Natalija el Hage

Instrumente studentischer Veranstaltungskritik
Konstruktion, Einsatz und Beispiele

Hefte zur Bildungs- und Hochschulforschung (14)
Arbeitsgruppe Hochschulforschung, Sozialwissenschaftliche Fakultät,
Universität Konstanz, Dezember 1995

Natalija el Hage

Instrumente studentischer Veranstaltungskritik
Konstruktion, Einsatz und Beispiele

Hefte zur Bildungs- und Hochschulforschung (14)
Arbeitsgruppe Hochschulforschung, Sozialwissenschaftliche Fakultät,
Universität Konstanz, Dezember 1995

Einleitung

In den letzten Jahren hat sich das Thema Lehrevaluation zu einem der wichtigsten Punkte auf der hochschulpolitischen Tagesordnung entwickelt. Eine große Anzahl an Projekten zur Beurteilung und Verbesserung der Lehre sind inzwischen auf den Weg gebracht. Eine Übersicht zu größeren Projekten der Lehrevaluation in Deutschland, zu den gesetzlichen Bestimmungen der Bundesländer und zu notwendigen Rahmenbedingungen ist im Heft 15 der Hefte zur Bildungs- und Hochschulforschung („Studienreform durch Lehrevaluation? Ansätze, Projekte und Verwendungsmöglichkeiten.“) zu finden.

Die unüberschaubare Anzahl von kleineren und größeren Vorhaben der Lehrevaluation, die häufig Daten der studentischen Veranstaltungskritik verwenden, wird von sehr unterschiedlichen Initiatoren und Initiatorinnen betrieben. Von ganzen Hochschulen, die mit externer Hilfe evaluiert werden, bis hin zu studentischen Fachschaften in einzelnen Fächern ist alles vertreten. So unterschiedlich wie die Betreiber und Betreiberinnen der Projekte sind, so unterschiedlich ist die Qualität der verwendeten Fragebogen zur studentischen Veranstaltungskritik.

Der vorliegende Bericht dient zum einen dazu, Prinzipien und Konstruktionskriterien für die Entwicklung und Verwendung von Instrumenten studentischer Veranstaltungs-kritik aufzuzeigen; zum anderen sollen in Deutschland entwickelte und verwendete Instrumente vorgestellt werden, die methodisch geprüft und sich praktisch bewährt haben. Veranschaulicht werden soll damit, daß eine sorgsame Fragebogenkonstruktion notwendig ist, zumal wenn dieses Beurteilungsverfahren dazu dienen soll, die Qualität verschiedener Veranstaltungen und Lehrender zu vergleichen.

Verschiedene Gründe sprechen für sorgfältig ausgearbeitete Fragebogen. Marsh (1982b) und Aleamoni und Thomas (1980) haben nachgewiesen, daß detaillierte und multidimensionale Fragebogen weniger anfällig für externe Einflüsse sind, verglichen mit Fragebogen, die sich auf nur wenige globale Items stützen. Über mögliche Faktoren, die sich auf die Ergebnisse studentischer Lehrbewertung auswirken können, wird in Heft 13 der Konstanzer Hefte zur Bildungs- und Hochschulforschung („Zur Validität studentischer Veranstaltungskritiken. Befunde empirischer Studien zu einem umstrittenen Verfahren.“) berichtet.

In diesem Heft sollen anhand bisheriger Untersuchungsergebnisse und -erfahrungen Kriterien zusammengestellt werden, die für die Erstellung solcher Fragebogen der studentischen Veranstaltungskritik beachtenswert sind. Da die meisten dieser Untersuchungen in den Vereinigten Staaten durchgeführt wurden, fehlen an einigen Punkten noch für die deutsche Hochschullandschaft notwendige Untersuchungen.

Im zweiten Teil dieses Bericht sind deutschsprachige Instrumente der Lehrbewertung dargestellt. Ausgewählt wurden (aufgrund der beschränkten Möglichkeiten) Fragebogen, die einerseits zu Forschungszwecken benutzt werden oder andererseits bei hochschulübergreifenden Evaluationsprojekten verwendet wurden. Dies soll keine Herabwürdigung oft sorgsam erstellter fach- oder hochschulspezifischer Instrumente sein. Es würde jedoch den vorhandenen Rahmen sprengen, die ganze Bandbreite der verwendeten Verfahren aufzuzeigen.

Für die Bereitschaft, uns diese Auskünfte zu geben, danken wir den Autoren der Instrumente sehr. Die Beschreibungen der von ihnen erstellten Verfahren werden in Kapitel 10 vorgestellt.

Dieser Bericht soll dazu beitragen, daß die Anzahl allzu spontan erstellter und deshalb methodisch meist mangelhafter Fragebogen abnimmt. Diese werden oft nur aufgrund unzureichend bekannter Alternativen zusammengestellt. Im vorliegenden Heft werden einige Kriterien aufgezeigt, die bei der Konstruktion und Auswertung von entsprechenden Fragebogen in Betracht gezogen werden sollten. Darüberhinaus bietet es erprobte Instrumente, die übernommen werden oder zur Orientierung bei der Entwicklung eigener Lösungen oder Kurzform-Fragebogen dienen können. Der Vorteil einer solchen Verwendung liegt nicht zuletzt darin, daß Vergleichsdaten vorliegen und herangezogen werden können.

Die Darstellungsweise dieses Berichts versucht, sowohl interessierten Anwendern und Anwenderinnen als auch methodisch Erfahrenen gerecht zu werden. Daher werden praktische Hinweise zur Zielgruppe, zur Durchführungsdauer und zum Auswertungsaufwand angeführt, aber auch entsprechende Gütekriterien werden soweit wie möglich genannt. Mit diesem Beitrag verbinden wir die Hoffnung, interessierten Anwendern und Anwenderinnen wichtige Informationen darüber zu vermitteln, was bei Fragebogen für Studierende zur Bewertung von Lehrveranstaltungen und Lehrenden zu beachten ist, damit möglichst zutreffende und zuverlässige Urteile erfaßt werden. Denn zur Verbesserung der Lehrqualität können sie letztlich eher beitragen, wenn sie selbst bestimmten Gütestandards genügen.

Konstanz, Dezember 1995 Natalija el Hage

Inhalt

Einleitung

In den letzten Jahren hat sich das Thema Lehrevaluation zu einem der wichtigsten Punkte auf der hochschulpolitischen Tagesordnung entwickelt. Eine große Anzahl an Projekten zur Beurteilung und Verbesserung der Lehre sind inzwischen auf den Weg gebracht. Eine Übersicht zu größeren Projekten der Lehrevaluation in Deutschland, zu den gesetzlichen Bestimmungen der Bundesländer und zu notwendigen Rahmenbedingungen ist im Heft 15 der Hefte zur Bildungs- und Hochschulforschung („Studienreform durch Lehrevaluation? Ansätze, Projekte und Verwendungsmöglichkeiten.“) zu finden.

Die unüberschaubare Anzahl von kleineren und größeren Vorhaben der Lehrevaluation, die häufig Daten der studentischen Veranstaltungskritik verwenden, wird von sehr unterschiedlichen Initiatoren und Initiatorinnen betrieben. Von ganzen Hochschulen, die mit externer Hilfe evaluiert werden, bis hin zu studentischen Fachschaften in einzelnen Fächern ist alles vertreten. So unterschiedlich wie die Betreiber und Betreiberinnen der Projekte sind, so unterschiedlich ist die Qualität der verwendeten Fragebogen zur studentischen Veranstaltungskritik.

Der vorliegende Bericht dient zum einen dazu, Prinzipien und Konstruktionskriterien für die Entwicklung und Verwendung von Instrumenten studentischer Veranstaltungs-kritik aufzuzeigen; zum anderen sollen in Deutschland entwickelte und verwendete Instrumente vorgestellt werden, die methodisch geprüft und sich praktisch bewährt haben. Veranschaulicht werden soll damit, daß eine sorgsame Fragebogenkonstruktion notwendig ist, zumal wenn dieses Beurteilungsverfahren dazu dienen soll, die Qualität verschiedener Veranstaltungen und Lehrender zu vergleichen.

Verschiedene Gründe sprechen für sorgfältig ausgearbeitete Fragebogen. Marsh (1982b) und Aleamoni und Thomas (1980) haben nachgewiesen, daß detaillierte und multidimensionale Fragebogen weniger anfällig für externe Einflüsse sind, verglichen mit Fragebogen, die sich auf nur wenige globale Items stützen. Über mögliche Faktoren, die sich auf die Ergebnisse studentischer Lehrbewertung auswirken können, wird in Heft 13 der Konstanzer Hefte zur Bildungs- und Hochschulforschung („Zur Validität studentischer Veranstaltungskritiken. Befunde empirischer Studien zu einem umstrittenen Verfahren.“) berichtet.

In diesem Heft sollen anhand bisheriger Untersuchungsergebnisse und -erfahrungen Kriterien zusammengestellt werden, die für die Erstellung solcher Fragebogen der studentischen Veranstaltungskritik beachtenswert sind. Da die meisten dieser Untersuchungen in den Vereinigten Staaten durchgeführt wurden, fehlen an einigen Punkten noch für die deutsche Hochschullandschaft notwendige Untersuchungen.

Im zweiten Teil dieses Bericht sind deutschsprachige Instrumente der Lehrbewertung dargestellt. Ausgewählt wurden (aufgrund der beschränkten Möglichkeiten) Fragebogen, die einerseits zu Forschungszwecken benutzt werden oder andererseits bei hochschulübergreifenden Evaluationsprojekten verwendet wurden. Dies soll keine Herabwürdigung oft sorgsam erstellter fach- oder hochschulspezifischer Instrumente sein. Es würde jedoch den vorhandenen Rahmen sprengen, die ganze Bandbreite der verwendeten Verfahren aufzuzeigen.

Für die Bereitschaft, uns diese Auskünfte zu geben, danken wir den Autoren der Instrumente sehr. Die Beschreibungen der von ihnen erstellten Verfahren werden in Kapitel 10 vorgestellt.

Dieser Bericht soll dazu beitragen, daß die Anzahl allzu spontan erstellter und deshalb methodisch meist mangelhafter Fragebogen abnimmt. Diese werden oft nur aufgrund unzureichend bekannter Alternativen zusammengestellt. Im vorliegenden Heft werden einige Kriterien aufgezeigt, die bei der Konstruktion und Auswertung von entsprechenden Fragebogen in Betracht gezogen werden sollten. Darüberhinaus bietet es erprobte Instrumente, die übernommen werden oder zur Orientierung bei der Entwicklung eigener Lösungen oder Kurzform-Fragebogen dienen können. Der Vorteil einer solchen Verwendung liegt nicht zuletzt darin, daß Vergleichsdaten vorliegen und herangezogen werden können.

Die Darstellungsweise dieses Berichts versucht, sowohl interessierten Anwendern und Anwenderinnen als auch methodisch Erfahrenen gerecht zu werden. Daher werden praktische Hinweise zur Zielgruppe, zur Durchführungsdauer und zum Auswertungsaufwand angeführt, aber auch entsprechende Gütekriterien werden soweit wie möglich genannt. Mit diesem Beitrag verbinden wir die Hoffnung, interessierten Anwendern und Anwenderinnen wichtige Informationen darüber zu vermitteln, was bei Fragebogen für Studierende zur Bewertung von Lehrveranstaltungen und Lehrenden zu beachten ist, damit möglichst zutreffende und zuverlässige Urteile erfaßt werden. Denn zur Verbesserung der Lehrqualität können sie letztlich eher beitragen, wenn sie selbst bestimmten Gütestandards genügen.

Konstanz, Dezember 1995 Natalija el Hage

Inhalt

1 Prinzipien der Entwicklung von Instrumenten

1.1 Unterschiedliche Anforderungen an Instrumente

Unter dem Begriff der Instrumente zur Lehrevaluation sind sehr unterschiedliche Verfahren zu finden. Dies können einerseits sehr elaborierte Instrumente sein, die sorgsam entwickelt wurden, andererseits Fragebogen, die fast beliebig zusammengestellt erscheinen. Es gibt sowohl spezielle Instrumente, die nur für bestimmte Fächer oder nur für bestimmte Kursarten gedacht sind, als auch Instrumente, die bereits im Routineverfahren eingesetzt werden. Instrumente studentischer Veranstal-tungskritik unterscheiden sich in Hinblick auf ihr Einsatzfeld, ihre Qualität und ihren Umfang. Auch in Deutschland existieren die verschiedensten Verfahren, von detailliert überprüften Instrumenten bis hin zu kurzfristig zusammengestellten Fragepaketen.

Bei der Planung von Projekten zur Lehrevaluation hat man daher die Wahl entweder auf ein vorhandenes Instrument zurückzugreifen, ein solches Instrument entsprechend den Bedürfnissen abzuwandeln oder aber einen eigenen Fragebogen zu entwickeln.

Bei der Frage, für welche Form man sich entscheidet, kommt es vor allem darauf an, zu welchen Zwecken die gesammelten Daten verwendet werden sollen. Wird nur ein Instrument benötigt, um Lehrenden eine kurze Rückmeldung zu geben, kann selbst ein einfaches Instrument mit relativ globalen Fragen nützlich sein. Ein solches Instrument sollte dann in erster Linie dazu dienen, eine breitere Datengrundlage für Gespräche über die Lehrqualität im Kurs bereitzustellen.

Wird jedoch angestrebt, die Lehrqualität in einem Fachbereich insgesamt zu erfassen und dafür Evaluationsdaten zu gewinnen, sollen Lehrende untereinander verglichen oder didaktische Maßnahmen geplant werden, so sind andere, anspruchsvollere Instrumente erforderlich. Für jede über das direkte “Kursfeedback“ hinausgehende Absicht ist es sinnvoll, komplexere Instrumente zu verwenden.

Eine sorgfältige Fragebogenkonstruktion ist besonders dann zwingend, wenn ein Instrument zum Vergleich zwischen Lehrenden und ihren Lehrleistungen dienen soll oder wenn die Befunde nicht nur zu einer besseren Kommunikation über die Lehrqualität führen sollen, sondern auch reale (z.B. finanzielle) Konsequenzen haben können.

Um den notwendigen Gütestandard zu erreichen, der vorauszusetzen ist, um Lehrende miteinander zu vergleichen, sind detailliertere und multidimensionale Fragebogen unerläßlich. Diese Instrumente sind weniger anfällig für mögliche externe Einflüsse. Fragebogen mit relativ wenigen globalen Items erreichen nicht das erforderliche Niveau an Zuverlässigkeit, Meßgenauigkeit und Objektivität. Wichtig ist deshalb bereits vor der Lehrbewertung zu klären und zu entscheiden, für welchen Zweck die Daten benötigt werden und wie sie verwendet werden sollen.

1.2 Aspekte der Fragebogenkonstruktion und der Anwendung

In den Vereinigten Staaten, die schon eine lange Tradition in der Lehrbewertung haben, existieren die unterschiedlichsten Untersuchungen zur Frage nach optimalen Kriterien für die Fragebogen studentischer Veranstaltungskritik. Im folgenden werden Ergebnisse derartiger Untersuchungen (vor allem aus den USA) vorgestellt. Daraus werden Kriterien abgeleitet, die für die Konstruktion von Fragebogen relevant sind, die einen weiterreichenden Anspruch haben, als Lehrenden eine kurze Rückmeldung zu geben, Fragebogen folglich, die eine möglichst hohe Qualität aufweisen sollen, um auch Veranstaltungen und Lehrende miteinander vergleichen zu können.

Vorab ist zu betonen, daß es sich um Konstruktionskriterien für Fragebogen zur studentischen Veranstaltungskritik handelt. Beurteilt wird mit ihnen nur die jeweilige Veranstaltung oder Veranstaltungsreihe. Aspekte der Lehre und der Lehrorganisation, die den Fachbereich oder die gesamte Hochschule betreffen, werden nicht weiter diskutiert. Zu vermuten ist, daß für diese Form der Evaluation zum Teil andere Kriterien eine Rolle spielen, daß z.B. für die Beurteilung der Lehrorganisation weitaus stärker auf bilanzierende und wertende (im Vergleich zu beobachtbaren) Daten zurückgegriffen werden kann als dies bei der Beurteilung einzelner Lehrveranstaltungen der Fall ist.

Die folgenden Kapitel integrieren jeweils thematisch zusammenhängende Untersuchungsergebnisse und umfassen auch Schlußfolgerungen für die Entwicklung entsprechender Fragebogen. Der Aufbau wurde so gewählt, daß die einzelnen Aspekte bei der Konstruktion von Instrumenten studentischer Veranstaltungskritik nach ihrer Wichtigkeit aufgeführt sind.

Die *Auswahl und Abgrenzung verschiedener Dimensionen*, die im nächsten Kapitel behandelt werden, stellen dementsprechend den wichtigsten "Merkposten" bei der Entwicklung dieser Verfahren dar. Die Multidimensionalität der Erfassungsmethode ist wohl der Grundstein für eine möglichst hohe Güte des Verfahrens.

Ein weiterer wichtiger Punkt, der die Qualität eines Fragebogens (vor allem seine Meßgenauigkeit) beeinflussen kann, ist die Klärung des *Vergleichsmaßstabs, den die Beurteilenden* bei der Beantwortung anlegen. Dieser Vergleichsmaßstab ist bei *der Formulierung der Fragen und Vorgaben*, d.h. bei der Formulierung der sog. Items, zu beachten.

Die daran anschließenden Kapitel behandeln Fragen, die den *Gesamtbogen*, die *Vergleichstabellen* und die *Ergebnisdarstellung* betreffen. Ein weiterer Abschnitt ist dem *Zeitpunkt und der Frequenz von Datenerhebungen* gewidmet.

Im Kapitel neun ist eine *zusammenfassende Übersicht* zu finden, die alle genannten Kriterien im Sinne eines "Merkblattes" auflistet, die für die Entwicklung, Gestaltung und Anwendung aufgrund vorliegender Untersuchungen und Erhebungen wichtig sind.

2 Auswahl und Abgrenzung verschiedener Dimensionen

Die Lehre ist ein vielschichtiger Vorgang mit vielen Aspekten und Facetten. Um sie möglichst effizient zu erfassen, ist daher ein multidimensionales Instrument notwendig. Aus diesem Grund werden nachfolgend Ansätze zur Ermittlung entsprechender Dimensionen vorgestellt und ihre Befunde diskutiert.

2.1 Multidimensionalität von Lehrbewertungen

Angesichts der Komplexität von Lehrleistungen ist als erstes zu klären, welche Dimensionen einer Veranstaltung bzw. des oder der Lehrenden abgebildet werden sollen.

Marsh und Hocevar (1984) führten eine großen Untersuchung (31.322 Studierende) zum Intrumen „Students´ Evaluation of Educational Quality“ (SEEQ) durch. Über den Vergleich von konvergenter und diskriminanter Validität zwischen den Faktoren versuchten sie, die Multidimensionalität der Lehre, bzw. ihre Wahrnehmung durch die Studierenden, zu bestätigen. Die konvergente Validität erwies sich als substantiell (r=.68), die diskriminante lag weit niedriger.

Die neun konzipierten Faktoren des SEEQ konnten bestätigt werden. Sie sind als Dimensionen der Lehrqualität zu verstehen. Dabei handelt es sich um die folgenden Dimensionen:

- Wert der erworben Kenntnisse (Learning/ Value),

- Enthusiasmus (Enthusiasm),

- Kursorganisation (Organziation),

- Verhalten zur Gruppe (Group Interaction),

- Zuwendung (Individual Rapport),

- Ermutigung (Breadth of Coverage),

- Fairness der Tests und Examina (Examinations),

- Fairness der Notenvergabe (Assignments),

- Kursschwierigkeit (Workload, Difficulty)

Dennoch werden die meisten Fragebogen zur Lehrevaluation nicht in diesem Sinne konzipiert bzw. ausgewertet (vgl. auch Scott, 1977, S. 746). Sie vernachlässigen die Dimensionalität der Lehre insgesamt oder lassen einzelne Dimensionen unberücksichtigt.

Problematisch an einer globalen Fragebogenkonstruktion für studentischen Veranstaltungskritiken ist die weitaus **höhere Anfälligkeit für Verzerrungen** durch äußere Einflüsse. Eine ganze Reihe von Faktoren steht immer wieder in Verdacht, die Lehrbewertung unsachgemäß zu beeinflussen (z.B. Geschlecht, rhetorische Fähigkeiten des/der Lehrenden). Nach dem bisherigen Forschungsstand (für eine Übersicht vgl. el Hage, 1995) ist davon auszugehen, daß relativ wenig verzerrende Auswirkungen externer Einflüsse nachzuweisen sind. Diese Einflußgrößen können, falls sie empirisch bestätigt werden, meist rechnerisch einkalkuliert und die Ergebnisse entsprechend bereinigt werden.

Es gibt demnach zwei Möglichkeiten, den Einfluß externer, möglicherweise verzerrender Faktoren auf die studentische Veranstaltungskritik zu minimieren: zum einen die multidimensionale Gestaltung der Fragebogen und zum anderen die Bereinigung gewonnener Ergebnisse. Letzteres ist, nach dem bisherigen Stand, hauptsächlich für den Faktor „Motivation und Interesse am Kursthema“ notwendig. Der Einfluß aller anderen Faktoren ist entweder vernachlässigbar gering oder durch die mehrdimensionale Gestaltung der Erhebungsinstrumente aufzufangen.

Ein weiterer Nachteil einer globalen Erfassung und Interpretation von Lehrleistungen (z.B. durch Bildung eines Durchschnittswertes) ist der erhebliche **Informationsverlust**, der dadurch hingenommen werden muß. Beide Punkte, Anfälligkeit für Verzerrungen und Informationsqualität, werden im folgenden diskutiert.

Aleamoni und Thomas (1980) gehen sehr differenziert der Frage nach, welche Unterschiede zwischen globalen und differenzierten Vorgaben bestehen. Die beiden Autoren untersuchten ob: a) Studentische Charakteristika höher mit generellen als mit speziellen Items korrelieren, b) die Beantwortung der generellen Items ein guter Prädiktor für die Beantwortung spezieller Items ist, und c) ob spezielle Items bessere Prädiktoren der Bewertung der Lehrenden für studentische Leistungen und Verhalten in einem Kurses sind als generelle Items. Insgesamt korrelierten die Autoren Daten von 7.242 beantworteten Bogen des Illinois Course Evaluation Questionares (CEQ).

Ergebnisse: Für alle externen Variablen mit Ausnahme von Kursgröße und Dozentenrang war die Korrelation mit generellen Items signifikant höher (1 Prozent Niveau) als mit speziellen Items. Keine der 56 Korrelationen mit den Subskalen des CEQ war größer als .19 (.013 -.183). Die Korrelationen mit den generellen Items betrug dabei r= .052 -.375. Die Korrelation zwischen der Fähigkeitsbewertung durch Lehrende und des CEQ war für die Subskalen in allen Fällen signifikant höher als für die generellen Items. Korrelationen zwischen generellen Items und Subskalen lagen zwischen .29-.49.

Die Schlußfolgerung von Aleamoni und Thomas (1980) ist eindeutig: Die Korrelationen zwischen externen Variablen und Lehrevaluation ist durch die Fragebogenkonstruktion bedingt. Lehrbewertungen sollten daher mit Subskalen arbeiten und nicht auf generellen Items beruhen.

Diese Folgerung, die in der Literatur immer wieder bestätigt wird (z.B. durch das Auffinden verschiedener Faktoren in der Lehrbewertung), ist bei der Konzeption von Fragebogen und ihrer späteren Auswertung zu berücksichtigen. Vergröbert lassen sich die Ergebnisse von Aleamoni und Thomas wie folgt zusammenfassen: Benutzt man globale Beurteilungen für Kurs und Lehrende, dann können in viel stärkerem Maße "Beliebtheit" oder allgemeine Akzeptanz des Lehrenden erfaßt werden, als dies bei spezielleren Beurteilungen einzelner Lehraspekte der Fall wäre.

Die Problematik des Einflusses externer Variablen auf die Lehrbewertung versucht auch **Marsh** (1982b) in das Bewußtsein der Anwendenden solcher Instrumente zu bringen, indem er in allen seinen Artikeln betont: "Vieles der Vieldeutigkeit in den Untersuchungen zur studentischen Lehrbewertung stammt von der unglückseligen Praxis anzunehmen, daß studentische Bewertungen eindimensional seien" (Marsh, 1982b, S. 278). Außer Marsh sind auch eine große Anzahl anderer Autoren und Autorinnen der gleichen Ansicht, z.B. DuCette und Jane Kenney (1982).

Bei der Interpretation der Ergebnisse ist ebenfalls eine differenzierte Herangehensweise zu beachten. Es ist weder methodisch zulässig, noch inhaltlich aufschlußreich, Durchschnittswerte über Dimensionen zu bilden, die wenig miteinander korrelieren sollen, d.h. unterschiedliche Sachverhalte darstellen.

Ein wesentliches Ziel studentischer Veranstaltungskritik liegt in der Verbesserung der Lehre und der Qualifizierung der Lehrenden. Um eine solche Verbesserung und Qualifizierung erreichen zu können, muß den Dozierenden jedoch eine differenzierte Rückmeldung gegeben werden. Dies ist bei globalen Informationen nicht der Fall, weil offen gelassen wird, in welche Richtung die Verbesserung gehen sollte. Zwar wird in globalen Fragebogen oft versucht, dieses Manko durch die nachdrückliche Aufforderung zu ergänzenden Kommentaren der Studierenden zu mildern, diese sind jedoch aufwendig auszuwerten und können eine quantifizierbare Rückmeldung nicht ersetzen.

Aus den vorliegenden Ergebnissen kann die **Schlußfolgerung** gezogen werden, daß die Wahrung der Multidimensionalität eine der wichtigsten Voraussetzungen für ein objektives und informatives Instrument der Lehrbewertung ist. Sie ist deshalb bei der Entwicklung des Instruments wie bei der Auswertung der erhobenen Daten gleichermaßen zu beachten.

2.2 Dimensionen der Veranstaltungsbewertung

Nach der Feststellung, daß Lehre multidimensionale Aspekte hat, stellt sich die Frage: Welche Aspekte können ermittelt werden? Dabei muß zum einen nach den Verfahren gefragt werden: Wie können solche Dimensionen ermittelt werden? Zum anderen nach der Übereinstimmung bisher in empirischen Untersuchungen gefundener Dimensionen: Wie können sie miteinander verglichen und aufeinander bezogen werden.

Wie können Dimensionen der Lehrbewertungen ermittelt werden?

Grundsätzlich ließe sich bei der Ermittlung von Lehrdimensionen eine induktive und eine deduktive Vorgehensweise unterscheiden. Eine **deduktive Handlungsweise** ist in diesem Fall kaum möglich, da, wie Schott bereits 1973 anmerkte, keine "Theorie des Unterrichts" existiert.

In den letzten zehn Jahren hat sich in diesem Punkt zwar einiges entwickelt, so wurde z.B. die kognitive Informationsverarbeitung aufgegriffen und im schulischen Lernen untersucht. Dennoch ist man in der Pädagogischen Psychologie von einer solchen "Theorie des Unterrichts" noch weit entfernt.

Weinert (1986) z.B. erhebt die Forderung: "So gesehen erschiene es ohne Nutzen, den Gegenstandsbereich der Pädagogischen Psychologie gegenüber der allgemeinen Lernforschung abzugrenzen. Was wir brauchen ist vielmehr eine theoretisch verzahnte Lern- und Instruktionspsychologie! Diese Forderung ist erst in Ansätzen erfüllt..." (S. 784). Als Grund für diesen Mangel sieht Weinert die "disfunktionale Arbeitsteilung" zwischen dem Auffinden elementarer psychischer Lernmechanismen und der angewandten pädagogischen Forschung.

Eine weitere Ursache für eine fehlende theoretische Grundlage im Bereich der Lehre spricht **Heckhausen** an: Die Bindung der Pädagogischen Psychologie an die Institution Schule und die "üblichen Schüler-Altersstufen" (Heckhausen, 1986, S.787). Um eine "Theorie des Unterrichts" an Hochschulen zu entwickeln sind mit Sicherheit bereits genügend Grundlagen vorhanden. Eine solche Theorie existiert jedoch noch nicht. Folglich können Items zwar aus verschiedenen Ansätzen "herausgefiltert", gleichwohl nicht systematisch abgeleitet werden.

In allen vorliegenden Untersuchungen liegt daher überwiegend eine **induktive Herangehensweise** vor. Verwandt wurden z.B. Flanagans "Critical incidents-technique", "manuelle" Klusterung von Dozenten-/ Dozentinnenbeschreibungen zu homogenen Kategorien, multidimensionale Skalierung u.ä.. Es wurden auch Sekundäranalysen bereits gefundener Faktoren durchgeführt und diese zu gemeinsamen Dimensionen zusammengeführt.

Dabei muß bedacht werden, daß die meisten gefundenen Dimensionen Ergebnisse von Faktorenanalysen an entwickelten oder bestehenden Fragebogen sind. Es ist also theoretisch möglich, daß die "im Umlauf befindlichen" Items nur ganz bestimmte Arten von Dimensionen erfassen und andere dabei vernachlässigen. Um so mehr verdeutlicht dies die Notwendigkeit einer Anwendung von pädagogischen, psychologischen und anderen Theorien auf die Hochschullehre. Dies wäre auch dienlich, um das Mißtrauen von professoraler Seite gegen solche Instrumente abzubauen.

Einige theoretische Überlegungen zum `Unterricht´ an Hochschulen

Bereits in der Studie von **Müller-Wolf und Fittkau** (1971) ist erkennbar, daß die Autoren sich nicht nur für das Aufdecken von Lehrdimenisionen interessieren, sondern auch für ihre Wirkung. "Der deutlichste gefundene Zusammenhang besteht zwischen einem aktiven, dynamischen, anregenden, somit `didaktisch effektiven´ Lehrverhalten (...) und einer `zufriedenen Stimulierung´ der Studenten im Rahmen ihres Studiums (...): r=.67. Doch in fast demselben Maße hängt die Zufriedenheit der Studenten mit der Lehrveranstaltung, ihre Stimulierung und ihr Gefühl der Attraktivität des Fachgebiets (...) von dem `emotional positiven, demokratischen Lehrverhalten´ des Hochschullehrers ab: r=.63" (S. 176). Die gefundenen Interaktionen können kaum schon als Theorieansätze gewertet werden. Dennoch sind sie als ein Auftakt dieser Forschung in Deutschland zu sehen.

Erste theoretische Überlegungen zur Hochschullehre hat in Deutschland **Schott** (1973) formuliert. Er unterscheidet dabei einen informationspsychologischen und einen motivationspsychologischen Gesichtspunkt. Zu den informationspsychologischen Aspekten gehören dabei z.B. Informationsvermittlung einerseits und Informationsaufnahme und Informationsverarbeitung andererseits. Zu den motivationspsychologischen Aspekt werden dagegen vor allem motivierende Handlungen des/ der Lehrenden gefaßt und ihre Wirkung auf die Studierenden.

Schotts Ableitungen aus der Allgemeinen Psychologie und der Motivationspsychologie und deren Anwendung auf die Lehre sind zwar interessant, führen aber nicht zu konkreten Ableitungen für die Lehre und bleiben, trotz ihrer Detailliertheit, abstrakt.

Dennoch sind die Schlußfolgerungen von Schott erste Ansätze für eine Theorie des ‘Hochschulunterrichts’. Ob es allerdings sinnvoll ist, in diesem Sinne weiter zu forschen, auch mit der ganz speziellen Zielsetzung der Wissensvermittlung an Hochschulen, bleibt noch zu diskutieren. Zwar könnten Fragebogen der studentischen Veranstaltungskritik dann aus entsprechenden Theorien abgeleitet werden. Es stellt sich jedoch die Frage, inwieweit diese sich von den vorhandenen unterscheiden würden.

Welche Dimensionen der Lehrbewertung sind zu beachten?

Die nachfolgend aufgelisteten acht Dimensionen der Lehrbewertung, die als typisch gelten dürfen, sind aus verschiedenen Studien gewonnen. Der Vergleich wurde nur über die Benennung der Dimensionen durchgeführt, was streng genommen nur als eine Nebeneinanderstellung gelten kann.

Um die Dimensionen, die z.T. mit identischen Begriffen übertitelt sind, wirklich vergleichen zu können, müßte man die einzelnen Items, die zu der entsprechenden Dimension zusammengefaßt wurden, vergleichen bzw. die Faktoren, die sich bei der gleichen Stichprobe in den verschiedenen Instrumenten ergeben, miteinander korrelieren.

Da es sich um Aspekte der Lehre handelt, die leichter beobachtbar und abgrenzbar sind als viele andere psychologische Konstrukte (z.B. Extraversion), läßt sich vermuten, daß die gefundenen Dimensionen, die gleichartig benannt wurden, im großen und ganzen auch vergleichbar sind.

Um eine Aussage zu den Dimensionen von Fragebogen machen zu können, wurde in drei Schritten vorgegangen. Im ersten Schritt wurden die gefundenen Dimensionen gemeinsam mit mehr oder weniger namensgleichen Dimensionen aus bisher genannten Studien zusammengestellt. Im zweiten Schritt wurden Studien hinzugezogen, deren Struktur mit den anderen nicht vergleichbar ist, die gefundenen Faktoren nicht kompatibel sind oder deren Dimensionen nicht auf einer Faktorenanalyse beruhen (z.B. Lita Linzer Schwarz, 1980). Im dritten Schritt wurden insgesamt ähnliche Dimensionen begründet zusammengefaßt.

Klar zu trennen sind zwei große Bereiche: "... die Faktorenanalyse zeigt, daß Studierende zwischen Geselligkeit und Qualifikationsaspekten der Lehre unterscheiden können" (Beatty und Zahn, 1990, S. 281). Während eine der beiden übergeordneten Dimensionen verhältnismäßig einfach zu dem Faktor "Zuwendung" zusammengezogen werden kann, sind für die Beschreibung der zweiten übergeordnete Dimension eine ganze Reihe von Faktoren notwendig.

Zusammenfassung möglicher Dimensionen

1. "**Zuwendung**": Freundlichkeit, achtet auf Studierende, freundlich und aufgeschlossen sein, Studierende achten/ nicht herabsetzen, Sensibilität für Reaktionen, demokratisch sein, Toleranz für einen anderen Blickwinkel, Zugänglichkeit für Kritik, Emotionales Klima im Kurs, Respekt, macht Mut, gleichberechtigter Austausch, Erreichbarkeit.

In diese Dimension wurden alle Faktoren eingeordnet, die eher auf die "mitmenschliche" Ebene bezogen werden können. Eine Zwitterrolle nimmt hier der Faktor "Diskussionsbereitschaft" ein, da er je nach Itemzusammensetzung und Formulierung eher als Merkmal positiven Verhaltens zu Studierenden oder aber als didaktisches Mittel gesehen werden kann.

2. "**Fairneß von Prüfungen und Benotungen**": Diese Dimension könnte im allgemeinen schon in der ersten miteinbegriffen werden. Da jedoch Examen und Noten in der Hochschule eine große Rolle spielen, gibt es genügend Rechtfertigungen, dies als eigene Dimension zu betrachten. Unter Umständen können Fragen zur Prüfungsrelevanz des Stoffes mit aufgommen werden, wenn dies als ein Maßstab für die Fairneß von Prüfungen definiert ist.

3. "**Kommunikationsfähigkeit**": Abstrakte und komplexe Ideen und Theorien darstellen können, verständlich darstellen können, gut erklären können, Tempo der Wissensvermittlung, Voraussetzungsadäquatheit, Erklärungsaktivität, Prägnanz, Verwendung von Hilfsmitteln, akustische Verständlichkeit und Zusammenfassungen geben. Die genannten Punkte machen eine Dimension deutlich, die vor allem in der Fähigkeit der Stoffvermittlung und ihrer Voraussetzungen (z.B. Umgang mit Hilfsmitteln) liegt.

4. "**Kurs- bzw. Stofforganisation**": Schwerpunktsetzung, Systematik, Zeitökonomie, Koordination und Vororientierung. Diese gut abzugrenzende Dimension ist in den meisten Fragebogen zu finden.

Mit der letzten Dimension wären die tendenziell leicht abzubegrenzenden Bereiche genannt. Dabei dürfen gerade für die Punkte drei und vier die notwendigen Voraussetzungen weder vergessen, noch als eigenständige Faktoren gesehen werden. Diese Voraussetzungen sind: Fachwissen und Vorbereitung.

Ohne ein elementares Fachwissen und eine grundlegende Vorbereitung werden keine guten Ergebnisse in den letzten beiden Dimensionen möglich sein. Die genannten Grundlagen sind nicht beobachtbar und können nur abgeleitet werden. Sie sollten daher auch nicht abgefragt werden, wenn es um die Beurteilung der tatsächlichen Lehrleistung in Veranstaltungen geht.

Die folgenden Dimensionen (fünf bis acht) sind schwerer abgrenzbar und greifen z.T. ineinander über.

5. "**Stimulierung**": Stimulation und Bestärkung eigenständigen Denkens, zufriedenstellende Stimulation vs. enttäuschte Stagnation, Motivierung durch Interesse, Entwicklung von Eigeninitiative vs. Desinteresse und begeistern können. Gerade in dieser Dimension wären weitere Forschungen notwendig. In der pädagogischen Psychologie gibt es z.B. Theorien zum sog. "Entdeckenden Lernen", daß u.a. die intellektuelle Neugier der Kinder stimulieren soll. Diese Theorien beziehen sich ausschließlich auf den Schulunterricht. Ähnliche Analysen wären auch für den Hochschulbereich notwendig.

6. "**Variabilität vs. Monotonie**": Dies ist ein Faktor, der weitreichendere Punkte beinhaltet als die Kurs- und Stofforganisation. Vermutlich ist es ein wichtiger Aspekt der Informationsaufnahme (Schott, 1973), da er gemeinsam mit anderen die Voraussetzung für das "Aufmerksam bleiben können" bildet und damit wiederum die Voraussetzung für die kontinuierliche Informationsaufnahme über einen längeren Zeitraum .

7. "**Enthusiasmus**": Darunter fällt auch Dynamik/ Charisma und Anziehungskraft des Vortragenden. Diese Dimension hat großen Einfluß auf die vorherigen beiden Dimensionen. Enthusiasmus für ein Fach ist mit Sicherheit eine gute Grundlage, um Studierende zu begeistern und intellektuelle Neugier zu stimulieren. Auf der anderen Seite wird es ein charismatischer Lehrer und eine dynamische Dozentin weitaus leichter haben, die Aufmerksamkeit wachzuhalten und weniger Monotonie aufkommen zu lassen. Wie die Fox-Studien nachweisen, behalten Studierende von einem enthusiastischen Vortrag auch mehr Informationen. Problematisch ist an dieser Dimension, daß sie von allen bisher genannten am ehesten auf eine Persönlichkeitseigenschaft deutet, und somit am wenigsten für den/ die Lehrende veränderbar oder trainierbar ist. Dennoch spielt dieser Faktor in der Lehre eine große Rolle (vgl. Murray, 1983).

8. "**Kurswert**": Wert des Kurses für die Studierenden, Relevanz des Kursmaterials, Prüfungsrelevanz und Theorie-Praxis-Verhältnis. Dieser Faktor beurteilt am stärksten den Kurs aus einer externen Perspektive, und ist somit auch ein erster Ansatz zu einer Kursreform. Stärker als bei den anderen Faktoren wird hier weniger die Qualität des/ der Lehrenden in den Mittelpunkt gestellt, als die Qualität und Relevanz des angebotenen Stoffes.

In diesen Bereich fallen grob zudem "Kursschwierigkeit" und "Arbeitsbelastung im Kurs". Da festgestellt wurde, daß diese beiden Aspekte kaum mit den Lehrbewertungen korrelieren, sollte überlegt werden, ob sie nicht als einfache Einschätzungen aus Feedbackgründen abgefragt werden können. Möglich wäre aber auch, sie so zu formulieren, daß sie zu anderen Dimensionen passen, z.B. kann die Einschätzung: "Der Kurs ist zu schwer" entweder der Kursorganisation oder der Kommunikations-fähigkeit des Lehrenden zugeordnet werden (bei entsprechender Umformulierung).

Schlußfolgerung

Lehre besteht aus verschiedenen Dimensionen. Es kann als erwiesen gelten, daß Studierende unterschiedliche Lehraspekte differenzieren können. Daß in einigen Studien wenige Faktoren gefunden werden, liegt vermutlich häufig an der zu geringen Itemzahl oder an zu wenig differenzierten Formulierungen (z.B. Mazer, 1977); oder aber an ungenügender Kenntnis der Testtheorie, die davon ausgeht, daß ein Merkmal sich aus verschiedenen Merkmalsfacetten zusammensetzt, und diese mit Hilfe von Items erfaßt werden, anstatt zu versuchen, eine Dimension mit einem Item zu erfassen, (wie z.B. Terry und McIntosh, 1988)

Fragebogen müssen somit so konstruiert sein, daß sie unterschiedliche Faktoren der Lehrqualität erfassen, da bei globalen Items eine weitaus höhere Anfälligkeit für externe Einflüsse besteht. Wie kann man die möglichen Faktoren ermitteln? Eine „Theorie des Unterrichts" an den Hochschulen existiert bestenfalls in Teilen. Daher ist ein stringent deduktives Vorgehen z.Z. nicht möglich.

Nach Durchsicht verschiedener empirischer Studien kristallisieren sich acht mögliche Faktoren heraus: Zuwendung, Faireß von Prüfungen und Benotungen, Kommunikationsfähigkeit, Kurs- bzw. Stofforganisation, Stimulierung, Variabilität vs. Monotonie, Enthusiasmus und Kurswert.

Diese acht Faktoren sind grundlegende und in den verschiedensten Fragebogen immer wieder zu findende Dimensionen. Die aufgeführten Dimensionen sollten jedoch nicht als gegeben akzeptiert werden. Weitere Untersuchungen in diesem Bereich -und vor allem eine Theoriebildung- wären sinnvoll.

Deutlich wird an den genannten Dimensionen, daß in den Vereinigten Staaten ganz überwiegend nur der einzelne Kurs und seine Aspekte bewertet werden. Notwendig wäre aber ebenfalls entweder eine Dimension "Fachbereichsbewertung" einzuführen, bei der Aspekte der Lehrorganisation, die den ganzen Fachbereich betreffen, bewertet würden, oder aber spezielle Fragebogen für die "Evaluation der Lehre eines Fachbereichs" zu entwickeln.

3 Vergleichsmaßstab der Beurteilenden

Die Evaluation der Lehre braucht, wie jede andere Bewertung, einen Vergleichsmaßstab, eine **Bezugsnorm**. Ohne vorgegebenen Vergleichsmaßstab verwenden Studierende einen eigenen Maßstab, der unterschiedlich ausgerichtet sein kann. Die Frage nach der "optimalen" Art der Bezugsnorm soll an dieser Stelle diskutiert werden. Ihre Beantwortung ist relevant, da die Formulierung von Items grundsätzlich auch bestimmte Vergleichsoptionen beinhaltet. Diese Optionen verschieben sich je nach gewählter Formulierung.

3.1 Idealnorm oder sozialer Vergleich

Dabei muß als erstes festgelegt werden, ob mit den Items Beobachtungen der Studierenden oder Wertungen erfaßt werden sollen. Items, die beobachtbare Elemente der Lehre erfassen sollen, führen insgesamt zu höheren Gütewerten des Instruments. Dennoch kann es an einigen Stellen sinnvoll sein, Items miteinzubeziehen, die wertende Urteile erfordern, entweder weil der erwünschte Sachverhalt nur so erfaßt werden kann (z.B. die Kursschwierigkeit) oder weil ein Interesse daran besteht, die Wirkung bestimmter Lehrelemente auf die Studierenden zu erfassen (z.B. inwieweit der Kurs bei der Vorbereitung auf die Prüfung nützlich war). Studierende würden in diesen Fällen, d.h. bei wertenden Beurteilungen, die **Selbstreferenz** als Vergleichsmaßstab anlegen.

Bei Items, die vor allem die Einschätzung von beobachteten Sachverhalten zum Ziel haben, sind zwei unterschiedliche Vergleiche möglich: der Vergleich mit einer Idealnorm und/ oder ein sozialen Vergleich. Die **Idealnorm** kann eine existierende Person sein, deren Unterricht als "ideal" empfunden wird oder aber, was wahrscheinlicher ist, die persönliche Vorstellung von einer optimalen Lehre. Der **soziale Vergleich** dagegen erfolgt durch den Bezug auf andere Lehrende der Hochschule oder des Fachbereichs. Durch entsprechende Formulierung kann sichergestellt werden, welcher Vergleich von den Studierenden durchgeführt wird.

Beispiel I: Bitte bewerten Sie den Dozenten A bei den entsprechenden Items auf der Antwortskala von 1 wie sehr gut bis 5 wie ungenügend. Bei dieser Formulierung ist nicht klargestellt, ob Studierende den Dozenten im Vergleich zu anderen Lehrenden bewerten oder ihn mit ihrer Idealvorstellung vergleichen.

Beispiel II: Bitte bewerten Sie die Dozentin B bei den entsprechenden Items im Vergleich zu anderen Lehrenden auf der Antwortskala von 1 wie weit überdurchschnittlich bis 5 wie weit unterdurchschnittlich.

Beispiel III: Bitte bewerten Sie das Tafelbild des Dozenten A. Bei diesem Beispiel ist nicht näher spezifiziert, welche Elemente des Tafelbilds wichtig sind und bewertet werden sollen. Daher ist zu vermuten, daß Studierende das Tafelbild des Dozenten A

nach ihren eigenen Vorstellungen von wichtigem und unwichtigem Aspekten beurteilen werden, da "Tafelbild" eine relativ breite Kategorie ist, die viel Interpretationsspielraum läßt.

Beispiel IV: Bitte bewerten Sie inwieweit das Tafelbild der Dozentin B übersichtlich ist. In diesem Fall ist bereits vorher bestimmt, daß die Übersichtlichkeit des Tafelbildes ein wichtiger oder entscheidender Punkt ist, den die Studierenden beurteilen sollen. Bei diesem Beispiel ist gut vorstellbar, daß Studierende die gleichen Kriterien an ein Tafelbild anlegen wie die Gestaltenden des Fragebogens, und auch, daß die Studierenden untereinander gleiche oder ähnliche Kriterien verwenden, da das zu Beurteilende relativ präzise gefaßt ist.

Levinthal, Lansky und Andrews (1971) z.B. gehen der Vermutung nach, daß Studierende ihre "maximalen Unterscheidungsabstände" strecken, um sie mit den Evaluationsschemata in Einklang zu bringen. Der Grund jedoch, daß Studierende ihr eigenes "Wertsystem" benutzen, liegt vor allem an der Formulierung der Items, wie im weiteren deutlich wird.

263 Studierende der Einführung in die Psychologie gaben in der Studie eine Gesamtbewertung ab und beantworteten einen Fragebogen. **Ergebnis des Vergleichs** zwischen den Bewertungen und den von den Studierenden genannten Ideal- (Soll-) Normen waren deutliche Unterschiede in der Wichtigkeitseinstufung der Items, z.B. wurde nur ein einziges Item einstimmig möglichst positiv gesehen, d.h. eine möglichst hohe Ausprägung als ideal angenommen, bei drei anderen Items wurde dagegen eine mittlere Ausprägung für ideal befunden, z.B. auch für das folgende Beispielitem: "Er war nachgiebig und flexibel."

Problematisch ist, daß die Autoren (bzw. die Fragebogenkonstrukteure) sich nicht bewußt gemacht haben, daß eine solche Ausprägung nicht an ihrem äußersten Ende, sondern eher im Mittelbereich als optimal betrachtet wird. Dadurch, daß weder das Wertesystem der Studierenden ermittelt, noch das Item aus einer Theorie abgeleitet wurde, kommt es häufig zu einer nicht akkuraten Interpretation durch Anwendende eines solchen Fragebogens, da in kurzen und globalen Fragebogen die Ergebnisse summiert werden, und somit implizit davon ausgegangen wird, daß ein hoher Wert auf einem Item gleichzeitig auch eine außergewöhnlich gute Lehrleistung bedeutet.

3.2 Einbeziehung von Soll- und Relevanzurteilen

Sommer und Petermann (1978) haben untersucht, inwieweit präskriptive Urteile den Informationsgehalt des Fragebogens nachweisbar erhöhen. Sommer und Petermann zielen einen Fragebogen an, der ein Modell zur Verknüpfung von Ist-, Soll- und Relevanzurteilen beinhaltet. Das Ist-Urteil bezieht sich dabei auf die Bewertung des momentanen Zustandes. Das Soll-Urteil hat die Funktion, die individuelle Bezugsnorm

deutlich zu machen. Das Relevanzurteil entspricht einem Gewichtsfaktor mit dem die Ist-Urteile multipliziert wurden. Je geringer die Relevanz eingeschätzt wird, desto stärker nähern sich die gewichteten Urteile an ihr arithmetischen Mittel Null. Dadurch schränkt sich die mögliche Varianz ein und wird korrelationsstatistisch unwirksam.

114 Studierende beantworteten den Fragebogen; vier Image-Analysen wurden durchgeführt:

1) Betrachtet man nur das Ist-Urteil führt dies zu zwei bipolaren und eindeutig interpretierbaren Faktoren mit 41 Prozent aufgeklärter Varianz.
2) Das Ist-Urteil gewichtet mit dem Relevanzurteil kommt zu drei Faktoren und 38 Prozent aufgeklärter Varianz. Die Faktoren sind jedoch unipolar und mit Ausnahme des ersten kaum zu benennen.
3) Nimmt man die Ist-Soll-Differenz, so ergibt sich ein starker Generalfaktor (38 Prozent der Varianz) und drei weitere kaum interpretierbare Faktoren. Insgesamt kommt es zu 59 Prozent aufgeklärter Varianz. Jedoch ist eine Vergröberung der Struktur zu bemerken.
4) Gewichtet man die Ist-Soll-Differenz mit dem Relevanzurteil, ergeben sich 4 Faktoren und 53 Prozent aufgeklärter Varianz. Die Faktorenstruktur ist gut interpretierbar und differenzierter als in den vorherigen Analysen.

Die Schlußfolgerung der Autoren lautet: Präskriptive Urteile beinhalten wichtige Zusatzinformationen, daher sollten Soll- wie auch Relevanzurteile in die Lehrbewertung miteinbezogen werden.

Leider findet sich im Artikel von Sommer und Petermann nicht ein einziges Beipielitem aus ihrem Fragebogen. Wenn man davon ausgeht, daß auch sie Items benutzen, deren Vergleichsnormen nicht eindeutig geklärt und deren optimale Ausprägungen nicht aufgeschlüsselt wurden, dann ist einsichtig, daß die Faktorenstruktur durch Einbezug von Soll-und Relevanzurteilen, somit durch die Klärung der Vergleichsnorm, differenzierter und deutlicher wird. Es ist dadurch möglich, einen höheren Informationsgehalt zur Ansicht der Studierenden über die Lehre zu erzielen.

Dieses Verfahren erscheint jedoch für die Lehrbewertung sehr aufwendig und sollte daher auf die Konstruktion entsprechender Instrumente beschränkt bleiben. Bei der Entwicklung von Instrumenten zur Lehrevaluation ist es notwendig, sowohl Studierende als auch Lehrende nach den für sie relevanten Aspekten zu befragen. Im späteren Einsatz bei der Lehrbewertung erscheint dieser Aufwand nicht nötig.

Es ist daher unbedingt darauf zu achten, daß bei den Itemformulierung allgemein akzeptierte Normen von Lehraspekten vorgegeben und in den Formulierungen zu erkennen sind. Wenn das Ziel der Fragebogenkonstruktion sein soll, einen Fragebogen

mit möglichst hohen Gütekriterien (im Sinne der Testtheorie) zu erhalten, so kann dies nur durch eine möglichst niedrige Subjektivität der Beurteilungen erreicht werden.

In einem solchen Falle sollten die Items, die von den Studierenden keine Beobachtungen sondern Wertungen erheben (d.h. als Bezugspunkt die Selbstreferenz haben), gesondert ausgewertet werden

Ein noch offen gebliebener Punkt ist die Frage nach dem Vorzug eines "idealen" oder eines "sozialen" Vergleichssystems. Dies läßt sich mit einem Zitat des Schulforschers Weiß beantworten: "Für eine Lehrveranstaltungskritik gib es gegenwärtig keine lehrzielorientierten Bezugsnormen. Jeder Studierende hat aber Erfahrungen mit verschiedenen Lehrern. Die Mängel des sozialen Maßstabs erscheinen mir bei der Beurteilung von Lehrleistungen geringer. Ich halte es für vertretbar, soziale Maßstäbe anzuwenden, etwa `im Vergleich zu anderen Lehrveranstaltungen überdurchschnittlich - durchschnittlich - unterdurchschnittlich´" (Weiß, 1989, S. 39).

Da im Moment für den Hochschulunterricht keine andere Bezugsnorm vorhanden ist, d.h. kein Kategoriensystem für gute Lehre existiert, erscheint es angebracht, die Studierenden aufzufordern, einen Vergleich mit anderen Lehrenden zu ziehen.

Schlußfolgerung

Es erscheint sinnvoll, bei der Konstruktion eines Fragebogens die Ansichten der Betroffenen zu erheben (subjektive Norm). Damit kann gewährleistet werden, daß zum einen die relevanten Aspekte der Lehrqualität im Instrument enthalten sind und zum anderen ihre unterschiedliche Gewichtung deutlich wird.

Bei der Datenerhebung jedoch sollte von den einzelnen Meinungen abstrahiert und ein möglichst „objektives Instrument“ angestrebt (objektive Norm) werden. Dies ist vor allem dann erforderlich, wenn Lehrende miteinander verglichen werden sollen, da für eine angestrebte hohe Meßgenauigkeit eine geringe Antwortvarianz Voraussetzung ist.

Sollen die Studierenden wertende Aussagen abgeben, so sind diese getrennt zu betrachten.

Zusammenfassend läßt sich sagen, daß es vorteilhaft ist, die angestrebten Vergleichsmöglichkeiten konkret vorzugeben. Ein sozialer Vergleich und genauer gefaßte Beschreibungskategorien sind dabei zweckmäßig.

4 Formulierung der Fragen und Vorgaben (Items)

Das wichtigste an einem Fragebogen sind die Fragen und Vorgaben, die sog. Items. An ihrer Qualität entscheidet sich die Validität der Befunde. Auf ihre Formulierung ist daher besondere Aufmerksamkeit zu verwenden.

An dieser Stelle sollen nicht allgemeine Kriterien der Itemformulierung wie z.B. Verständlichkeit, Vermeidung von Antworttendenzen etc. aufgeführt werden, wie sie in entsprechenden Lehrbüchern zu finden sind. Vielmehr soll auf jene wichtigen Punkte verwiesen werden, die speziell bei der Formulierung von Items (Fragen und Vorgaben) einer Lehrbewertung zu beachten sind.

4.1 Fragen und Vorgaben sollen von Beobachtbarem ausgehen

Häufig sind in Fragebogen zur Lehrbewertung Items zu finden, die nach nicht beobachtbaren Elementen fragen. Dieses Problem ist, trotz vieler „Spontanerhebungen“ zur Veranstaltungskritik, in Deutschland nicht so häufig anzutreffen wie in den U.S.A. Dennoch schleichen sich in viele hiesigen Fragebogen ebenfalls solche Items ein.

Alle folgenden Beispiele sind der Dokumentation zur Evaluation der Lehre (Hochschul-Informations-System, 1992, Teil 1) entnommen: "Wie gut ist der Professor vorbereitet?" und "Wie ist die fachliche Kompetenz des Dozenten?" sind zwei typische Fragen. Beide angesprochenen Sachverhalte sind jedoch für die Studierenden nicht direkt wahrnehmbar, sondern erschließen sich ihnen durch verschiedene im Kurs wahrgenommene Begebenheiten. Bei der Frage nach der Güte der Vorbereitung der Lehrenden z.B. kann aus der Beantwortung von Fragen, zur Strukturierung der Vorlesung, dem berühmten "roten Faden" oder ähnlichem auf die vermutete Vorbereitung geschlossen werden.

Dies erscheint an zwei Punkten problematisch, wenn die „Lehrleistung“ erfaßt werden soll. Zum einen kann es vorkommen, daß sich Lehrende gut vorbereiten, dies aber nicht vermitteln können. Zum anderen verleitet eine solche Fragestellung jeden Studierenden zu seiner eigenen Interpretation von "gut vorbereitet". Die Folgen beschreiben Thorndike und Hagen (1969): "Bewertungen haben die größte Chance genau zu sein für die Aspekte, die nach außen hin gezeigt werden .... Die Erfahrung hat gezeigt, daß diese valider bewertet werden" (S. 428).

Die Einsicht, daß die Antworten der Studierenden subjektiv sein könnten, ist auch bei vielen Fragebogengestaltenden vorhanden. Um das Problem zu umgehen und Lehrenden nichts Falsches zu unterstellen, sind daher immer wieder Items der folgenden Art zu finden: "Wie wichtig, glaubst Du, ist dem Dozent der Lehrauftrag?", "Wie beurteilst Du das Engagement des Professors im Seminar?", "Zeigt der Professor starken persönlichen Einsatz, oder wirkt er eher lustlos?", "Mein Eindruck ist, daß die Einstellung des Dozenten zu den Studenten positiv ist.", "Der Dozent aktualisiert meiner Meinung nach sein Fachwissen." (Hervorhebung durch die Autorin.)

Studentische Veranstaltungskritik sollte keine “Meinungsumfrage“ sein, sondern ein Beurteilungsinstrument im Sinne einer einschätzenden Beobachtung. Der direkte Nutzen einer Information sowohl für den administrativen Gebrauch als auch für Verbesserungsvorschläge zur Lehre ist äußerst gering, wenn die Frage z.B. nach dem "Engagement im Seminar" nicht näher definiert wird.

Wie eng sollen Items gefaßt werden?

Bei der Formulierung von Items zu Lehrbewertung stellt sich die Frage, wie nah an Beobachtbarem diese abgefaßt werden sollen. Lehrbewertungsbogen unterscheiden sich maßgeblich von üblichen Beobachtungsmethoden durch die Breite der Beobachtungseinheiten. Dennoch ist es für die Güte eines Fragebogens zur studentischen Veranstaltungskritik wichtig, **Items möglichst nahe an Beobachtbarem** zu verfassen. Wie genau also müssen Itemeinheiten formuliert werden?

Ein trefflicher Hinweis ist bei Faßnacht (1979) zu finden: "Die Frage nach der Auflösung ist eine Frage nach dem Zweck ... Wir können ja versuchen, einen bayerischen Hackbraten ebenso fein aufzuschneiden wie Bündner Fleisch..." (S. 75). Man kann davon ausgehen, daß für eine Lehrevalution relativ grobe Kategorien ausreichen würden. Allerdings ist dabei zu beachten, daß je ungenauer die Einheiten gewählt werden, desto eher dies zu niedrigeren Gütekriterien führen kann und praktische Folgerungen erschwert werden.

Ein **optimierter Mittelweg** muß somit zwischen der Forderung nach möglichst genauer Verhaltensbeschreibung und dem begrenzten Umfang eines Beurteilungsinstruments gefunden werden. Dieser muß jeweils für den einzelnen Fragebogen neu "bestimmt" werden.

4.2 Verantwortlichkeit des oder der Lehrenden

Es müßte selbstverständlich sein, daß nur Aspekte in eine Lehrbewertung einfließen, die von den Lehrenden tatsächlich beeinflußt werden können. Ein in überraschend vielen Fragebogen auftauchendes Item ist folgendes: "Konnte Dich der Dozent für den Stoff motivieren?" Alle Lehrenden werden die Erfahrung gemacht haben, daß es nicht einzig und allein von ihnen abhängt, ob einzelne Studierende sich für bestimmte Themen motivieren lassen.

Dieses Problem wird bereits in der vermutlich ersten deutschen Untersuchung zur studentischen Veranstaltungskritik von **Müller-Wolf und Fittkau** (1971) deutlich: "Die "Entwicklung von Eigeninitiative im behandelten Fachgebiet" (...) hängt hingegen nur in schwächerem Maße mit "didaktisch effektiver Aktivität" (...) und "positivem demokratischem Lehrverhalten" ($r = .44$ u. $r = .41$) des Hochschullehrers zusammen. Hier zeigen sich die Grenzen der Auswirkungen des Lehrverhaltens und der Einfluß eigenständiger Entschlüsse der Studierenden" (S. 177).

Ähnlich liegt die Problematik bei der Frage nach einer ausreichenden Raumgröße. Oder aber wenn Assistenten/ Assistentinnen für die Konzeption der Vorlesung bewertet werden, für die sie nicht verantwortlich sind.

Sinnvoll können solche Fragen nur erhoben werden, wenn sie als zusätzliche Variablen verwandt werden, aber nicht in eine Wertung für die entsprechenden Lehrenden miteinbezogen werden, die zum Vergleich mit anderen Lehrenden dienen soll.

4.3 Relevanz von Items für die Lehrbewertung

Lehre hat verschiedenste Aspekte, die man mit einem Fragebogen erfassen kann. Jedoch unterscheiden nicht alle Aspekte die Stärken und Schwächen der Lehrenden voneinander. Ein gutes Beispiel dafür ist die Kursschwierigkeit, die in verschiedenen Untersuchungen wenig bis kaum mit anderen Dimensionen der Lehrbewertung korreliert und wenig die Gesamtbewertung der Lehrenden beeinflußt. Ein solches Item ist somit für die Lehrbewertung nicht maßgeblich. Nur Items, die zwischen den Gruppen besserer und schlechterer Lehrender unterscheiden, sollten in die vergleichende Bewertung eingehen.

Dabei erhebt sich auch die Frage, wie mit Items umgegangen wird, die in einigen Kursen wichtige, in anderen dagegen keine Rolle spielen. Eine Lösungsmöglichkeit stellt das **Cafeteria-Prinzip** dar. Dieses Prinzip (vgl. Webler, 1992, S. 150 ff.) bietet die günstige Möglichkeit, ohne größere Schwierigkeiten Fragebogen abzuwandeln, um sie für spezielle Kurse passend zu gestalten. Vorteilhaft dabei ist, daß ein Pool bereits formulierter und evaluierter Items existieren würde, die einen geringeren Aufwand bei der Fragebogenzusammenstellung ermöglichen.

Allerdings muß, um die Vergleichbarkeit zu gewährleisten, ein bestimmter Teil an Items identisch sein. Möglich wäre, "**Itempakete**" für verschiedene Kursarten, Fächer u.ä. zusammenzustellen. Eher aus "pädagogischen Gründen" wäre es sinnvoll, zusätzlich einige Items von Lehrenden der jeweiligen Kurse nennen zu lassen. Webler (1992) schlägt ein `Zwiebelmodell´ vor, mit einem Standardteil, zusätzlichen fach- bzw. lokalspezifischen Fragen und wenigen offenen Fragen.

Letzterer Punkt ist relativ wichtig, da immer wieder Lehrende vermuten, daß sie besser bewertet würden, wenn sie selbst die Items formulieren oder zumindest zusammenstellen könnten. **Mishra** (1979) untersuchte diese Vermutung an 76 Kursen (mit 1650 Studierenden) in denen 50 Lehrende bewerten wurden. Der dabei benutzte Itempool enthielt 227 Items zu verschiedenen Aspekten der Lehrbewertung. Die 27 Items aus diesem Pool, die sowohl von Studierenden als auch von Lehrenden hoch bewertet wurden, stellte Mishra zu einem Fragebogen zusammen. Die Dozenten und Dozentinnen wurden aufgefordert, vor der Verteilung der Bögen diese Items aufmerksam durchzulesen und die wichtigsten Items für ihren Kurs zu identifizieren.

Mishra stellte fest, daß Lehrende signifikant besser auf den von ihnen genannten Items bewertet wurden. Korrelationen zwischen dem, was ein Lehrender wichtig findet, und dem, wo er gut bewertet wird, sind logischerweise hoch. Dies kann jedoch entweder dissonanztheoretisch erklärt werden: In den Lehraspekten, die man als wichtig empfindet, wird man sich auch um gute Leistungen bemühen, um eine Dissonanz zu vermeiden. Oder es ist als selbstwertdienliche Kausalbeziehung zu sehen, d.h. Lehrdimensionen, die gut beherrscht werden, werden auch als wichtig empfunden.

Eine andere Sache dagegen ist es, wenn Items nicht von einzelnen Lehrenden als "wichtiger" betrachtet werden, sondern entweder eine allgemeine Übereinstimmung besteht oder eine theoretische Begründung vorhanden ist, warum manche Aspekte der Lehre wichtiger sind als andere und somit stärker gewichtet werden sollten.

Schlußfolgerung

Instrumente der studentischen Veranstaltungskritik sollten sich aus Items zusammensetzen, die möglichst von Beobachtbarem ausgehen, damit die Beurteilungen der Studierenden eindeutig sein können. Dabei muß ein akzeptabler Mittelweg zwischen der Genauigkeit der Itemformulierung und der Fragebogenlänge gefunden werden.

Die Fragen müssen des weiteren so gestellt sein, daß es für die Studierenden deutlich ist, ob sie Lehrende an ihren Idealvorstellungen oder an anderen Lehrenden vergleichen sollen. Es muß für die Studierenden fest umrissen sein, welchen Vergleichsmaßstab (Anker) sie bei der Beantwortung derselben anlegen sollen.

Weiter sollten Lehrende nur in Dimensionen bewertet werden, für die sie verantwortlich sind und die gute und weniger gute Lehrende tatsächlich unterscheiden.

Es sollte die Möglichkeit bestehen, allgemein interessierende Fragen, wie nach der Grundmotivation für das Fach oder der Überfüllung der Veranstaltung, zu stellen. Jedoch sollten diese Fragen klar abgegrenzt von den Lehrbeurteilungen sein und nicht in Maßen zur Lehrleistung oder Lehrqualität einbezogen werden.

Weiter ist vorab zu klären, ob alle Aspekte gleich wichtig für die Beurteilung der Qualität einer Veranstaltung oder der Lehrleistung eines/ einer Lehrenden sind, und somit gleichbedeutend in die Bewertung einbezogen oder verschieden gewichtet werden.

5 Zur Gestaltung des Gesamtbogens

An dieser Stelle sollen Fragen der Instrumentengestaltung, die nicht einzelne Items oder Dimensionen, sondern den gesamten Fragebogen betreffen, behandelt werden.

5.1 Fragebogenlänge

Die einzige vorliegende Untersuchung zu dem nicht unwichtigen Punkt der Fragebogenlänge wurde von Meredith (1981b) erstellt. Im Anschluß an einen Fragebogen beantworteten 4.877 Studierenden eine gestellte Zusatzfrage nach der bevorzugten Fragebogenlänge. Genannte Antworten variierten zwischen 1-75 Items; 50 Prozent der Studierenden nannten 20-40 Items. Durchschnittlich wurden 25 Items genannt. Meredith schließt daraus: "Evaluationsskalen von 15 bis 35 Items... sind für Studierende akzeptabel" (S. 490).

Um zu beantworten, wie lang ein Fragebogen zur Lehrbewertung sein darf, müssen verschiedene Aspekte beachtet werden. Natürlich ist allein schon nach Testgütekriterien ein längerer Fragebogen vorteilhaft. Man ist jedoch bei der Beantwortung darauf angewiesen, daß Studierende ihn möglichst sorgfältig ausfüllen. In welchem Maße dies geschieht, hängt von der Motivation der Studenten und Studentinnen ab. Und: Motivation ist beeinflußbar!

Für die praktische Anwendung wäre es denkbar, daß am Beginn einer institutionalisierten Evaluation mit sorgfältigen Erklärungen und Einführungen relativ umfangreiche Fragebogen einsetzbar wären. Für regelmäßige Durchführungen müßten sie dann vermutlich gekürzt werden (jedoch so, daß die Multidimensionalität nicht darunter zu leiden hat), um ein Routineinstrument zu erhalten.

5.2 Anzahl der Antwortpunkte (Ratingpunkte)

Auch zur Frage der Zahl der Antwortpunkte (Ratingpunkte) liegt bedauerlicherweise nur eine Studie vor. Lily **Neumann & Neumann** (1981) werteten einen von 500 Studierenden ausgefüllten Fragebogen aus acht geisteswissenschaftlichen Kursen aus. Dieser enthielt Items zur "Zufriedenheit mit verschiedenen Kursaspekten". Die Studierenden beantworteten den Fragebogen in sechs randomisiert ausgesuchten Gruppen. Variiert wurde dabei einzig die Länge der Ratingskala, die 2,3,4,5,7 oder 10 Punkte enthielt.

Das Wissenschaftspaar konnte feststellen, daß alle Studierenden die komplette Skala bei den wenigen Punkten (2/3) nutzten, 80 Prozent dies bei den mittleren (4/5) taten und nur 60 Prozent machten von dieser Möglichkeit bei den beiden langen Skalen (7/10) Gebrauch. Lily Neumann und Neumann kommen zu der Schlußfolgerung, daß die beiden kurzen Skalen zu besseren Ergebnissen führen, jedoch zu wenig

Differenzierungsmöglichkeiten bieten. Die beiden langen Varianten weisen eine Vermeidung der Extreme auf und erscheinen ihnen zu verschwommen. Sie vermuten einige Schwierigkeiten bei der Abgrenzung der Beurteilungen. Sie argumentieren schließlich für eine 5-Punkt-Skala.

Die Anzahl der Punkte auf der Rating-Skala ist auch im Hinblick auf die Bezugsnorm der Lehrbewertung bedeutsam. Würde man mit "Idealnormen" arbeiten, so müßte definiert werden, was die einzelnen Ratingpunkten zu bedeuten hätten: z.B. 1 - optimal, 2 - nur in Einzelpunkten verbesserungsbedürftig, 3 - verbesserungsbedürftig, usw., damit die Befragten nicht den einzelnen Ratingpunkt anders interpretieren. Benutzt man dagegen eine "Sozialnorm", so müßten die Ratingpunkte als Vergleiche formuliert werden, z.B. im Vergleich zu anderen Lehrenden: 1 - weit überdurchschnittlich, 2 - überdurchschnittlich, 3 - durchschnittlich, usw..

Nach der Untersuchung von Lily Neumann & Neumann (1981) ist zu bezweifeln, daß bei langen Skalen die einzelnen Ratingpunkte unterschieden werden können, was aber auch von der Prägnanz der Formulierung abhängig ist. Bei der Sozialnorm ist es dagegen unwahrscheinlich, daß mehr als fünf Ratingpunkte unterscheidbar sind.

Auch das Skalenniveau sollte nicht vergessen werden. Wie solche Instrumente im allgemeinen, verfügen Fragebogen zur Lehrevaluation vermutlich in den seltensten Fällen über Intervallskalenniveau. Auf Basis des Ordinalskalenniveaus ist es zwar möglich zu ermitteln, daß Lehrende z.B. eine weniger gut aufgebaute Kursstruktur haben, aber nicht, um wieviel schlechter diese im Vergleich zu anderen ist. Auch aus diesem Grund erscheint es sinnvoll, die Antwortstufen nicht allzu sehr auszu-differenzieren.

5.3 Instruktion

Lehrbewertungen können mit ganz unterschiedlichen Zielsetzungen durchgeführt werden. Im weiteren wird der Frage nachgegangen, ob der spätere Verwendungszweck, wenn er den Studierenden bekannt ist, ihre Bewertung beeinflußt.

Orpen (1980) untersuchte, ob 326 Studierende der Einführung in die Wirtschaftswissenschaften sich von der Art der Einleitung beeinflussen ließen. Ihnen wurde entweder mitgeteilt, daß der Fragebogen nur Feedbackzwecken dienen sollte oder daß er Auswirkungen auf Beförderung, Festanstellung und Gehalt haben würde. Die Ergebnisse zeigen signifikant bessere Bewertungen (1%-Niveau), wenn die Auswertung der Fragebogen administrativen Zwecken dienen.

Pasen et al. (1978) untersuchten zwei Kurse mit 85 Studierenden, die Kurse der Einführung in der Psychologie bewerteten. Das benutzten Instrument, die “Endeavor Instructional Rating Card“, weist zwei Faktoren auf: "Fach-/ Lehrkenntnis" und "Verhältnis zu den Studierenden". Verglichen wurde die Bewertung zweier Dozenten A und B bei variierter Instruktion (Feedbacknutzen und administrativer Nutzen).

Ergebnis: Ein signifikanter Einfluß konnte nur auf den Faktor "Verhältnis zu den Studierenden" ermittelt werden, lag somit auf der "affektiven Ebene".

Laura A. Driscoll und Goodwin (1979) kamen bei einer Studie an 92 Kursen zu folgendem Ergebnis: Lehrende erhielten signifikant höhere Werte bei den Instruktionen: "Administrative Zwecke" und "Kursverbesserung". Es konnten Anzeichen ermittelt werden, daß von der Instruktion eher kleine und graduierte Kurse betroffen waren.

In der Studie von **Blount et al.** (1978) analysierten die Autoren Daten von 40 Dozenten und Dozentinnen, die mit dem “Illinois Course Evalution Questionaire“ (CEQ) bewertet wurden. Dabei wurden in der Instruktion folgende Nutzungszwecke genannt: Feedback, Administrative Zwecke und Informationsmöglichkeit Studierender in den nächsten Semestern. An die Studierenden wurde randomisiert der CEQ mit den variierten Instruktionen verteilt. Lehrende erhielten signifikant unterschiedliche Werte entsprechend den verschiedenen Instruktionen.

Aleamoni und Pamela Z. Hexner (1980) analysierten zwei vergleichbare Gruppen eines Geschichtskurses beim gleichen Dozenten. In zwei Durchläufen bewerteten ihn 164 bzw. 80 Studierende mit dem Course Evaluation Questionaire (CEQ/ 50 Items). Im ersten Durchlauf wurde der CEQ nur mit einigen Standardformulierungen ausgeteilt, im zweiten Durchlauf wurden die Zwecke des CEQ (Bestimmung von Gehaltserhöhung, Beförderung etc.) zusätzlich verbal erläutert. Es konnte festgestellt werden, daß beim zweiten Durchlauf die Bewertung auf allen sechs Subskalen signifikant höher war.

Insgesamt kann als gut abgesichert gelten, daß die Instruktion, bzw. der dort genannte Verwendungszweck die Lehrbewertung signifikant beeinflußt. Um eine Vergleichbarkeit der Ergebnisse zu ermöglichen, ist die Empfehlung Laura Driscoll und Goodwinn (1979) zu berücksichtigen, die zu einer Standardisierung raten.

5.4 Auftraggeber

Mit dem Problem unterschiedlicher Instruktionen verwandt ist die Frage nach dem Einfluß verschiedener Auftraggeber. Das heißt: Beeinflußt es Studierende, ob eine Lehrevaluation von Kursleitern und -leiterinnen, der Fachschaft oder z.B. der Verwaltung durchgeführt wird?

Bedauerlicherweise gibt es dazu nur eine einzige Untersuchung. **Orpen** (1980) kam zu der Feststellung, daß die Ergebnisse in den Lehrbefragungen deutlich besser sind (1%-Niveau), wenn die Befragung als von Lehrenden initiiert dargestellt wurde, im Unterschied zu einer Einführung, in der das Studierendenparlament als Auftraggeber genannt wurde.

Es ist anzunehmen, daß sich Studierende in den U.S.A. und Deutschland nicht so grundlegend unterscheiden und daher auch hier ähnliche Effekte bei unterschiedlichen Instruktionen möglich sind. Es ist allerdings zu vermuten, daß die Rangreihe der Lehrenden in beiden Fällen die gleiche sein dürfte, falls vergleichbare Dimensionen verwandt wurden.

5.5 Anonymität

Erstaunlich wenige Untersuchungen wurden zur Anonymität von Fragebogen gemacht. In der Umfrage von **Penfield** (1978) wurde ermittelt, daß eine zufriedenstellende Anonymität offensichtlich nicht immer erreicht werden konnte. Nur ein Drittel der Studierenden empfand die Anonymität als meistens gesichert, 54 Prozent waren der Ansicht, daß dies nur manchmal der Fall sei.

Dabei kommt erschwerend hinzu, daß es sich um Lehrbewertungen handelt, bei denen die Namen nicht angegeben wurden, d.h. die Studierenden nur Angst haben mußten, daß der/ die Lehrende z.B. ihre Schrift erkennen könnte (siehe auch Discoll und Godwinn, 1979).

Orpen (1980) konnte jedoch keine Bewertungsunterschiede zwischen namentlich und anonym beantworteten Fragebogen finden.

Schlußfolgerung

Als groben Richtwert für die Fragebogenlänge können nach Meredith (1981b) 15-35 Items gelten. Ansonsten ist die Fragebogenlänge vom Verwendungszweck abhängig. Ein häufig eingesetztes Routineinstrument muß natürlich entsprechend kürzer sein als eine umfassende einmalige Grund-Datensammlung.

Nach Lily Neumann und Neumann (1981) ist eine 5-Punkte Rating-Skala zu empfehlen. Allerdings sind auch Instrumente mit Antwortmöglichkeiten von einer groben 3-Punkte Skala bis zu einer typischen 6-Punkte Skala (entsprechend des Notensystems) oder bis hin zu einer 7-Punkte-Skala verwendbar.

Will man vergleichbare Ergebnisse erreichen, so erscheint es angebracht, standardisierte Instruktionen zu verwenden. Für die Verantwortlichkeit der Studierenden spricht, daß Ergebnisse der Veranstaltungskritik bei schwerwiegenderen Konsequenzen positiver ausfallen. Entsprechend sollte man auch beim Vergleich von Ergebnissen unterschiedlicher Auftraggeber vorsichtig sein.

Viele Faktoren sprechen für eine möglichst anonyme Datenerhebung. Auf die Ergebnisse scheint sich dies jedoch nicht auszuwirken (Orpen, 1980). Diese Diskussion scheint in Deutschland auch schon abgeschlossen noch bevor sie geführt wurde: Es werden nur anonyme Daten verwendet.

6 Vergleichstabellen nach Veranstaltungsart und Fächern

Lehrende sollten vergleichen können, wie sie in der Gegenüberstellung mit ihren Kolleginnen und Kollegen bewertet werden. Will man diesen Vergleich nicht nur am jeweiligen Fachbereich oder an der jeweiligen Hochschule belassen, sind langfristig Vergleichstabellen für das entsprechende Instrument notwendig.

Patricia A. Cranton und Schmith (1986) sprechen sich für ein Normsystem aus, das die unterschiedlichen Kurscharakteristika miteinbezieht. Diese vorgeschlagene unterschiedliche Eichung bezieht sich vor allem auf die bereits erwähnten unterschiedlichen Bewertungen, z.B. für große vs. kleine Klassen, für verschiedene Veranstaltungsformen und für alle weiteren externen Kriterien, die nachweislich einen Einfluß auf die Lehrbewertungen haben.

Jedoch ist fraglich, ob man die sehr unterschiedlichen Fächer einheitlich in Bezug auf die Qualität der Lehre vergleichen kann. Im folgenden wird der Frage nachgegangen, ob die Wichtigkeit unterschiedlicher Lehraspekte über verschiedene Fächer divergiert.

In der Untersuchung "Studenten und Hochschullehrer" des Ministeriums für Wissenschaft und Kunst (Baden-Württemberg) werden in verschiedenen Fächern unterschiedlich eingeschätzt: Wichtigkeit des Sprechstundenbesuchs, Inanspruchnahme und Einschätzung der Nützlichkeit von Studien- und Fachschaftsberatung, unterschiedliche Akzeptanz studentischer Kritik und unterschiedliche Studienmotivationen. Schon allein zahlenmäßig ist das Betreuungsverhältnis unterschiedlich (Gleich et al., 1982).

Ähnliches wird in dem Bericht: "Studiensituation und studentische Orientierung an Universitäten und Fachhochschulen" des Bundesministerium für Bildung und Wissenschaft (1988) deutlich. Studierende verschiedener Fächer unterscheiden sich in Bezug auf Geschlechterverteilung, soziale Herkunft und Berufserfahrung (Peisert et al., 1988), Prozentsatz an Zweitstudierenden, in der Schule belegte Leistungskurskombinationen, d.h. in Bezug auf unterschiedliche schulische Vorerfahrung und Motive der Studienfachwahl.

Außer den Studierendenpopulationen unterscheiden sich auch die Fächer voneinander, bzw. werden als unterschiedlich empfunden in Bezug auf: die unterschiedliche Beteiligung an Forschungsprojekten und unterschiedlichen Praxisbezug bzw. den Wunsch danach, Regelungsdichte, Studieraufwand, empfundenes Klima, Notenstandards, Berufsaussichten, politisches Interesse und alternative oder konventionelle Orientierung. Beim letzteren wurde auch die Auswirkung auf das Studium überprüft: Konventionell Orientierte legen deutlich mehr Wert auf eine hohe soziale Position und ein gutes Einkommen und auf eine leistungsbezogene Studienhaltung.

Obwohl Zahlen einer Erhebung aus den Jahren 1984/ 85 benutzt wurden, kann man davon ausgehen, daß sich in den letzten zehn Jahren die grundlegenden Tendenzen, vor allem in Bezug auf die unterschiedlichen Anforderungen in den Fächern, nicht geändert haben.

Steffani Engler (1991) ermittelte Unterschiede bei Studierenden der Fächer Erziehungswissenschaften, Rechtswissenschaften, Elektrotechnik und Maschinenbau, vor alem bezüglich Zeitbudget und Lebensstil. Hinweise für die Fächerunterschiede sind auch in der Untersuchung von **Andrea Frank** (1991) zu finden.

Marquez, David und Dorfman (1979) hatten bei ihrer Studie eine ähnliche Zielsetzung. Mögliche Differenzen in der Bewertung von Studierenden verschiedener akademischer Disziplinen sollten gefunden werden. Die Studierenden wurden auch nach ihrem Konzept einer `effektiven Lehre´ befragt. Interviewt wurden vierzig Studenten und vierzig Hochschullehrer, gleichermaßen verteilt auf insgesamt vier verschiedene Fakultäten. Von den befragten 80 Personen wurden nur bei acht systematische Variationen gefunden. Diese waren abhängig von der Fakultätszugehörigkeit, die bei der Beschreibung des fiktiven Lehrenden genannt wurde.

Vorliegende Untersuchungen lassen die Vermutung zu, daß Studierende verschiedener Fächer ihre Lehrenden nach unterschiedlichen Ansprüchen und Kriterien bewerten. Leider ist es nicht möglich, eindeutig zu sagen, ob sich die Zugehörigkeit zu einem Fach in der Beurteilung von Lehrenden niederschlägt. Man kann nur vermuten, daß dies bei der studentischen Veranstaltungskritik einen möglicher Einflußfaktor darstellt, obwohl die Studie von Marquez, David und Dorfman (1979) dagegen spricht.

Schlußfolgerung

Insgesamt scheinen in diesem Bereich möglicher systematischer Einflußfaktoren auf die Lehrbewertung und deren Berücksichtigung in „Vergleichstabellen“ weitere Untersuchungen erforderlich. Es kann für mögliche Vergleichstabellen jedoch festgestellt werden, daß sie vermutlich für verschiedene Veranstaltungsarten, u.U. auch für verschiedene Semester oder Kursgrößen, unterschiedlich sein müßten. Kursarten- und Fächerunterschiede könnten ohne größere Bedenken mit Hilfe unterschiedlicher Itemzusammenstellung (Cafeteria-Prinzip) aufgefangen werden. Ob Vergleichstabellen dennoch nach Fächern angelegt werden sollten, muß noch genauer geprüft werden.

Bisher scheint nur von der Projektgruppe um Prof. Diehl in Gießen (siehe Diehl in diesem Heft, S. 36f.) geplant, eine Datenbank aufzubauen. Diese soll sich aus den Evaluationsdaten speisen, die mit den von ihm entwickelten Instrumenten (Fragebogen zur Beurteilung von Seminaren mit Referaten [VBREF] und Fragebogen zur Beurteilung von Vorlesungen [VBVOR]) gewonnen werden.

7 Ergebnisdarstellung und ihre Vorbedingungen

Einige Vorbedingungen, wie z.B. reliable und valide Fragebogen, wurden schon erwähnt. Um aber verwertbare Ergebnisse darstellen zu können, müssen zusätzlich gewisse Standards der Beteiligung und des Rücklaufs beachtet werden.

Dabei sind vor allem zwei Prozentzahlen wichtig: Zum einen der Anteil Studierender, der im Laufe des Semesters den entsprechenden Kurs verließ, die sog. *Drop-out-Quote*; zum anderen der Anteil Studierender, der den Fragebogen ausgefüllt zurückgegeben hat, nachdem er im Kurs verteilt wurde, die sog. *Rücklaufquote.* Außer diesen Vorbedingungen sollen an dieser Stelle außerdem Kriterien der Ergebnisdarstellung genannt und die Frequenz der Datenerhebung besprochen werden.

7.1 Abnehmen der Teilnehmerzahl (Drop-out-Quote)

Im Vergleich zur Rücklaufquote gibt es zu diesem Thema keinerlei Untersuchungen im Bereich der Lehrevalution. Es konnten nicht einmal Überlegungen oder Ansätze zu Hypothesen gefunden werden. Daher sollen im weiteren nur Fragen, die auch bei allen Längsschnittstudien eine Rolle spielen, aufgeworfen und keine Konklusionen vorgestellt werden.

Das Phänomen des „Kursschwundes“ ist allen Mitgliedern und Beobachtenden von Hochschulen bekannt und in nahezu allen Fachbereichen zu finden: Am Beginn des Semesters sind die Veranstaltungen oft brechend voll, je weiter das Semester voranschreitet, um so weniger Teilnehmer und Teilnehmerinnen sind in den Kursen zu finden. Ohne sich allgemein mit den Gründen zu beschäftigen, die zu diesem Phänomen führen, stellt sich die Frage, ob sich dies auf die Qualität der Lehrbewertung auswirkt: Sind die Bewertungen der Studierenden, die nach und nach aus dem Kurs "abwandern" mit denen, die bleiben, vergleichbar?

Wenn es Unterschiede gibt, sind bestimmte Gesetzmäßigkeiten zu finden? Sind diese Gesetzmäßigkeiten auf die Qualität der Lehre zu beziehen? Sind die gefundenen Muster selbst schon Ausdruck von Mängeln der Lehrqualität? Findet man Unterschiede zwischen den "Anwesenden" und den "Dropouts", wie soll man damit umgehen?

Es wäre nützlich, entweder vorhandene Erhebungen zur Lehrevaluation (soweit möglich) heranzuziehen oder gesonderte Untersuchungen vorzunehmen, um diese Fragen empirisch fundiert beantworten zu können.

7.2 Beteiligung an der Befragung (Rücklaufquote)

Die Rücklaufquote ist ein Merkmal, das in den einzelnen Studien häufiger genannt wird. Dennoch liegt nur eine einzige Untersuchung zur Auswirkung unterschied-licher Rücklaufquoten vor.

Rindermann (in diesem Heft, S. 46) konnte eine relativ hohe Genauigkeit und Stabilität von Lehrbewertungen bereits bei einer Mindestzahl von zehn Studierenden ermitteln.

McBean und Lennoy (1985) berichten von großen Anstrengungen im Herbstsemester 1982 an der Universität von Waterloo, eine überdurchschnittlich hohe Rücklaufquote zu erzielen. Untersucht wurde, welche Auswirkung unterschiedliche Rücklaufquoten für die Ergebnisse haben. Randomisierte Untergruppen wurden gebildet, die 80 Prozent, 68 Prozent, 50 Prozent und 25 Prozent Rücklauf simulierten.

Die wichtigsten Ergebnisse waren: Ist die Stichprobe viel kleiner (geringer als 25 Prozent) als die Ursprungsgruppe, so besteht Potential für eine erhebliche Datenungenauigkeit, jedoch nur in Gruppen, die weniger als 30 Mitglieder haben. Ist die Gruppe dagegen größer als 30, so ist der Fehler bis zu einer Rücklaufquote von 50 Prozent nicht sehr groß. Die Autoren schließen daraus, daß in großen Gruppen (>30) eine 50 Prozent Antwortrate eine genügend präzise Auskunft der Lehrbewertung gibt, während in kleinen Gruppen (<30) ungefähr 80 Prozent ihre Bögen abgeben müssen, um den gleichen Vertrauensgrad zu erreichen.

Problematisch an diesen Feststellung ist die Ausgangsstichprobe von randomisierten Subgruppen. Es müßte untersucht werden, ob die Ergebnisse von Studierenden, die Lehrbeurteilungen abgeben, anders ausfallen, als die von Studierenden, die sie nicht abgeben.

7.3 Ergebnisdarstellung: Profile und Verteilungen

Aus der Forderung nach multidimensional gestalteten Fragebogen zur Lehrevaluation ergibt sich die Konsequenz einer differenzierten Darstellung unterschiedlicher Dimensionswerte. Diese einzelnen Werte sollten in keinem Fall zu einem bloßen Gesamtwert zusammengezogen werden, der für einen Vergleich herangezogen werden würde. Bei einem solchen Verfahren bleibt nicht nur eine Unmenge an Informationen ungenutzt und gleiche Gesamtwerte können sich aus völlig unterschiedlichen Einzelergebnissen zusammensetzten, sondern es ist auch methodisch unzulässig, Dimensionen, die möglichst niedrig miteinander korrelieren, d.h. verschiedene Sachverhalte erfassen, einfach zu summieren.

Die Ergebnisse der Lehrbewertung sollten nach ihren einzelnen Dimensionen getrennt aufgeführt und als *Profile* dargestellt werden. Die einzelnen Markierungen zwischen den Extrempolen abgetragen, werden miteinander verbunden und zeigen anschaulich positive und negative “Ausschläge“ in der Lehrevaluation einzelner Lehrender.

Außer methodischen Gründen sprechen noch andere Gründe praktischer Art für diese Form der Darstellung. Erstens muß dem Lehrenden oder der Lehrenden deutlich werden, wo genau ihre “Schwächen“ liegen, um diese verbessern zu können. Zweitens

erscheint es für den einzelnen/ die einzelnen motivierender, anstelle eines möglichen schlechten Gesamtergebnisses ein Profil zu erhalten, das vermutlich nur in den allerseltensten Fällen keine positiven `Ausschläge´ enthält und damit auch auf "Stärken" verweist.

Der Zweck der Datenerhebung hat nicht nur Auswirkungen auf die Wahl eines Instruments zur Lehrbewertung, sondern auch auf die Art der Ergebnisdarstellung. Wird angestrebt, Lehrenden eine möglichst detaillierte Rückmeldung über ihre Veranstaltung zu geben, so sind differenziertere und individuellere Auswertungen notwendig, als wenn ein allgemeiner Vergleich über Lehrende und Lehrveranstaltungen angestrebt wird.

Schlußfolgerung

Es wäre an der Zeit, eine Untersuchungen unter Studierenden durchzuführen, die jene Gründe und Motive herauszufinden versucht, die dazu führen, eine Veranstaltung nicht weiter zu besuchen. Zeitmangel, verschobene Prioritätensetzung, Kurs wurde kurzfristig aus Neugier beleg; viele Antworten scheinen denkbar. Darunter auch Urteile, die u.U. dazu führen müßten einen Kurs neu zu gestalten, weil sie sich auf Mängel des Kurses beziehen.

Bei der Auswertung von Fragebogen sollte eine akzeptable Rücklaufquote vorliegen. Dies wären nach Rindermann (in diesem Heft, S. 49) mindestens 10 Studierende. Man kann sich aber auch an den Ergebnissen von McBean und Lennoy (1985) orientieren, die bei großen Kursen (>30) eine notwendige Rücklaufquote von 50 Prozent ermittelten, um keine Ergebnissverzerrung zu riskieren, bei kleinen Kursen (<30) sollten dagegen eine Rücklaufquote von 80 Prozent erreicht werden.

Nachdem ein sorgsam konstruiertes Instrument zur Datenerhebung verwandt wurde, sollte auf die Ergebnissdarstellung ähnlich viel Wert gelegt werden. Nur eine differenzierte Präsentation (z.B in Profilform und mit Verteilungen) kann zu einer optimalen Nutzung der Daten führen.

8 Zeitpunkt und Frequenz der Beurteilung

Der Zeitpunkt und die Häufigkeit studentischer Lehrbeurteilungen sind maßgeblich durch den Zweck, dem sie dienen sollen, zu bestimmen. Über „optimale“ Zeitpunkte und Frequenzen liegen zwar keine Untersuchungen vor, aber einige Überlegungen und Anhaltspunkte zu dieser Frage sollen vorgelegt werden.

8.1 Zeitpunkt der Befragung

Bevor nicht geklärt ist, ob Studierende, die bis zum Kursende im Unterricht verbleiben, gleiche Bewertungsmuster wie Kursabrecher und -abrecherinnen abgeben, erscheint es unzureichend, nur am Ende eines Kurses Daten zu erheben. Es sollte mindestens ein weiterer Erhebungszeitpunkt feststehen.

Um einen günstigen Zeitpunkt zu finden, müßte geklärt werden, ab welcher Semesterwoche sich Studierende ein Bild von den Lehrleistungen ihres Dozenten/ihrer Dozentin bilden können und ob an diesem Zeitpunkt die Anwesensheitsquote noch hoch genug ist.

8.2 Frequenz der Befragung

Ein weiterer Grund für Beurteilungen in der Semestermitte sind mögliche Verbesserungen für den weiteren Kursverlauf, da der Lehrende/die Lehrende das Feedback nutzen kann, um Veränderungen noch vorzunehmen.

Sturm (1991) unterscheidet dabei zwischen Prozeß- und Gesamtevaluation. Bei der Prozeßevaluation für Feedbackzwecke kann eine möglichst hohe Beurteilungsfrequenz sinnvoll sein, die z.B. auch durch am Ende einer Stunde abzugebende Kurzkommentare (siehe Ritter, 1992) zu erreichen ist. Für eine Gesamtevaluation, z.B. für administrative Zwecke, können jährliche bzw. zweijährliche Bewertungen, die sich grundsätzlich aus mehreren Kursen zusammensetzen sollten, ausreichen, da davon auszugehen ist, daß Lehrfähigkeit eine relativ stabile Fertigkeit darstellt.

Betrachtet man die Gesamtevaluation über einen gewissen Zeitraum für einen Kurs, so ergibt sich wiederum eine Prozeßevaluation, die als Grundlage zu einer Neustrukturierung dienen kann.

Schlußfolgerung

Insgesamt ist die vorzusehende Beurteilungsfrequenz letztlich vom Zweck der Veranstaltungskritik her zu bestimmen. So wäre es sinnvoll, bei neu konzipierten Kursen oder Veranstaltungen, die ein Lehrender/ eine Lehrende zum ersten mal anbietet, bereits während des Kursverlaufs ein Feedback zu ermöglichen.

Diese Form von Feedback könnte z.B. im Sinne einer Prozeßevaluation Instrumente verwenden, wie z.B. die Kurzkommentare, die Ritter (1992) vorschlägt. Bei der Einführung einer solchen Prozeßevaluation ist sowohl die Kapazität der Hochschule und ihre Änderungsbereitschaft als auch die Motivation der Studierenden zu beachten.

Bei bereits etablierten Kursen und erfahrenen Lehrenden erscheint eine abschließende Gesamtevaluation am Ende jedes Kurses ausreichend. Diese Rückmeldung sollte zwar relativ kurz und leicht auszuwerten sein, aber doch so gestaltet werden, daß die Lehrenden entsprechende Umgestaltungen ihrer Kurse vornehmen können.

Weiter sollte eine umfassendere Erhebung alle zwei bis drei Jahre erfolgen, die dann zusätzlich Elemente der Lehrorganisation umfassen müßte. Dies kann z.B. in Form von Lehrberichten geschehen und in diesem Rahmen veröffentlicht werden.

9 Zusammenfassende Übersicht: Empfehlungen

Die vorgestellten Punkte sind gedacht als Ansatz zur Formulierung verbindlicher Kriterien für die Erstellung und Verwendung von Fragebogen der studentischen Veranstaltungskritik. Gerade im Bereich der Itemformulierung, der möglicherweise unterschiedlichen Vergleichstabellen und der Frequenz von Datenerhebungen sollten allerdings weitere Untersuchungen durchgeführt bzw. in der Praxis erprobt werden, um die Erkenntnislücken zu schließen. Dennoch lassen sich aufgrund empirischer Untersuchungen und praktischer Erprobungen eine Reihe von Empfehlungen zusammenstellen, an denen sich die Entwicklung, der Einsatz und die Verwendung von Instrumenten zur Bewertung der Lehre richten sollten.

Die aufgeführten Kriterien beziehen sich auf Fragebogen zur Bewertung von Veranstaltungen und Veranstaltungsreihen, die so konstruiert sein sollen, daß sie einen Vergleich zwischen den Lehrenden ermöglichen. Für eine direkte Rückmeldung an den Veranstaltungsleiter/ die Veranstaltungsleiterin am Ende eines Kurses reichen weitaus einfachere Instrumente aus.

Soll neben der Bewertung einzelner Veranstaltungen und Lehrenden auch eine Beurteilung der Lehre im gesamten Fachbereich vorgenommen werden, so sind zusätzliche Instrumente bzw. zusätzliche Fragen nötig, die stärker auf die Beurteilung der Lehrorganisation fokussieren. Vor der Erhebung von Daten sollte daher die Nutzung der Datenverwendung feststehen. Erst nachdem eine Entscheidung über die Richtung und den Zweck der Lehrevaluation getroffen ist, sollte die Festlegung auf ein Instrument oder eine Kombination verschiedener Verfahren getroffen werden.

Abgrenzung verschiedener Dimensionen

⇨ Fragebogen sind mehrdimensional zu konstruieren. Jede Dimension sollte von mehreren Items erfaßt werden.

⇨ Wichtige Dimensionen, die nicht ausgelassen werden sollten: Zuwendung, Fairneß von Prüfungen und Benotungen, Kommunikationsfähigkeit, Kurs- bzw. Stofforganisation, Anregung, Variabilität vs. Monotonie, Enthusiasmus und Kurswert.

⇨ Zusätzliche Aspekte erheben, die Lehrende oder Studierende interessieren (z.B. Kursschwierigkeit) oder für weitere wissenschaftliche Untersuchungen interessant sind (z.B. Geschlecht).

⇨ Faktoren erfassen, die einen möglichen Einfluß auf die Lehrbewertung haben, um sie in der Darstellung kontrollieren bzw. ausweisen zu können.

Formulierung der Fragen und Items

⇨ Möglichst von Beobachtbaren ausgehen. Soweit wie möglich konkrete Items verwenden und globale vermeiden.

⇨ Vergleichsmaßstab auch bei der Itemformulierung beachten. Die optimale Ausprägung eines Lehraspektes sollte am Anfang oder Ende der Ratingskala liegen.

⇨ In die Wertung nur Merkmale einbeziehen, für die Lehrende verantwortlich sind.

⇨ Vorab entscheiden, ob die unterschiedlichen Items unterschiedlich gewichtet werden sollen

Gesamtbogen

⇨ Itemzahl von Verwendungszweck abhängig. 15-35 Items empfehlenswert.

⇨ Bei der Beurteilungsskala 5er Punktsystem sinnvoll, aber auch 3-7 Punkte sind je nach Zweck vertretbar.

Anweisung (Intstruktion) und Durchführung

⇨ Anweisung (Instruktion) standardisieren, sowohl hinsichtlich der Zwecksetzung als auch des Bewertungsbezuges (Anker).

⇨ Studentische Veranstaltungskritiken anonym durchführen.

Auswertung

⇨ Aspekte, die außerhalb der Lehrdimensionen liegen (z.B. Kursschwierigkeit), sollten nicht in die Lehrbewertung eingehen.

⇨ Faktoren, die einen Einfluß auf die Lehrbewertung haben, müssen entsprechend konkretisiert (ausgewiesen oder verrechnet) werden. Nach dem heutigen Kenntnisstand wäre dies hauptsächlich der Faktor „Motivation und Interesse am Kursthema“.

Vergleichstabellen

⇨ Vergleichstabellen existieren noch nicht. Ein erster Ansatz ist von Diehl (Gießen) geplant.

⇨ Bei der Erstellung ist es sinnvoll, nach verschiedenen Veranstaltungsarten, u.U. auch nach der Semesterzahl der Studierenden, nach Kursgröße, vielleich auch nach Fächern zu unterteilen.

Ergebnisdarstellung

⇨ Drop-out-Quote der Veranstaltung beachten.

⇨ Eine grobe Richtlinie für den Rücklauf stellen mind. 10 Studierende dar. Ansonsten sollte die Rücklaufquote der ausgegebenen Bogen in kleinen Kursen (< 30) mindestens 80 Prozent betragen, in großen Kursen (> 30) mindestens 50 Prozent.

⇨ Daten nicht als simple Gesamtnoten darstellen, sondern differenziert (z.B. als Profile) ausweisen.

⇨ Nicht nur Mittelwerte, sondern auch Streuungen bzw. Verteilungen miteinbeziehen.

Frequenz der Beurteilung

⇨ Bei neuen Kursen/ Lehrenden sollten *Kurz*rückmeldungen bereits während des Kursverlaufs erfolgen (Prozessevaluation).

⇨ Bei etablierten Veranstaltungen ist eine Gesamtevaluation am Ende bzw. in der Mitte der Veranstaltung ausreichend.

⇨ Alle 2-3 Jahre sollte eine ausführliche Evaluation, die auch die Lehrorganisation beinhaltet, durchgeführt werden (z.B. in Form einen Lehrberichts).

Neben den genannten relativ konkreten Kriterien, die sich direkt auf die Entwicklung von Fragebogen der studentischen Lehrveranstaltungskritik beziehen, gibt es einige übergreifende Punkte, die hier nur kurz erwähnt werden sollen. So ist es für die Evaluation der Lehrqualität an einer Hochschule auf Dauer sinnvoll, einen Pool von Items zu sammeln, aus dem je nach Bedarf für einzelne Lehrveranstaltungen ein spezifischer Bogen für dessen Beurteilung zusammengestellt werden kann.

Diese Empfehlung liegt der Gedanke des Cafeteria-Prinzips zugrunde. Dabei werden für die Lehrbewertung einem standardisierten Kern, der für die Beurteilung aller Lehrveranstaltungen verwendet wird, fach- oder kursspezifische Fragen hinzugefügt. Dies hat den Vorteil, daß man sowohl ein erprobtes als auch ein flexibles Instrument zu Verfügung hat. Des weiteren werden Verfahren eher akzeptiert, wenn die Lehrenden sie selbst wenigstens zum Teil den Bedürfnissen ihrer Veranstaltung und ihrem Informationsbedarf anpassen können.

Das Thema Akzeptanz der Lehrevaluation durch die Lehrenden spielt auch bei der Art und Weise der Datenauswertung und ihrer Präsentation eine Rolle. Der vorliegende Bericht geht vor allem auf die Konstruktionskriterien von Instrumenten studentischer Veranstaltungskritik ein. Dabei werden auch Punkte der Auswertung genannt. Nicht näher wird jedoch der Bereich der Umsetzung und Durchführung von Projekten der Lehrevaluation beschrieben. Aus den bisher gewonnen Erfahrungen erscheint es sinnvoll bereits vor der Erhebung eine gewisse Akzeptanz der Verfahren zu erreichen. Noch wichtiger ist allerdings zu klären, welchem Zweck die Daten dienen, wie sie ausgewertet und veröffentlicht werden sollen.

Die Beurteilung von Lehrveranstaltungen kann auch Elemente einer Personalbeurteilung haben oder zumindest so aufgefaßt werden. Damit ist sie ein sensibles Instrument, mit dem in einer gewissen Rücksichtnahme verfahren werden sollte. Rücksicht gegenüber den Beurteilten und ihrem Bedürfnis nach Fairness, manchmal auch nach Diskretion, aber auch mit Rücksicht auf die Studierenden und deren Bedürfnis nach Information und Transparenz, aber auch nach Veränderung.

10 Instrumente studentischer Veranstaltungskritik

Der Empfehlung folgend, sich bei der Entwicklung von Fragebogen zur Lehrbewertung an methodisch geprüften und praktisch erprobten Instrumenten zu orientieren bzw. diese einzusetzen, werden im folgenden einige vorgestellt. Wissenschaftler, die durch entsprechende Veröffentlichungen auf sich aufmerksam gemacht haben, wurden gebeten, das von ihnen entwickelte Verfahren selbst vorzustellen.

Die Gliederung für die Darstellung der Fragebogen sollte mit einer kurzen Beschreibung des Instruments beginnen. Dabei soll deutlich werden, um welche Art von Verfahren es sich handelt und wie es konstruiert wurde. Weiter interessieren die zu messenden Dimensionen und die Zielgruppe, für die die Verwendung des Instruments gedacht ist. Außerdem wurden die Autoren um Hinweise zur Durchführung und Auswertung der Fragebogen gebeten; danach wurde die Frage nach der Form der Ergebnisdarstellung gestellt und gefragt, ob daran gedacht ist, Vergleichstabellen zu bilden.

Um die Qualität eines Fragebogens beurteilen zu können, sind neben der Art der Fragebogenerstellung auch die Kennzahlen der Gütekriterien aussagekräftig. Diese können mit einer ganzen Reihe mehr oder weniger aufwendiger Verfahren ermittelt werden. Es wurden nicht bei allen Verfahren die entsprechenden Gütekriterin ermittelt. Diese Art der Qualitätsbestimmung wird auch erst dann wirklich relevant, wenn es beim Vergleich zwischen Lehrenden um tatsächliche Auswirkungen gehen wird.

Die Wissenschaftler sollten mögliche angestrebte Weiterentwicklungen der von ihnen konstruierten Instrumente erläutern und gegebenenfalls besondere Hinweise zur Verwendung ihres Verfahrens mitteilen.

Schließlich wird den Lesern und Leserinnen weitere Literatur zu dem jeweiligen Instrument genannt und die Kontaktadresse für den Bezug des entprechenden Instruments mitgeteilt.

Die Instrumente werden in alphabetischer Reihenfolge ihrer Konstrukteure aufgeführt. Abschließend werden noch kurz die Verfahren des Projektes „Lehrevaluation“ in Mannheim dargelegt.

Überblick über die Instrumente

10.1 Fragebogen zur Beurteilung von Vorlesungen (VBVOR) von J. M. Diehl

I. Kurze Beschreibung des Instruments

Der VBVOR ist ein Fragebogen zur studentischen Beurteilung von Hochschulvereranstaltungen, in denen die Vermittlung des Stoffes ausschließlich durch den Dozenten bzw. die Dozentin erfolgt (Veranstaltungen vom Typ Vorlesung). Das aus 16 Items sowie Zusatzfragen bestehende Instrument stellt eine neukonstruierte Kurzform des von Diehl & Kohr (1977) entwickelten VBPSYCH (40 Items) dar. Der ursprüngliche Itempool wurde aus einer Sammlung (freier) studentischer Aussagen zu Vorlesungen im Fach Psychologie gewonnen. Durch Faktorenanalysen und nachfolgende Itemanalysen enstanden dann die Lang- bzw. Kurzform.

Mit den vier Skalen zu je vier Items (und vierstufigem Antwortformat) werden folgende Veranstaltungsaspekte erfaßt:

[1] Relevanz und Nützlichkeit der Veranstaltungsinhalte
[2] Verhalten des Dozenten (der Dozentin) gegenüber den Veranstaltungsteilnehmern und -teilnehmerinnen.
[3] Angemessenheit von Schwierigkeit und Umfang der Veranstaltungsinhalte
[4] Methodik und Aufbau der Veranstaltung.

Vom Fragebogen existiert eine männliche und eine weibliche Form, bei denen vom Dozenten bzw. der Dozentin gesprochen wird.

Der Fragebogen wird bereits an mehreren Psychologie-Fachbereichen der Bundesrepublik eingesetzt (teilweise von Seiten der Fachschaften). Er hat sich gleichermaßen in anderen Fachrichtungen bewährt (z.B. Wirtschaftswissenschaften, Pädagogik, Medizin).

II. Durchführung und Auswertung

Die Beantwortung der 16 Items sowie der vorgegeben Zusatzfragen dauert 5-10 Minuten. Für die Eingabe und Auswertung des Daten existiert ein DOS-Programm, das aufgrund seiner einfachen Bedienung und Eingabemasken auch von EDV-Ungeübten schnell gehandhabt werden kann. Nach Einarbeitung liegt die Eingabedauer für einen Fragebogen bei 1-2 Minuten.

III. Ergebnisdarstellung

Nach Dateneingabe liefert das Programm unmittelbar eine achtseitige Auswertung der Veranstaltung, die folgendes enthält:

- Itemtexte (skalenweise geordnet) mit Antwortverteilungen
- Durchschnittliche Skalenwerte
- Verteilung der Skalenwerte (grafisch)
- Interkorrelation der Skalen
- Verteilung der Antworten bei den Zusatzfragen.

Die Daten einer Veranstaltung werden zugleich als ASCII-File abgespeichert und können dadurch mit Statistik-Programmen weiter analysiert werden. Das Anlegen einer Datenbank mit Veranstaltungen erfolgte bisher (aus Zeitgründen) nicht.

IV. Gütekriterien

Die Skalen besitzen faktorielle Gültigkeit. Die Item-Trennschärfen sowie die innere Konsistenz der Skalen können als gut bezeichnet werden. Die Werte von Cronbachs alpha sind im einzelnen: Skala 1 (.84), Skala 2 (.81), Skala 3 (.81) und Skala 4 (.89).

V. Angestrebte Weiterentwicklung

Es ist geplant, das Auswertungsprogramm derart zu erweitern, daß es auch benutzer-formulierte Zusatzfragen analysieren kann.

VI. Besondere Hinweise zur Verwendung des Verfahrens

Manual und Fragebogenexemplare sind als ASCII-Dateien auf einer Diskette erhältlich, die auch das Auswertungsprogramm enthält. Diese Diskette darf beliebig kopiert und weitergegeben werden. Anwender(innen) des VBVOR werden gebeten, sich beim Autor registrieren zu lassen. Dies soll die Sammlung von Evaluationsdaten für eine (spätere) VBVOR-Datenbank erleichtern.

10.2 Fragebogen zur Beurteilung von Seminaren mit Referaten (VBREF) von J. M. Diehl

I. Kurze Beschreibung des Instruments

Der VBREF ist eine Fragebogen zur studentischen Beurteilung von Hochschulveranstaltungen, in denen die Stoffvermittlung (überwiegend) durch studentische Referate erfolgt. Das aus 30 Items sowie Zusatzfragen bestehende Instrument stellt eine noch nicht veröffentlichte Neukkonstruktion dar. Der ursprüngliche Itempool wurde aus einer Sammlung (freier) studentischer Aussagen zu Referatveranstaltungen im Fach Psychologie gewonnen. Durch Faktorenanalysen und nach-folgende Itemanalysen enstand anschließend die endgültige Form des Fragebogens.

Mit den sechs Skalen zu je fünf Items (und vierstufigem Antwortformat) werden folgende Veranstaltungsaspekte erfaßt:

[1] Qualität der Referatsvorträge
[2] Verhalten des Dozenten (der Dozentin) gegenüber den Veranstaltungsteilnehmern und -teilnehmerinnen.
[3] Angemessenheit von Schwierigkeit und Umfang der Veranstaltungsinhalte
[4] Relevanz und Nützlichkeit der Veranstaltungsinhalte
[5] Umfang der Frage- und Diskussionsmöglichkeiten
[6] Auswahl und Zusammenhang der Referatsthemen

Vom Fragebogen existiert eine männliche und eine weibliche Form, bei denen vom Veranstaltungsleiter bzw. der Veranstaltungsleiterin gesprochen wird. Der Fragebogen wird bereits an mehreren Psychologie-Fachbereichen der Bundesrepublik eingesetzt (teilweise von Seiten der Fachschaften). Er hat sich gleichermaßen in anderen Fachrichtungen bewährt. (z.B. Ernährungswissenschaften, Pädagogik).

II. Durchführung und Auswertung

Die Beantwortung der 30 Items sowie der vorgegeben Zusatzfragen dauert etwa 10 Minuten. Für die Eingabe und Auswertung der Daten existiert ein DOS-Programm, das aufgrund seiner einfachen Bedienung und Eingabemasken auch von EDV-Ungeübten schnell gehandhabt werden kann. Nach Einarbeitung liegt die Eingabedauer für einen Fragebogen bei etwa zwei Minuten.

III. Ergebnisdarstellung

Nach Dateneingabe liefert das Programm unmittelbar eine elfseitige Auswertung der Veranstaltung, die folgendes enthält:

- Itemtexte (skalenweise geordnet) mit Antwortverteilungen,
- Durchschnittliche Skalenwerte,
- Verteilung der Skalenwerte (grafisch),
- Interkorrelation der Skalen,
- Verteilung der Antworten bei den Zusatzfragen.

Die Daten einer Verstaltung werden zugleich als ASCII-File abgespeichert und können dadurch mit Statistik-Programmen weiter analysiert werden. Das Anlegen einer Datenbank mit Veranstaltungen erfolgte bisher (aus Zeitgründen) nicht.

IV. Gütekriterien

Die Skalen besitzen faktorielle Gültigkeit. Die Item-Trennschärfen sowie die innere Konsistenz der Skalen können als gut bezeichnet werden. Die Werte von Cronbachs alpha sind im einzelnen: Skala 1 (.84), Skala 2 (.81), Skala 3 (.88), Skala 4 (.74), Skala 5 (.85) und Skala 6 (.79).

V. Angestrebte Weiterentwicklung

Es ist geplant, das Auswertungsprogramm derart zu erweitern, daß es auch benutzer-formulierte Zusatzfragen analysieren kann.

VI. Besondere Hinweise zur Verwendung des Verfahrens

Manual und Fragebogenexemplare sind als ASCII-Dateien auf einer Diskette erhältlich, die auch das Auswertungsprogramm enthält. Diese Diskette darf beliebig kopiert und weitergegeben werden. Anwender(innen) des VBVOR werden gebeten, sich beim Autor registrieren zu lassen. Dies soll die spätere Sammlung von Evaluationsdaten für eine VBREF-Datenbank erleichtern.

Literatur und Kontaktadresse

Diehl, J.M. & Kohr, H.U. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. Psychologie in Erziehung und Unterricht, 24, 61-75.

Diehl, J.M. (1983). Untersuchungen zur Einstellung von Psychologiestudenten gegenüber Statistik und Statistik-Veranstaltungen. Berichte aus dem Fachbereich 06 Psychologie, Universität Gießen.

Diehl, J.M. (1994). Fragebogen zur studentischen Evaluation von Hochschulveranstaltungen. Manual zum VBVOR und VBREF. Fachbereich 06 Psychologie (Abteilung Methodik), Universität Gießen.

Prof. Dr. Joerg M. Diehl
Justus-Liebig-Universität Gießen
Fachbereich 06 Psychologie
Otto-Behaghel-Str. 10
D-35394 Gießen
Tel (0641) 702-5448 Fax (0641) 702-3811
Email diehl@psychol.uni-giessen.de

10.3 Evaluation der Lehre an der Ruhr-Universität Bochum: Fragebogen für Vorlesungen (I. Studierendenbogen, II.Dozent(inn)enbogen) von H. Kromrey

Universitätskommission für Lehre, Studium und Studienreform, Wissenschaftliche Leitung: Prof. Dr. Helmut Kromrey

I. Kurze Beschreibung des Instruments

Befragt wurden alle Hörer/innen in Vorlesungen; erhoben wurde in einer Stichtagswoche ("Normalwoche" in der Mitte des Semesters) in allen Vorlesungen einer Fakultät. Parallel füllte jede Lehrperson einen Dozent(inn)enFragebogen aus.

Der Studierendenbogen enthält Fragen zur Stellung der/des Studierenden im Studiengang, zu den Teilnahmegründen sowie zu den Erwartungen an die Vorlesung, zu studentischen Beurteilungen über Didaktik, Stoffmenge und Schwierigkeitsgrad, über Vorkenntnisse und Selbsteinschätzung des Lernerfolgs, den eigenen Arbeitsaufwand, Rahmenbedingungen der Vorlesung sowie das Auftreten und Verhalten der Lehrperson. Im Dozent(inn)enbogen sind ähnliche Themenbereiche aus der Perspektive der Lehrperson einzuschätzen.

Die Items wurden anhand der Durchsicht hochschuldidaktischer Literatur sowie bereits vorliegender Fragebogen zusammengestellt; Bewertungen werden auf einer 5-Punkte-Ratingskala abgefragt. Das endgültig eingesetzte Instrument ist das Ergebnis mehrerer Pretest-Durchgänge: Studierende unterschiedlicher Fakultäten wurden gebeten, den Bogen auszufüllen; anschließend wurde in einem mündlichen Nach-Interview Frage für Frage die Semantik aus dem Verständnis des Ausfüllenden sowie das Ausfüll-Verhalten erfaßt. Bei Verständnis-Differenzen wurden die Befragten um (für sie eindeutigere) Alternativ-Formulierungen gebeten.

Neben dem Vorlesungsbogen existieren spezielle Instrumente für Seminare sowie für Veranstaltungen mit Übungscharakter. Sie unterscheiden sich vom Vorlesungsbogen insbesondere im didaktischen Teil sowie bei den Fragen nach dem studentischen Teilnahmeverhalten.

II. Hinweise zur Durchführung und Auswertung

Der Studierendenbogen ist im Einvernehmen mit der jeweiligen Lehrperson ca. 20-30 Minuten vor Schluß der Vorlesung von einer studentischen Hilfskraft zu verteilen, von den Teilnehmern unmittelbar auszufüllen und anschließend von der Hilfskraft wieder einzusammeln. Die Lehrperson verläßt die Veranstaltung und füllt den eigenen Bogen in ihrem Büro aus. Dadurch soll die Anonymität der Befragung nicht nur gesichert, sondern auch sichtbar gemacht werden. Für die Gültigkeit der Befragungsergebnisse ist

es unabdingbar, daß möglichst alle Teilnehmer den Bogen ausfüllen; ein Austeilen der Bögen, verbunden mit dem Appell, ihn ausgefüllt in der nächsten Woche wieder mitzubringen oder ihn in einem "Evaluationsbüro" abzugeben, führt zu extrem selektivem Rücklauf und inhaltlich verzerrten Resultaten.

Ausgewertet wird der Studierendenbogen a) je Vorlesung als Rückmeldung an die jeweilige Lehrperson sowie b) kumuliert über alle Vorlesungen als zusammenfassende Information für die Fakultät. Der Dozent(inn)enbogen kann nur fakultätsweise ausgewertet werden.

III. Ergebnisdarstellung

Die Daten werden schnellstmöglichst erfaßt, so daß eine Häufigkeitsauszählung noch vor Ende des Semesters der jeweiligen Lehrperson zur Verfügung gestellt werden kann. Die Auszählung orientiert sich an den Kategorien des Fragebogens und enthält neben den (absoluten und relativen) Häufigkeiten der betreffenden Veranstaltung als Vergleich die kumulierten Werte der Fakultät. Da sich gezeigt hat, daß die studentischen Urteile innerhalb der Vorlesungen sehr heterogen ausfallen, wird auf die Berechnung von Durchschnittswerten verzichtet, d.h. es wird die informationsreichere Dokumentation der Verteilung der Antworthäufigkeiten beibehalten.

IV. Gütekriterien

In abschließenden Auswertungen wurde der Einfluß von Bias-Variablen überprüft. Ergebnis: Je globaler studentische Beurteilungen erfragt werden (Beispiel: Hat die Vorlesung bisher Ihre Erwartungen insgesamt erfüllt?), desto geringer ist ihr Bezug zum Lehrverhalten der jeweiligen Dozentin bzw. des jeweiligen Dozenten. Im Vordergrund stehende Einflußgrößen sind: das "Image" des zu vermittelnden Stoffs (beliebt/unbeliebt), der Grad der Freiwilligkeit des Veranstaltungsbesuchs (Pflichtveranstaltung mit Klausur versus Wahlveranstaltung), das "mitgebrachte" persönliche Interesse der/des Studierenden. Werden Durchschnittsbeurteilungen je Lehrveranstaltung gebildet, erklären diese - von der Lehrperson kaum beeinflußbaren - Faktoren den größten Teil der Variation zwischen den Veranstaltungen.

V. Besondere Hinweise zur Verwendung des Verfahrens

Teilnehmerbeurteilungen (insbesondere globale Evaluations-Statements) eignen sich zwar als Indikatoren des Stimmungsbilds in der Veranstaltung, nicht jedoch als Maße der Qualität der Lehre und führen bei einem direkten Vergleich zwischen Lehrveranstaltungen zwangsläufig zu Fehlschlüssen. Teilnehmerbefragungen sind jedoch - bei großen Lehrveranstaltungen - ein durch andere Instrumente nicht zu ersetzendes Verfahren der Rückmeldung an die Lehrperson (nur in dieser Funktion ist im übrigen die studentische Veranstaltungskritik in früheren Jahrzehnten diskutiert worden).

VI. Angestrebte Weiterentwicklung

Aufbauend auf den Erfahrungen mit Hörer(innen)befragungen an der Ruhr-Universität Bochum (über 10.000 ausgewertete Bögen) sowie vergleichbaren Ansätzen in anderen Universitäten wird ein explizit auf die Funktion "studentische Rückmeldung" optimiertes Instrument entwickelt und seit drei Semestern testweise am Soziologischen Institut der Freien Universität Berlin eingesetzt.

10.4 Fragebogen zur studentischen Rückmeldung in Lehrveranstaltungen von H. Kromrey

I. Eingangsbefragung, II. Befragung zur abschließenden Evaluation,
Institut für Soziologie an der Freien Universität Berlin,
Projekt 'Qualität von Lehre und Studium', Leitung: Prof. Dr. Helmut Kromrey

I. Kurze Beschreibung des Instruments

Ziel der Befragung ist a) die Bereitstellung von Informationen an die Lehrperson über die Teilnehmer in ihrer Veranstaltung, b) die Rückmeldung der Lehrbeurteilungen durch die Studierenden am Semesterschluß.

Entsprechend enthält der Bogen zur Eingangsbefragung ausschließlich Angaben über die ausfüllende Person: Stellung im Studium, zeitliche Belastung durch Erwerbstätigkeit und andere außeruniversitäre Verpflichtungen, Besuchsgründe, Erwartungen/Befürchtungen, Informationsquellen für die Entscheidung zum Besuch der Veranstaltung, subjektive Anforderungen an die Didaktik, Vorkenntnisse, beabsichtigte Teilnahmefrequenz, geplanter Arbeitsstil und Arbeitsaufwand für die Veranstaltung.

Der Bogen zur abschließenden Evaluation greift einen Teil dieser Fragen wieder auf (Auswirkungen der außeruniversitären Belastungen auf das Studium, erfüllte/nicht erfüllte Erwartungen, bestätigte/nicht bestätigte Befürchtungen, Beurteilung der Didaktik) und ergänzt die Items um weitere evaluative Komponenten: Beurteilung der Veranstaltungsstruktur sowie Arbeitsmaterialien, der Beratung/Betreuung durch die Lehrperson, des Auftretens der Lehrperson, des Stoffumfangs und seines Schwierigkeitsgrads, des wahrgenommenen eigenen Lernerfolgs sowie der Abstimmung der Veranstaltung mit anderen Studieninhalten.

Die Items wurden anhand der Durchsicht der Didaktik-Literatur, eigener vorhergehender Lehrevaluations-Befragungen sowie bereits vorliegender Fragebogen zusammengestellt; Bewertungen werden überwiegend auf 5-Punkte-Ratingskalen abgefragt.

Das endgültig eingesetzte Instrument ist das Ergebnis mehrerer Pretest-Durchgänge: Studierende wurden gebeten, den Bogen auszufüllen; anschließend wurde in einem mündlichen Nach-Interview Frage für Frage die Semantik aus dem Verständnis des Ausfüllenden sowie das Ausfüll-Verhalten erfaßt. Bei Verständnis-Differenzen wurden die Befragten um (für sie eindeutigere) Alternativ-Formulierungen gebeten.

II. Hinweise zur Durchführung und Auswertung

Die Rückmeldung soll die Lehrperson in die Lage versetzen , ihre Veranstaltung nicht nur langfristig den Erfordernissen des Studiengangs anzupassen, sondern sie auch kurzfristig möglichst "zielgruppengerecht" durchzuführen. Zu diesem Zweck müssen die Informationen über die Teilnehmer möglichst schnell zur Verfügung stehen. Die Eingangsbefragung wird daher bereits in der zweiten Veranstaltung des Semesters durchgeführt (Verteilen des Fragebogens ca. 20-30 Min. vor Schluß, sofortiges Ausfüllen und anschließendes Einsammeln), sofort für die EDV aufbereitet und ausgezählt. Spätestens in der vierten Semesterwoche kann die Lehrperson die Ergebnisse mit den Studierenden diskutieren und ggf. noch Anpassungen des Lehrprogramms vornehmen.

Die abschließende Evaluation erfolgt in der vorletzten Semesterwoche (Durchführung wie bei der Eingangsbefragung). Da die Studierenden gebeten werden, sowohl auf dem Eingangs- wie auf dem Abschlußbogen eine anonyme Personenkennung einzutragen (die beiden letzten Buchstaben des Vornamens der Mutter sowie des Vaters, die ersten beiden Ziffern des eigenen Geburtsdatums), können die Daten der Anfangs- und Abschlußerhebung personenweise zusammengeführt werden. So ist in der Gesamtauswertung für die Lehrveranstaltung nicht nur ein Vergleich der Häufigkeitsverteilungen (Anfang / Schluß) möglich, sondern auch die individuelle Veränderung des Meinungsbildes nachvollziehbar.

III. Ergebnisdarstellung

Die Daten werden veranstaltungsweise aufbereitet und ausgezählt sowie darüber hinaus auf Institutsebene kumuliert. Dadurch sind auch veranstaltungsübergreifende Auswertungen (z.B. Cluster- und Faktorenanalysen) möglich.

IV. Gütekriterien

Über die mehrfach pretest-gestützte Gestaltung des Fragebogens und der Items (vgl. Punkt I) hinausgehend wurden spezifische Gütekriterien-Analysen nicht durchgeführt. Bei dem Fragebogen handelt es sich nicht um ein "Meß"-Instrument (im engeren Sinne) zur Ermittlung von "Lehrqualität", sondern um einen breit angelegten - in der Diskussion mit den Lehrveranstaltungsteilnehmern zu validierenden - Informationsinput für die Qualitätsentwicklung und Qualitätssicherung in der Lehre.

V. Angestrebte Weiterentwicklung

Keine. Die Befragungen werden je nach Wunsch der Lehrperson als "Einmal-Erhebungen" mit möglichst individuellem Zuschnitt auf den eigenen Informationsbedarf durchgeführt oder in mehreren Semestern wiederholt.

VI. Besondere Hinweise zur Verwendung des Verfahrens

Die Akzeptanz dieser Form von studentischer Rückmeldung ist bisher sowohl auf seiten der Lehrenden als auch der Studierenden auf große Akzeptanz gestoßen. Mehr als 75 Prozent der ausfüllenden Studierenden plädieren für die Wiederholung solcher Befragungen. In Nachfolgediskussionen mit Lehrenden haben diese die Verwertbarkeit und den Nutzen der erhaltenen Informationen als hoch eingeschätzt. Allerdings ist der damit verbundene Arbeitsaufwand recht hoch. Für kleine Lehrveranstaltungen empfiehlt sich daher eher der Rat an die Lehrenden, das Thema Lehre und Didaktik unmittelbar mit den Teilnehmern (durchaus angeleitet durch die Fragen der Rückmeldebögen) zu diskutieren.

Literatur und Kontaktadresse

H. Kromrey: Studentische Vorlesungskritik. Empirische Daten und Konsequenzen für die Lehre. In: Soziologie, Heft 1/1993, S. 39-56

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

H. Kromrey: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Ph. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung, Münster 1995 (2.), S. 105-128

H. Kromrey: Studentische Befragungen zu Lehre und Studium. Von der Lehrevaluation zur Qualitätsentwicklung. Langfassung eines Vortrags auf dem Kolloqium zur 'Evaluation der Lehre' an der TU Dresden am 19.5.1995; veröffentlicht in einem Tagungsband, hrsg. von der TU Dresden.

Kontaktadresse:

Prof. Dr. Helmut Kromrey

Freie Universität Berlin

Institut für Soziologie

Babelsberger Str. 14-16

10715 Berlin

10.5 Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE) von H. Rindermann und M. Amelang (1994)

I. Beschreibung des Instruments und der Konstruktion

Der HILVE stellt ein multidimensionales Instrument zur Evaluation von Lehrveranstaltungen dar. Auf 14 Skalen (37 vorgegebene Items) können Studierende differenziert zur Lehrveranstaltung und zur Lehre des Dozenten Stellung nehmen. Objekte der Beurteilung sind das *didaktische Lehrverhalten des Dozenten* (z.B. Die Dozentin/Der Dozent kann Kompliziertes verständlich machen), das *soziale Lehrverhalten* (z.B. Die Dozentin/Der Dozent ist kooperativ), *Veranstaltungsaspekte* (Anforderungen, Interessantheit der Veranstaltung), die *studentischen Beiträge* (Referate, Beteiligung), die *studentische Vor- und Nachbereitung* der Veranstaltung sowie das *Veranstaltungsthema*. In zwei Skalen (quantitativer und qualitativer *Lernumfang* sowie *Allgemeinbeurteilung*) können der Erfolg und summativ die Qualität der Veranstaltung eingeschätzt werden.

Die einzelnen Skalen umfassen jeweils zwei bis vier vorgegebene Items. Die Antwortskala variiert zwischen trifft nicht zu (1) und trifft zu (7). Fünf freie Items erlauben der Lehrkraft, veranstaltungsspezifische Fragen zu stellen. Diese können aus einer angebotenen Itemliste ausgewählt oder selbständig formuliert werden. Offene Fragen (was ist besonders positiv?, was weniger?, Verbesserungsvorschläge?, Anmerkungen zum Instrument) ermöglichen den Veranstaltungsteilnehmern, frei formulierte Kommentare zu schreiben. Diese offenen Fragen sollen helfen, durch Items des Instruments nicht abgedeckte, aber subjektiv bedeutsame Aspekte erfassen zu können. Zudem können den Verbesserungsvorschlägen Anregungen für die Seminargestaltung entnommen werden.

Frei wählbare Items und die fakultativen Blöcke *Referate*, *Beteiligung* und *Diskussion* erlauben ein veranstaltungsadaptives Vorgehen. Das Instrument ist für alle gängigen Veranstaltungsformen geeignet. Nicht geeignet ist es für Praktika (z.B. im Labor), Seminare mit überwiegender Gruppenarbeit oder reine Referatverlesestunden.

Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation wurde bisher in sozialwissenschaftlichen (Psychologie, Soziologie, Pädagogik, Ethnologie etc.), sprachwissenschaftlichen (Romanistik) und medizinischen Studienfächern an drei verschiedenen Hochschulen eingesetzt (N=6.400 Studierende). Es kann sowohl von Studierenden als auch von geschulten Fremdgutachtern eingesetzt werden.
Das Inventar wurde anhand bewährter Testkonstruktionsverfahren erstellt. Ausgangspunkt der Itemauswahl waren Befragungen von Lehrenden und Studierenden, anhand welcher Kriterien sie die Lehre beurteilen würden und welche Aspekte aus ihrer Sicht gute Lehre charakterisieren würde. Darüber hinaus wurden englisch- und deutschsprachige Lehrinventare als weitere Itemgenerierungsquelle herangezogen (z.B. Marshs SEEQ). Dieser Itempool wurde nach erhobenen Veranstaltungsbeurteilungen

durch statistische Analysen (Objektivität, Reliabilität, Dimensionalität etc.) und unter Zuhilfenahme von inhaltlichen Kriterien (Feedbackrelevanz, Multiperspektivität der Lehre) auf 37 Items des HILVE reduziert. Zum Schluß wurde der HILVE Lehrenden und Fachschaften zur Überprüfung der studienfachspezifischen Geeignetheit vorgelegt. Das Konstruktionsverfahren kann als iterativ statistisch-inhaltlich bezeichnet werden. Ein Manual (Rindermann & Amelang, 1994) beschreibt Konstruktion und Gütekriterien und schlägt Anwendungsmöglichkeiten vor.

II. Hinweise zur Durchführung und Auswertung

Lehrevaluationen sollten *innerhalb* der Veranstaltungen vorgenommen werden, weil bei veranstaltungsexternem Vorgehen die Teilnahmequote zu gering ist. Die Inventare sollten 15 bis 10 Minuten vor regulärem Veranstaltungsende ausgeteilt werden, damit genug Zeit für Instruktion, Austeilung, Bearbeitung (ca. 8 bis 10 Minuten) und Einsammlung der Bögen zur Verfügung steht. Sollte der Dozent die Bögen selbst einsammeln, sollte Anonymität durch eine bereitgestellte Kiste gewährleistet werden, in die die Studenten die Bögen einwerfen.

Um Auswirkungen auf laufende Veranstaltungen erzielen zu können, sollten Lehrevaluationen Mitte des Semesters vorgenommen werden. Spätestens nach zwei Wochen müßte das Feedback (Auswertung) den Lehrenden vorliegen. In einer veranstaltungsinternen Besprechung können dann die Resultate besprochen werden, Ursachen kritischer Beurteilungen eruiert und Verbesserungsvorschläge gemeinsam erarbeitet werden. Es empfiehlt sich eine zusätzliche Begleitung durch hochschuldidaktische Beratung.
Die Auswertungsdauer je Fragebogen hängt von der Auswertungsmethode ab: Der HILVE wird in Varianten angeboten, die sowohl eine maschinelle als auch eine manuelle Auswertung erlauben. Bei Vorhandensein eines Lesegerätes können in kurzer Zeit große Zahlen an Bögen gelesen und verarbeitet werden. Freie (handschriftliche) Antworten müssen durch Hilfskräfte oder den Dozenten selbst ausgewertet werden (Gruppierung, Auszählung, Feedbackformulierung). Bei universitätsweiten Erhebungen empfiehlt sich die Einrichtung einer zentralen Evaluations- und Hochschuldidaktikstelle, die professionell Auswertung und Beratung der Dozenten vornimmt.

III. Ergebnisdarstellung

In der Handanweisung des HILVE ist eine Feedbackform beschrieben und abgebildet. Dozenten sollten unstandardisierte Rohwerte (Mittel und Streuungen, je nach Interesse auch Verteilungen) als auch normierte Item- und Dimensionswerte rückgemeldet werden. Normen (s. Tabellen in der Handanweisung) bieten ein Bezugssystem zur relativen Interpretation der eigenen Resultate an. Die Normen wurden für Vorlesungen und Seminare sowie für Studienfächer separat berechnet.

Die Ergebnisse können graphisch (Histogramme, Liniendiagramm) aufbereitet werden. Hierfür empfehlen sich gängige Programme (Harvard Graphics, SAS, SPSS etc.).
In einer Datenbank werden Resultate gesammelt und aufbereitet, um für wissenschaftliche Reanalysen sowie für fachspezifische Normierungen ausreichende Stichprobengrößen erzielen zu können. Die Autoren bitten, neue HILVE-Daten an die unten genannte Adresse (Rindermann) zu schicken.

IV. Gütekriterien

Für den HILVE wurden die klassischen Gütekriterien ermittelt (ausführliche Darstellung s. Rindermann, 1996). Berichtet werden an dieser Stelle Dimensionswerte für Veranstaltungsmittel, jeweils für alle HILVE-Skalen zusammen und für die Dozenten- bzw. Lehreffektivitätsskalen (ohne studentische Skalen und Thema) getrennt in Klammern.

Die mittleren *Trennschärfen* der Items betragen r=.78 (.80) und können als sehr gut bezeichnet werden.

Die *Interraterreliabilität* liegt bei 20 Studierenden in der Höhe von r=.86 (.90), bei 10 Studierenden beträgt sie r=.77 (.80). Die *innere Konsistenz* (Cronbachs-α) beträgt r=.89 (.90). Die *Retestreliabilität* (Mitte-Ende des Semesters) liegt bei r=.71 (.72). In einer Vorstudie ohne Aufforderung zur Feedbackbesprechung innerhalb der Veranstaltung liegt die Retestreliabilität bei r=.90 (.92). Die Meßgenauigkeit der Dimensionen kann als gut bezeichnet werden, die Stabilität erreichte aber in der Hauptstudie aufgrund der Rückmeldeprozedur (Besprechung) relativ geringe Werte. Studentische Rohwerte sind weniger zuverlässig als Veranstaltungsmittel.

Die *Urteilerübereinstimmungen* einzelner studentischer Rohwerte sind gering (r=.27 bzw. .31 in Dozenten- und Lehreffektivitätsskalen), sie liegen aber in der Höhe, wie sie von studentischen Beurteilerübereinstimmungen aus der internationalen Forschung und aus der Forschung zum Peer-Review-Verfahren bei der Beurteilung eingereichter Manuskripte durch Fachgutachter mitgeteilt werden (s. Rindermann, 1996). Veranstaltungsmittelwerte ($N \geq 10$ Studenten) zeichnen sich jedoch durch eine hohe Interraterreliabilität aus, d.h. würde eine andere studentische Stichprobe die gleiche Veranstaltung beurteilen, würden deren Resultate zu r=.77-.90 mit den Werten der ersten Stichprobe korrelieren. Veranstaltungsbeurteilungen müssen sich deshalb auf Mindeststichprobenumfänge von $N \geq 10$ Studierenden stützen. In diesem Falle können Evaluationsresultate als relativ meßgenau und repräsentativ bezeichnet werden.

In *Faktorenanalysen* von Beurteilungen durch Studierende, Dozenten und Fremdurteiler konnten vier Faktoren zweiter Ordnung extrahiert werden: Dozenten-, Lehreffektivitäts-, studentische Selbstbeurteilungs- und Anforderungsskalen können bei den drei Urteilergruppen faktorenanalytisch getrennt werden.

Die *Validität* wurde mittels des Vergleichs der studentischen Urteile mit denjenigen von geschulten Fremdurteilern überprüft. Die mittleren Korrelationen betragen r=.59 (.49) und liegen damit in für Validitätskoeffizienten günstiger Höhe. In Analysen zur Untersuchung des Einflusses von Biasvariablen (Variablen, die das studentische Urteil verzerren ohne mit der realen Lehre zusammenzuhängen) konnten individuelle studentische Charakteristika als Biasvariablen ausgeschlossen werden. Leistungsniveau (Abiturnote, veranstaltungsspezifische Klausurresultate), Geschlecht, Alter, Studienmotivation, Studienzufriedenheit usw. korrelieren wenig oder nicht mit der Beurteilung der Lehre. Zusammenhänge konnten jedoch beim Besuchsgrund und dem Vorinteresse (beide retrospektiv eingeschätzt) in der Höhe von r=.10 bis .30 beobachtet werden. Studenten, die eine Veranstaltung aus Interesse besuchen, beurteilen diese etwas besser als Kommilitonen, die diese wegen eines Scheines oder aus Pflichtgründen besuchen. Zum Thema (interessant, relevant) können hohe Korrelationen beobachtet werden, allerdings kann auch gute Lehre für Themen interessieren – die Stärke des jeweiligen Bedingungs-Wirkungs-Verhältnisses ist nicht abschließend geklärt.

Vergleiche von Evaluationsergebnissen, die aus *verschiedenen Veranstaltungen eines Dozenten* gewonnen wurden mit Resultaten aus *Veranstaltungen gleichen Themas, die verschiedene Dozenten hielten*, ermöglichen einen Rückschluß auf die Generalisierbarkeit von Evaluationen. Hierbei ließ sich nachweisen, daß Dozentenvariablen weitgehend dozentenbezogen, studentische oder Themenvariablen jedoch mehr veranstaltungsspezifisch beurteilt werden.

V. Angestrebte Weiterentwicklung

Weiterentwicklungen und Forschungsarbeiten sollen um drei Schwerpunkte kreisen:

1. Erprobungen in verschiedenen Fächern sollen die Berechnungen von weiteren fachspezifischen Normen erlauben.
2. Weitere frei wählbare Items sollen behavioral formuliert und deren Auswirkung auf den didaktischen Prozeß und die Veranstaltungsbeurteilung untersucht werden.
3. Die Gestaltung eines effektiven Rückmelde- und Beratungsprozesses soll untersucht werden.

VI. Besondere Hinweise zur Verwendung des Verfahrens

Generell ist zu empfehlen, Evaluationen in mehreren verschiedenen Veranstaltungen (Mindeststichprobengröße jeweils N≥10 Studierende) vorzunehmen, falls es Ziel ist, die Lehrqualität eines Dozenten zu bestimmen. Der Einfluß eventuell auftretender Störvariablen kann so minimiert werden. Bei systematischen Abweichungen durch

verschiedene studentische Stichprobenzusammensetzungen können Korrekturen vorgenommen werden.

Der HILVE-Fragebogen ist auch in einer maschinell lesbaren Form auslieferbar. Bei maschineller Auswertung können in kurzer Zeit fehlerarm und kostengünstig große Mengen von HILVE-Bögen ausgewertet werden. Dies erlaubt durch eine kurzfristige Rückmeldung des Feedbacks und ergänzende Beratung, Lehrevaluationresultate für die Optimierung der laufenden Veranstaltung einzusetzen.

Literatur und Kontaktadressen

Rindermann, H. & Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE)*. Heidelberg: Asanger.
ISBN: 3-89334-277-X

Das Manual wurde veröffentlicht im Asanger-Verlag:

Roland Asanger Verlag GmbH, Rohrbacherstr. 18, 69115 Heidelberg; Tl. 06221/183104, Fax 06221/160415

Weitere Literatur:

Rindermann, H. & Amelang, M. (1994). Entwicklung und Erprobung eines Fragebogens zur studentischen Veranstaltungsevaluation. *Empirische Pädagogik, 8*(2), 131-151.

Rindermann, H. (1996/im Druck). Verbesserung der Lehre durch Veranstaltungsevaluation? In E. Witruk (Hrsg.), *Tagungsband der 5. Fachgruppentagung Pädagogische Psychologie*. Leipzig.

Rindermann, H. (1996/im Druck). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen: Analysen der Validität und zu Auswirkungen ihres Einsatzes anhand des HILVE*. Landau: Verlag Empirische Pädagogik.

Adressen der Autoren:

Dipl.-Psych. Heiner Rindermann
Institut für Pädagogische Psychologie und Empirische Pädagogik
Leopoldstr. 13,
D-80802 München
E-mail: Rinderma@Mip.Paed.Uni-Muenchen.De

Prof. Dr. Manfred Amelang
Psychologisches Institut der Universität Heidelberg
Hauptstr. 47-51
D-69117 Heidelberg

10.6 Bielefelder Fragebogen zur Bewertung von Lehrveranstaltungen von W. D. Webler, IZHD/Arbeitsgruppe Studentische Veranstaltungskritik

I. Kurze Beschreibung des Instruments

Maschinenlesbarer Fragebogen mit 6 Varianten. Die Items wurden nach Sichtung internationaler und deutscher Fragebogen nach einem eigenen Modell guter Lehre generiert. Die Varianten erfassen alle Arten von Lehrveranstaltungen. Die Langfassung A (geistes- und sozialwissenschaftliche Seminare) hat 74 Items und 2 offene Fragen. Die Versionen für Vorlesungen sowie die für (naturwissenschaftliche) Übungen und Praktika sind kürzer.

Eine ab 1994 eingesetzte, um ein Drittel gekürzte Fassung hat (trotz Maschinenlesbarkeit) 6 ad hoc formulierbare Fragemöglichkeiten und 2 offene Fragen. Das Instrument wurde nach einem Pilotversuch (1991 in der Fakultät für Soziologie) ab 1992 in 11 von 13 Fakultäten flächendeckend in allen Lehrveranstaltungen (mehr als 700) eingesetzt. Ab 1994 ging der Einsatz zurück (Abnutzungseffekte wegen zu häufigen Einsatzes) und wurde z.T. durch andere schriftliche/mündliche Methoden der Rückmeldung ersetzt.

Eine Variante des Fragebogen-Instruments wurde vom IZHD für Veranstaltungsbewertungen im Rahmen des Lehrbericht-Projekts an Fachbereichen der hessischen Hochschulen ebenfalls in einer großen Zahl unterschiedlicher Lehrveranstaltungen eingesetzt.

II. Hinweise zur Durchführung und Auswertung

Die Dauer zum Ausfüllen des Fragebogens beträgt je nach Version 10-20 Minuten. Ausgefüllt wird in der Mitte des Semesters in allen Veranstaltungen einer Semesterwoche, während einer laufenden Veranstaltung. Die Auswertung erfolgt durch Belegleser und ein Auswertungsprogramm. Der Belegleser liest 3.000 doppelseitige Fragebogen pro Stunde.

III. Ergebnisdarstellung

Die Ergebnisse werden pro Veranstaltung in absoluten Zahlen, Balkendiagrammen und Mittelwerten ausgedruckt (z.B. auf Overheadfolien zur Diskussion in der Lehrveranstaltung; „Bielefelder Diskursmodell“). Eine Datenbank wird nicht angelegt (vereinbarte Randbedingung).

IV. Gütekriterien

Keine Analysen nach Gütekriterien durchgeführt.

V. Angestrebte Weiterentwicklung

Zur Zeit nicht beabsichtigt.

VI. Besondere Hinweise zur Verwendung des Verfahrens

Hinweise zur Verwendung des Verfahrens sind in folgenden Berichten enthalten:

W.-D. Webler u.a.: Lehrberichte. Empirische Grundlagen, Indikatorenauswahl und Empfehlungen. (Bundesministerium für Bildung und Wissenschaft (Hg.): Schriftenreihe Studien zu Bildung und Wissenschaft, Bd. 107). Bad Honnef 1993. Anhang mit verschiedenen Fragebogen-Varianten. Bonn 1993, Anhang.

W.-D. Webler: Evaluation der Lehre: Praxiserfahrungen und Methodenhinweise. In: Hellmut Winkler (Hg.): Qualität der Hochschulausbildung. (Wiss. Zentrum für Berufs- und Hochschulforschung der Universität GH Kassel; Werkstattberichte 40), Kassel 1993.

Adresse des Autors:

Dr. W.-D. Webler

Universität Bielefeld

Interdisziplinäres Zentrum für Hochschuldidaktik

Postfach 100131

D-33501 Bielefeld

Tel.: 0521/106-4680

Fax-Nr.: 0521/1066034

10.7 Fragebogen zur „Evaluation der Lehre“ an der Universität Mannheim von H. D. Daniel

Das Evaluationsprojekt „Evaluation der Lehre“ wurde aufgrund eines Antrags des Mannheimer AStAs Ende 1992 begonnen und vom Landesministerium Baden-Württemberg gefördert. Die Senatskommission für Lehre begleitete das Projekt, dessen Leitung in den Händen von Dr. Hans-Dieter Daniel lag.

Das große und umfassende Projekt an der Universität Mannheim hatte zum Ziel eine Vollerhebung im betriebswirtschaftlichen Grundstudium (BWL I und BWL II) durchzu-führen. Später erfolgten weitere Erhebungen zur Lehrbewertung in den Geisteswissenschaften. Alle beteiligten Lehrenden nahmen auf freiwiliger Basis an der Lehrbewertung teil.

Entsprechend der Art der Veranstaltung (Vorlesungen, Übungen, Tutorien) wurden von Daniel veranstaltungsspezifische Fragebogen entwickelt. Außerdem wurde ein allgemeiner Fragebogen ausgearbeitet, der veranstaltungsübergreifende Informationen und einige Angaben zur Person abfragte.

Die Bearbeitungsdauer betrug aufgrund der relativ umfassenden Fragebogen 25-30 Minuten, maximal nach 45 Minuten konnte die Erhebung abgeschlossen werden.

Folgende Fragebogen wurden entwickelt:

Fragebogen zu den Lehrveranstaltungen in „Grundzüge der Betriebswirtschaftslehre“ 1992/93 (Allgemeiner Teil)

Fragebogen zu den Vorlesungen in „Grundzüge der Betriebswirtschaftslehre“ 1992/93

Fragebogen zu den Übungen in „Grundzüge der Betriebswirtschaftslehre“ 1992/93

Fragebogen zu den Tutorien in „Grundzüge der Betriebswirtschaftslehre“ 1992/93

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 (Allgemeine Fragen)

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 - Fragebogen für Einführungsveranstaltungen -

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 94 - Fragebogen für Vorlesungen -

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 94 - Fragebogen für Seminare -

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 94 - Fragebogen für sprachpraktische Übungen -

Hörerbefragung in der Fakultät für Sprach- und Literaturwissenschaft 1993/94 94 - Fragebogen für Tutorien -

Evaluation der Lehre, Fragebogen für Einführungsveranstaltungen (Anglistisches Seminar, KurzFragebogen)

Fragebogen für Studierende des Diplomstudienganges „Philologie mit wirtschafts-wissenschaftlicher Qualifikation“ mit dem Kernfach Slavistik

Eine Überprüfung der Gütekriterien dieser Instrumente wurde nicht durchgeführt.

Das Projekt wurde, wie bereits erwähnt, von einer Senatskommission für Lehre betreut, die die Endfassung der jeweiligen Fragebogen verabschiedete. Folge dieser Konstellation sind ausgesprochen umfangreiche Fragebogen, die vor allem zu wissenschaftlichen Zwecken oder zur Itemrequirierung geeignet erscheinen. Wird eine Vollerhebung geplant, so ist mit der Notwendigkeit von viel Personal zu rechnen.

Literatur

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

Aleamoni, L. M. und Thomas, G. (1980). Differential Relationships of Student, Instructor, and Course Characteristics to General and Specific Items On a Course Evaluation Questionnaire. *Teaching of Psychology*, Bd.7 (4), S.233-235

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

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

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

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

Daniel, H.D (1995). Das Modellprojekt „Evaluation der Lehre“ an der Universität Mannheim. Teil 2: Statistische Auswertung von Befragungen in Lehrveranstaltungen. In: Mohler, P. (Hrsg.) Universität und Lehre. *Ihre Evaluation als Herausforderung an die Empirische Sozialforschung*. Waxmann Verlag: Münster

Daniel, H.D (1995). Bewertung der Lehre aus Sicht der Studierenden und Absolventen. In D.Müller-Böling (Hrsg.), *Qualitätssicherung in Hochschulen* (S.160-185). Gütersloh: Verlag Bertelsmann Stiftung

Daniel, H.D (1995). Der Berufseinstieg von Betriebswirten. *Personal - Zeitschrift für Human Resource Management*, 47 (10), S.429-499

Daniel, H.D (1995). Die Wahl von Studienort und Hochschule. In E. Dichtl & M. Lingenfelder (Hrsg.), *Effizient studieren: Wirtschaftswissenschaften* (S.3-22). Wiesbaden: Gabler

Daniel, H.D (1995). Datenbedarf zu und Anfertigung von Lehrberichten. Erfahrungsbericht aus Baden-Württemberg. Statistisches Bundesamt Wiesbaden (Hrsg.), *Effizienzbemessung der Hochschulausbildung auf statistischer Grundlage* (S.62-65). Stuttgart: Metzler-Poeschel (Schriftenreihe „Spektrum Bundesstatistik“, Band 7)

Daniel, H.D (1995). Ist wissenschaftliche Leistung in Forschung und Lehre meßbar? *Universitas -Zeitschrift für interdisziplinäre Wissenschaft*, 50 (585), S. 205-209

Diehl, J.M. & Kohr, H.U. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. Psychologie in Erziehung und Unterricht, 24, 61-75.

Diehl, J.M. (1983). Untersuchungen zur Einstellung von Psychologiestudenten gegenüber Statistik und Statistik-Veranstaltungen. Berichte aus dem Fachbereich 06 Psychologie, Universität Gießen.

Diehl, J.M. (1994). Fragebogen zur studentischen Evaluation von Hochschulveranstaltungen. Manual zum VBVOR und VBREF. Fachbereich 06 Psychologie (Abteilung Methodik), Universität Gießen.

Driscoll, Laura A. und Goodwin, W. L. (1979). "The Effects of Varying Information About Use and Disposition of Results on University Students´ Evaluation of Faculty and Courses". In: American Educational Research Journal, Bd. 16, S. 25-37

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

Englert, Steffani, (1993). *Fachkultur, Geschlecht und soziale Reproduktion*. Weinheim: Deutscher Studien Verlag

el Hage, Natalija (1995). Zur Validität studentischer Veranstaltungskritiken. Befunde empirischer Studien zu einem umstrittenen Verfahren. Hefte zur Bildungs- und Hochschulforschung, Bd. 13

Faßnacht, G. (1979). „Systematische Verhaltensbeobachtung“. München: ernst Reinhardt Verlag

Frank, Andrea (1991). Wodurch ist der Hochschullehrer/ die Hochschullehrerin wichtig für das studentische Lernen?. In: Webler, W. und Otto, H. (Hrsg.) *Der Ort der Lehre in der Hochschule.Lehrleistungen, Prestige und Hochschulwettbewerb*. Weinheim: Deutscher Studienverlag

Gleich, J. M., Meran, G. und Bargel, T. (1982). *Studenten und Hochschullehrer*. Serie: Bildung in neuer Sicht, Hrsg.: Ministeriums für Wissenschaft und Kunst (Ba-Wü), Bd. 48

Heckhausen, H. (1986). Die Pädagogische Psychologie vor neuen Herausforderungen. In: Weidemann, B. und Krapp, A., *Pädagogische Psychologie*. S. 786-788

Hochschul-Informations-System (1992). *Aktuelle Aktivitäten an deutschen Hochschulen*. Dokumentation, Teil 1

Kromrey, H. Studentische Vorlesungskritik. Empirische Daten und Konsequenzen für die Lehre. *Soziologie*, Heft 1/1993, S. 39-56

Kromrey, H. Wie erkennt man "gute Lehre"? (1994). Was studentische Vorlesungsbefragungen (nicht) aussagen. *Empirische Pädagogik*, 2, S. 153-168

Kromrey, H. (1995). Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Ph. Mohler (Hg.). *Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung*, Münster (2.), S. 105-128

Kromrey, H. (1995). *Studentische Befragungen zu Lehre und Studium. Von der Lehrevaluation zur Qualitätsentwicklung.* Langfassung eines Vortrags auf dem Kolloqium zur 'Evaluation der Lehre' an der TU Dresden am 19.5.1995; veröffentlicht in einem Tagungsband, hrsg. von der TU Dresden.

Levinthal, C. F., Lansky, L. M. und Andrews, O. E. (1971). Student Evaluations of Teacher Behaviors as Estimations of Real-Ideal Discrepancies: A Critique of Teacher Rating Methods. *Journal of Educational Psychology*, Bd. 62 (2), S. 104-109

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

Marquez, T. E., Lane, D.M. und Dorfman, P. W. (1979). Toward the Development of a System for Instructional Evaluation: Is there Consensus Regarding What Constitutes Effective Teaching? *Journal of Educational Psychology*, Bd. 71(6), S. 840-849

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

Marsh, H. W. und Hocevar, D. (1984). The Factorial Invarance of Student Evaluations of College Teaching. *American Educational Research Journal*, Bd. 21 (2), S.341-366

Mazer, G. E. (1977). Evaluating the Evalutions: A Factor Analysis of Student Ratings. *Counselor Education and Supervision*, Bd.17, S.6-11

McBean, E. A. und Lennoy, W. C. (1985). Effect of Survey Size on Student Ratings of Teaching. *Higher Education*, Bd. 14, S. 117-125

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

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

Müller-Wolf, H.-M. und Fittkau, B. (1971). Lehrverhalten von Hochschullehrern und seine Bedeutung für Einstellungen und Verhalten von Studenten. *Zeitschrift für Entwichlungspsychologie und Pädagogische Psychologie*, Bd. 3 (3), S.165-180

Murray, H.G. (1983). Low-Interference Classroom Teaching Behaviors and Student Ratings of College Teaching Effectiveness. *Journal of Educational Psychology*, Bd. 75 (1), S. 138-149

Neumann, Lily und Neumann, Y. (1981). Comparison of Six Lenghts of Rating Scales: Students´Attitudes Toward Instrucion. *Pschological Reports*, Bd. 48, S. 399-404

Orpen, C. (1980). The Susceptibility of Student Evaluation of Lecturers to Situation Variablbes. *Higher Education*, Bd. 9, S. 293-306

Pasen, R., Frey, P., Menges R. und Rath, G. (1978). Different Administrative Directions and Student Ratings of Instruction: Cognitive Versus Affective Effects. *Research In Higher Education*, Bd. 9, S.161-167

Penfield, D: A. (1978). Student Ratings of college Teaching: Rating the Utility of Rating Forms. *The Journal of Educational Research*, Bd. 72, S. 19-23

Preisert, H., Bargel, T. und Framheim, Gerhild (1988). Studiensituation und studentische Orientierung an Universitäten und Fachhochschulen In*: Schriftreihe: Studien zu Bildung und Wissenschaft*, Bd. 59, Hrg. Bundesminister für Bildung und Wissenschaft

Rindermann, H. & Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE)*. Heidelberg: Asanger.

Rindermann, H. & Amelang, M. (1994). Entwicklung und Erprobung eines Fragebogens zur studentischen Veranstaltungsevaluation. *Empirische Pädagogik, 8*(2), 131-151.

Rindermann, H. (1996/im Druck). Verbesserung der Lehre durch Veranstaltungsevaluation? In E. Witruk (Hrsg.), *Tagungsband der 5. Fachgruppentagung Pädagogische Psychologie*. Leipzig.

Rindermann, H. (1996/im Druck). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen: Analysen der Validität und zu Auswirkungen ihres Einsatzes anhand des HILVE*. Landau: Verlag Empirische Pädagogik.

Ritter, U. P. (1992). Studentische Partizipation bei der Evaluation von Lehrveranstaltungen. *Das Hochschulwesen*, Bd. 2, S.87-90

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

Scott, C.S. (1977). Student Ratings and Instructor-Defined Extenuationg Circumstances. *Journal of Educational Psychology*, 69, S. 744-747

Sommer, J. und Petermann, F. (1978). Deskriptive und präskriptive Aussagen in einem Fragebogen zur Beurteilung akademischer Lehrveranstaltungen. *Zeitschrift für Entwicklungspsychologie und Pädagogischer Psychologie*, 10 (4), S.336-346

Sturm, M. (1991). Lehrveranstaltungsbewertung als Grundlage für die kontinuierliche Verbesserung der universitären Lehre. *Zeitschrift für Hochschuldidaktik*, 1-2, S.16-34

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

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

Thorndike, R. L. und Hagen, Elizabeth (1969*). Measurement and Evaluation in Psychology and Education*. New York: John Wiley & Sons, Inc.

Webler, W.-D. (1992). Evaluation der Lehre - Praxiserfahrungen und Methodenhinweise. In: Grühn, D. und Gattwinkel, H. (Hrsg.*) Evaluation von Lehrveranstaltungen*. S. 143-161, Berlin. Zentrale Universitäts-Druckerei

W.-D. Webler u.a. (1993). Lehrberichte. Empirische Grundlagen, Indikatorenauswahl und Empfehlungen. (Bundesministerium für Bildung und Wissenschaft (Hg.): Schriftenreihe Studien zu Bildung und Wissenschaft, Bd. 107). Bad Honnef.

Webler, W.-D. (1993). Evaluation der Lehre: Praxiserfahrungen und Methodenhinweise. In: H. Winkler (Hg.): Qualität der Hochschulausbildung. (Wiss. Zentrum für Berufs- und Hochschulforschung der Universität GH Kassel; Werkstattberichte 40). Kassel.

Weinert, F. E. (1986). Lernforschung als eine zentrale Aufgabe der Pädagogische Psychologie, In: Weidemann, B. und Krapp, A., *Pädagogische Psyschologie*. S.783-786, Urban Schwarzberg: Pchychologie Verlags Union,

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

ISSN 1616-0398