wird die fehlende „theoretische Fundiertheit der erfassten Lehreffektivitätskonstrukte“ (Astleitner/Krumm 1991: 242) festgehalten. Hierbei sollte eine Berücksichtigung entsprechender Informationsverarbeitungsprozesse ebenso wenig unterbleiben wie die Vernachlässigung entsprechender Theorien zur Befragungssituation.

Wenn man diese Faktoren jedoch berücksichtigt, erscheint es sehr wohl möglich, mit Hilfe studentischer Einschätzungen einige wichtige Informationen zu erhalten, denn nicht nur Cashin (1995: 6) fasst den entsprechenden Forschungsstand wie folgt zusammen: „In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or need to control; probably more so than any other data used for evaluation. Nevertheless, student ratings are only one source of data about teaching and must be used in combination with multiple sources of data if one wishes to make a judgement about all of the components of college teaching. Further, student ratings are data that must be interpreted“.

6. Die Lehrevaluation an der Fakultät für Verwaltungswissenschaft der Universität Konstanz

Aufgrund der institutionellen Verankerung dieses Projekts über die fakultätsinterne Evaluation der Lehre ist es sicher verständlich, dass hier etwas ausführlicher auf die an der Fakultät für Verwaltungswissenschaft der Universität Konstanz stattgefundenen Lehrevaluationen eingegangen wird.[48] Mit Hilfe dieser Befragung sind auch einige tiefergehende Analysen möglich, da die entsprechenden Datensätze analysierfähig vorliegen. Im Anhang A (vgl. A 38) ist das entsprechende Instrument aus dem Sommersemester 1999 dokumentiert, das eine an einigen Stellen verbesserte Version des seit 1997 eingesetzten Fragebogens darstellt. Wenn man nun diesen Bogen mit den anderen Evaluationsinstrumenten in den Anhängen A und B sowie der auf den vorhergehenden Seiten dargestellten theoretischen und methodischen Diskussion vergleicht, so kann man zuerst knapp festhalten, dass sich aus diesem Vergleich heraus keine Veränderungsnotwendigkeit des bislang eingesetzten Instruments ergibt.

6.1 Eine kurze Einschätzung des Instruments

Die Fragen erheben neben dem zu einer Rückmeldung an den jeweiligen Dozenten beziehungsweise Dozentin unabdingbaren Veranstaltungstitel den Studiengang und die Semesterzahl der Studierenden, einige Einschätzungen über den oder die Lehrenden sowie über einige Charakteristika der Lehrveranstaltung selbst. Darüber hinaus wird der eigene Arbeitsaufwand erfragt. Schließlich gibt es mit Hilfe mehrerer offener Fragen die Möglichkeit, besonders positive, aber natürlich auch negative Punkte und Verbesserungsvorschläge zu benennen. Für Tutorien, die in der Regel andere Funktionen erfüllen sollen, existiert ein gesonderter Fragebogen (vgl. auch hierzu A 38).

Der Bogen der Fakultät für Verwaltungswissenschaft stellt somit fast einen Standardfragebogen zur Lehrveranstaltungsevaluation dar. Die Dimensionen sind klar getrennt und vor allem auch durch Studierende beantwortbar.[49] Aufgrund der organisatorischen Umsetzung der jeweiligen Lehrevaluation stellt sich jedoch auch hier die bereits mehrfach in diesem Bericht diskutierte Frage nach der Selektivität der Ergebnisse. Aufgrund der nicht unerheblichen Kosten und des ebenfalls nicht zu unterschätzenden organisatorischen und zeitlichen Arbeitsaufwandes ist es zu verstehen, dass eine Lehrevaluation nur einmalig am Ende eines Semesters

---

[48] Die hier zu berücksichtigenden studentischen Lehrevaluationen fanden alle vor der Umorganisation der Fakultäten an der Universität Konstanz statt, so dass hier wirklich nur über die Fakultät für Verwaltungswissenschaft berichtet wird.

[49] Als einzige Ausnahme könnte die Frage nach der Kompetenz des Dozenten beziehungsweise der Dozentin genannt werden. Einerseits ist unklar, ob die soziale oder die fachliche Kompetenz (oder eine Mischung) gemeint ist und andererseits erscheint es zumindest diskussionswürdig, ob Studierende wirklich die fachliche Kompetenz der jeweils Lehrenden einschätzen können.

durchgeführt wird; im Interesse eines vollständigeren Bildes über die Einschätzung der Lehrsituation durch die Studierenden wäre es jedoch wünschenswert, auf während des Semesters derartige Einschätzungen vornehmen zu lassen. Diese mehrmalige oder zumindest eben zweimalige Erhebung der studentischen Urteile hätte zudem den Vorteil, dass die entsprechenden Ergebnisse bereits in den laufenden Lehrveranstaltungen berücksichtigt werden könnten.

Ein tiefergehender Vergleich der Lehrevaluation an der Fakultät für Verwaltungswissenschaft der Universität Konstanz mit den anderen dokumentierten Fragebögen zeigt, dass in Konstanz fast gänzlich auf die Erhebung soziodemographische Faktoren und Hintergrundsvariablen verzichtet wird. So wichtig diese Informationen für eine Erklärung der unterschiedlichen Einschätzungen der Lehrsituation durch einzelne Studierende vielleicht sein mögen, für die konkrete Lehrevaluation sind sie nicht notwendig. In Anbetracht des damit verbundenen Erhebungsaufwands erscheint ein derartiger Verzicht verstehbar und sinnvoll.

Hinsichtlich der Verbreitung der Ergebnisse der jeweiligen Lehrevaluationen (vgl. Fakultät für Verwaltungswissenschaft 1999b) ist daran zu erinnern, dass die einfache Wiedergabe von Mittelwerten aufgrund der sehr unterschiedlichen Teilnehmerzahl der einzelnen Veranstaltungen nur sehr vorsichtig zu interpretieren sind. Gerade für den meist unterstellten Wirkmechanismus ist es zudem nochmals festzuhalten, dass insbesondere die offenen Antworten, soweit sie eine ernsthafte Beschäftigung der Studierenden darstellen, weitervermittelt werden sollten, um konkrete Kritik und Verbesserungsvorschläge zu kommunizieren.

6.2 Itemanalyse des Lehrevaluationsinstruments der Fakultät

Stellt man aus den bisher vorhandenen Lehrevaluationen einen gepoolten Datensatz mit allen 4.921 Beobachtungen zusammen, kann eine Itemanalyse über alle bisherigen Lehrevaluationen durchgeführt werden.

Eine Faktorenanalyse dieses gepoolten Datensatzes (n=3.590) der Bewertungsfragen zeigt einen einzigen Faktor mit fast 60 Prozent erklärter Varianz. Entsprechend läßt sich in einer Regression fast 75 Prozent der Varianz der zusammenfassenden Bewertung durch die Einzelbewertungen erklären. Berücksichtigt man zusätzlich den Arbeitsaufwand in Stunden, die Gewinnung neuer Erkenntnisse und die Einschätzung des Anforderungen, so erhöht sich zwar die erklärte Varianz in keiner Weise, es zeigt sich aber für „neue Erkenntnisse“ und „Anforderungen“ ein Effekt. Eine etwas genauere Analyse zeigt, dass die Studenten sowohl zu hohe als auch zu niedrige Anforderungen negativ sanktionieren. Berücksichtigt man entsprechende Dummyvariablen in der Regression, so ist nur der Effekt der zu hohen Anforderungen signifi-

kant (vgl. Tabelle 6.1).[50] Ebenso muß festgehalten werden, dass die zusammenfassende Bewertung des Dozenten um so schlechter ausfällt, desto höher die Zahl der Teilnehmer ist. Dies gilt auch bei Konstanthaltung aller anderen Variablen.

Tabelle 6.1: Regression zur Erklärung der zusammenfassenden Bewertung (gemessen in Form einer Schulnote) ($r^2 = 0{,}753$)

Variable	β-Koeffizient
Bewertung Fachkompetenz (Schulnoten)	0,206*
Bewertung Vorbereitung (Schulnoten)	0,199*
Bewertung Eingehen auf Fragen (Schulnoten)	0,226*
Bewertung Strukturierung einzelner Stunden (Schulnoten)	0,211*
Bewertung Zusammenhang zwischen einzelnen Stunden (Schulnoten)	0,291*
wenig neue Erkenntnisse	0,040*
Anzahl der Stunden zur Vor- und Nachbereitung	0,009
zu hohe Anforderungen: ja	0,041*
zu geringe Anforderungen: ja	0,012
Anzahl der Studierenden in der Veranstaltung	0,049*

n = 3.004; * = p < 0,05

Ebenso bemerkenswert ist eine Reliabilitätsschätzung der aus den Einschätzungsvariablen gebildeten Skala. Es zeigt sich ein Alpha von 0,85, ein für nur 6 Items in den Sozialwissenschaften außergewöhnlich hoher Wert. Erwähnenswert ist hierbei, dass die Fachkompetenz des Dozenten mit dem Skalenwert am schwächsten kovariiert (vgl. Tabelle 6.2).

[50] Aufgrund der Abhängigkeiten der einzelnen Beobachtungen wurde hier eine Regression unter Berücksichtigung der natürlichen Klumpung (innerhalb der Lehrveranstaltungen) mit robusten Standardfehlern gerechnet.

Tabelle 6.2: Reliabilitätsschätzung

Item	Item-Test-Korrelation	Item-Rest-Korrelation	Inter-Item-Kovarianz	Alpha nach Löschung
Fachkompetenz	0,616	0,472	0,419	0,855
Vorbereitung	0,753	0,618	0,358	0,827
Eingehen auf Fragen	0,712	0,548	0,369	0,844
Strukturierung einzelner Stunden	0,800	0,668	0,322	0,820
Zusammenhang zwischen einzelnen Stunden	0,785	0,653	0,330	0,822
zusammenfassende Bewertung	0,904	0,847	0,320	0,786

Insgesamt kann festgehalten werden, dass das in der Fakultät verwendete Instrument sowohl konsistente als auch reliable Messungen erlaubt. Aus psychometrischer Sicht bedarf das Instrument keiner Veränderungen. Sollen allerdings die Ergebnisse zukünftig erklärt werden, so sollten Fragen nach der Veranstaltungsart und nach Veranstaltung im Grund- oder Hauptstudium dem Datensatz hinzugefügt werden. Für weitere Analysen ist die zukünftige Berücksichtigung eines (eventuell anonymen) Personenidentifiers unverzichtbar.

7. Forschungsansätze und methodische Probleme potentieller Validierungsstrategien

Zielsetzung dieses letzten Abschnittes ist es, die bereits gerade vorgestellten verschiedenen Validierungsversuche noch einmal kurz zu betrachten und vor allem auf ihre praktische Umsetzung und Durchführbarkeit hin zu untersuchen. Hierbei sollen nicht noch einmal die verschiedenen Ergebnisse der unterschiedlichen Forschungsgruppen vorgestellt werden (vgl. hierzu Abschnitt 5 und die dort vorgestellte Literatur). Vielmehr sollen einige grundlegende Probleme derartiger Untersuchungen und damit der entsprechenden Validierungsstrategien besprochen werden.

Ausgangspunkt der Diskussion ist dabei die Feststellung, dass die Zielsetzung der meisten Studiengänge die Vermittlung einer entsprechenden Wissensgrundlage zur erfolgreichen beruflichen oder wissenschaftlichen Tätigkeit sein soll. Bereits oben wurde als erste und vielleicht naheliegendste Validierungsstrategie deshalb die Operationalisierung der Lehreffektivität durch die erzielten Noten diskutiert. Die meisten der an Universitäten durchgeführten Kurse entziehen sich jedoch durch ihren Aufbau einer entsprechenden Bewertung, denn nur sogenannte ‚multiple-section courses' (Marsh 1984; Cashin 1995) können hier sinnvollerweise untersucht werden. Eine naheliegende Fortführung des hier zugrundeliegenden Gedankens ist nun jedoch, den wirklichen praktischen Erfolg bei der entsprechenden Zielerreichung zu untersuchen. Nun sind die Ziele universitärer Ausbildung selten wirklich trennscharf und explizit zu finden. Mindestens zwei entsprechende, sicherlich nicht immer dekkungsgleiche Ziele lassen sich unterstellen: Einerseits die Vermittlung eines entsprechenden fachlichen und methodischen Wissens, das zu einem eigenständigen wissenschaftlichen Arbeiten notwendig ist, andererseits die Vermittlung praktisch auf dem Arbeitsmarkt relevanter Erfahrungen, die zu einer raschen und guten Plazierung auf dem jeweiligen Berufsfeld führen sollte.

Zu der ersten hier angerissenen Strategie sind bislang keine Forschungsarbeiten bekannt. So bleibt unklar, welche Studiengänge im allgemeinen und speziell welche einzelnen Lehrveranstaltungen zu einer entsprechenden wissenschaftlichen Qualifikation führen. Aufgrund der Heterogenität zwischen, vor allem aber auch innerhalb der einzelnen Forschungsgebiete erscheint es auch kaum erwartbar, hier einheitliche Kriterien aufstellen zu können. Ein derartiges Vorgehen zur validen Bestimmung der entsprechenden Lehrqualität scheint von vornherein zum Scheitern verurteilt zu sein. Bereits bei der allgemeinen Darstellung der Evaluationsforschung (vgl. Abschnitt 2) wurde hervorgehoben, dass ohne eine klare Zieldefinition eine Evaluation letztlich nicht möglich ist. Gerade an dieser einheitlichen Zieldefinition hinsichtlich der wissenschaftlichen Qualifikation fehlt es aber vielen Evaluationsversuchen an der Hochschule.

Bedenkenswerter erscheint hingegen die Strategie, die berufsqualifizierende Funktion der Ausbildung zu berücksichtigen. Auf diesem Gedanken beruht auch die Überlegung, entsprechende Absolventen eines Studienganges zur Evaluation heranzuziehen (vgl. oben Abschnitt 5.2). So reizvoll ein entsprechendes Untersuchungsdesign nun denn auch sein mag, die praktische Umsetzung erweist sich meist als sehr komplex. Anhand der an der Universität Konstanz durchgeführten Absolventenbefragung lassen sich die verschiedenen Probleme deutlich machen (Kreuter/Kopp 2000).

- Bei der Befragung von Absolventen ist mit einem großem Adressenproblem zu rechnen. Der jeweilige Adressenbestand der Hochschulverwaltung veraltet rasch und zu einem großem Umfang. Die jeweils möglichen Strategien (Kreuter/Kopp 2000) setzen einen großen Ressourcenaufwand voraus, der wohl nur selten durchführbar ist.
- Selbst wenn man auf aus methodischen Gründen auf die retrospektive Einschätzung einzelner Veranstaltungen verzichtet und stattdessen nur faktische Veranstaltungsbesuche erheben will, stößt dies rasch an Grenzen. Ab einer gewissen Zeitdauer sind auch derartige Informationen nicht mehr valide zu erheben und damit entsprechende Berechnungen für die Übertrittswahrscheinlichkeit in das Berufsleben nicht mehr vertretbar.

Wenn man, was aus verschiedenen hier skizzierten methodischen Gründen angeraten erscheint, letztlich darauf verzichtet, die Einschätzungen der – ehemaligen – Studierenden über einzelne Lehrveranstaltungen bei der Analyse des Übergangs in das Berufsleben zu berücksichtigen, hat man damit natürlich auch das Untersuchungsziel geändert: Hier wird nun nicht mehr versucht, das studentische Rating anhand externer Faktoren zu validieren; vielmehr wird die Wichtigkeit einzelner Bereiche des Studiums für diesen Prozeß untersucht. Bislang sind keine Studien bekannt, die diese beiden Aspekte wieder miteinander verkoppeln, also die objektive Bedeutung einzelner Fachbereiche oder Wissensgebiete beim Übergang in das Erwerbsleben mit der subjektiven Qualität dieser Lehrveranstaltungen zu verbinden. Das hierfür notwendige Untersuchungsdesign müßte zudem die verschiedenen Arbeitsmärkte und Fachgebiete und deren Heterogenität abbilden; so ist es durchaus denkbar, dass selbst der Besuch einer schlechten Lehrveranstaltung in einem wichtigen Fachbereich für den Arbeitsmarkt bedeutsamer ist als die Teilnahme an verschiedenen sehr gut durchgeführten Veranstaltungen zu eher randständigen Themen.

Eine denkbare Alternative ist es nun, mit Hilfe eines umfassenden Panels die Ausbildungswege, vor allem aber auch die spätere berufliche Karriere der Studierenden einer Fakultät zu untersuchen. Ein solches Vorgehen ist zwar aus methodischer Sicht sicherlich eindeutig zu präferieren und würde auch zu gut interpretierbaren Ergebnissen führen, die praktische Umsetzung, die finanzielle und personelle Belastung und nicht zuletzt der relativ lange Zeitrahmen und die verschiedensten auch datenschutzrechtlichen Bedenken, den eine derartige Untersu-

chung einnehmen würde, lassen eine Validierungsstudie mit Hilfe eines Panels jedoch als kaum realisierbar erscheinen.

Dennoch erscheint es möglich, mit Hilfe einer Absolventenbefragung wenigstens ansatzweise einige Informationen über die spätere Bedeutung einzelner Studienbereiche zu erfahren. Hierzu sind zwei Wege denkbar.

- Einerseits ist es denkbar den Einstieg in das berufliche Leben zu untersuchen. Hierbei können die Übergangswahrscheinlichkeit in das Berufsleben mit Hilfe sogenannter ereignisdatenanalytischer Verfahren untersucht werden, Hierdurch kann man z.B. feststellen, ob es statistisch interpretierbare Effekte auf die Übergangswahrscheinlichkeit in das Berufsleben in Abhängigkeit vom Themengebiet der Diplomarbeit gibt.[51]
- Andererseits kann versucht werden, die Verwendung von Wissensbeständen, Kenntnissen und Fähigkeiten in den verschiedenen Arbeitsmarktsegmenten zu untersuchen.[52]

Wie immer nun aber auch die konkreten Ergebnisse dieser Absolventenbefragungen aussehen möchten, um sie als Validierung entsprechender studentischer Lehrveranstaltungsbewertungen einsetzen zu können, muß man wie erwähnt diese beiden Untersuchungen miteinander verkoppeln. Dies ist nur durch die Zusammenführung der Ergebnisse auf der Ebene individueller Datenbestände möglich. Zumindest bislang gehört dies aber einerseits weder zum Standard studentischer Evaluationen der Lehre noch andererseits zur Zielsetzung entsprechender Untersuchungen über den Verbleib der Absolventen einzelner Studiengänge. Aufgrund der mit beiden Untersuchungsarten in der Praxis verbundenen Probleme und Schwierigkeiten erscheint es auch wenig wahrscheinlich, dass sich diese Situation ändern wird. Die grundsätzliche Kopplung – etwa im Rahmen einer Dauerbeobachtung der Studierenden und ihrer späteren Karriere – erscheint aufgrund der damit verbundenen Durchsetzungsprobleme wenig wahrscheinlich.

Trotz dieser pessimistischen Einschätzung der Realisierbarkeit entsprechender Validierungsbemühungen kann insgesamt kein negatives Urteil über die Leistungsfähigkeit und Gültigkeit studentischer Lehreinschätzungen gefällt werden. Zwar ist es wohl kaum möglich, die Umsetzung der grundlegenden Ziele einer universitären Ausbildung mit Hilfe der skizzierten Ansätze zu überprüfen, die Ergebnisse der Forschung (vgl. Absatz 5.2) zeigen jedoch, dass studentische Urteile hinsichtlich einiger Dimensionen sicherlich ein relativ gutes Instrument darstel-

[51] Ohne hier genauer auf derartige Berechnungen am Beispiel der Konstanzer Absolventenbefragung eingehen zu können (vgl. Kreuter/Kopp 2000: 28ff), lässt sich durch entsprechende Berechnungen zeigen, dass es bei der statistischen Kontrolle einiger Kovariate sehr wohl Studienbereiche gibt, die den Berufseinstieg erleichtern oder erschweren.

[52] Eine derartige Untersuchung bei Kreuter/Kopp (2000: 37ff) ausführlich dargestellt.

len. Ob die einzelnen Sitzungen gut strukturiert sind und ob der jeweilige Dozent oder Dozentin vorbereitet wirkt, läßt sich mit Hilfe studentischer Einschätzungen sicherlich relativ zuverlässig und valide erheben. Ob neuere wissenschaftliche Entwicklungen erfasst werden, ein Feld umfassend abgedeckt ist und der Kurs insgesamt relevante Wissensbereiche abdeckt, sind allerdings Fragen, die sicherlich nicht in derartige Befragungen gehören.

Die entscheidende Frage bezieht sich aber auf die Validität der Evaluation für die über die unmittelbare Lehrsituation hinausgehenden Aspekte universitärer Lehre. Ob zu einer verstärkten Beschäftigung mit einem Thema oder zu selbständigem Denken angeregt wurde, lässt sich nicht innerhalb einer Lehrveranstaltung ermitteln. Langfristige Effekte lassen sich halt nur langfristig untersuchen. Ohne langfristige externe Validierungen kann die Frage nach der Qualität universitärer Lehre nicht beantwortet werden.

8. Schlußbemerkung

Der Versuch, auch in Hochschulen Qualitätssicherung und Evaluation einzuführen, kann zweifellos als eine der wichtigeren Veränderungen in der institutionellen Organisation der Universität der letzten Jahre bezeichnet werden. Trotzdem wird es noch eine Weile dauern bis die Lehren aus der allgemeinen Evaluationsforschung auch in diesem Bereich allgemein bekanntes Wissen sein wird.

Eine grundlegende Erkenntnis der Literatur zur allgemeinen Evaluationsforschung ist es, dass ohne eine klare Festlegung der Ziele einer bestimmten Maßnahme letztlich nicht möglich ist, den Grad der Zielerreichung festzulegen. Gerade im Bereich der Universität aber erscheint eine Übereinkunft, was unter guter Lehre verstanden werden soll, nicht immer gegeben. Nicht zuletzt aus diesem Grunde können Evaluationen ab und an als ‚publikumswirksame Selbstprofilierung' (Brandstädter 1990: 224) oder gar als reiner Aktionismus erscheinen. Nun gibt es aber gerade in dem Bereich der Hochschulbildung ausreichend Gründe, Maßnahmen zur Qualitätssteigerung und damit eben auch Evaluationen durchzuführen (Wissenschaftsrat 1997). Auch wenn bei diesem Maßnahmen ein ganzer Kanon von Mitteln zum Einsatz kommt, so spielen studentische Einschätzungen der Lehrqualität doch fast immer eine Rolle – und dieser Einbezug der Studierenden und ihrer Meinung über die einzelnen Lehrveranstaltungen kommt zumindest in der öffentlichen Diskussion eine wichtige Rolle zu.[53]

Ausgangspunkt des hier vorgestellten Forschungsprojektes war jedoch weniger eine inhaltliche Bewertung dieser Bemühungen oder ein allgemeines Urteil über die aktuelle Lehrsituation, sondern vielmehr das Vorhaben, die bislang bei der internen Evaluation von Lehrveranstaltungen durch Studierende zum Einsatz kommenden Befragungsinstrumente kritisch zu hinterfragen und dabei zu prüfen, „ob das Instrument prinzipiell zur Evaluation von Lehrveranstaltungen geeignet erscheint" (Fakultät für Verwaltungswissenschaft 1999a: 2). Wenn man auf diese Frage in einem Satz eingehen will, so lautet die vielleicht im Nachhinein nur wenig erstaunliche Antwort, dass man dies so einfach nicht beantworten kann. Die im Mittelpunkt dieses Berichtes stehende Darstellung der verschiedenen Befragungsinstrumente (vgl. die Anhänge A und B) zeigt, dass in diesem Bereich eine sehr große Heterogenität herrscht. Von geprüften und getesteten Instrumenten, die zu mehreren Zeitpunkten eines Semesters zum Einsatz kommen bis hin zur einmaligen am Ende einer Veranstaltung eingesetzten kurzen Statementbatterie findet sich fast alles. Je nachdem ergeben sich aus den einzelnen Ergebnissen

[53] Die hier vertretene skeptische Position bezieht sich darauf, dass zwar in der Öffentlichkeit beziehungsweise in der veröffentlichten Meinung das Bild der Hochschule stark durch die Lehrveranstaltungen und ihre Ausgestaltung geprägt wird, dass innerhalb der Universität jedoch sich nur wenige institutionelle Anreize für ein starkes Engagement in der Lehre finden. Berufungen und Karrieren erfolgen wohl immer noch aufgrund von Publikationen und weniger aufgrund der Lehrleistung.

natürlich auch sehr unterschiedliche Konsequenzen: Gerade einmalige Befragungen in einem Teil der Lehrveranstaltungen am Ende eines Semesters ergeben sicherlich kein wirklich realistisches Bild der Lehrsituation. Wie hier diskutiert, können verschiedene Selektions- oder Auswahlprozesse zu einer – so ist zumindest zu vermuten: wohl insgesamt – positiven Verzerrung führen. Zumindest um hier ein besseres Bild zu erzielen, erscheint auch eine Untersuchung der entsprechenden No-Shows und Veranstaltungsabbrecher dringend geboten.

Die Analyse der Konstanzer Evaluationen der letzten Semester zeigt darüber hinaus, dass die diesem Vorgehen zugrundeliegende theoretische Annahme einer allmählichen Verbesserung der Lehre aufgrund der vorgenommenen Veranstaltungskritik letztlich wohl nicht haltbar ist. Die – allerdings insgesamt sehr positive – Einschätzung der Lehrqualität durch die Studierenden schwankt wohl eher zufällig als dass sie dieser Gesetzmäßigkeit folgt. Es zeigt sich auch in diesem Feld, dass ohne eine genaue theoretische Modellierung des Lehr- und Lernprozesses nur schwer instrumentelle Eingriffe möglich und erfolgreich sind.

Mit besonders großen praktischen Probleme zu kämpfen haben Versuche, studentische Lehrbewertungen anhand externer Faktoren zu validieren. Auch hier ist häufig unklar, welche Faktoren denn als Indikator einer erfolgreichen Ausbildung dienen sollen.[54] Methodisch letztlich sinnvolle Langzeitstudien scheinen aufgrund der Vielzahl damit verbundener organisatorischer und anderer Probleme wohl realistischerweise nur schwer durchführbar.

Dies bedeutet jedoch nicht, dass derartige Befragungen insgesamt nur wenig Sinn machen würden. Man sollte sich nur von Beginn an darüber im Klaren sein, welche Aufgaben die studentische Bewertung von Lehrveranstaltungen erfüllen kann und auf welche Fragen hierdurch sicherlich keine Antworten gefunden werden können. Die studentische Bewertung von Lehrveranstaltungen kann dann durchaus hinsichtlich einiger wichtiger Dimensionen ein bedeutsamer Rückkopplungsmechanismus sein. Konkret auf die Studiums- und Veranstaltungssituation bezogene Fragen scheinen zuverlässig und valide beantwortbar. Problematischer erscheint jedoch die Interpretation dieser Ergebnisse. An universitäre Lehrveranstaltungen werden selbst bei einem einheitlichen Studiengang sehr unterschiedliche Anforderungen gestellt, die teilweise einfach nicht alle gleichzeitig erfüllbar sind. Was dann unter guter Lehre zu verstehen ist, ist schließlich keine mit Hilfe methodischer Kriterien beantwortbare Frage.

---

[54] Zu einer ähnlichen Einschätzung kommt Gold (1996: 149), wenn er schreibt: „Das Validieren eines Inventars zur Lehrqualität setzt eine theoretische Begründung der Kriterien guter Lehrveranstaltungen voraus. Der Validitätsdiskussion müßte also eine Kriteriendiskussion vorangehen“. Es bedarf jedoch eines gewissen Optimismus zu glauben, dass man hier zu einem einfachen und einheitlichen Ergebnis über die Kriterien guter Lehre gelangen kann.

9. Empfehlungen

Eine isolierte Verbesserung des Instruments wird wenig zu einer Verbesserung der Evaluation beitragen. Langfristig wird eine Evaluation des Instruments nur dann eine wirkliche Evaluation und keine Meinungsumfrage darstellen, wenn die wirklich relevanten externen Validierungen vorgenommen werden. Konkret bedeutet dies

1. Es muß regelmäßig eine Befragung der Absolventen der Fakultät (einschließlich aller Abbrecher) nach einem Jahr erfolgen. Der Katalog der Fragen sollte sich auf die Dauer bis zum Jobantritt und die üblichen Bewertungskriterien des Jobs beziehen.
2. Hierzu ist der Aufbau einer regelmäßigen Adressenpflege notwendig.
3. In größeren Intervallen (z.B. alle 3 Jahre) sollte eine Befragung aller Absolventen stattfinden.
4. Hierzu ist es notwendig einer Stelle der Fakultätsinfrastruktur, z.B. innerhalb des Dekanats, die inhaltliche Verantwortung für diese Befragungen zuzuteilen.
5. Zusätzlich sollten die akademischen Karrieren aller Doktoranden der Fakultät über die Datenbanken SOLIS und SSCI verfolgt werden. Deren Zitatshäufigkeit sollte ebenfalls Bestandteil der Evaluation werden.
6. Das eigentliche Evaluationsinstrument sollte minimal um einige Fragen erweitert werden. Hierzu gehören zunächst die formalen Merkmale der Veranstaltung (Pflichtkurs, Wiederholungskurs, Teilnehmerzahl, Räume) sowie vor allem Fragen zur wahrgenommenen Motivation der Studierenden, der wahrgenommenen Qualität studentischer Beiträge (Referate etc. soweit dies in der Veranstaltung möglich war) und zum Interaktionsstil des Dozenten mit den Studierenden.[55] Im Anhang E findet sich eine Version des bisherigen Evaluationsinstruments der Fakultät für Verwaltungswissenschaft der Universität Konstanz, das um solche Fragen ergänzt (aber noch nicht getestet) wurde.

---

[55] In dem Heidelberger Inventar (Rindermann 1996a) finden sich beispielsweise folgende Fragen:

Dimension	Frage
Motivation der Studierenden	Grund für den Besuch der Veranstaltung (Mehrfachantwort möglich) Pflichtveranstaltung, Schein, Prüfungsrelevant wegen der Dozentin/des Dozenten aus Interesse, Thema und/oder ...[offene Antwort möglich]
Qualität studentischer Beiträge	Die Referate der Studierenden sind interessant Die Diskussionen in der Veranstaltung sind produktiv
Interaktionsstil des Dozierenden	Der Dozent ist im Umgang mit den Studierenden freundlich und aufgeschlossen
Relevanz der Veranstaltung	Die Veranstaltung fördert mein Interesse am Studienfach Der Besuch der Veranstaltung lohnt sich

7. Um die methodischen Probleme unterschiedlichen Antwortverhaltens in verschiedenen Veranstaltungen kontrollieren zu können und vor allem um für die langfristige Evaluation der Veranstaltungen die Zuordnung der Studenten zu den Veranstaltungen verfügbar zu machen (auf das Prüfungssystem muß aus rechtlichen Gründen hier verzichtet werden), ist die Verwendung einer eindeutigen Identifikationsnummer unverzichtbar.
8. Aus rein pragmatischen Gründen empfiehlt sich die Verwendung der Matrikelnummer. Dies setzt voraus, dass die Rohdaten nicht mehr den Dozenten zur Verfügung gestellt werden und die Studenten die Notwendigkeit dieser Maßnahme einsehen.
9. Es muß rechtlich geklärt werden, ob die Abiturnote, die Ergebnisse der Kurse und vor allem die Ergebnisse der Diplomprüfungen sowie der Diplomarbeit dem Evaluationsdatensatz hinzugefügt werden dürfen. Nur unter Verwendung dieser Daten ließe sich ein methodisch einwandfreies Evaluationssystem der tatsächlichen Ergebnisse der Lehre aufbauen.
10. Die Beantwortung der Lehrevaluation sollte nicht mehr durch Paper-Pencil-Methoden erfolgen, sondern in Form eines Internet-Surveys. Nach Einloggen mit Matrikel und Passwort (die Verwendung des Prüfungsamtpasswortes bietet sich an) wird der Fragebogen bearbeitet. Dieses Vorgehen bietet eine Reihe von Vorteilen: Druck- und Organisationskosten verschwinden, die Kooperation der Lehrenden ist nicht notwendig, die Ergebnisse liegen maschinenlesbar vor, die Auswertung kann automatisch erfolgen, eine studentische Responsequote ist berechenbar. Der einzige Nachteil liegt in der unverzichtbaren Notwendigkeit eines Personenidentifiers.
11. Die Entwicklungskosten für ein solches Systems dürften in der Größenordnung von ca. 10.000 DM liegen und sich daher schon nach 2-3 Jahren amortisieren.
12. Nach unserem Wissen wäre dies zumindest für die Bundesrepublik ein einmaliges Evaluationskonzept, dass sowohl die langfristige Entwicklungsplanung der Fakultät stützen würde als auch durch die Verfügbarkeit der Ergebnisse während des Semesters eine Rückkopplung noch innerhalb einer Veranstaltung erlauben würde.[56]
13. Sollten sich die Absolventenbefragungen oder die Ergebnisse des Prüfungssystems aus technischen, rechtlichen oder sonstigen Gründen dem Evaluationsdatensatz der Fakultät nicht zufügen lassen,. würde die Lehrevaluation auf eine reine Meinungsumfrage in der Lehrveranstaltung reduziert. In diesem Fall würden wir die Einstellung der formalen Lehrevaluation und die Rückkehr zu den qualitativen Evaluationen in der Eigenverantwortung der Lehrenden empfehlen.

[56] Zwar gibt es einige Ansätze zur Online-Lehrevaluation (zum Beispiel an der Wirtschaftsuniversität Wien, der Universität Greifswald und der Technischen Universität Hamburg-Harburg), diese sind aber nicht Bestandteil eines umfassenden Lehrbewertungskonzepts, sondern lediglich die online-Form einer herkömmlichen Befragung, die allenfalls um ein halbautomatisches Auswertungsmodul erweitert wurden.

Literatur

Abrami, Philip C., d'Apollonia, Sylvia, Cohen, Peter A., 1990: Validity of Student Ratings of Instruction: What we Know and What We Do Not. Journal of Educational Psychology 82: 219-231.

Alkin, Marvin C., Daillak, Richard, White, Peter, 1979: Using Evaluations. Does Evaluation Make a Difference? Beverly Hills/London: Sage.

Altrichter, Herbert, Schratz, Michael, 1992: Hohe Schulen am Prüfstand. Einleitende Überlegungen zur Evaluation und Entwicklung von Universitäten. S. 7-30 in: Herbert Altrichter, Michael Schratz (Hg.): Qualität von Universitäten. Evaluation: Impulse für Innovation? Innsbruck/Wien: Österreichischer StudienVerlag.

Anderson, Gary, 1998: Fundamentals of Educational Research. 2nd edition. London: Falmer Press.

Arbeitsgruppe Bildungsbericht am Max Planck-Institut für Bildungsforschung, 1994: Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick. Vollständig überarbeitete und erweiterte Neuausgabe. Reinbek: Rowohlt.

Astleitner, Hermann, 1991: Studentische Einschätzungen von universitärem Lehrverhalten: Das Problem impliziter Theorien. Psychologie, Erziehung, Unterricht 38: 116-122.

Astleitner, Hermann, Krumm, Volker, 1991: Studentische Einschätzungen als hochschuldidaktische Evaluationsmethode: Validitätsprobleme? Zeitschrift für Hochschuldidaktik 15: 241-255.

Austin, Robert D., 1996: Measuring and Managing Performance in Organizations. New York: Dorset House Publishing.

Balke, Stefan, Stiensmeier-Pelster, Joachim, Welzel, Andreas, 1991: Auswirkungen der Spiegel-Rangliste westdeutscher Universitäten auf die Wahl des Studienortes. S. 307-316 in: Wolff-Dietrich Webler, H. U. Otto (Hg.): Der Ort der Lehre in der Hochschule: Lehrleistungen, Prestige und Hochschulwettbewerb. Weinheim : Deutscher Studien-Verlag.

Baumgartner, Peter, 1999: Evaluation mediengestützten Lernens. Theorie - Logik - Modelle. S. 63-99 in: Michael Kindt (Hg.): Projektevaluation in der Lehre. Multimedia an Hochschulen zeigt Profil(e). Münster/New York: Waxmann.

Basow, Susan A., Howe, Karen G., 1987: Evaluations of College Professors: Effects of Professors‘ Sey-Type, and Sex, and Students’ Sex. Psychological Review 60: 671-678.

Becher, Gerhaad, Kuhlmann, Stefan, 1995: Evaluation of Technology Policy Programmes in Germany. Dordrecht/Boston/London: Kluwer.

Berendt, Brigitte, Stary, Joachim (Hg.), 1993: Evaluation zur Verbesserung der Lehre und weitere Maßnahmen. Weinheim: Deutscher Studienverlag.

Bowen, William G., Bok, Derek, 1998: The Shape of the River. Long-Term Consequences of Considering Race in College and University Admissions. Princeton: Princeton University Press.

Bortz, Jürgen, Döring, Nicola, 1995: Forschungsmethoden und Evaluation. Zweite vollständig überarbeitete und aktualisierte Auflage. Berlin/Heidelberg: Springer.

Brandtstätter, Jochen, 1990: Evaluationsforschung: Probleme der wissenschaftlichen Bewertung von Interventions- und Reformprojekten. Zeitschrift für Pädagogische Psychologie 4: 215-227.

Braskamp, L. A., Ory, J. C., 1994: Assessing faculty work: Enhancing individual and instituional performance. San Francisco: Jossey-Bass.

Bülow-Schramm, Margret, 1994: Planen - beurteilen – analysieren - anwenden. Einführung in die Evaluation der Lehre. Handbuch der Hochschullehre: Informationen und Handreichungen aus der Praxis für die Hochschullehre 1994: D1.1: 1-30.

Bülow-Schramm, Margret, Carstensen, Doris (Hg.), 1995: Frischer Wind für Evaluation? Chancen und Risiken von peer review an deutschen Universitäten. Hochschuldidaktische Arbeitspapiere Nr. 28. Interdisziplinäres Zentrum für Hochschuldidaktik der Universität Hamburg. Hamburg.

Carstensen, Doris, Reissert, Reiner, 1995: Interne und externe Evaluation – Modell und Praxis – Eine Zwischenbilanz aus der Sicht des HIS. Kurzinformationen A16/95 des Hochschul-Informations-Systems. Hannover.

Carstensen, Doris, 1997: Wirksamkeit der internen und externen Evaluation von Lehre und Studium. Mit Berichten aus den Ländern Bremen, Flandern, Niederlande, Niedersachsen und Nordrhein-Westfalen. Bericht der Arbeitsgruppe Evaluation und Leistung. HIS-Kurzinformation 12/97. Hannover.

Cashin, William E., 1988: Student Ratings of Teaching: A Summary of the Research. IDEA Paper No. 20. Center for Faculty Evaluation and Development. Kansas State University. Unpublished paper. Manhattan, Kansas.

Cashin, William E., 1990: Student Ratings of Teaching: Recommendations for Use. IDEA Paper No. 22. Center for Faculty Evaluation and Development. Kansas State University. Unpublished paper. Manhattan, Kansas.

Cashin, William E., 1995: Student Ratings of Teaching: The Research Revisited. IDEA Paper No. 32. Center for Faculty Evaluation and Development. Kansas State University. Unpublished paper. Manhattan, Kansas.

Cashin, William E., 1996: Developing an Effective Faculty Evaluation System. IDEA Paper No. 33. Center for Faculty Evaluation and Development. Kansas State University. Unpublished paper. Manhattan, Kansas.

Chen, Huey-Tsyh, 1990: Theory-Driven Evaluations. Newbury Park: Sage.

Clarke, Alan, Dawson, Ruth, 1999: Evaluation Research. An Introduction to Principles, Methods and Practive. London/Thousand Oaks: Sage.

Cohen, Peter A., 1981: Student Ratings of Instruction and Student Achievment: A Meta-analysis od Multisection Validity Studies. Review of Educational Research 51: 281-309.

Cook, Thomas D:, Aplleton,. Hilary, Conner, Ross F., Shaffer, Ann, Tamkin, Gary, Weber, Stephen J., 1975: „Sesame Street“ Revisited. New York: Russell Sage Foundation.

Cooley, William W., Lohnes, Paul R., 1976: Evaluation Research in Education. New York: Irvington Publishers.

Dahrendorf, Ralf, 1965: Arbeiterkinder an deutschen Universitäten. Tübingen: Mohr Siebeck.

Daniel, Hans-Dieter, 1998: Studentische Beurteilung von Lehrveranstaltungen – Anlage, Durchführung und Ergebnisse eines Modellprojekts an der Universität Mannheim. S. 79-104 in: Hochschulrektorenkonferenz (Hg.): Evaluation und Qualitätssicherung an den Hochschulen in Deutschland – Stand und Perspektiven. Beiträge zur Hochschulpolitik6/1998. Bonn.

Diehl, Joerg M., 1996: Studentische Evaluation von Hochschulveranstaltungen. Ein Kommentar zu Kromrey. Zeitschrift für Pädagogische Psychologie 10: 167-170.

Dorn, Heinz, 1984: Evaluierung – Prüfverfahren des Bundesrechnungshofs? Erfolgskontrolle durch den Bundesrechnungshof. S. 463-469 in: Gerd-Michael Hellstern, Hellmut Wollmann (Hg.): Handbuch zur Evaluierungsforschung. Band 1. Opladen: Westdeutscher Verlag.

Endruweit, Günter, 1992: Programmevaluation als Laienspiel. Bemerkungen über Meinungsforschung, Sozialforschung und Pfusch bei Studentenbefragungen. Soziologie 1992: 107-115.

Esser, Hartmut, 1995: Lehrevaluation – No Shows, Karteileichen, Schleifendreher. Deutsche-Universitäts-Zeitung 1995 (Heft 18): 22-25.

Fakultät für Verwaltungswissenschaft, 1999a: Evaluierung von Forschung und Lehre: Überprüfung und Weiterentwicklung von Evaluierungsinstrumenten. Antrag an das Ministerium für Wissenschaft, Forschung und Kunst. Unveröffentlichtes Manuskript. Konstanz.

Fakultät für Verwaltungswissenschaft der Universität Konstanz, 1999b: Lehrevaluation Universität Konstanz WS 98/99. Verwaltungswissenschaft. Gesamtauswertung. Unveröffentlichtes Manuskript. Konstanz.

Fallon, Daniel, 1998: Kein Geheimnis und nicht mühsam. Evaluation von Forschung, Lehre und Service in den USA. Forschung & Lehre 8/98: 403-405.

Feger, Hubert, 1992: Vergleichende Bewertung von Lehrveranstaltungen. Anmerkungen zur Methodik. Arbeitspapier der Freien Universität Berlin. Manuskript. Berlin.

Feldman, K. A., 1989: The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education 30: 583-645.

Fend, Helmut, 1982: Gesamtschule im Vergleich: Bilanz der Ergebnisse des Gesamtschulversuchs. Weinheim/Basel: Beltz.

Fend, Helmut 1998: Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung. Weinheim/München: Juventa.

Freeman, Harvey R., 1994: Student Evaluations of College Instructors: Effects of Type of Course Taught, Instructor Gender and Gender Role, and Student Gender. Journal of Educational Psychology 86: 627-630.

Freeman, Howard E., Solomon, Marian A., 1981: The Next Decade in Evaluation Research. S. 12-26 in: Robert A. Levine, Marian A. Solomon, Gerd-Michael Hellstern, Hellmut Wollmann (Hg.): Evaluation Research and Practice. Beverly Hills: Sage.

Frey, Bruno S., 1990: Ökonomie ist Sozialwissenschaft. Die Anwendung der Ökonomie auf neue Gebiete. München: Vahlen.

Gelfert, Hans-Dieter, 1992: Kriterien für eine gute akademische Lehre. Eine Erwiderung. Das Hochschulwesen 40: 139-140.

Gerlich, Peter, 1993: Hochschule und Effizienz. Anstöße zur universitären Selbstreflexion. Wien: PassagenVerlag.

Giesen, Thomas, 1995: Professoren zum TÜV? Lehrevaluation zwischen gesellschaftlichen Ansprüchen und informationeller Selbstbestimmung. S. 37-44 in: W. Schmitz (Hg.): Evaluation der Lehre. Dresden: Dresdner Universitätsverlag.

Gold, Andreas, 1996: Können Studierende die Qualität der Lehre beurteilen? Einige Anmerkungen zu Rindermanns Antwort an seine Kritiker. Zeitschrift für Pädagogische Psychologie 10: 147-150.

Gräf, Lorenz, 1991: Fragwürdige Experten. Sekundäranalyse der Spiegel-Untersuchung zur Qualität westdeutscher Universitäten. Soziologie 1991: 69-85.

Grühn, Dieter, 1992: Freie Universität Berlin – Diskussion über Qualität der Lehre in Gang gekommen: Hohe Beteiligung an der Studentenbefragung. Das Hochschulwesen 2/1992: 92-93.

Heid, Helmut, 2000: Die Messbarkeit menschlichen Handelns. Evaluation – ein Begriff und dessen Bedeutung. Neue Zürcher Zeitung Nr. 216 vom 16. und 17. September 2000: 101-102.

Hellstern, Gerd-Michael, Wollmann, Hellmut, 1984: Evaluierung und Evaluierungsforschung – ein Entwicklungsbericht. S. 17-93 in: Gerd-Michael Hellstern, Hellmut Wollmann (Hg.): Handbuch zur Evaluierungsforschung. Band 1. Opladen: Westdeutscher Verlag.

Helmke, Andreas, 1996: Studentische Evaluation der Lehre – Sackgassen und Perspektiven. Zeitschrift für Pädagogische Psychologie 10: 181-186.

Herbst, Jürgen, 1991: Lehrleistungen als Kriterium des Hochschulwettbewerbs – die US-amerikanische Erfahrung. S. 295-306 in: Wolff-Dietrich Webler, H. U. Otto (Hg.): Der Ort der Lehre in der Hochschule: Lehrleistungen, Prestige und Hochschulwettbewerb. Weinheim : Deutscher Studien-Verlag.

Hansen, Ursula, Hennig-Thurau, Thorsten, Wochnowski, Holger, 1997: TEACH-Q: Ein valides und handhabbares Instrument zur Bewertung von Vorlesungen. Die Betriebswirtschaft 57: 376-396.

Hochschulrektorenkonferenz (Hg.), 1995: Europäische Pilotprojekte für die Qualitätsbewertung im Bereich der Hochschulen. Bundesrepublik Deutschland. Nationaler Bericht. Dokumente zur Hochschulreform, 105/1995. Bonn.

Hochschulrektorenkonferenz (Hg.), 1998: Evaluation. Sachstandsbericht zur Qualitätsbewertung und Qualitätsentwicklung in deutschen Hochschulen. Dokumente und Informationen 1/1998. Bonn.

Holtkamp, Rolf, Schnitzer, Klaus (Hg.), 1992: Evaluation des Lehrens und Lernens. Ansätze, Methoden, Instrumente. Evaluationspraxis in den USA, Großbritannien und den Niederlanden. Hannover: HIS.

House, Ernest R., 1993: Professional Evaluation. Social Impact and Political Consequences. Newbury Park: Sage.

Huber, Michael, 1999: Universitätsentwicklung durch Verhandlungen. Können Universitäten so lernen? S. 471-483 in: Claudia Honegger, Stefan Hradil, Franz Traxler (Hg.): Grenzenlose Gesellschaft? Verhandlungen des 29. Kongresses der DGS, des 16. Kongresses der ÖGS, des 11. Kongresses der SGS in Freiburg 1998. Opladen: Leske + Budrich.

Jencks, Christopher, 1992: Rethinking Social Policy. Race, Poverty, and the Underclass. New York: Basic Books.

Keil-Slawik, Reinhard, 1999: Evaluation als evolutionäre Systemgestaltung. Aufbau und Weiterentwicklung der Paderborner DISCO (Digitale InfraStruktur für COmputergestütztes kooperatives Lernen). S. 11-36 in: Michael Kindt (Hg.): Projektevaluation in der Lehre. Multimedia an Hochschulen zeigt Profil(e). Münster/New York: Waxmann.

Kellermann, Paul, 1992: Ranking- und Review-Verfahren in den Vereinigten Staaten. Das Hochschulwesen 40: 126-131.

Kieser, Alfred, 1998: Going Dutch – Was lehren niederländische Erfahrungen mit der Evaluation universitärer Forschung? Die Betriebswirtschaft 58: 208-224.

Krempkow, René, 1998: Ist ‚gute Lehre' meßbar? Die Verwendung studentischer Lehrbewertungen zur Darstellung der Lehrqualität und weitere Maßnahmen. Das Hochschulwesen 46: 195-199.

Krais, Beate, 1983: Bildung als Kapital. Neue Perspektiven für die Analyse der Sozialstruktur? S. 199-220 in: Reinhard Kreckel (Hg.): Soziale Ungleichheit. Sonderheft 2 der Sozialen Welt. Göttingen: Schwartz.

Kreuter, Frauke, Kopp, Johannes, 2000: Absolventenbefragung des Fachbereichs Politik- und Verwaltungswissenschaft der Universität Konstanz. Abschlußbericht des Forschungsprojektes „Evaluation der Arbeitsmarktorientierung und des Arbeitsmarkterfolges durch eine Absolventenbefragung". Konstanz: Konstanzer Online-Publikations-System (KOPS) 2000 (http://www.ub.uni-konstanz.de/kops/volltexte/ 2000/521).

Kreuzer, Markus, 1999: Evaluation von Lehrveranstaltungen an Universitäten. Methodische Aspekte der Evaluation, Evaluationsmodelle für Lehre und Lehrveranstaltungen sowie Formulierungen eines Designs zur Lehrveranstaltungsevaluation. Diplomarbeit im Fach Soziologie der Universität Marburg. Unveröffentlichtes Manuskript. Marburg.

Kromrey, Helmut, 1993a: Studentische „Lehrevaluation“ oder (nur) „Teilnehmerbefragungen“ in Lehrveranstaltungen? Methodische Probleme bei der Bewertung von Lehrqualität. S. 43-51 in: Heinz O. Gralki, Dieter Grühn, Heidemarie Hecht (Hg.): Evaluation schafft Autonomie. Perspektiven der Lehrbewertung an Hochschulen. Berlin: FU Dokumentationsreihe.

Kromrey, Helmut, 1993b: Lehrevaluation darf nicht auf Umfragen reduziert werden. Zur Aussagekraft von Lehrveranstaltungsbefragungen. Mitteilungen des Hochschullehrerverbandes 4: 268-271.

Kromrey, Helmut, 1994a: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragungen von Vorlesungsteilnehmern. S. 105-128 in: Peter Ph. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderungen an die Empirische Sozialforschung. Münster/New York: Waxmann.

Kromrey, Helmut, 1994b: Wie erkennt man „gute Lehre“? Was studentische Vorlesungsbefragungen (nicht) aussagen. Empirische Pädagogik 8: 153-168.

Kromrey, Helmut, 1995a: Evaluation. Empirische Konzepte zur Bewertung von Handlungsprogrammen und die Schwierigkeit ihrer Realisierung. Zeitschrift für Sozialisationsforschung und Erziehungssoziologie 15: 313-336.

Kromrey, Helmut, 1995b: Buchbesprechung von H. Rindermann und M. Amelang: Das Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE). Zeitschrift für Pädagogische Psychologie 9: 221-224.

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

Künzel, Rainer, 1997: Evaluation als Instrument der Selbststeuerung autonomer Hochschulen. S. 74-80 in: Qualitätssicherung in Lehre und Studium. Niedersächsische Erfahrungen im internationalen Vergleich. Dokumentation zum Symposium der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA) am 22. und 23. Mai 1997 an der Universität Hannover. Schriftenreihe „Evaluation der Lehre“ 2/97 der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA). Hannover.

Lamnek, Siegfried, 1990: Zur Professionalität der Studie: „Welche Uni ist die beste?“. Soziologie 1990: 91-100.

Lohnert, Beate, Rolfes, Manfred, 1997: Handbuch zur Evaluation von Lehre und Studium an Hochschulen. Ein praxisorientierter Leitfaden. Schriftenreihe „Evaluation der Lehre“ 3/97 der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA). Hannover.

Marsh, Herbert W., 1980: The Influence of Student, Course, and Instructor Characteristics on Evaluations of University Teaching. American Educational Research Journal 17: 219-237.

Marsh, Herbert W., 1982: SEEQ: A Reliable, Valid, and Useful Instrument for Collecting Students‘ Evaluations of University Teaching. British Journal of Educational Psychology 52: 77-95.

Marsh, Herbert W., 1984: Student‘s Evaluation of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. Journal of Educational Psychology 76: 707-754.

Marsh, Herbert W., 1987: Students‘ Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. International Journal of Educational Research 11: 253-388.

Marsh, Herbert W., Roche, Lawrence, 1993: The Use of Student‘s Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness. American Educational Research Journal 30: 217-251.

Mayer, Thomas, 1992: Die besten Universitäten im Presseurteil. Eine Attacke wider die Trägheit. S. 151-161 in: Herbert Altrichter, Michael Schratz (Hg.): Qualität von Universitäten. Evaluation: Impulse für Innovation? Innsbruck/Wien: Österreichischer StudienVerlag.

Mußgnug, Reinhard, 1992: Gefährden Lehrevaluationen die Freiheit der Wissenschaft? Mitteilungen des Hochschulverbandes 4: 253-256.

Nystroem, Astrid, Dickenberger, Dorothee, 2000: Modulares Instrumentarium zur Lehrevaluation. Universität Mannheim. Version II. Arbeitspapier des Projekt Evaluation der Lehre. Unveröffentlicht. Mannheim.

Palandt, Klaus, 1997: Qualitätssicherung der Lehre in politischer Verantwortung. S. 19-25 in: Qualitätssicherung in Lehre und Studium. Niedersächsische Erfahrungen im internationalen Vergleich. Dokumentation zum Symposium der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA) am 22. und 23.

Mai 1997 an der Universität Hannover. Schriftenreihe „Evaluation der Lehre“ 2/97 der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA). Hannover.

Rau, Einhard, 1996: Evaluation der Hochschullehre. Eine kommentierte Bibliographie. Frankfurt/Berlin/Bern: Peter Lang.

Rauch, Martin, 1995: „Nun evaluiert mal schön!“ Evaluation von Hochschuleinrichtungen als Beitrag zur Qualität von Lehre und Forschung. Handbuch der Hochschullehre: Informationen und Handreichungen aus der Praxis für die Hochschullehre 1995: D1.3: 1-14.

Rindermann, Heiner, 1996a: Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen. Landau: Verlag Empirische Pädagogik.

Rindermann, Heiner, 1996b: Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf Kritik an der Lehrevaluation. Zeitschrift für Pädagogische Psychologie 10: 129-145.

Rindermann, Heiner, Amelang, Manfred, 1994a: Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE). Handanweisung. Heidelberg: Asanger Verlag.

Rindermann, Heiner, Amelang, Manfred, 1994b: Zur Validität der Beurteilungen von Lehrveranstaltungen durch Studierende. Manuskript für den 39. Kongreß der Deutschen Gesellschaft für Psychologie in Hamburg vom 25. Bis 29.September 1994 in Hamburg. Manuskript. Heidelberg.

Ritter, Ulrich Peter, 1993: Evaluation: erstes Scheitern und Perspektiven. S. 180-186 in: Brigitte Berendt, Joachim Stary (Hg.): Evaluation zur Verbesserung der Lehre und weitere Maßnahmen. Weinheim: Deutscher Studienverlag.

Rossi, Peter H., Freeman, Howard E., 1993: Evaluation. A Systematic Approach. 5th edition. Newbury Park/London: Sage.

Rost, Jürgen, 1996: Lehrbuch Testtheorie, Testkonstruktion. Bern: Huber.

Schencker-Wicki, Andrea, 1996: Evaluation von Hochschulleistungen. Leistungsindikatoren und Performance Measurements. Wiesbaden: Deutscher Universitätsverlag.

Scheuch, Erwin K., 1990: Wie gut sind unsere Hochschulen? Soziologie 1990: 73-90.

Schnell, Rainer, Hill, Paul B., Esser, Elke, 1999: Methoden der empirischen Sozialforschung. 6. Auflage. München/Wien: Oldenbourg.

Scholz, Oskar Berndt, 1995: Evaluation von Lehrveranstaltungen durch Studenten. Nachlese einer empirischen Erhebung. Forschung & Lehre 9/95: 497-501.

Schwermer, Rolf, 1999: Kritische Anmerkungen zur „Evaluation in der Lehre“. S. 57-61 in: Michael Kindt (Hg.): Projektevaluation in der Lehre. Multimedia an Hochschulen zeigt Profil(e). Münster/New York: Waxmann.

Sechrest, Lee, Figueredo, Aurelio Jose, 1993: Program Evaluation. Annual Review of Psychology 44: 645-674.

Seidel, Hinrich, 1997: Zwei Jahre flächendeckende Evaluation von Lehre und Studium in Niedersachsen. S. 9-17 in: Qualitätssicherung in Lehre und Studium. Niedersächsische Erfahrungen im internationalen Vergleich. Dokumentation zum Symposium der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA) am 22. und 23. Mai 1997 an der Universität Hannover. Schriftenreihe „Evaluation der Lehre“ 2/97 der Zentralen Evaluationsagentur der niedersächsischen Hochschulen (ZevA). Hannover.

Sohr, Sven, 1993: Studentische Evaluation: Ein geistiger Schildbürgerstreich? S. 168-172 in: Brigitte Berendt, Joachim Stary (Hg.): Evaluation zur Verbesserung der Lehre und weitere Maßnahmen. Weinheim: Deutscher Studienverlag.

Stamm, Margrit, 1998: Qualitätsevaluation und Bildungsmanagement im sekundären und tertiären Bereich. Aarau/Frankfurt/Salzburg: Sauerländer.

Statistisches Bundesamt (Hg.), 1997: Datenreport 1997. Zahlen und Fakten über die Bundesrepublik Deutschland. Bonn: Bundeszentrale für politische Bildung.

Sturm, Michael, 1994: „Die Reise ins Ich ...“ – Sebstevaluation im Hochschulunterricht. Handbuch der Hochschullehre: Informationen und Handreichungen aus der Praxis für die Hochschullehre 1994: D1.2: 1-28.

Süllwold, Fritz, 1992: Universitäre Lehre. Welche Realität wird bei der Beurteilung von Hochschullehrern durch Studierende erfaßt? Mitteilungen des Hochschulverbandes 40: 321-322.

Teichler, Ulrich, 1992: Evaluation von Hochschulen auf der Basis von Absolventenbefragungen. Erfahrungen und Überlegungen aus der Bundesrepublik Deutschland. S. 79-102 in: Herbert Altrichter, Michael Schratz (hg.): Qualität von Universitäten. Evaluation: Impulse für Innovation? Innsbruck/Wien: StudienVerlag.

Van de Vijver, Fons, Leung, Kwok, 1997: Methods and Data Analysis for Cross-Cultural Research. Thousand Oaks: Sage.

Webler, Wolff-Dietrich, 1991: Kriterien für gute akademische Lehre. Das Hochschulwesen 39: 243-249.

Webler, Wolff-Dietrich, 1992: „Kriterien“ in der Diskussion. Antwort auf Gelfert. Das Hochschulwesen 40: 185-187.

Webler, Wolff-Dietrich, 1995: Das Modell eines Lehrberichts über die Evaluation von Lehre und Studium und erste Ergebnisse. Das Hochschulwesen 45: 258-266.

Webler, Wolff-Dietrich, 1996: Qualitätssicherung in Lehre und Studium an deutschen Hochschulen. Zeitschrift für Sozialisationsforschung und Erziehungssoziologie 16: 119-148.

Willems, J., Gijselaers, W., de Bie, D., 1994: Qualitätssorge in der Lehre. Leitfaden für die studentische Lehrevaluation. Neuwied: Luchterhand.

Wissenschaftsrat, 1997: Empfehlungen zur Stärkung der Lehre in den Hochschulen durch Evaluation. S. 55-104 in: Wissenschaftsrat (Hg.): Empfehlungen und Stellungsnahmen 1996. Band I. Bonn.

Wottawa, Heinrich, Thierau, Heike, 1998: Lehrbuch Evaluation. 2. Vollständig überarbeitete Auflage. Bern/Göttingen: Huber.

Zifonun, Natalie, 1999: Studierende im Wintersemester 1998/99. Wirtschaft und Statistik 6/1999: 505-510.

Anhänge:

A: Ein Überblick über die zur Evaluation der Lehre in sozialwissenschaftlichen Studiengängen eingesetzten Instrumente

B: Einige weitere in der Literatur zu findende Instrumente zur Evaluation der Lehre

C: Dokumentation des Befragungsmaterials (Fragebogen, Anschreiben)

D: Konstanzer Lehrevaluation Entwicklungsprofile

E: Ein erweitertes Instrument zur Lehrevaluation im Fachbereich für Politik- und Verwaltungswissenschaft

Anhang A: Ein Überblick über die zur Evaluation der Lehre in sozialwissenschaftlichen Studiengängen eingesetzten Instrumente

Liste der dokumentierten Erhebungsinstrumente

A 1: Universität Bayreuth, Kulturwissenschaftliche Fakultät
A 2: Freie Universität Berlin, Institut für Soziologie
A 3: Universität Bremen, Institut für empirische und angewandte Soziologie
A 4: TU Dresden, Institut für Soziologie
A 5: Albert-Ludwigs-Universität Freiburg, Institut für Soziologie
A 6: Martin-Luther-Universität Halle-Wittenberg, Institut für Soziologie
A 7: Universität Leipzig, Fakultät für Sozialwissenschaften, Institut für Soziologie
A 8: Otto-von-Guericke-Universität Magdeburg, Institut für Soziologie
A 9: Ludwig-Maximilians-Universität München, Institut für Soziologie
A 10: Universität Rostock, Institut für Soziologie
A 11: Universität Trier, Fachbereich IV
A 12 TU Berlin, Fachbereich 7
A 13: Ruhr Universität Bochum, Fakultät für Sozialwissenschaft
A 14: Technische Universität Chemnitz, Philosophische Fakultät
A 15: Universität Essen, Fachbereich 1
A 16: Georg-August-Universität Göttingen, Sozialwissenschaftliche Fakultät
A 17: FernUniversität-Gesamthochschule Hagen, Institut für Politikwissenschaft
A 18: Universität Hildesheim, Institut für Sozialwissenschaften
A 19: Universität Gesamthochschule Kassel, Fachbereich 5
A 20: Pädagogische Hochschule Ludwigsburg, Fakultät für Erziehungs- und Gesellschaftswissenschaften
A 21: Technische Universität München, Fakultät für Wirtschafts- und Sozialwissenschaften
A 22: Universität Paderborn, Sozialwissenschaften
A 23: Universität Potsdam, Wirtschafts- und Sozialwissenschaftliche Fakultät
A 24: Universität Regensburg, Philosophische Fakultät III: Geschichte, Gesellschaft, Geographie
A 25: Universität des Saarlandes, Saarbrücken, Fachbereich 6: Sozial und Umweltwissenschaften
A 26: Pädagogische Hochschule Schwäbisch Gmünd, Fakultät I: Fachbereich Soziologie/ Politikwissenschaft
A 27: Universität Stuttgart, Fakultät Geschichts-, Sozial- und Wirtschaftswissenschaften

A 28: RWTH Aachen, Philosophische Fakultät (Fachbereich 7)
A 29: Technische Hochschule Darmstadt, Institut für Politikwissenschaft
A 30: Universität Dresden, Philosophische Fakultät, Institut für Politikwissenschaft
A 31: Universität Gießen, Institut für Politikwissenschaft
A 32: Ernst-Moritz-Arndt-Universität Greifswald, Institut für Politikwissenschaft
A 33: Friedrich-Schiller-Universität Jena, Institut für Politikwissenschaft
A 34: Otto-von-Guericke-Universität Magdeburg, Institut für Politikwissenschaft
A 35: Ludwig-Maximilians-Universität München, Centrum für angewandte Politikforschung, Geschwister-Scholl–Institut für Politische Wissenschaft
A 36: Universität Rostock, Institut für Politik- und Verwaltungswissenschaften
A 37: Eberhard-Karls-Universität Tübingen, Institut für Politikwissenschaft
A 38: Universität Konstanz, Fakultät für Verwaltungswissenschaft

B: Einige weitere in der Literatur zu findende Instrumente zur Evaluation der Lehre

Liste der dokumentierten Erhebungsinstrumente

B 1: Heidelberger Inventar zur Lehr-Veranstaltungs-Evaluation (aus Rindermann 1996a: 285ff)

B 2: ProfiLe (Professionalisierung individueller Lehre), Projekt an der Universität Mannheim (aus Nystroem/Dickenberger 2000)

B 3: TEACH-Q (aus Hansen/Hennig-Thurau/Wochnowski 1997)

B 4: SEEQ (Students‘ Evaluations of Educational Quality) (aus Marsh 1982)

C: Dokumentation des Befragungsmaterials (Fragebogen, Anschreiben)

D: Konstanzer Lehrevaluation Entwicklungsprofile

E: Ein erweitertes Instrument zur Lehrevaluation im Fachbereich für Politik- und Verwaltungswissenschaft