Maren Hiltmann

Die Erfassung des Konstruktes 'gute Lehre' in Fragebogen-Verfahren

Mainzer Beiträge zur Hochschulentwicklung, Bd. 6

Hrsg.: Zentrum für Qualitätssicherung und -entwicklung (ZQ)

Maren Hiltmann

Die Erfassung des Konstruktes 'gute Lehre'
in Fragebogen-Verfahren

Mainzer Beiträge zur Hochschulentwicklung, Bd. 6

Hrsg.: Zentrum für Qualitätssicherung und -entwicklung (ZQ)

Mainz 2002

(Zugleich als Diplomarbeit an der Fakultät für Psychologie
der Ruhr-Universität Bochum)

ISBN: 3-935461-04-6
ISSN: 1616-5799

Vorwort

Mit den Mainzer Beiträgen zur Hochschulentwicklung ist neben der Dokumentation von Evaluationsergebnissen beabsichtigt, Raum für die kritische Auseinandersetzung mit evaluationsrelevanten Themen sowie für die Darstellung von Arbeiten aus der Hochschul- und Bildungsforschung zu bieten. Sie stehen grundsätzlich auch Autorinnen und Autoren offen, die außerhalb der Universität Mainz tätig sind und in entsprechenden Bereichen forschen.

Der vorliegende Band 6 der Schriftenreihe von Maren Hiltmann befasst sich mit dem Konstrukt ‚gute Lehre'. Hierzu nimmt sie im Rahmen ihrer Diplomarbeit, die von Prof. Rosemann und Prof. Wottawa (Ruhr-Universität Bochum) betreut wurde, Bezug auf die Forschungslage zur Veranstaltungsevaluation und geht auf die Differenzen zwischen den unterschiedlichen Traditionen der Lehrveranstaltungsbewertung in den USA und Deutschland ein. Von besonderem Interesse ist hierbei die Analyse von Fragebögen im Hinblick auf die jeweils verwendeten impliziten oder expliziten Dimensionen für die Erfassung guter Lehre.

Die Ergebnisse der Untersuchung sind nicht zuletzt deshalb interessant, als sie dazu beitragen könnten, die Diskussion um Kriterien guter Lehre wieder aufzugreifen. Dies scheint um so notwendiger, als die Vorstellungen über gute Lehre, damit aber auch ein Verständnis über die Aufgaben und Ziele von Hochschulen – insbesondere auch den Stellenwert von Forschung und Lehre – zwar latent Gesprächsgegenstand sind, aber nur selten in Hochschulen und im Umfeld von Hochschulen systematisch erfasst werden.

Für den Kontext der Evaluation kommt die Autorin zu einer aus meiner Sicht angemessenen Einschätzung des Stellenwertes von Lehrveranstaltungsbefragungen, indem sie empfiehlt, die im Rahmen von Evaluationen eingesetzten Verfahren nicht kontrovers, sondern integrativ zu diskutieren. Dieser, der Perspektiven- und Methodenvielfalt verpflichtete Ansatz, so sind sich Praktiker nicht nur im Bereich der Hochschulevaluation einig, bietet angemessene Zugänge, die dem zunehmenden Charakter von Evaluation als Organisationsentwicklung entgegenkommen.

Dr. Uwe Schmidt

(Zentrum für Qualitätssicherung und -entwicklung)

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

Danksagung

An dieser Stelle möchte ich allen herzlich danken, die zum Gelingen dieser Arbeit beigetragen haben.

Insbesondere danke ich dem akademischen Betreuer der Arbeit, Herrn Prof. Dr. Bernhard Rosemann, für die wohlwollende Unterstützung, seine konstruktiven Anregungen und die mir gewährten Freiräume. Ebenso danke ich Herrn Prof. Dr. Heinrich Wottawa für seine Gutachtertätigkeit.

Mein besonderer Dank gilt den beteiligten Kodierern, ohne die diese Arbeit nicht möglich gewesen wäre. Er gilt weiterhin all jenen, die mir ermöglicht haben, für ein Dreivierteljahr an der University of Georgia, Athens, USA, zu studieren. Viele Anregungen und Überlegungen aus dieser Zeit flossen in die vorliegende Arbeit ein.

Ferner danke ich meinen Eltern sowie Till Mettig, Thomas Ernst, Kerstin Koch, Katja Bürgerhoff und Sandra Calabrese für die allzeit hilfreichen Anregungen und aufmunternden Worte.

Bochum, im Oktober 2001

Maren Hiltmann

1. Einleitung

In einer Pressemitteilung des Bundesministeriums für Bildung und Forschung sowie des Bundesministeriums des Inneren vom 30. Mai 2001 heißt es: „Kabinett beschließt neues Dienstrecht für Professoren (...) mit ausgeprägtem Leistungsbezug“ (Aktenzeichen 80/2001). Es soll die sogenannte ‚Bezahlung nach Leistung‘ eingeführt werden, eine Kombination aus einem Mindestgehalt und einem variablen Anteil, „der sich unter anderem aus der Bewertung von Leistung in Lehre und Forschung oder der Studienbetreuung zusammensetzt“. Als eine Konsequenz dieser Reform stünden künftig regelmäßig Bewertungen von Lehr- und Forschungsleistungen an.

Dies lässt danach fragen, welche Bewertungsverfahren vorliegen, welche Kriterien herangezogen werden könnten und welche Messgenauigkeit und Zuverlässigkeit die Verfahren aufweisen. Der Gesetzesentwurf überrascht nicht, da seit Beginn der 1990er Jahre die Lehrevaluation auch in Deutschland zu einem viel und oftmals heftig diskutierten Dauerthema geworden ist. Zahlreiche einschlägige Magazine veröffentlichten Hochschulrankings und andere populärwissenschaftlich orientierte Artikel zur Situation an deutschen Hochschulen (z. B. „Welche Uni ist die beste?“ *Der Spiegel Nr. 50*, 1989). Initiiert durch die Aufmerksamkeit der Öffentlichkeit drang die Evaluation der Lehre dann schnell auch in Form umfangreicher Modellprojekte an die Hochschulen (vgl. zusammenfassend Reissert, 1992). Inzwischen ist die Evaluation der Lehre bereits in den Gesetzen einiger Bundesländer verankert (z. B. Nordrhein-Westfalen). Der neue Gesetzesentwurf zum Dienstrecht geht jedoch einen deutlichen Schritt weiter: mit einem leistungsabhängigen Vergütungsanteil auf Basis der Evaluation von Lehr- und Forschungsleistungen führte die Evaluation für die Lehrenden erstmalig zu individuellen Konsequenzen.

An amerikanischen Universitäten gibt es dagegen seit Beginn des 20. Jahrhunderts die ungebrochene Tradition der Evaluation von Lehr- und Forschungsleistungen. Dort liegen inzwischen vielerorts elaborierte Evaluationsmodelle vor, deren Ergebnisse vielfach für Bleibe- und Berufungsverhandlungen genutzt werden.

Die vorliegende Arbeit greift die Evaluation von Lehrleistungen, die sogenannte *Evaluation der Lehre*, heraus. „Die Evaluation von Lehre, insbesondere durch die Verwendung von Fragebogen, ist ein Thema, das Stellungnahmen, Kontroversen und Verwirrung erzeugt“ – so lautet eine Charakterisierung der amerikanischen Evaluationsforscher Braskamp und Ory aus dem Jahre 1994 (S. 1). Wie im Zitat angeklungen, spielt insbe-

sondere die Befragung von Studierenden mittels Fragebogen als eine Form der Lehrevaluation eine herausragende Rolle. Fragebogenverfahren als Bestandteil von Evaluationsmodellen gibt es beinahe ausnahmslos an jeder amerikanischen Hochschule (Arreola, 2000). In Deutschland ist die studentische Lehrevaluation ein – gleichermaßen umstrittenes – Thema. Die Charakterisierung der Autoren Braskamp und Ory beschreibt damit auch die deutschen Verhältnisse zutreffend. In Deutschland gibt es ebenso unzählige Stellungnahmen, Kontroversen und mancherorts sicherlich auch Verwirrung zu Rolle, Qualität und Nutzen von Fragebogenverfahren zur Lehrevaluation.

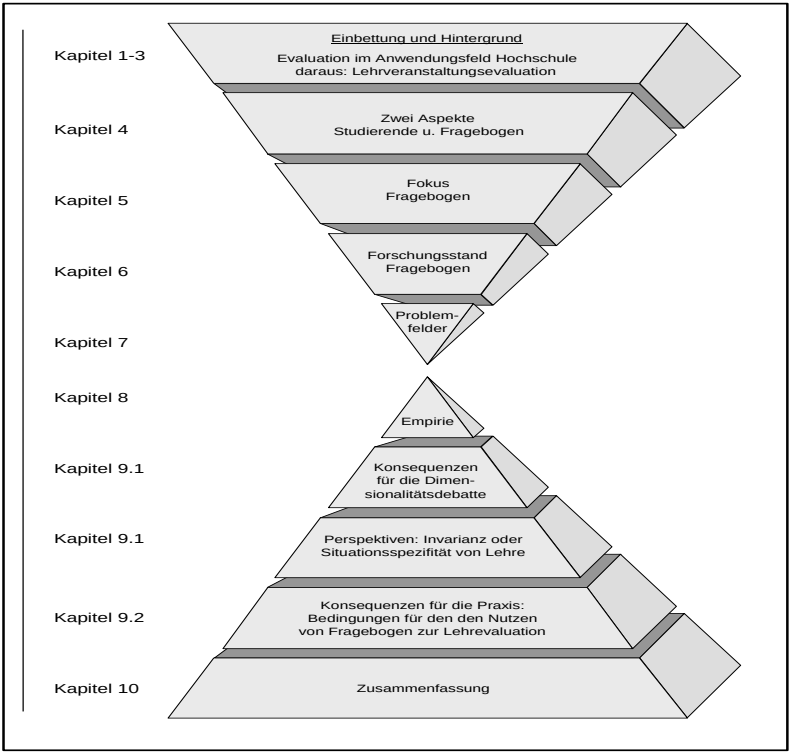
Diese Arbeit möchte daher einen Beitrag zur Transparenz der Aktivitäten und Verfahren im Bereich der studentischen Lehrevaluation aus wissenschaftlicher Sicht leisten. Dabei sollen insbesondere Ergebnisse und Diskussionen aus der langen Tradition der Lehrevaluation(sforschung) in den USA einbezogen werden, da diese bisher wenig rezipiert worden sind (el Hage, 1996; Rindermann, 2001).

Zunächst erfolgt die Klärung einiger grundlegender Begriffe und die Einordnung der studentischen Lehrevaluation in das weite Feld der Evaluation im Anwendungsbereich Hochschule (Kapitel 2). Danach werden die historischen Entwicklungen der Evaluation der Lehre in den USA ebenso wie in Deutschland dargestellt (Kapitel 3). Wenn es um Evaluation der Lehre als studentische Lehrevaluation geht, dann sind besonders zwei Aspekte von zentraler Bedeutung: die Frage nach den Studierenden als gewinnbringende Informationsquelle und die Frage nach der Methodik, also den in der studentischen Lehrevaluation eingesetzten Fragebogen. Eine kurze Darstellung dieser beiden Aspekte erfolgt in Kapitel 4. Die Auseinandersetzung mit den Instrumenten, den verschiedenen Fragebogenverfahren, die zur Befragung von Studierenden eingesetzt werden, steht im Zentrum dieser Arbeit. In Kapitel 5 erfolgt daher ein Überblick über Aufbau, Form und Zielsetzungen solcher Verfahren. Kapitel 6 vertieft die Diskussion über die Qualität dieser Verfahren in Form einer intensiven Auseinandersetzung mit dem aktuellen wissenschaftlichen Forschungstand der deutschsprachigen und anglo-amerikanischen Literatur. Daraus werden Problemfelder abgeleitet, die für die intendierte Erfassung des Konstruktes ‚gute Lehre' relevant sind (Kapitel 7). Schließlich wird ein zentrales Problem herausgegriffen, nämlich die Frage der Dimensionalität von Fragebogenverfahren als Hinweis auf die Dimensionalität des Konstruktes, und dies mündet in den empirischen Teil dieser Arbeit ein (Kapitel 8). Darin wird auf sprachlicher Ebene der Items untersucht, welche Dimensionen in aktuellen deutschen und amerikanischen Lehrevalua-

tionsfragebogen anzutreffen sind, mit welcher Häufigkeit sie auftreten und wie sich diese Häufigkeiten verteilen. Eine Diskussion des Dargestellten in Hinblick auf mögliche Konsequenzen für die Dimensionalitätsdebatte sowie für die weitere Forschung bzw. Praxis erfolgt in Kapitel 9. Schließlich beinhaltet Kapitel 10 eine Zusammenfassung der Arbeit.

Die nachfolgende Grafik (S. 4) gibt den Aufbau und die Gliederung der vorliegenden Arbeit wieder.

Abb. 1. Aufbau und Gliederung der vorliegenden Arbeit



2. Evaluation

Wie eingangs erwähnt, will sich die Arbeit mit der Evaluation von Lehre auseinandersetzen. Die Arbeit verfolgt zunächst das Ziel, die Aktivitäten und Entwicklungen von Lehrevaluation aus wissenschaftlicher Sicht darzustellen und zu diskutieren. Zuerst soll daher der Begriff «Evaluation» näher beleuchtet werden, wozu im Folgenden einige verwandte Begriffe, Definitionsansätze und abschließend allgemeine Kennzeichen von Evaluation kurz vorgestellt werden (Abschnitt 2.1). Evaluation findet in vielen Anwendungsfeldern statt. Das Anwendungsfeld, in dem sich die Thematik dieser Arbeit bewegt, ist der Hochschulbereich. Um die Evaluation von Lehre in das komplexe Anwendungsfeld der Hochschule einzuordnen, werden im Weiteren drei grundlegende Gliederungsaspekte dargestellt (Abschnitt 2.2).

2.1 Der Evaluationsbegriff

Unter dem Begriff «Evaluation» werden teilweise höchst unterschiedliche Auffassungen subsumiert. Äußerst unscharf ist zudem seine Abgrenzung zu Begriffen wie Qualitätskontrolle, Qualitätssicherung, Wirkungskontrolle, oder einfach nur Bewertung. Evaluation kann sich auf verschiedene Gegenstandsbereiche wie Personen, Programme, Maßnahmen, Wirkungen, oder Institutionen beziehen. Ein spezifisches Methodeninstrumentarium für «Evaluationsmethodik» ist noch nicht etabliert. In der Geschichte der psychologischen Forschungsmethodik begann sich die Evaluationsmethodik Anfang des 20. Jahrhunderts als ein neuer, zunehmend eigenständig werdender Strang herauszubilden. Doch erst seit dem letzten Drittel des 20. Jahrhunderts verknüpft er sich explizit mit dem Oberbegriff der Evaluationsmethodik (Sprung & Sprung, 2000).

Wer nach Definitionen des Begriffs «Evaluation» sucht, wird schnell auf die Vielfältigkeit dieses Begriffs stoßen. Daher erscheint es ohne weitere Präzisierung wenig sinnvoll, von *der* Evaluation zu sprechen. Die Vielfalt führt dazu, dass sich der Begriff Evaluation "prinzipiell einer abstrakten, die Wirklichkeit gleichzeitig voll umfassenden Definition" entzieht (Wottawa & Thierau, 1998, S. 13). Die Autoren halten es daher für zweckmäßiger, allgemeine Kennzeichen wissenschaftlicher Evaluation herauszuarbeiten. Tabelle 1 führt die von Wottawa und Thierau herausgearbeiteten Kennzeichen auf. Die dort genannten Kennzeichen stellen auch das Verständnis von Evaluation dar, das dieser Arbeit zugrunde liegt.

Tab. 1. Allgemeine Kennzeichen wissenschaftlicher Evaluation (nach Wottawa & Thierau, 1998)

Kennzeichen wissenschaftlicher Evaluation (w.E.)	
Planungs- und Entscheidungshilfe	w.E. umfasst Tätigkeiten, die unter dem Aspekt der Bewertung von Handlungsalternativen stehen.
Ziel- und Zweck-orientierung	w.E. hat zum Ziel, praktische Maßnahmen zu überprüfen, zu verbessern oder über sie zu ent-scheiden.
Wissenschaftsbezug	w.E. passt sich dem aktuellen Stand wissenschaftlicher Techniken und Forschungsmethoden an.

In der Literatur finden sich neben den Ansätzen zur Definition von Evaluation bzw. den allgemeinen Kennzeichen von Evaluation weitere Ausdifferenzierungen und Fachtermini. Sie treten oft in Form von Begriffspaaren auf, die bestimmte Arten oder Formen von Evaluation beschreiben. Eine ausführliche Übersicht findet sich bei Wottawa und Thierau (1998). Zwei dieser vielfältigen Unterscheidungsdimensionen von Evaluation sollen hier für die weitere Diskussion kurz eingeführt werden. Sie beziehen sich auf unterschiedliche Vorgehensweisen im Evaluationsprozess:

- summative vs. formative Evaluation

  Bei einer summativen Vorgehensweise steht eine zusammenfassende Bewertung, meist globaler Art, im Vordergrund. Sie ist im Sinne einer Bilanz-Evaluation oder Qualitätsfeststellung zu verstehen (Diagnoseabsicht). Im Unterschied dazu strebt die formative Evaluation eine schrittweise Stabilisierung oder Verbesserung im Sinne einer Gestaltungsevaluation oder Qualitätssicherung an (Optimierungsabsicht).

- interne vs. externe Evaluation (auch Selbst- vs. Fremdevaluation)

  Bei internen Evaluationen handelt es sich um Vorhaben, bei denen die Mitglieder einer Institution, die Gegenstand der Evaluation ist, den Evaluationsprozess selbst durchführen. Alternativ dazu ist die externe Evaluation zu sehen, bei der eine Trennung von Beurteilern und "Betroffenen" zugrunde liegt. Externe Beurteiler führen die Evaluation durch.

2.2 Evaluation im Anwendungsfeld Hochschule

Für das weitere Verständnis der Zusammenhänge soll der Gesamtkontext, die Evaluation im Anwendungsfeld Hochschule, näher betrachtet werden. Zur Gliederung des Feldes bieten sich drei zentrale Aspekte an, die im Folgenden erörtert werden. Erstens ist der Gegenstandsbereich von Evaluation im Kontext von Hochschule näher zu bestimmen. *Was* kann evaluiert werden? Zweitens ist von Bedeutung, *wozu* oder warum etwas evaluiert werden soll. Als dritter Aspekt ist schließlich auch von Bedeutung *wie*, also auf welche Art und Weise bzw. mit welchen Methoden evaluiert werden soll.

2.2.1 Gegenstandsbereich — was evaluieren

Webler (1996) nennt in einem Überblick über Probleme hinsichtlich der Qualität von Lehre und Studium an deutschen Hochschulen drei Problemfelder, welche er auf drei verschiedenen Ebenen ansiedelt. Diese eigenen sich als Gliederung für die unterschiedlichen Gegenstandsbereiche von Evaluation an Hochschulen. So kann Evaluation (1) auf der Ebene eines ganzen Fachbereiches oder ganzer Studiengänge, (2) auf einer individuellen Ebene bzw. der Ebene der einzelnen Lehrveranstaltungen und (3) auf der Ebene der Rahmenbedingungen stattfinden.

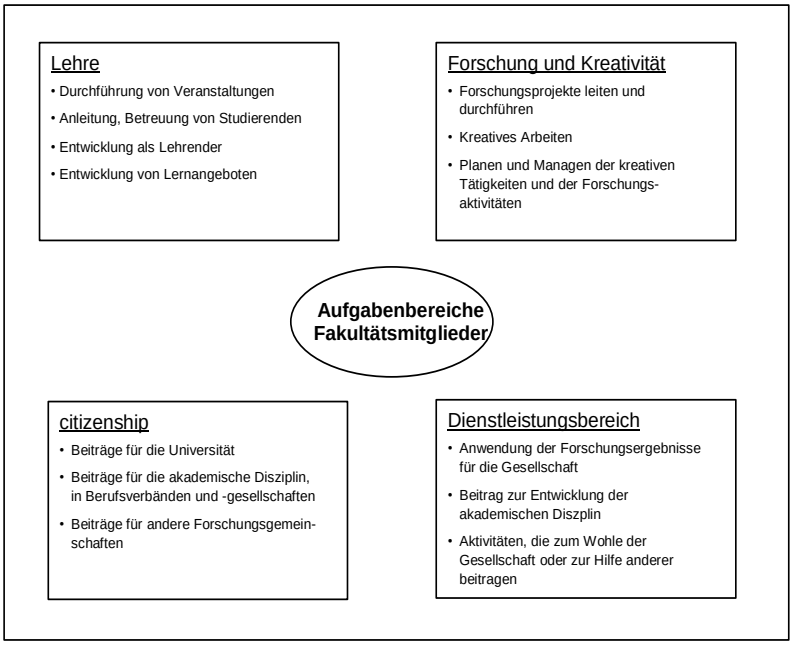
Evaluation auf Fakultäts-, Fachbereichs- oder Institutsebene

Die Gesamtheit der Aktivitäten der Lehrenden an einer Fakultät hat viele Facetten. Verschiedene amerikanische Autoren haben versucht, die jeweiligen Facetten zu identifizieren und zu beschreiben. Eine der differenziertesten Übersichten findet sich bei Braskamp und Ory (1994). Die Autoren gliedern das Geschehen an einer Fakultät in die vier Bereiche

- Lehre (teaching)
- Forschung und Kreativität (research and creative acitivity)
- Dienstleistung (practice and professional service)
- Citizenship

Die Tätigkeiten oder Aufgaben, die diese Bereiche umfassen, werden anhand der weiteren Gliederungspunkte aufgeschlüsselt. Zusammenfassend werden sie in Abbildung 2 vorgestellt.

Abb. 2. Aufgabenbereiche von Fakultätsmitgliedern (in Anlehnung an Braskamp & Ory, 1994)



Andere Klassifikationen versuchen, die Komplexität der Aufgabenfelder ähnlich zu beschreiben (z. B. Arreola, 1995, 2000). Bereits an der vorgestellten Gliederung von Braskamp und Ory ist nachzuvollziehen, dass sie die realen Verhältnisse nur vereinfachend wiedergeben kann. Die mit einer Klassifikation verbundene Informationsreduktion ist sicherlich Aufgabe einer sinnvollen Beschreibung. Gleichwohl wird deutlich, dass es beinahe unmöglich sein dürfte, alle Aspekte der täglichen Arbeit von Lehrenden an einer Fakultät zu erfassen. Folglich ist es äußerst schwierig, die Arbeit an einer Fakultät in ihrer Vollständigkeit und Komplexität angemessen zu bewerten. Weiterhin wird den in der Abbildung 2 aufgeführten Aufgabenbereichen über verschiedene Personengruppen (Dekane, Lehrende, Studierende) oder Hochschulformen (Fachhochschule, Universität) hinweg nicht die gleiche Bedeutung zugemessen. Gerade hier dürften auch zwischen amerikanischen und deutschen Verhältnissen große Unterschiede bestehen.

Vor der Durchführung einer Evaluation sollte daher genau festgelegt werden, welche Aspekte als bedeutsam gelten und somit Gegenstand der Evaluation sein sollen. Nach Arreola (2000) ist es ratsam, dass jeder Fachbereich institutionelle sowie personale Zielvorstellungen formuliert, anhand derer dann die Evaluationsbereiche ausgewählt werden. Werden diese in einem weiteren Schritt noch gewichtet, entsteht das sogenannte «faculty role model», das als Grundlage für den weiteren Evaluationsprozess dienen soll.

Vor diesem Hintergrund wird deutlich, dass die Evaluation der Lehre nur einen Aspekt eines viel komplexeren Aufgabengefüges erfasst. Evaluation von Lehre stellt immer nur einen Teilbereich dar, wenn auch einen wesentlichen – so die Einschätzung vieler Autoren (beispielsweise Bülow-Schramm & Reissert, 1993).

Das Aufgabenfeld Lehre umfasst mehre Aspekte (vgl. Abb. 2), weshalb eine weitere Differenzierung erforderlich ist. Ähnlich wie Webler (1996) unterscheidet Rindermann (1999b) zwischen Lehrevaluation und Lehrveranstaltungsevaluation. *Lehrevaluation* bezieht sich auf die Evaluation der Lehre insgesamt inklusive der Ausbildungsinhalte und Studienbedingungen. *Lehrveranstaltungsevaluation* bezeichnet dagegen die Evaluation konkreter Lehrveranstaltungen, z. B. Vorlesungen oder Seminare. Um diese geht es im nächsten Abschnitt.

Evaluation von Lehrveranstaltungen

Die Durchführung von Lehrveranstaltungen ist eine zentrale Aufgabe aus dem Bereich der Lehre und somit ein naheliegender Ansatz zur Evaluation der Lehre. Gilt doch die Handlungsebene der Lehrveranstaltung als "die Keimzelle" für die Qualität der Ausbildung (Bülow-Schramm & Reissert, 1993). Als Anfang der 1990er Jahre Diskussionen um die Qualität der Lehre an deutschen Hochschulen aufkamen, war es daher nicht verwunderlich, dass die ersten Projekte zur Lehrevaluation genau auf dieser Ebene ansetzten.

Die in dieser Arbeit zu diskutierenden Fragestellungen gehen deswegen ebenfalls von der Ebene der einzelnen Lehrveranstaltung aus. Wenn im Folgenden auch auf die Ebene der Lehrveranstaltungsevaluation fokussiert wird, so bleibt dennoch als Rahmen für diese Arbeit zu beachten, dass in Anbetracht der Komplexität des Lehrbetriebs "die Lehrveranstaltungskritik nur ein kleiner Baustein im Rahmen einer umfassenderen Evaluation sein kann" (Preißer, 1993, S. 515). Eine Verkürzung der Eva-

luation der Lehre auf Lehrveranstaltungsevaluation ist deshalb zu Recht zu kritisieren.

Die Evaluation von Rahmenbedingungen

Die Rahmenbedingungen als Gegenstand von Evaluationen werden selten als eigenständiger Evaluationsbereich aufgeführt. Eine Ausnahme bilden Westermann, Spies, Heise und Wollburg-Claar (1998), die einen eigenen Fragebogen zur Beurteilung von Studienbedingungen (FB-ST) neben oder unabhängig von einem Fragebogen zur Beurteilung einer Lehrveranstaltung (FB-LV)[1] entwickelten.

Bei der Evaluation von Rahmenbedingungen kann zwischen *materiellen und immateriellen Rahmenbedingungen* für gute Lehre unterschieden werden. Erstere beziehen sich beispielsweise auf das Angebot und den Zustand von Gebäude und Räumen (Belüftung, Akustik, Ausstattung, Größe), auf die Qualität der zur Verfügung stehenden Sachmittel (Bibliotheksbestand, technische Hilfsmittel) oder auf die Verfügbarkeit von Hilfen bei der Vorbereitung und Durchführung von Veranstaltungen (Assistenten, Hilfskräfte). Nach Webler (1996) sind jedoch als Voraussetzung für gute Lehre die immateriellen Rahmenbedingungen wichtiger. Diese beziehen sich in erster Linie auf den Stellenwert der Lehre für die Anerkennung des beruflichen Erfolges oder die Rolle der Lehre im Habilitations- und Berufungsverfahren. Ebenso hält Preißer (1993) Probleme bei immateriellen Aspekten für zentral, wie beispielsweise die institutionellen und organisatorischen Rahmenbedingungen, die den Lehraspekten vorangehen. Aufgezählt werden hier u.a. undurchschaubare Prüfungsordnungen, Bürokratisierung der Hochschulen, asymmetrische Kommunikationsstrukturen und fehlendes Managementwissen in der Verwaltung. Ausführliche Diskussionen dieser und verwandter Aspekte finden sich auch bei Webler und Otto (1991) sowie bei Enders und Teichler (1995).

2.2.2 Zielsetzungen von Evaluation — wozu evaluieren

Evaluationen werden aus verschiedensten Gründen durchgeführt. Hauptziele von Evaluation sind nach Callahan (1993, zitiert nach Rindermann, 1996b): Informationsgewinnung, Bereitstellung von Entscheidungsgrundlagen und Einführung von Veränderung oder Verbesserung.

[1] Beide Fragebogen dienen explizit nur der Erfassung der Studien- und Lehrveranstaltungs*zufriedenheit*, da diese Form von Befragung deutlich von einer umfassenden Evaluation unterschieden werden soll (Westermann et al., 1998).

Evaluation folgt damit einer Trias aus Wahrnehmung — Bewertung — Entscheidung. In der Literatur der Lehrveranstaltungsevaluation wird insbesondere das Kriterium der Ziel- und Zweckorientierung von Evaluation (vgl. Abschnitt 2.1) weiter aufgeschlüsselt. So lassen sich auf einer konkreteren Ebene mehrere Ziele von Evaluation unterscheiden. In der Folge führen die verschiedenen Zielvorstellungen zu unterschiedlichen Evaluationsmodellen (s. Tab. 2).

Tab. 2. Evaluationsmodelle und Zielsetzungen (vgl. Gralki & Hecht, 1992, S. 101)

Evaluationsmodell	**Ziele**
Qualifikationsmodell	Optimierung der Lehrqualifikation des Dozen-ten, der Lehrveranstaltung und/oder von Studierenden
Transparenzmodell	Stärken- und Schwächenanalyse auf Ver-anstaltungs-, Fach- oder Universitätsebene
Kommunikationsmodell	Anregung oder Intensivierung der Diskussion zwischen Lehrenden und Studierenden in der Veranstaltung und/oder im Fachbereich
Steuerungsmodell	Informationsgrundlage und Steuerungsinstru-ment bei Entscheidungen über Mittelvergabe, Ausstattung, Bewerbungen, Personalbeurtei-lung
Forschungsmodell	z. B. die Evaluation von Weiterbildungsmaß-nahmen oder Lehr/Lern-Prozessen

Evaluationen können darüber hinaus auch (gewollt oder ungewollt) als Mittel zur Erfüllung rechtlicher Vorschriften, als Rechtfertigung für bereits gefallene Entscheidungen, als Marketinginstrument oder als Taktik bzw. Ablenkungsmittel fungieren (Rindermann, 1996b).

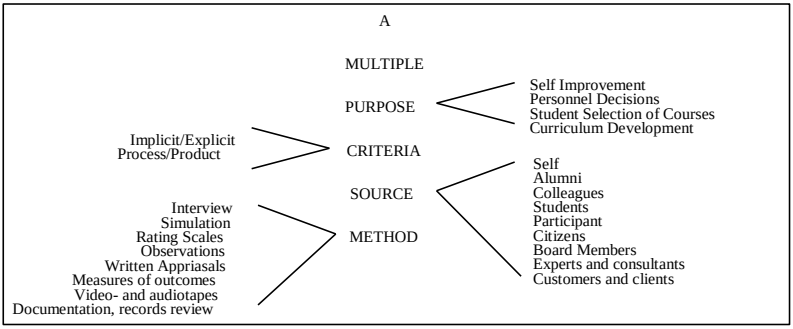
2.2.3 Der multiple Evaluationsansatz — wie evaluieren

Nachdem Gegenstand und Ziel einer Evaluation festgelegt sind, müssen die Informationsquellen und Methoden der Datenerhebung bestimmt werden. An dieser Stelle soll es nicht um eine breite Diskussion verschiedener Möglichkeiten gehen, sondern vielmehr darum, ein grundlegendes Prinzip zu verdeutlichen. Diese Prinzip taucht in der anglo-amerikani-

schen Literatur immer wieder auf und mündet in einem speziellen Evaluationsansatz.

Nach dieser Sichtweise soll die Evaluation nach einem multiplen Ansatz erfolgen. Der Ansatz fordert daher, in Abhängigkeit von den Zielen bzw. dem Zweck einer Evaluation, verschiedene Kombinationen aus Kriterien, Datenquellen und Methoden zu wählen. Hinter der Idee, dass auch die Evaluation von Lehre multipel erfolgen soll steht die Auffassung, dass Informationen, die auf Basis einer einzelnen Methode (z. B. Fragebogen) anhand einer einzigen Quelle (z. B. Studierende) gewonnen wurden, nicht genügen, um die Qualität der Lehre eines Fakultätsmitgliedes oder eines Fachbereiches angemessen beurteilen zu können (vgl. Braskamp et al., 1984, S. 7). Es sollten immer verschiede Kombinationen aus Datenquellen und Methoden in einen umfassenden Evaluationsprozess einfließen. Zahlreiche Experten der anglo-amerikanischen Evaluationsforschung für den Hochschulbereich vertreten diesen Ansatz (vgl. Arreola, 2000; Braskamp et al., 1984; Cashin, 1988, 1995, 1996; Centra, 1993). Abbildung 3 zeigt diesen Ansatz, wie er in den Veröffentlichungen von Braskamp und Mitarbeitern erscheint.

Abb. 3. Der multiple Ansatz der Lehrevaluation (in Anlehnung an Braskamp et al., 1984, S. 30 bzw. Braskamp & Ory, 1994, S. 82)



Natürlich ist es aus praktischer und ökonomischer Sicht nicht möglich, Daten von allen Informationsträgern mit mehreren, jeweils verschiedenen, Methoden zu erheben. Nach Centra (1993) gewährleistet ein multipler Ansatz jedoch, dass schon die Verwendung einiger verschiedener Informationsquellen und Evaluationsmethoden die Qualität der Evaluation erhöht. Ähnlich argumentieren Braskamp et al. (1984) mit dem Hinweis, dass die Integration der verschiedenen Schritte der Datensammlung den

Evaluationsprozess fairer, glaubwürdiger und vollständiger macht. Bei dieser Sichtweise geht es darum, die Einzelinformationen zusammenzufügen und sich ein vollständigeres (Ab-)Bild zu machen. Leitbegriffe, die dieses Verständnis charakterisieren könnten wären Methodenvielfalt und Triangulation. Nach Meinung amerikanischer Autoren scheint der multiple Ansatz der bisher geeignetste Weg, dem Ziel einer zweckorientierten, validen, zuverlässigen, glaubwürdigen und in ihren sozialen Folgen wünschenswerten Evaluation der Lehre nahe zukommen (Braskamp & Ory, 1994).

2.3 Fazit

Für eine Auseinandersetzung mit der Evaluation von Lehre wurde bisher in den Evaluationsbegriff eingeführt und das komplexe Feld der Evaluation im Hochschulbereich durch drei Aspekte gegliedert. Daraus bleibt festzuhalten:

Evaluation stellte sich als ein vielfältiger Begriff dar, weswegen auf eine allgemein verbindliche Definition verzichtet wurde – zugunsten der bei Wottawa und Thierau (1998) aufgeführten allgemeinen Kennzeichen von Evaluation. Weiterhin wurden drei Aspekte vorgestellt, die dazu dienten, Evaluation im Anwendungsfeld Hochschule zu strukturieren. Dabei handelte es sich um die Aspekte Gegenstand, Zielsetzung und Vorgehensweise (multipler Ansatz) von Evaluation. Sie bilden den Hintergrund für die weitere Entwicklung der Arbeit, wobei der Gegenstandsbereich auf die Ebene der Lehrveranstaltung eingegrenzt wurde.

Für ein tieferes Verständnis der Diskussion um die Lehrveranstaltungsevaluation sind die verschiedenen Traditionen und Kontexte dieser Thematik von Bedeutung. Die USA verfügen im Gegensatz zu Deutschland über eine sehr lange Tradition von Lehrevaluation; deren Ergebnisse wurden in Deutschland bislang allerdings kaum rezipiert (Rindermann, 2001). Sie flossen an vielen Stellen in diese Arbeit ein, weswegen im nächsten Kapitel die historischen Entwicklungen und Einflüsse von Lehrevaluation beider Kulturräume, dem amerikanischen und dem deutschen, vorgestellt werden.

3. Historische Entwicklungen und Einflüsse der Lehrevaluation

Die Wurzeln der Evaluation von Lehre liegen nach Centra (1993) im mittelalterlichen Europa, wo eine der ersten formalen Formen der Evaluation der Lehre an den damaligen Universitäten stattfand: Zum einen gab es ein Komitee aus Studierenden, die regelmäßig darüber berichteten, ob der Professor die Behandlung verschiedener Themen zu festgelegten Terminen einhält oder ob er in Verzug gerät. Zum anderen gab es eine (indirekte) Evaluation der Leistung der Professoren, da ihr Gehalt von der Anzahl der zahlenden Studenten abhing. Von ähnlichen Beispielen berichten auch von Friedeburg (1996) und Webler (1996).

Obwohl die ersten Wurzeln von Lehrevaluation in Europa zu finden sind, liegen die *Ursprünge der modernen Ära von Lehrevaluation* in Amerika. Dort begann die wissenschaftliche Auseinandersetzung Anfang des 20. Jahrhunderts. Nach Marsh (1987) gab es vor dieser Zeit nur einige, vereinzelte Publikationen, die sich mit studentischer Lehrevaluation beschäftigen. Seit den Anfängen der Lehrevaluation in den späten 1920er Jahren fand eine kontinuierliche Entwicklung und Ausdifferenzierung des Forschungsfeldes statt. Centra (1993) unterscheidet vier Phasen in der Geschichte der modernen Lehrevaluation: eine frühe Phase von 1920-1960; die 1960er Jahre; die goldene Ära der Evaluationsforschung in den 1970er Jahren und die letzte Phase von 1980 bis zur Gegenwart.

Vorab mögen die verschiedenen Gründungsdaten wissenschaftlicher Evaluationsgesellschaften als interessanter Hinweis für die unterschiedlichen Entwicklungen in Deutschland, Europa und den USA dienen: Die Gründung der Deutschen Gesellschaft für Evaluation im Jahre 1997 erfolgte relativ spät; etwas früher, im Jahre 1994, wurde die Europäische Gesellschaft für Evaluation (European Evaluation Society) ins Leben gerufen. Im Vergleich dazu gründete sich die American Evaluation Association bereits 1986, und dies durch Zusammenschluss der schon seit längerem existierenden „Evaluation Research Society“ (Ohio) und des „Evaluation Network“ (Michigan).

Insbesondere in Deutschland war die Diskussion um Lehrevaluation lange Zeit kein eigenständiges Thema und wurde in seiner Geschichte immer wieder mit anderen Themen verwoben. In einer internationalen Vergleichsstudie aus dem Jahre 1998 bekam Deutschland dann schließlich neben Rußland den letzten Platz in der Lehrevaluation zugewiesen (Rindermann, 2001).

3.1 Die Anfänge von 1920 bis 1960

Herman H. Remmers und Kollegen von der Purdue University publizierten 1927 den wahrscheinlich ersten studentischen Evaluationsfragebogen, die *Purdue Rating Scale for Instructors* (McKeachie, 1990). Remmers ist in Amerika als der Pionier oder Vater der studentischen Lehrevaluation in die Geschichte der Evaluationsforschung eingegangen (Marsh, 1987). Sein Verdienst sind die ersten umfangreichen Studien zu Fragen der Reliabilität studentischer Lehrevaluation mittels Fragebogen (vgl. Remmers, 1928, 1930, 1931, 1934; Remmers & Brandenburg, 1927). Remmers und Kollegen führten u.a. systematische Befragungen von Studenten und Absolventen durch und verglichen die Reliabilitäten des Verfahrens für die verschiedenen Daten. Weiterhin untersuchten sie den Zusammenhang zwischen der Bewertung der Studienleistung durch den Dozenten und den studentischen Bewertungen für diesen Kurs bzw. diesen Dozenten. Neben der Purdue University hatten auch die Harvard University, die University of Washington, die University of Texas sowie andere Hochschulen Mitte der 1920er Jahre begonnen, erste Maßnahmen studentischer Lehrevaluation einzuführen (s. Marsh, 1987).

Die Intention für die Entwicklung von Fragebogen zur Evaluation einer Lehrveranstaltung entstammte der Vorstellung, durch die Veröffentlichung der Ergebnisse den Studierenden eine Hilfestellung für ihre Kurswahl zu bieten (Ory, 1990). Auf die gesamte Anzahl der Universitäten gesehen, wurden die Instrumente zur Lehrveranstaltungsevaluation in der frühen Entwicklungsperiode nur sehr selten eingesetzt, obwohl Remmers und Kollegen den Einsatz solcher Instrumente propagierten. Zugleich vertraten Remmers und Kollegen anfänglich die zurückhaltende Position (vgl. Brandenburg & Remmers, 1927), dass das Verfahren nicht als Maß für Lehreffektivität eingesetzt werden könne und auch nicht für Personalentscheidungen herangezogen werden sollte. Remmers löste sich später von dieser zurückhaltenden Position und schlussfolgerte, basierend auf mehr als 20 Jahren Forschung: "No research has been published invalidating the use of the student opinion as one criterion of teacher effectiveness" (Remmers, 1958 zitiert nach Marsh, 1987, S. 259).

3.2 Die Entwicklungen in den 1960er Jahren

In den Studentenprotesten der 1960er Jahre wurde gefordert, die Lehre an den Universitäten zu verbessern. In Deutschland mündeten diese Proteste in eine breite und langanhaltende Hochschulreformdiskussion, die jedoch erst in den 1970ern ihre ersten Früchte trug. Anders in den USA: Dort

war die Evaluation von Vorlesungen und Seminaren eine direkte Reaktion auf die Forderungen der Studierenden einzugehen und ihnen Gehör zu schenken (Centra 1993). Viele amerikanische Institutionen folgten den Vorreitern wie der Harvard University und begannen mit der Entwicklung eigener Instrumente. Doch die Qualität der damaligen Verfahren war oftmals fragwürdig. "Insbesondere in den 1960ern sind Kollegen mit schlecht konstruierten Evaluationsverfahren durchgekommen." (vgl. Centra, 1979, S.2).

Die von den amerikanischen Studenten eingeforderte Verbesserung der Lehre drang erst nach und nach in Köpfe der Lehrenden und Verantwortlichen; ganz allmählich machte sich ein Umschwung bemerkbar - weg von einer Art Alibifunktion hin zu einer veränderten Haltung: zur fundierten Konzeption von Instrumenten sowie zur zunehmend häufigeren Durchführung von Befragungen. Denn die Administratoren begannen sich für die Ergebnisse der Befragungen zu interessieren. Manchmal mündeten die Ergebnisse bereits in Personalentscheidungen ein (Ory, 1990). Die Durchführung von Befragungen stellte aber auch in dieser Phase noch immer eine unregelmäßige Maßnahme dar und war durchweg freiwillig. Insgesamt jedoch hatte sich das Interesse für Lehrveranstaltungsevaluation enorm gesteigert und viele Universitäten strebten den dauerhaften Einsatz dieser Verfahren an. Diese Stimmungslage war schließlich der Ausgangspunkt für eine neue amerikanische Forschungswelle in den 1970er Jahren (vgl. Centra, 1993).

3.3 Die Entwicklungen in den 1970er Jahren

Die 1970er Jahre gelten als die goldene Ära amerikanischer Lehrveranstaltungsevaluation. In dieser Phase wurden Studierende als eine der wichtigsten Quellen im Evaluationsprozess betrachtet. Eine breit angelegt Studie von Centra aus dem Jahre 1979 ergab, dass die Wichtigkeit studentischer Lehrevaluation aus Sicht der Dekane neben der Evaluation durch Peers und ihrem eigenen Urteil unter den ersten drei Rängen rangiert. Viele umfassend angelegte Studien untersuchten eine Bandbreite von Fragestellungen hinsichtlich Konzeption, Validität, Reliabilität, systematischer Verzerrungen und Nützlichkeit der Verfahren. Die Untersuchungsergebnisse wiesen überwiegend positive, die Verfahren und ihren Einsatz unterstützende Befunde auf (Centra, 1979, 1993).

In die 1970er Jahre fällt auch die Gründung des *Joint Committee on Standards for Educational Evaluation* (1975). Diese inzwischen weit verbreiteten und in Amerika allgemein akzeptierten Evaluations-Standards be-

schreiben die erwünschten Qualitätsmerkmale von Evaluation. Sie dienen als Regelwerk für die Beurteilung von Evaluationsleistungen und ggf. notwendigen Sanktionierungen. Ein weiterer Punkt, der vermutlich zu der Blüte studentischer Lehrevaluation geführt haben dürfte, war der zu verzeichnende enorme Kostenanstieg im Bereich der höheren Bildung (Braskamp & Ory, 1994). Durch finanzielle Probleme bedrängt, sahen viele Administratoren einen Ausweg in der systematischen Evaluation, deren Ergebnisse z. B. in Personalentscheidungen mit einflossen (Seldin, 1984).

Ende der 1970er Jahre war die Einführung von Evaluationsverfahren in Amerika weit voran geschritten. Die überwiegende Mehrheit der Universitäten setzte Ende der 1970er Jahre Fragebogenverfahren zur Evaluation von Lehrveranstaltungen ein. Zugleich waren die 1970er Jahre auch die Zeit, in der der Einsatz der Fragebögen zur Evaluation von Lehrveranstaltungen in Amerika kritisch diskutiert wurde. So empfahlen Evaluationsexperten der Zeit wie zum Beispiel K. O. Doyle oder John A. Centra den Einsatz der Verfahren, obwohl sie zugleich auch auf die Grenzen und Gefahren eines Fragebogeninstruments hinwiesen (vgl. Doyle, 1975; Centra, 1979).

In der *Bundesrepublik Deutschland* setze die Beschäftigung mit studentischer Evaluation von Lehrveranstaltungen erst in den 1970er Jahren in nennenswertem Umfang ein (Diehl, in Druck). Erste Fragebogenverfahren wurden entwickelt und überprüft (z. B. Basler, 1978; Diehl & Kohr, 1977; Kleine & Merkens, 1979; Müller-Wolf, 1977; Sommer & Petermann, 1978).

Die Kritik an der Lehre wurde allerdings vor dem Hintergrund einer Umstrukturierung des gesamten Universitätsbetriebes diskutiert. Verfahren, Methoden und Instrumente zur Überprüfung der universitären Ausbildung wurden deshalb vor allem unter den Gesichtspunkten „Wie wird gelehrt und gelernt“ betrachtet (vgl. Hochschuldidaktische Arbeitspapiere z. B. von Bülow, 1977, Meyer-Althoff, 1978). Studienreform und Hochschuldidaktik waren die Schlagworte der Zeit. „Lehrevaluation diente u.a. dazu, die Qualität der Ausbildung zu hinterfragen und um die studienreformtechnischen Ansätze zu rechtfertigen und abzusichern“ (Bülow-Schramm & Reissert, 1993, S. 399). Doch rückten diese Ansätze der inhaltlichen und didaktischen Studienreform in Deutschland angesichts anderer Probleme während der 1980er in den Hintergrund (s.u.).

3.4 Aktuelle Entwicklungen von 1980 bis zur Gegenwart

In den USA setzten sich die Entwicklungen aus den vorangegangenen Dekaden auf dem Gebiet der Lehrevaluation kontinuierlich fort. Es wurde vor allem angestrebt, die teils heterogenen Forschungsergebnisse zu bündeln und gesicherte Erkenntnisse in eine sinnvolle Praxis konsequent umzusetzen. Folgende Titel wichtiger amerikanischer Publikationen stehen exemplarisch für die Einstellung der Zeit : „Handbook of Teacher Evaluation“ von Millman (1981), „Evaluating Teaching Effectiveness“ von Braskamp, Brandenburg und Ory (1984) sowie „Changing Practice in Faculty Evaluation“ von Seldin (1984). Ebenso waren die 1980er Jahre die Zeit der ersten großen Literatur Reviews und der Metaanalysen (Cohen, 1980, 1981; Marsh, 1984, 1987).

Die meisten Anwender sahen in den Forschungsergebnissen die Reliabilität und Validität der Verfahren als ausreichend gesichert und betrachteten die Befragung Studierender durch Fragebogenverfahren als ein sinnvolles Instrument im Evaluationsprozess (Ory, 1990). Einen Punkt, der die Haltung der Anwender ebenfalls mitgeprägt haben dürfte, formuliert Eble (1984). Nach Meinung des Autors lag der Grund für den Einsatz von Fragebogen weniger in der Tatsache, dass die Administratoren von der Qualität des Verfahrens vollständig überzeugt gewesen wären. Vielmehr führt er den hohen Verbreitungsgrad darauf zurück, dass die schriftlichen Ergebnisse einer solchen Befragung unter Umständen sogar im Gericht angeführt werden können. Ende der 1980er Jahre gab es nur noch einzelne Universitäten in Nordamerika, die keine studentischen Lehrevaluationen durchführten (Abrami, 1989).

Seit den 1990er Jahren ist die Erhebung von Daten zur Qualität der Lehre an amerikanischen Colleges und Universitäten zumeist verpflichtend. Die studentische Lehrevaluation mittels Fragebogen wird als valider, wenn auch nur als *ein* Indikator im Evaluationsprozess betrachtet. Nichts desto weniger ist die Evaluation von Lehrveranstaltungen durch die Befragung Studierender nach wie vor ein Thema wissenschaftlicher Auseinandersetzung (beispielsweise Abrami, 2001; Arreola, 2000; Greenwald, 1997; Hoyt & Pallett, 1999; McKeachie, 1997; Marsh & Roche, 1997). So werden beispielsweise die folgenden Fragen gegenwärtig diskutiert: Inwiefern können die Ergebnisse studentischer Lehrevaluation auch sinnvoll für Personalentscheidungen verwendet werden? Unter welchen Bedingungen führen Rückmeldungen zu Verhaltensmodifikationen?

In der Bundesrepublik beherrschten vornehmlich „Oberflächenphänomene“ (Webler, 1996) die hochschulpolitischen Debatten der 1980er

Jahre. Überfüllungsprobleme, Ressourcenprobleme aufgrund immer knapper werdender Haushaltsmittel und zu lange Studienzeiten setzten die Akzente. Die Diskussion der Lehrqualität geriet in den Hintergrund. Das änderte sich nahezu schlagartig, als Ende 1989 das Magazin DER SPIEGEL eine Umfrage unter Studierenden zur Studiensituation mit abschließendem Hochschulranking veröffentlichte. Neuauflagen und weitere Umfragen dieser Art folgten (SPIEGEL, 1993; STERN, 2001; FOCUS, 1993) und wurden für ihre methodischen Mängel vielfach kritisiert (z. B. Gräf, 1991; Kriz, 1995; Lamnek, 1990; Neidhardt, 1990, 1991). Seither „ist die Debatte um die Lage der Lehre an deutschen Hochschulen nicht mehr nur ein periodisch wiederkehrendes Thema des Feuilletons, sondern Gegenstand verschiedenster hochschulpolitischer Initiativen und vieler wissenschaftlicher Versuche zur Evaluation der Lehre geworden“ (Hornbostel & Daniel, 1995, S. 29).

Seit der ersten Hälfte der 1990er Jahre führten Wissenschafts- und Interessenverbände Tagungen mit dem Ziel durch, die Lage zu diagnostizieren, sprachen Empfehlungen aus und setzen teils umfangreiche Modellprojekte und Förderungsprogramme an. Im Unterschied dazu war die Einführung von überwiegend einheitlichen Evaluationsmaßnahmen in den europäischen Nachbarländern Frankreich, den Niederlanden und Großbritannien ein eigenständiges Thema – bereits in den 1980er Jahren. Die Maßnahmen in Deutschland erreichten nach Webler (1996) finanzielle Volumina von 1-17 Mio. DM. Übersichten zu den zahlreichen Initiativen und den detaillierten Entwicklungen geben Webler (1992, 1996) sowie verschiedene Autoren im Band 4 der Beiträge zur Hochschulforschung (1993).

Bisher handelt es sich bei all diesen Initiativen um Modelle mit „Vorreiterfunktion“; zu den langfristigen Wirkungen gibt es noch keine systematischen Analysen. Festzustellen ist immerhin eine deutlich gestiegene Problemwahrnehmung im Bereich der Lehre. Des Weiteren gibt es erste Ansätze, diese Initiativen zu bündeln, sie in einheitliche Modelle zu integrieren und in Evaluationsverbünden zusammenzuschließen[2]. Welche Rolle die Befragung Studierender zu ihrer Einschätzung der Qualität der Lehre in den verschiedenen Formen von Evaluation, z. B. in Lehrberichten[3] oder den Modellen der internen und externen Evaluation[4] einnimmt,

[2] Beispiele hierfür sind der *Verbund Norddeutscher Universitäten*, die *Zentrale Evaluationsagentur der niedersächsischen Hochschulen* (ZEvA), das *Bündnis für Lehre* der Universitäten Mannheim und Heidelberg oder fest integrierte Evaluationszentren bzw. -büros (Mainz, Halle).

[3] Möglichen Ziele und Konzeptionen von Lehrberichten erörtert Habel (1995).

ist nicht einmal annähernd geklärt (exemplarisch Habel, 1995). Einzig in Nordrhein-Westfalen sind studentische Lehrveranstaltungsbewertungen bereits verpflichtend (Webler, 1996). Manche Autoren sind der Ansicht, dass sich die Hochschulen der Lehrevaluation auf Dauer nicht mehr entziehen können (Diehl, in Druck). Dies wäre spätestens mit der Einführung des neuen Besoldungsgesetzes für Professoren erreicht, da es explizit eine Bewertung der Lehrleistung vorsieht. Für die wissenschaftliche Auseinandersetzung in Deutschland hält Diehl (in Druck) insgesamt fest, dass die Anzahl der Originalarbeiten und Übersichten zwar zunehmend wachse, aber bislang weit unter der des anglo-amerikanischen Raumes liegt.

3.5 Zusammenfassender Vergleich: Entwicklungen in den USA und in Deutschland

Die amerikanische Geschichte der Entwicklung der Fragebogen zur Lehrveranstaltungsevaluation kann auf eine langjährige, ungebrochene Forschungstradition zurückblicken. Seit den 1920er Jahren ist sie kontinuierlich systematisiert und verstärkt worden. Amerikanische Lehrevaluationsforschung lässt sich charakterisieren als ein "steady move to more care in gathering data, more attention to ruling out prejudice and subjectivity, and more involvement of those actually affected by the process" (Eble, 1984, S. 96). Heute ist die Evaluation von Lehrveranstaltungen durch Befragung der Studierenden in den USA fest etabliert, vielerorts verpflichtend.

Im Unterschied zu den europäischen Nachbarländern war die deutsche Geschichte der Lehrevaluationsforschung von jeher in die Diskussion um Reformen der Hochschulen eingebettet (Webler, 1996). Der aktuelle Stand der Entwicklungen zur Lehrevaluation in Deutschland ließe sich kennzeichnen mit Pluralismus, Heterogenität und Unübersichtlichkeit – sowohl in den Forschungsaktivitäten als auch in den Maßnahmen an den jeweiligen Hochschulen.

---

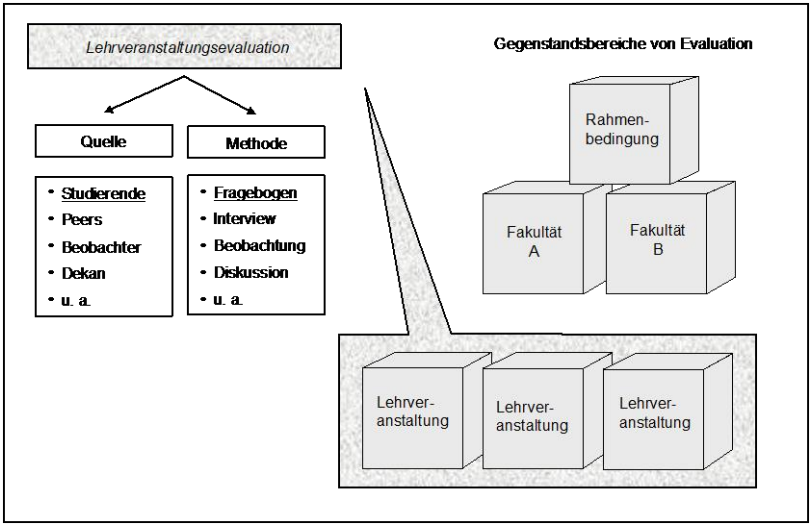
[4] Über verschiedene Modelle der internen und externen Evaluation geben Reissert (1994) sowie Carstensen & Reissert (1997) Auskunft.

4. Zwei Aspekte studentischer Veranstaltungsbeurteilungen

Der multiple Ansatz der Lehrevaluation (vgl. Abschnitt 2.2.3) weist unter anderem auf die Unterscheidung der beiden Aspekte «Quelle» und «Methode» hin. Postuliert wird, dass multiple Verknüpfungen von Quellen, Methoden und Kriterien die Güte eines Evaluationsprozesses erhöhen. Für die verschiedenen Kombinationen stehen mehrere Methoden und mehrere Quellen zur Auswahl (vgl. Abbildung 3).

Betrachtet man die Ebene der Lehrveranstaltungen als Gegenstand von Evaluation, so erwecken diesbezügliche Veröffentlichungen oftmals den Eindruck, dass Lehrveranstaltungsevaluation mit der Befragung von Studierenden via Fragebogen gleichzusetzen sei. Dabei handelt es sich um eine Reduzierung, die zu Recht von einigen Autoren kritisiert wird (vgl. Kromrey, 2001). Die nachfolgende Abbildung veranschaulicht die erwähnte Fokussierung auf studentische Lehrveranstaltungsevaluation via Fragebogen.

Abb. 4. Quellen und Methoden für die Evaluation von Lehrveranstaltungen – die Fokussierung auf studentische Lehrveranstaltungsevaluation via Fragebogen.



Bei der Fokussierung auf Lehrveranstaltungsevaluation wird häufig nicht deutlich genug zwischen den beiden Aspekten «Quelle» und «Methode» unterschieden. Die Wichtigkeit der Differenzierung soll an einem einfachen Beispiel erläutert werden: Es ist theoretisch möglich, dass Studierende als Quelle für die Evaluation von Lehrveranstaltung von Bedeutung sind, dass aber die Fragebogenmethode Schwächen aufweist – oder umgekehrt.

Die beiden nächsten Abschnitte widmen sich daher überblicksartig der Frage, welche Bedeutung Studierende als Quelle für die Erhebung von Evaluationsdaten haben und welche Methoden studentischer Veranstaltungskritik eingesetzt werden können.

4.1 Studierende als Informationsquelle

Eine naheliegende Informationsquelle für die Evaluation von Lehrveranstaltungen sind Studierende. Offensichtlich ist, dass Studierende als Adressaten der Lehre eine einzigartige Quelle darstellen. Sie haben mehr als jede andere Gruppe (Absolventen, andere Fakultätsmitglieder) die Möglichkeit, das Lehrgeschehen zu verfolgen. Auffällig sind aber auch die über Jahre hinweg im anglo-amerikanischen wie auch im deutschen Sprachraum immer wieder auftauchenden Zweifel an der Fähigkeit der Studierenden, Veranstaltungscharakteristika und Charakteristika der Lehrenden zu beschreiben und zu bewerten. Diese Zweifel stammen aus allen Interessengruppen mit Ausnahme der Studierenden selbst: Professoren, Forscher und Administratoren bringen ihre Vorbehalte seit jeher in Artikeln und Kommentaren zum Ausdruck (z. B. Detchen 1940; Kromrey, 1995, 1996a).

Jedoch ist nach Braskamp und Ory (1994) die Glaubwürdigkeit der Studierenden hinsichtlich ihrer Beschreibungs- und Urteilsfähigkeit keine Frage, die sich allgemein mit "ja" oder "nein" beantworten ließe. Statt dessen gehe es darum, zu identifizieren und zu differenzieren, in welchen Bereichen Studierende glaubwürdige und zuverlässige Beschreibungen bzw. Bewertungen oder Einschätzung vornehmen können.

Einen solchen Ansatz beschreibt Arreola (2000). Er präsentiert eine 12-Felder Matrix, die im Sinne des multiplen Evaluationsmodells als Hilfsmittel genutzt werden kann, um die Datenquellen auszuwählen, die für die Evaluation einzelner oder mehrerer Rollenkomponenten von Lehre herangezogen werden können. Die 12 Felder-Matrix ergibt sich aus der Kombination von drei möglichen Informationsquellen (Studierende,

Peers, Dekan) sowie aus vier von Arreola definierten Rollenkomponenten von Lehre (vgl. Arreola, 2000, S. 11ff.):

- Fachliche Expertise (content expertise) – Fähigkeiten, Wissen und Erfahrung des Lehrenden in Bezug auf das jeweilige Fachgebiet
- Informationsübermittlung / Lehrkompetenz (instructional delivery skills) – die Fähigkeiten des Lehrenden im Rahmen der Interaktion mit den Studierenden
- Fähigkeiten zur Planung und zum Aufbau einer Lehrveranstaltung – die technischen Fähigkeiten des Lehrenden
- Kursmanagement – die bürokratischen Fähigkeiten des Lehrenden, eine Veranstaltung zu führen und zu managen.

Wie aus dieser Matrix hervorgeht (vgl. Abb. 5), sind Studierende die einzige Quelle, die sinnvollerweise zur Erfassung des Prozesses der Informationsvermittlung (instructional delivery) genutzt werden sollte.

Abb. 5. Ressourcen-Identifikations-Matrix zur Evaluation verschiedener Komponenten von Lehre (Arreola, 2000, S. 28).

	Source		
Role components	*Students*	*Peers*	*Department Head*
Instructional Delivery Skills	Yes	No	No
Instructional Design Skills	Yes	Yes	No
Content Expertise	No	Yes	Yes
Course Management	No	No	Yes

Neben Arreola bieten auch Braskamp und Ory (1994) sowie Cashin (1989) teilweise sehr elaborierte Übersichten an. Gemeinsam ist allen Autoren, dass sie in Studierenden eine wichtige Quelle für die Erfassung bestimmter Aspekte des Lehrens sehen. Perspektiven ihrer Wahrnehmung bzgl. dieser Aspekte können in einigen Fällen durch keine andere Quelle abgedeckt werden. Studierende als Informationsquelle sollten deshalb in

keiner Lehrevaluation fehlen (Braskamp & Ory, 1994; Hoyt & Pallett, 1999).

In Deutschland gibt es bisher keine Auflistungen dieser Art. Die Diskussion um die Urteilskompetenz der Studierenden wird in Kapitel 6 noch einmal im Zusammenhang mit der Frage der Homogenität der Urteile der Studierenden aufgegriffen. Vorab bleibt festzuhalten, dass obwohl die Urteilskraft der Studierenden im Sinne einer objektiven Wahrnehmung von einigen Autoren grundsätzlich bezweifelt wird (vgl. Rosemann, 1999), Studierende dennoch als wichtige Beteiligte im Evaluationsprozess gelten. (vgl. Diehl, 1989; vgl. auch Webler, 1993).

4.2 Fragebogen und andere Methoden studentischer Veranstaltungskritik

Zur Erfassung der studentischen Einstellungen zur universitären Lehre bieten sich theoretisch mehrere Methoden an. Zu nennen sind beispielsweise Interviewtechniken, Diskussionen, die schriftliche Beantwortung offener Fragen, Beobachterrollen und Beobachtungsprotokolle etc. (vgl. Richter, 1994). Bei Cashin (1989) werden Studierende zum Beispiel als geeignete Quelle bei der Methode der Fragebogenerhebung, des Interviews oder der kurzen schriftlichen Befragung aufgeführt.

Die Evaluationstechniken weisen jeweils unterschiedliche methodische Niveaus auf und bilden insgesamt ein vielfältiges Methodenspektrum für unterschiedlichste Zielsetzungen (Braskamp & Ory, 1994; Richter, 1994; Webler, 1993). Ein ausführlicher Vergleich der verfügbaren Techniken der (nicht nur studentischen) Lehrevaluation wäre sicherlich lohnend, sprengte jedoch den Rahmen dieser Arbeit.

Meist unweigerlich verknüpft mit dem Ziel der Berücksichtigung der studentischen Einschätzungen ist die Methode des Fragebogenverfahren. Viele Autoren kritisieren diese "Verengung der 'Evaluation' der Lehre auf die reine Durchführung von Lehrveranstaltungskritiken sowie die damit einhergehende Konzentration der Diskussion auf Befragungs- und Auswertungstechniken" (Preißer, 1993, S. 508).

Wegen der hohen Augenscheinplausibilität und des hohen Verbreitungsgrades der Methode der studentischen Lehrveranstaltungsevaluation *mittels Fragebogen*, soll diese in der vorliegenden Arbeit kritisch diskutiert werden. Dazu werden im nächsten Kapitel zunächst einige allgemeine Aspekte von Fragebogenverfahren zur Lehrveranstaltungsevaluation vorgestellt.

5. Fragebogenverfahren zur Lehrevaluation

5.1 Überblick

In den anglo-amerikanischen Ländern sind Fragebogenverfahren zur Lehrveranstaltungsevaluation fest etabliert und eine der häufigsten Strategien zur Lehrevaluation (Ellis, 1993; Seldin, 1984). Seldin (1995) berichtet, dass über 85 Prozent der Colleges und Universitäten in den USA Fragebögen zur Lehrveranstaltungsevaluation an Studierende austeilen. Seither hat diese Anzahl ständig zugenommen. Heute existieren nur noch wenige Universitäten in den USA, die dieses Verfahren nicht einsetzen. Auch in Deutschland ist die Befragung Studierender mittels Fragebogen die gängigste Form der Lehrevaluation (Bülow-Schramm & Reissert, 1993) bzw. „das einzige Verfahren, das zur Bewertung der Lehrqualität eines Dozenten tatsächlich an vielen Universitäten eingesetzt wird" (Schweer, 2001, S. 159). Dieser hohe Verbreitungsgrad – auch wenn Deutschland noch immer hinter den USA weit zurückliegt – erscheint plausibel, sobald man die Vorteile eines Fragebogenverfahrens betrachtet:

- Fragebogenkonstruktionen sind relativ ökonomisch und gut „machbar" im Sinne einer „technischen Leichtigkeit" (Bülow-Schramm & Reissert, 1993)
- Der Aufwand für die Durchführung einer Befragung im Vergleich zu anderen Verfahren (etwa Interviewstudien) ist gering. Handelt es sich um eine internetbasierte Befragung, reduziert sich der Aufwand insbesondere bzgl. der weiteren Datenverarbeitung nochmals erheblich. Bei einer solchen Online-Evaluation geben die Studierenden ihre Einschätzungen in einen Home-PC oder einen Computer auf dem Campus ein.
- Fragebogen sind kostengünstig und schnell auszuwerten, insbesondere wenn es sich um maschinenlesbare Bögen handelt.
- Daten aus Fragebogenverfahren sind bei geschlossenen Itemformaten leicht quantifizierbar und stellen somit ein (oft gewähltes und allgemein akzeptiertes) Mittel der Informationsverdichtung dar.

Neben diesen der Methode inhärenten Vorteilen dürfte ein weiterer Grund für die „Beliebtheit" von Evaluation mittels Fragebogen sein, dass sich Ergebnisse gut für werbewirksame Publikationen heranziehen lassen – in den USA etwa für die Werbung neuer Studierender, in Deutschland zur Erfüllung der geforderten Rechenschaft gegenüber der Öffentlichkeit. So interessierten sich beispielsweise für eines der ersten Evaluationsprojekte in Mannheim neben anderen Hochschulen auch „Presse, Rundfunk und

Fernsehen“ (Klein, 1993, S. 431). Vor diesem Hintergrund wundert es nicht, dass die ersten Pilotprojekte zur Lehrevaluation aus den 1990ern gerade bei der Erhebung via Fragebogen ansetzten (s.a. Abschnitt 2.4).

Doch gerade in Anbetracht des hohen Verbreitungsgrades gilt zu fragen: Sind Fragebogenerhebungen in der Lage, die mit ihnen verknüpften Ziele und Erwartungen zu erfüllen? Welche Zielsetzungen können mit dem Einsatz studentischer Lehrevaluation verfolgt werden? Stehen qualitativ gute Instrumente zur Verfügung?

Nur vor dem Hintergrund der Klärung dieser und ähnlicher Fragen ist es möglich, Sinn oder Unsinn des Einsatzes von Fragebogen in der studentischen Lehrevaluation zu bestimmen. Die nächsten Abschnitte versuchen daher, anhand verschiedener Kriterien einen Überblick über die „heterogene Landschaft“ der Verfahren zu geben.

5.2 Aspekte von Fragebogen zur Lehrevaluation

Die Fragebogenverfahren, die zur Lehrevaluation eingesetzt werden, sind heterogen. Es gibt weder im anglo-amerikanischen noch im deutschsprachigen Raum einen allgemein anerkannten Standard. Das Resultat ist eine inhomogene Landschaft unterschiedlichster Instrumente. Mögliche Unterscheidungsaspekte für einen fundierten Überblick sind die verschiedenen Zielsetzungen, der formale Aufbau, die theoretische Konzeption und die psychometrischen Eigenschaften der Verfahren.

5.2.1 Zielsetzungen und Funktionen

Im Rahmen der in Kapitel 1 vorgestellten Evaluationsmodelle lassen sich die in der Literatur genannten Einsatzzwecke für Lehrevaluationsfragebogen übersichtlich gliedern. Die folgende Tabelle gibt mögliche Zielsetzungen und die dazu korrespondierenden Evaluationsmodelle an. Letztere bestimmen das jeweilige Verständnis vom Zweck der Evaluation (s. Tab. 3, S. 27).

Während der Einsatz zu all diesen Zwecken im anglo-amerikanischen Raum anzutreffen ist, findet der Einsatz von Fragebogen in Deutschland überwiegend im Rahmen eines Feedback-Ansatzes statt (Bülow-Schramm & Reissert, 1993).

Viele Autoren deutschsprachiger Verfahren wenden sich explizit gegen die Verwendung ihrer Instrumente im Rahmen eines Steuerungsmodells

oder eines Qualifikationsmodells zu Vergleichszwecken (z. B. Diehl, 1998; Reischmann, 1995; Todt & Götz, 1997). Weiterhin von Bedeutung ist, dass nicht jeder Fragebogen die in Tabelle 3 genannten Ziele gleichermaßen gut erfüllen kann (Cashin, 1990). Die Verfahren sollten nur zweckgebunden eingesetzt werden gemäß der Ausrichtung für die sie konzipiert sind (Marsh, 1984).

Tab. 3. Zielsetzungen für den Einsatz von Fragebogen zur Lehrevaluation und korrespondierende Evaluationsmodelle

Ziele von Fragebogen zur Lehrevaluation Marsh (1991)	**Evaluationsmodell**
Informationsgrundlage für Personalentscheidungen	Steuerungsmodell
Feedback an den Veranstaltungsleiter, Stärken und Schwächen der Veranstaltung	Kommunikations-, Transparenz- und Qualifikationsmodell
Ergebnis oder Prozessbeschreibung über den Lehr/Lern-Prozess zu Forschungszwecken	Forschungsmodell
Informationen für Studierende als Hilfsmittel für die Wahl von Veranstaltungen	Transparenzmodell

Diese Ausführungen machen deutlich, wie wichtig es ist, Ziele und Erwartungen, die mit dem Einsatz eines Fragebogens verbunden sind, *vor* dessen Einsatz klar und deutlich an alle Beteiligten zu kommunizieren. Nur so kann Akzeptanz für den Einsatz von Fragebögen gewonnen werden.

5.2.2 Formaler Aufbau

Ein Fragebogen zur Lehrveranstaltungsevaluation besteht aus einem Itemset, durch das Einstellungen, Einschätzungen oder Wahrnehmungen der Studierenden hinsichtlich verschiedener Aspekte der Lehre, der lehrenden Person, den von der lehrenden Person verwendeten Materialien, der eigenen Person oder der Umgebungsbedingungen bzgl. einer Lehrveranstaltung erfasst werden sollen.

Eine im amerikanischen Sprachraum oft benutzte Klassifizierung unterscheidet zwischen drei Haupttypen von Fragebogenverfahren zur Lehrevaluation oder den *student ratings of instruction* (vgl. Braskamp et al.,

1984; Braskamp & Ory, 1994). Auch deutschsprachige Autoren nehmen Begriffe dieser Kategorisierung auf (vgl. Bülow-Schramm & Reissert, 1993; Webler, 1993). Es kann zwischen der Omnibus Form, dem Cafeteria System und der Ziel-basierten Form differenziert werden.

- Die (multidimensionale) *Omnibus Form* beinhaltet ein definiertes Itemset, das verschiedene Hauptmerkmale oder Hauptbereiche von Lehre repräsentiert. Items dieser Verfahren wurden meist durch statistische Analysemethoden in verschiedene Cluster gruppiert und Dimensionen des Konstruktes "Lehre" zugeordnet. Typische Dimensionen sind «Organisation» oder «Kommunikation». Eine Variante der Omnibus Form sind *multiple Standardformen*, wobei für verschiedene Veranstaltungstypen fest definierte Omnibusformen mit jeweils veranstaltungsspezifischen Items existieren (z. B. eine Variante für Seminare, eine andere für Vorlesungen).
- Das *Cafeteria System* stellt einen umfangreichen Itempool dar, aus dem die Lehrenden die ihrer Meinung nach relevanten Items auswählen, in einem Fragebogen zusammenstellen und zur Evaluation ihrer Veranstaltung einsetzen.
- Bei der *Ziel-basierten Form* repräsentieren die Items im Vorfeld festgelegte Veranstaltungsziele wie beispielsweise den Erwerb von Faktenwissen oder die Entwicklung spezifischer Fähigkeiten. Studierende bewerten dann ihren durch die Veranstaltung erzielten individuellen Fortschritt auf den verschiedenen Zielsetzungsdimensionen.

Die Anzahl der Items pro Verfahren variiert erheblich. So umfasst der Fragebogen zur Beurteilung von Vorlesungen (VBVOR) von Diehl (1998) 16 Items. Rindermann (1996b) berichtet von Verfahren mit Umfängen von bis zu 206 Items. Werden die Items verschiedenen Faktoren oder Skalen zugeordnet, handelt es sich um ein multidimensionales Verfahren. Die meisten Instrumente beinhalten daneben auch einige sogenannte globale Items[5].

In der Regel überwiegen geschlossene Itemformate, d.h. dass die Teilnehmer ihre Reaktionen auf einer meist 3 bis 7-stufigen Antwortskala festhalten. Die meisten Verfahren stellen zusätzlich ein bis zwei offene Fragen oder lassen Raum für Anmerkungen.

[5] Ein Beispiel für ein globales Item ist die Frage: „Wie würden Sie die Qualität dieser Lehrveranstaltung insgesamt beurteilen?"

5.2.3 Konzeption

Für die Konzeption eines Lehrevaluationsfragebogens bieten sich verschiedene Möglichkeiten zur Generation von Items an. Tabelle 4 gibt in Anlehnung an die Überblicksliteratur von Marsh (1987) und Rindermann (2001) Auskunft über typische Vorgehensweisen.

Tab. 4. Ansätze zur Itemgenerierung für Lehrevaluationsfragebogen

	Ansätze zur Itemgenerierung für Lehrevaluationsfragebogen
1.	*Synkretistischer-Ansatz* Orientierung an bereits existierenden Instrumenten, Adaption eines Instrumentes
2.	*Befragungs-Ansatz* Befragung Studierender und / oder Lehrender als (Praxis-)Experten nach Kennzeichen des „idealen" Dozenten, der „idealen" Veranstaltung oder direkt nach Kennzeichen „guter Lehre"
3.	*Theoretischer-Ansatz* Ableitung aus hochschuldidaktischen Theorien / Theorien über den Lehr- bzw. Lernprozess / Modellen der Unterrichtsforschung
4.	*Experten-Ansatz (Literaturschau)* Verwendung von in der Literatur genannten Konzepten und Ideen, die andere Experten für sinnvoll halten
5.	*Konsens-Ansatz* Gremien, Fachschaften oder andere Gruppierungen einigen sich intern auf einen (meist ad-hoc erstellen) Fragebogen und dessen Einsatz
6.	*Autoren-Ansatz* In Eigenregie entwickeltes Verfahren anhand impliziter oder expliziter Annahmen

In der Regel besteht die Konzeption aus der Kombination von zwei oder mehreren der genannten Vorgehensweisen. In den meisten Fällen werden diese Verfahren weiterhin mit faktorenanalytischen Methoden kombiniert, um die endgültige Auswahl der Items sowie ihre Zuordnung zu Skalen zu bestimmen. Nach Rindermann (2001, S. 61) stellt die synthetische Vorgehensweise „das Mittel der Wahl zur Entscheidung eines solchen Instrumentes dar". Diese Vorgehensweise „berücksichtigt verschiedene Quellen bei der Itementwicklung, zur Generierung und anschließenden

Überprüfung wird ein theoretisch-empirisches Vorgehen gewählt. Ein multifaktorielles Modell der Lehrveranstaltungsqualität sollte den konzeptionellen Rahmen des Entwicklungsverfahrens bilden“ (Rindermann, S. 2001, S. 60).

Es existieren keine exakten Zahlen oder Schätzungen über die Häufigkeiten der einzelnen Ansätze. Jedoch lassen sich verschiedene „Trends“ ausmachen. Dem von Rindermann beschrieben Königsweg dürften bisher nur wenige beschritten haben. So gab es nach Marsh (1984) im angloamerikanischen noch keine und im deutschsprachigen Raum nur vereinzelte Bemühungen, Fragebogen zur Lehrevaluation gezielt auf theoretischer Basis oder zumindest unter Verwendung theoretisch basierter Lehr-/Lern-Konstrukte zu entwickeln. Als positive Ausnahmen im deutschen Sprachraum seien die Verfahren von Rindermann und Amelang (1994) sowie von Westermann et al. (1998) genannt. Besonders beliebt zu sein scheint der synkretistische Ansatz – oft in Kombination mit Befragungen. So werden in der Regel Items aus vorhandenen Verfahren und Itempools herangezogen, wobei deren Items wiederum häufig aus Umfrageergebnissen unter Studierenden und Dozierenden generiert wurden.

5.2.4 Allgemeines zu den psychometrischen Eigenschaften

Evaluation beinhaltet mehr als nur die Erfassung von Daten oder den Vorgang des Messens; sie hat auch immer etwas mit Bewertung zu tun und ist somit ziel- und zweckorientiert (Wottawa & Thierau, 1998). Evaluationsaktivitäten können nur dann zu wirksamen Veränderungen führen, wenn die im Evaluationsprozess verwendeten Verfahren den allgemeinen methodischen Standards genügen. Ansonsten stellten Entscheidungen, die auf Basis statistisch ungenügender Daten fußen, keinen Fortschritt dar gegenüber subjektiven Entscheidungen. Dabei spielt es keine Rolle, ob es sich um die Entscheidung handelt, die Ergebnisse von Lehrevaluationen im Rahmen eines Steuerungsmodells für Personalentscheidungen zu nutzen, oder um Entscheidungen im Rahmen eines Feedbackmodells, in dem auf Basis der Daten beispielsweise festgelegt wird, welche Stärken und Schwächen einem Lehrenden zurückgemeldet werden sollen. In beiden Fällen hätte die Evaluation das Zielkriterium der Handlungsoptimierung bzw. Übelminimierung (vgl. Wottawa & Thierau, 1998) verfehlt, sofern die Entscheidungsbasis, die Daten, mangelnde Qualität aufwiesen.

Für die Verwendung von Fragebogen heißt dies, dass ihr praktischer und theoretischer Nutzen insbesondere auch von dem Ausmaß abhängt, in dem die Instrumente die Gütekriterien erfüllen (Arreola, & Aleamoni,

1990; Abrami, d'Apollonia & Rosenfield, 1996). Lehrevaluationsfragebogen sollten objektiv, messgenau und zuverlässig sein. El Hage (1996) argumentiert, dass für eine direkte Rückmeldung der Evaluationsergebnisse an Dozierende einfachere Instrumente ausreichten. Lediglich für die vergleichende Bewertung seien höhere Ansprüche an Validität und Messgenauigkeit zu stellen. Dem muss entgegnet werden, dass auch eine Rückmeldung auf Daten basieren sollte, die hinreichend genau erhoben wurden und auch tatsächlich den interessierenden Untersuchungsgegenstand (hier die Lehre) erfassen. Fragebogen müssen sowohl für summative als auch formative Zwecke reliabel und valide sein; wenn nicht die Lehrqualität adäquat beschrieben wird, wäre auch die Verwendung der Daten für Rückmeldungen nicht sinnvoll, argumentiert Rindermann (2001).

Die Gütekriterien, die in Bezug auf Fragebogen zur Lehrevaluation diskutiert werden, seien hier kurz vorgestellt, ehe dann in Kapitel 6 der Forschungsstand ausführlicher diskutiert wird.

Objektivität

Unter Objektivität eines Fragebogens – allgemeiner eines Tests – versteht man das Ausmaß, in dem die Ergebnisse des Tests unabhängig vom Untersucher sind. In der Regel wird zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterschieden (Fisseni, 1997).

Objektivität in der Durchführung liegt vor, wenn die Ergebnisse einer Lehrveranstaltungsevaluation nicht durch zufällige oder systematische Verhaltensvariationen des Untersuchers während der Testdurchführung beeinträchtigt werden. Objektivität hinsichtlich der Auswertung ist gegeben, wenn diese nach festgelegten Regeln vorgenommen wird. Interpretationsobjektivität betrifft den Grad, in dem die Interpretation des Testergebnisses von der Person des Interpretierenden unabhängig ist.

Im Idealfall sollte das Messergebnis hinsichtlich der drei Objektivitätsaspekte unabhängig vom Untersucher sein. Nach Wottawa (1993) ist diese völlige Unabhängigkeit von der Person des Untersuchers aber nicht zu erreichen.

Reliabilität

Die Reliabilität ist eines der zentralen Testgütekriterien und charakterisiert einen Test unter dem Aspekt der Präzision (Fisseni, 1997). Sie gibt die Zuverlässigkeit eines Tests im Sinne der Genauigkeit an, mit der er ein bestimmtes Merkmal misst. Generell wird zwischen verschiedenen

Aspekten der Reliabilität unterschieden: Retestreliabilität, Testhalbierungsreliabilität, Konsistenzanalyse (Teilung des Tests in mehr als zwei Hälften) und Paralleltestreliabilität. Für die Reliabilität von Fragebogen, die zur Lehrevaluation eingesetzt werden, liegen typischerweise Konsistenz- und Retestreliabilitätskoeffizienten vor.

Validität

Nach Lienert und Raatz (1994) gibt die Validität eines Tests den Grad der Genauigkeit an, mit dem dieser Test das Merkmal, das er messen oder vorhersagen soll, tatsächlich misst oder vorhersagt. In Bezug auf die Validität von Fragebogen zur Lehrevaluation heißt die Frage, ob die Verfahren überhaupt in der Lage sind, die Lehrleistung eines Dozenten oder die Lehrqualität einer Lehrveranstaltung zu erfassen. Dabei gibt es zwei Interpretationen dieser Frage (Rindermann, 2001): In einer sehr weit gefassten Auffassung von Validität ziele die Frage darauf, ob die eingesetzten Verfahren die Lehrqualität *aus studentischer Sicht* adäquat erfassten, also die Meinungen der Studierenden unverzerrt widerspiegelten. In einer engeren Sicht der Frage gälten die Instrumente als valide, wenn sie das *Geschehen einer Lehrveranstaltung* angemessen beschrieben. Abrami, d'Apollonia und Cohen (1990) differenzieren an dieser Stelle noch weiter. Die Daten bzw. die Urteile könnten einmal als valide gelten, wenn sie das Geschehen im Sinne des *Prozesses* adäquat beschrieben, und in einem anderen Sinne, wenn sie das *Produkt* des Lehrgeschehens, den Lernerfolg, ausreichend gut vorhersagten.

Die Bestimmung der Validität von Fragebogen zur Lehrevaluation gilt als schwierig (Rindermann, 2001). Um die Frage nach der Validität genauer abschätzen zu können, ist es notwendig, zwischen verschiedenen Arten von Validität zu unterscheiden bzw. zu versuchen, über verschiedene Zugänge die unterschiedlichen Validitätsaspekte abzudecken. Dabei werden in der Gliederung von Greenwald (1997) drei Aspekte unterschieden: Inhaltsvalidität, Kriteriumsvalidität und Konsequenzvalidität. Die folgende Übersicht zu den häufig diskutierten Kernfragen der Validität von Lehrevaluationsfragebogen gibt einen Überblick für die weitere Diskussion des Forschungsstandes (vgl. Tabelle 5).

Tab. 5. Validitätsaspekte von Fragebogen zur Lehrevaluation (adaptiert von Greenwald, 1997)

Validitätsaspekt	Fokus
Inhaltsvalidität	Über welche Komponenten ist das Konstrukt 'gute Lehre' in Fragebogen repräsentiert? (Fragen der Struktur und der Dimensionalität)
Kriteriums-validität	Wie gut korrelieren Ergebnisse einer studentischen Lehrveranstaltungsevaluation mit anderen Lehreffektivitäts-Indikatoren (Frage nach konvergenten und diskriminanten Beziehungen zu anderen Konstrukten)
Konsequenz-Validität	In welcher Weise können Ergebnisse einer studentischen Lehrevaluation genutzt werden, um einen sinnvollen Beitrag zum pädagogischen System zu leisten. (Frage nach der Brauchbarkeit oder Nützlichkeit)

Zu dem etwas weniger gebräuchlichen Begriff der Konsequenz-Validität sei angemerkt, dass hier das Ausmaß an Einfluss näher bestimmt wird, das die Ergebnisse auf den Lehr- und Lernprozess haben (vgl. auch Shulman, 1993 in Braskamp & Ory, 1994). Ebenfalls in diese Kategorie gehört eine weiter gefasste Definition von Validität, wonach sich Validität auch auf die Integrität (integrity) und Angemessenheit (appropriateness) der Schlussfolgerungen bezieht, die auf Basis der gewonnenen Fragebogendaten getroffen werden (vgl. Braskamp & Ory, 1994). Zu Recht weisen einige Autoren darauf hin, dass - unter der Annahme des Vorhandenseins reliabler und valider Daten - nicht garantiert ist, dass diese auch angemessen interpretiert und effektiv verwendet werden (s. auch McKeachie, 1997; Franklin & Theall, 1990; Theall & Franklin 1990a).

Unterscheidung von Verfahren anhand ihrer methodologischen Qualität

Hinsichtlich ihrer psychometrischen Eigenschaften lassen sich existierende Verfahren in zwei Gruppen aufteilen (Marsh, 1984). Die Gruppe der "ill-designed forms" repräsentiert schlecht konstruierte, aus dem Boden gestampfte ad-hoc Verfahren, die den wissenschaftlichen Gütekriterien der Objektivität, der Reliabilität und der Validität nicht standhalten. Die andere Gruppe der "well-designed forms" vereinigt dagegen die Verfahren, die nach wissenschaftlichen Gütekriterien konstruiert wurden.

Zum Zwecke der fundierten Auseinandersetzung mit der methodischen Qualität von Fragebogen zur Lehrveranstaltungsevaluation werden im nächsten Kapitel die wichtigsten Befunde der entsprechenden amerikanischen und deutschen Forschung dargestellt.

6. Forschungsstand zur methodischen Qualität der Verfahren

Nach Feuerstein (1997) weisen viele Evaluationsinstrumente mangelnde methodische Qualität auf. Das mag daran liegen, dass viele der verfügbaren Verfahren bereits bei der Konzeption nicht mit testtheoretischen Vorgehensweisen erstellt wurden (ill-designed forms, Marsh, 1984). Doch wie ist die methodische Qualität für die anhand von Gütekriterien konstruierten Verfahren (sogenannte well-designed questionnaires) zu beurteilen?

Für die Diskussion des Forschungsstandes in Bezug auf well-designed Fragebogen muss insbesondere auf Veröffentlichungen aus dem anglo-amerikanischen Raum zurückgegriffen werden. Für den deutschen Sprachraum liegen im Wesentlichen nur zwei Monographien vor, die sich dezidiert mit den Gütekriterien vorliegender Instrumente befassen (el Hage, 1996; Rindermann, 2001). Beide Autoren verweisen aber ebenfalls darauf, dass sie sich vielfach auf die Studien und Metaanalysen des anglo-amerikanischen Raumes stützen. Vorwegnehmend sei bereits an dieser Stelle darauf hingewiesen, dass in Deutschland nicht bei allen Verfahren die entsprechenden Gütekriterien ermittelt werden (El Hage, 1996). Statt dessen werden überwiegend Probleme und Gefahren studentischer Veranstaltungskritik diskutiert, was zu Lasten von Studien geht, die sich konkret-empirisch mit der Konstruktion, den Gütekriterien und den Einsatzmöglichkeiten von Instrumenten befassen (Diehl, in Druck).

6.1 Objektivität

Studien, die sich explizit mit der Objektivität von Fragebogen zur Lehrevaluation beschäftigen, sind der Verfasserin der Arbeit nicht bekannt. Durchführungs- und Auswertungsobjektivität können aber als gewährleistet angesehen werden, wenn Empfehlungen zur Handhabung und Auswertung der Instrumente vorhanden sind und berücksichtigt werden bzw. die Durchführung der Erhebungen von geschulten Personen vorgenommen wird. Zu dieser Einschätzung kommt auch el Hage (1996) mit der Äußerung, die Erhebungen studentischer Aussagen zur Lehre entsprächen „in dem Maße der Forderung nach ‚Objektivität', in dem sie standardisiert erhoben und ausgewertet wurden" (el Hage, 1996, S. 36)[6].

[6] Gelegentlich wird an dieser Stelle die grundsätzliche Frage erhoben, ob Studierende „objektive" Aussagen über Lehre treffen können bzw. ausreichende Urteilskompetenz besitzen. Dieses Problem wird ausführlicher in Kapitel 6 in Zusammenhang mit der Reliabilität von Fragebogen erörtert.

Die Interpretationsobjektivität als dem Grad der Eindeutigkeit, mit der die verschiedenen Anwender dem gleichen numerischen Wert die gleiche Merkmalsausprägung zuordnen (Fisseni, 1997), ist sicherlich schwieriger zu erzielen. Aus gleichen Auswertungsergebnissen verschiedener Probanden müssten gleiche Schlüsse gezogen werden (Lienert & Raatz, 1994, S. 8). Für den Fall der Lehrveranstaltungsevaluation wäre dies erreicht, wenn verschiedene Hochschuldidaktiker oder Evaluationsexperten aus dem Ergebnisprofil eines Dozierenden die gleichen Stärken und Schwächen abläsen und gleiche oder zumindest sehr ähnliche Schlüsse bezüglich der Lehrqualität dieser Person zögen.

6.2 Reliabilität

In den meisten Studien werden zur Reliabilitätsabschätzung Koeffizienten für die Konsistenz oder die Retestreliabilität angegeben. Jeder dieser Koeffizienten könnte auf Basis der individuellen Daten der einzelnen Urteile (Rohwerte) erfolgen oder auf Basis der Veranstaltungsmittelwerte.

Konsistenz: Skalenhomogenität

Bei der Angabe von Koeffizienten zur internen Konsistenz handelt es sich meist um Skalenhomogenitäts-Koeffizienten (Cronbach's α). Die dadurch erfasste Messgenauigkeit bezieht sich auf die Güte der einzelnen Skalen eines multidimensional aufgebauten Fragebogens zur Lehrveranstaltungsevaluation. Die Werte liegen für Veranstaltungsmittelwerte in der Regel höher als für Koeffizienten auf Basis von Rohwerten.

Rindermann nennt im Schnitt r = .88 für Veranstaltungsmittel und r = .81 für individuelle Werte über alle Skalen[7] (Rindermann, 2001). Wenn man die Angaben von Rindermann etwas näher betrachtet, dann beziehen sich diese auf recht wenige Studien und auf beide Kulturräume gemeinsam. Die Werte werden in der letzten Spalte der Tabelle 6 wiedergegeben. Um ein genaueres Bild zu bekommen, wurden auf Basis von Rindermanns Angaben (Rindermann, 2001, Tab. 7.7, S. 137) die Werte für beide Kulturräume getrennt berechnet sowie die Anzahl der zugrundeliegenden Studien angegeben. Insgesamt können nach Fisseni (1997) die Skalenhomogenitäten als mittlere Reliabilität eingestuft werden.

[7] Rindermann (2001) konnte zeigen, dass die Werte für Dozenten- und Lehreffektivitätsskalen höher liegen.

Tab. 6. Mittlere Skalenhomogenitäten auf Basis von Rindermann (2001, S. 137)

Rohwerte			**Veranstaltungsmittel**		
deutsch	Andere	gesamt	deutsch	andere	gesamt
r = .81	r = .79	r = .81	r = . 86	r = .93	r = .88
N* = 10	N = 1	N = 11	N = 3	N = 1	N = 4

Anmerkung. *Anzahl der Studien, auf denen die Angaben basieren.

Retestreliabilität (Stabilität)

Über die Methode der Testwiederholung kann man die Messgenauigkeit eines Tests im Sinne der Stabilität der Einschätzungen über die Zeit hinweg bestimmen. Die Testwiederholungsreliabilität oder auch Retestreliabilität kann als Korrelationskoeffizient für die Übereinstimmung der zu zwei verschiedenen Messzeitpunkten bzgl. desselben Instrumentes gewonnenen Rohwerte ermittelt werden. Wenn Daten eines Fragebogens zur Lehrveranstaltungsevaluation stabil sind, bedeutet das, dass Studierende das Verhalten und die Eigenschaften einer Lehrperson, die mit diesem Fragebogen erfasst werden, über die Zeit hinweg ähnlich beurteilen.

Marsh (1984) führt verschiedene Studien an, die die Variation der Antworten Studierender über die Zeit hinweg untersuchten, und schlussfolgert, dass amerikanische Fragebogenverfahren zur Lehrveranstaltungsevaluation über die Zeit hinweg stabil sind. Als Stabilitätskoeffizient für Fragebogen zur Lehrveranstaltungsevaluation berichten Braskamp und Ory (1994) eine mittlere Korrelation von r = .83. Die bei Rindermann (2001) genannten Werte liegen dagegen etwas niedriger (vgl. Tab. 7).

Tab. 7. Mittlere Retestreliabilitäten auf Basis von Rindermann (2001, S. 137)

Rohwerte			**Veranstaltungsmittel**		
Deutsch	Andere	gesamt	deutsch	andere	gesamt
r = .64	r = .61	r = .63	r = .81	r = .72	R = .78
N* = 3	N = 1	N = 4	N = 2	N = 1	N = 3

Anmerkung. *Anzahl der Studien, auf denen die Angaben basieren.

Homogenität bzw. Heterogenität der studentischen Urteile

Ein in Deutschland intensiv diskutierter Aspekt der Genauigkeit von Fragebogen zur Lehrveranstaltungsevaluation betrifft die Frage nach der Homogenität oder Heterogenität der Urteile der Studierenden. Die Konsistenz ist hoch, wenn verschiedene Studierende der gleichen Veranstaltung zu homogenen Einschätzungen auf einem gegebenen Item tendieren. Folglich wird die Heterogenität der Urteile innerhalb derselben Veranstaltung als Zeichen für die Ungenauigkeit des Urteils gewertet, letztlich als Hinweis auf die mangelnde Objektivität der Urteilenden. Umgekehrt verhält es sich allerdings für die Heterogenität der Urteile der Studierenden zwischen *verschiedenen Veranstaltungen*. Hier ist Heterogenität wünschenswert, ja geradezu notwendig für eine sinnvolle Interpretation der Evaluationsergebnisse. Weil sich gerade in Deutschland an diesem Thema eine Debatte entzürnt hat, seien die Haltungen und Argumente zusammenfassend vorgestellt.

Die Kritiker führen an: „Die Uneinigkeit der Studierenden darüber, was als gutes Lehrverhalten empfunden wird, ist außerordentlich groß." (Kromrey, 1996b, S. 155). Und ebenso Rosemann (1999, S. 39) „Zahlreiche empirische Untersuchungen (...) belegten in der Tat die große Urteilsheterogenität bei studentischen Bewertungen identischer Veranstaltungen derselben Dozenten." Schweer und Rosemann (1995) führen an, dass selten danach gefragt wurde, warum oder auf welche Weise diese Differenzen zustande kommen. Rosemann und Schweer (1996a) sowie Rosemann (1999) sehen in der Urteilerheterogenität keine Fehlerquelle, die es statistisch zu korrigieren gelte, sondern das Ergebnis wechselseitiger Wahrnehmungsprozesse, die Wirkung komplexer kognitiver Schemata und der jeweiligen Erwartungssysteme (ähnlich auch Weiss, 1991). Die Schlüsse, die aus diesem Sachverhalt gezogen werden, sind einheitlich –

unabhängig, ob es sich aus Sicht der Autoren um zu korrigierende Fehlerquellen oder um Abbilder sozialer Wirklichkeit handelt.

Für Schweer und Rosemann (1995) weisen die Befunde „auf die Notwendigkeit hin, hinsichtlich der Erfassung der Qualität der Lehre über ein geeigneteres Meßinstrument nachzudenken“ (S. 195) bzw. werde „über eine wie immer zu definierende Lehrqualität oder Lehrkompetenz nichts Verläßliches gesagt“ (Rosemann, 1999, S. 45). Kromrey kritisiert unter anderem immer wieder die Praxis der Mittelwertbildung oder der Zusammenfassung der nicht unabhängigen Urteile auf einzelnen Dimensionen zu einem Globalmaß (vgl. Kromrey, 1994, 1996b). Die Bildung von Mittelwerten sei unangemessen, da diese nur dann sinnvoll zu interpretieren seien, wenn die Daten tatsächlich eine zentrale Tendenz zur Mitte bei nur mäßiger Streuung aufwiesen. Seine Position wendet sich gegen diese und andere falsche Modellannahmen, etwa die Reduzierung des sozialwissenschaftlich komplexen Evaluationsmodells zu „bloßer Umfrageforschung“. „Die begriffliche Gleichsetzung des Einsatzes von Umfrageforschung zum Zwecke von Evaluation und Qualitätsentwicklung in Lehre und Studium mit Evaluation verursacht unnötige Verwirrung“ (Kromrey, 2001, S. 33). Im Unterschied zu Rosemann und Schweer (1996b) sieht er den Nutzen des Einsatzes von Lehrevaluationsfragebogen noch vergleichsweise positiv: „und sie [Umfrageforschung] kann als solches ein wichtiger Baustein in einem Konzept von Evaluation und Qualitätsentwicklung sein“ (Kromrey, 2001, S. 33).

Die Entgegnungen beziehen sich auf die Methoden, die zur Demonstration der Heterogenität verwendet wurden. So wenden sich beispielsweise Diehl (1996) oder Rindermann (1997b, 1998c) gegen die Verwendung von graphischen Methoden, sogenannten „Schnittmusterbogen“ (Diehl) oder „beliebig definierbaren Profilen“ (Rindermann) sowie gegen die Vermischung von Daten aus verschiedenen Veranstaltungen. Alleinig korrelationsstatistische Berechnungen der Beurteilerübereinstimmung seien angemessen (Rindermann, 1998c, 2001).

Ein weiterer Aspekt wird in unmittelbarer Folge angeführt. Die Analyseeinheit, auf deren Basis die Homogenität ermittelt werden soll, sei für die Diskussion von entscheidender Bedeutung. Während sich Kromrey (s.o.) ja explizit gegen die Verwendung von Mittelwerten ausspricht, verweist Rindermann (1996a,c) auf die gängige Praxis der Mittelwertbildung in den USA. Seiner Meinung nach ist es von großer Bedeutung, ob sich die Angaben auf die Übereinstimmung der individuellen Urteile oder auf Veranstaltungsmittelwerte beziehen. Folgt man dieser Unterscheidung, stellt sich das Bild differenzierter dar:

Rindermann und andere Autoren schließen sich der vielfach geäußerten Kritik der Einschätzung der Heterogenität an – sofern es um die individuelle Beurteilerübereinstimmung geht (Diehl, 1996; Gold 1996; Rindermann, 2001). Die Konsistenz individueller Urteile ist gering und liegt nach Rindermann (2001) zwischen r = .18 und r = .30, woraus der Autor in Übereinstimmung mit Kritikern schließt, dass „studentische Einzelangaben nicht als Maß zur Bestimmung der Lehrqualität herangezogen werden dürfen“ (S. 134)[8]. Für Skalen sind höhere Übereinstimmungswerte zu erzielen, wobei Kursskalen übereinstimmender beurteilt werden als Selbsteinschätzungsskalen, insbesondere Kursmittel sind messgenauer (Rindermann, 2001).

Der Grund ist folgender: Während sich in den Einzelurteilen die Streuungen zwischen Studierenden und zwischen Veranstaltungen als zwei verschiedene Varianzquellen überlagern (Feldman, 1977), ist dies für Veranstaltungsmittelwerte weniger der Fall. Weiterhin hat die Anzahl der in die Berechnung eingegangenen Urteile einen systematischen Einfluss auf die Höhe des Koeffizienten. Je mehr Studierende in einer Veranstaltung sind, um so höher ist die Übereinstimmung. Praktisch bedeutet das, dass Veranstaltungen mit höherer Teilnehmerzahl zuverlässigere Daten liefern (Cashin, 1988).

Die Diskussion schließt bei Rindermann (2001) in Anlehnung an amerikanische Autoren damit, dass den Urteilen der Studierenden bei 15 bis 20 Personen und mehr je Veranstaltung vertraut werden könne. In diesem Falle würden Interrater-Reliabilitäten von r = .70 bis r = .90 erzielt, für Veranstaltungen ab 25 Personen sogar von r = .90 und höher (Rindermann, 2001). „Die Urteilerübereinstimmungen sind hiermit groß genug, um für wissenschaftliche Untersuchungen und den praktischen Einsatz zuverlässige Veranstaltungsmittelwerte zu erzielen.“ (Rindermann, 1996a, S. 151).

Ähnlich berichtet Cashin (1988, 1995) bezüglich des IDEA-Systems mittlere (Median) „Intraclass-Korrelationen“[9] zwischen r = .69 für 10

[8] Interessant ist, dass die geringen individuellen Übereinstimmungen immer noch den Werten entsprechen, die für Gutacher-Übereinstimmungen bei wissenschaftlichen Publikationen anzutreffen sind (im Mittel r = .27), woraus wiederum der Schluss gezogen wird, dass eine geringe Urteilerübereinstimmung weitgehend unabhängig sei von der Urteilsperson (Student oder Wissenschaftler) und dem Gegenstand der Beurteilung (Lehre oder wissenschaftliches Manuskript) (vgl. el Hage, 1996; Rindermann, 2001).

[9] Eine Intraclass-Korrelation ist ein Index, der die Variation in den Antworten zwischen verschiedenen Kursen oder Veranstaltungen mit der Variation in den Antworten innerhalb der gleichen Veranstaltung vergleicht (Cashin, 1995).

Beurteilende und r = .91 für 40 Beurteilende. Andere, gut konstruierte Verfahren wiesen gleich hohe oder höhere Homogenitäts-Koeffizienten auf. Marsh (1984) berichtet beim SEEQ von Koeffizienten in ähnlicher Höhe (zwischen r = .60 für 10 Beurteilende und r = .95 für 50 Beurteilende). Er ist sogar der Auffassung, dass bei ausreichender Klassenstärke, die Reliabilität von Veranstaltungsmittelwerten vergleichbar sei mit den Reliabilitätsmaßen der "best objective tests" (Marsh, 1984).

Tab. 8. Angaben zu Beurteilerübereinstimmungen

Rindermann (2001, HILVE)			**Cashin (1995, IDEA)**		
individuell	bei Veranstaltungsmitteln		bei Veranstaltungsmitteln		
2 Urteiler	≤ 10 Urteiler	≤ 25 Urteiler	10 Urteiler	20 Urteiler	40 Urteiler
.29	.80	.89	.69	.83	.91

Abschließend sei angemerkt, dass inzwischen die Empfehlung ausgesprochen wird, für große Veranstaltungen getrennte Resultate für sich signifikant unterscheidende Untergruppen (etwa Studierende verschiedener Fachrichtungen) zurückzumelden. Diese Empfehlung (z. B. Rindermann, 1997a) geht sicherlich auf das in einer einfachen Datensimulation aufgezeigte Problem zurück. Süllwold (1992) schließt aus der Datensimulation, dass aufgrund der Konfundierung der Daten von nicht homogenen Beurteilungsgruppen „Erhebungen über die Beurteilung von Hochschullehrern durch Studierende in fundamentaler Weise fehlerhaft [sind]“ (Süllwold, 1992, S. 34). Mittelwerte über die Gesamtgruppen sind in diesem Falle inhaltsleer.

Zusammenfassend für die allgemeine amerikanische Auffassung zur Reliabilität von Fragebogenverfahren zur Lehrevaluation formulieren Abrami et al. (1996, S. 224) wie folgt: "The reliability of student ratings is not a contested issue." Amerikanische Forscher stellen die Reliabilität von Fragebogen zur Lehrveranstaltungsevaluation nicht mehr länger in Frage. Die zeigt auch deutlich eine von Centra oftmals angeführte Analogie zum berühmten Ausspruch von Abraham Lincoln über die Grenzen, Menschen „an der Nase herumzuführen“ (vgl. Centra, 1993). Bezüglich studentischer Lehrevaluation ist nach Centra zuverlässig zu sagen, dass Lehrende alle Studierende für einige Zeit an der Nase herum führen können; dass sie einige Studierende über lange Zeit an der Nase herumführen

können; dass sie aber kaum alle Studierenden dauerhaft an der Nase herum führen können (vgl. Centra, 1993, S. 60).

Wenn auch die Debatte über die Homogenität bzw. Heterogenität studentischer Urteile bisweilen weiten Raum in den deutschen Diskussionen einnimmt, so lässt sich insbesondere im Unterschied zu den Validitätsdebatten die deutsche Auffassung mit el Hage (1996) zusammenfassend charakterisieren: „Die Diskussion um Reliabilität wird bei der Veranstaltungskritik von Studierenden eher am Rande geführt, da viele Kritiker und Kritikerinnen von einer nicht gegebenen Validität ausgehen.“ (el Hage, 1996, S. 36).

6.3 Validität

Seit den 1970er Jahren bekundeten amerikanische Forscher ihre Zweifel an der Validität von Fragebogenverfahren, die zur Lehrveranstaltungsevaluation eingesetzt werden. Eine Fülle von Studien beschäftigte sich seither mit den verschiedenen Aspekten der Validität. Die Veröffentlichungen erreichten einen Höhepunkt in den frühen 1980er Jahren. In den 1990ern sank die Frequenz der Veröffentlichungen auf ihren niedrigsten Stand (vgl. Greenwald, 1997). Auch in Deutschland gilt die Validität von Fragebogen zur Lehrveranstaltungsevaluation als das am meisten umstrittene Thema. Eine Abschätzung zum Stand der Forschung zur Validität erfolgt anhand der bereits erwähnten Aspekte der Validität (vgl. Kap. 5.2.4).

Die *Konstruktvalidität* eines Tests: "Aufgrund theoretischer - sachlogischer und begrifflicher - Erwägungen und anhand von sich daran anschließenden empirischen Untersuchungen [wird] entschieden, ob ein Test ein bestimmtes Konstrukt zu erfassen vermag" (Lienert & Raatz, 1994, S. 11). Forscher versuchen, die Konstruktvalidität eines Verfahren über seine Einbettung in ein nomologisches Netzwerk theoretisch verwandter oder theoretisch entfernter Konstrukte zu beschreiben (vgl. Fisseni, 1997; Marsh, 1984). Die Konstruktvalidierung eines Tests ist ein enges Wechselspiel zwischen Theorie und Empirie, für das sich logische Analysen, empirisch-korrelationsstatistische und experimentelle Ansätze anbieten.

Die beiden im Folgenden diskutierten Aspekte der Inhaltsvalidität und der Kriteriumsvalidität können als spezielle Formen der Konstruktvalidität aufgefasst werden (Messick, 1988 in Lienert & Raatz, 1994).

Inhaltsvalidität – logische Validität

Nach Fisseni (1997) ist Inhaltsvalidität dann gegeben, wenn der Inhalt der Test-Items das Zielmerkmal hinreichend genau definiert und somit das Konstrukt hinreichend gut repräsentiert. Dies steht für inhaltliche Übereinstimmung einer empirischen Messung mit einem logischen Messkonzept, weswegen man gelegentlich auch von logischer Validität spricht. Ein Fragebogen zur Lehrevaluation besitzt demzufolge inhaltliche oder logische Gültigkeit, wenn seine Fragen eine inhaltlich repräsentative Auswahl aller möglichen Fragen darstellen, durch welche 'gute Lehre' operationalisiert werden kann (Abrami, 1985). Bereits an dieser Stelle wird deutlich, dass das zentrale Problem das der Definition des Zielmerkmals ist — also die Definition von 'guter Lehre'.

Bei der *Inhaltsvalidität* von Fragebogenverfahren zur Lehrevaluation handelt es sich um eine theoretische Form der Validität, für die es keine einheitliche Maßzahl gibt. Nach Abrami und d'Apollonia (1990) kann eine inhaltliche Validität eines Fragebogens durch Expertenurteil bestimmt werden. Ein Fragebogen ist in dem Grad inhaltlich valide, in dem er Items umfasst, die die Experten des Feldes für angemessen halten. Die "inhaltliche Validität wird einem Test in der Regel durch ein Rating von Experten als 'Konsens von Kundigen' zugebilligt" und das Ergebnis der Expertenbefragung "unter Verzicht auf einen numerischen Validitätskennwert im Testmanual mitgeteilt". Es handelt es sich also um eine "nur durch psychologische Einsicht begründbare Annahme" (s. für beide Zitate Lienert & Raatz, 1994, S. 11 bzw. S. 225).

In der Praxis liegen unterschiedlichste Verfahren vor, die jeweils zur Evaluation von Lehrveranstaltungen eingesetzt werden. Alle Verfahren intendieren, in irgendeiner Weise, 'gute Lehre' zu messen. Insbesondere unterscheiden sich die Fragebogen darin, wieviele und welche Dimensionen des Konstruktes 'gute Lehre' im Fragebogen repräsentiert sind, um das Konstrukt zu erfassen. Die Differenzen über die angemessene Form der Erfassung gehen sogar soweit, dass nicht einmal Einigkeit darüber besteht, ob es sinnvoll ist, das Konstrukt mehrdimensional zu erfassen, oder ob nicht eine unidimensionale Einschätzung in Form eines globales Urteils oder einer Allgemeinbewertung ausreichend ist. Die erste Form des Aufbaus wird von Marsh und Kollegen gefordert (vgl. Marsh, 1984); die zweite Position dagegen vertreten Abrami und Kollegen (beispielsweise d'Apollonia & Abrami, 1997). An dieser Stelle ist festzuhalten, dass die inhaltlichen Strukturen von Fragebogen zur Lehrevaluation höchst unterschiedlich sind, was auf eine mangelnde Inhaltsvalidität der Verfahren schließen lässt. Die Probleme der Inhaltsvalidität von Fragebo-

gen zur Lehrevaluation bzw. die Probleme einer Definition und der Dimensionalität des Konstruktes 'gute Lehre' werden ausführlicher in Kapitel 7 diskutiert.

Kriterienbezogene Validität – empirische Validität

Die kriterienbezogene Validität wird vorwiegend auf empirischem Weg ermittelt, weswegen sie gelegentlich auch als empirische Validität bezeichnet wird. Für diese Form der Validität spielen zwei Grundüberlegungen eine wichtige Rolle (s. Fisseni, 1997; Greenwald, 1997). Korrelieren Scores für theoretisch verwandte Konstrukte hoch mit den Scores des Konstruktes 'gute Lehre', liegt konvergente Validität oder Übereinstimmungsvalidität vor. Gleichzeitig sollten Scores des Konstruktes 'gute Lehre' gar nicht oder nur sehr schwach mit Scores theoretisch entfernter Konstrukte korrelieren (diskriminante Validität). Zusammenfassend werden diese Studien auch als „Multitrait-Multimethod Studien“ bezeichnet (vgl. Marsh, 1984).

Zur Bestimmung der *konvergenten Validität* von Verfahren zur Lehrevaluation wird von Forschern oftmals ein „multisection course design“ verwendet. In verschieden (Parallell-)Veranstaltungen gleichen Inhalts, gleicher Zielvorgaben mit Studierenden des gleichen Leistungsniveaus, aber mit unterschiedlichen Dozierenden, werden Daten via Fragebogen zur Lehrveranstaltungsevaluation erhoben. Zusätzlich wird als eines der wichtigsten Außenkriterien ein Leistungsmaß (z. B. Note in der Abschlussklausur) erhoben. Das Leistungsmaß wird als Operationalisierung des Lernfortschritts (als einem theoretisch verwandten Konstrukt) verwendet. Ziel der Forscher unter Zuhilfenahme eines solchen Designs ist es, eine Beziehung aufzuzeigen zwischen den durchschnittlichen Bewertungen der Lehre via Fragebogen und den durchschnittlichen Leistungen der Studierenden. Ein hoher positiver Zusammenhang wird als Indiz für die Konstruktvalidität des Fragebogens gewertet.

Mehrere Metaanalysen haben versucht, die zahlreichen Ergebnisse von Einzelstudien zusammenzufassen (s. Abrami, d'Apollonia, & Cohen, 1990; Abrami et al., 1996; Cohen, 1981, 1983; Dowell & Neal, 1982; Marsh, 1984, 1987). In der aktuellsten (1996) resümieren Abrami et al. unter Berücksichtigung anderer Metaanalysen: "Collectively, the results of the reviews suggest that some specific rating dimensions, as well as student global ratings, are moderately correlated with student learning in multisection college courses. On average, there exists a reasonable, but far from perfect, relationship between some student ratings and learning" (Abrami et al., 1996, S. 238).

Für Deutschland liegen bisher keine metaanalytischen Befunde vor. Meist untersuchen einzelne Arbeiten den Zusammenhang zwischen studentischen Lehrbeurteilungen und Wissenszuwachs als Außenkriterium. Exemplarisch sei die dazu kritisch urteilende Studie von Rosemann und Schweer (1996a) angeführt. Während signifikante Korrelationen für die Einschätzung des persönlichen Lernerfolges und der Evaluation der Lehrveranstaltung auf drei Skalen gefunden werden konnte, war dies für Zusammenhänge mit der Klausurleistung nicht der Fall (vgl. Tab. 9). Ebensowenig war ein bedeutsamer Zusammenhang zwischen Klausurleistung und subjektiv eingeschätztem Lernerfolg zu ermitteln (r = .13).

Tab. 9. Korrelationen zwischen drei Beurteilungsdimensionen und Lernerfolgsmaßen (Rosemann & Schweer, 1996a, S. 178)

	Subjektiver Lernerfolg Einschätzung des persönlichen Wissenszuwachs	**Objektiver Lernerfolg Klausurleistung**
Unterstützendes Dozentenverhalten	.23**	.12
Didaktisches Geschick	.37**	-.04
Transparenz der Leistungsanforderung	.38**	.00

Anmerkung. *N = 167, **p<.01.

Auch Rindermann (2001) konnte keine substantiellen Zusammenhänge zwischen den Skalen des HILVE und Klausurresultaten aus 8 verschiedenen Veranstaltungen finden, berichtet aber von mittleren Validitäten (Übereinstimmungen von im Schnitt r = .52 auf Basis von – so ist zu vermuten – amerikanischen Metaanalysen, S.177).

Auf Zusammenhänge des studentischen Urteils mit anderen möglichen Außenkriterien – wie beispielsweise den Selbstbeurteilungen der Lehrenden, den Urteilen von Beobachtern oder von Absolventen – sei hier nur zusammenfassend verwiesen. Für den amerikanischen Sprachraum gibt Cashin (1995) einen fundierten Überblick. El Hage (1996) fasst den Stand aus deutscher Sicht zusammen: „Die Übereinstimmung zwischen der Selbstbeurteilung der Lehrenden bzw. von Beobachtungsergebnissen und dem studentischen Urteil kann als relativ hoch betrachtet werden“ (el Hage, 1996, S. 48). Dies steht in Übereinstimmung mit den bei Rindermann (2001) angeführten Befunden.

Die Frage nach der *diskriminanten Validität* von Verfahren zur Lehrveranstaltungsevaluation ist zugleich die Frage nach potenziellen *Verzerrungsvariablen*. Die Faktoren, die einen systematischen Einfluss auf die Bewertung der Lehrveranstaltung haben, werden in der anglo-amerikanischen Literatur unter dem Stichwort "*bias*" diskutiert. Fehlt ein systematischer Zusammenhang zwischen der Bewertung einer Lehrveranstaltung und möglichen Bias-Faktoren (wie zum Beispiel Motivation der Studierenden oder Interesse am Thema der Veranstaltung), dann wird dies als Indikator für die Validität des Instrumentes bzw. der Urteile gewertet. Die Vermutung, dass Verzerrungsvariablen vorliegen und die Bewertung der Veranstaltung beeinflussen können, hat viele Betroffene bewegt (exemplarisch Whitley, 1984). Eine Reihe potenzieller Variablen werden in diesem Zusammenhang immer wieder diskutiert und intensiv erforscht. Marsh (1987) präsentiert eine Liste von Variablen, von denen Dozierende *vermuten*, sie übten einen systematischen Einfluss auf die Bewertung von Lehrveranstaltungen durch Studierende aus. Die Dozierenden nannten nach Häufigkeiten:

Schwierigkeit der Veranstaltung	(72%)
grading leniency[10]	(68%)
Popularität des oder der Dozierenden	(63%)
Interesse der Studierenden am Veranstaltungsthema	(62%)
Arbeitsaufwand	(60%)
Teilnehmerzahl	(60%)
Grund für den Besuch der Veranstaltung	(55%)
Leistungsniveau der Studierenden (GPA)[11]	(53%)

Durch das sorgfältige Studium der Veröffentlichungen zum Thema lässt sich dieses Bild nach Marsh jedoch nicht bestätigen. Weiterhin sei die Forschung zu möglichen Störvariablen selbst verzerrt (biased).

Arreola (2000) spricht gar von Mythen, die sich in den Köpfen festgesetzt hätten. Dem entgegen setzt er nach intensiver Durchsicht der Literatur der anglo-amerikanischen Forschung, dass bei gut konstruierten Verfahren:

---

[10] Diese Hypothese besagt, dass Studierende aufgrund guter Bewertungen ihrer eigenen Leistung tendenziell auch gute Beurteilungen abgeben.

[11] Die Abkürzung GPA steht für „graduate point average“.

- sich Dozierende nicht durch die Vergabe von guten Noten die Gunst der Studierenden "erkaufen" könnten,
- geringe Teilnehmerzahlen nicht automatisch zu besseren Bewertungen der Veranstaltung führen, noch dass hohe Teilnehmerzahlen automatisch in schlechteren Bewertungen resultieren,
- Studierende mit niedrigerem Leistungsniveau nicht strenger oder härter bewerten als Studierende mit höherem Leistungsniveau,
- Studierende in Pflichtveranstaltungen nicht dazu tendierten, schlechtere oder härtere Bewertungen vorzunehmen als Studierende im Wahlbereich,
- das Geschlecht der Dozierenden keinen Einfluss hat,
- Veranstaltungen am frühen Vormittag im Allgemeinen nicht schlechter beurteilt werden als Nachmittagveranstaltungen.

In einem Review neueren Datums halten d'Apollonia und Abrami (1997) für den Praxisbezug fest, dass die Bewertungen nicht übermäßig durch externale Faktoren wie zum Beispiel Charakteristika der Studierenden, des Kurses oder des Lehrer beeinflusst werden. Die zusammenfassende Einschätzung der Autoren Marsh und Roche (1997) scheint deshalb den Stand der anglo-amerikanischen Forschung zum Thema „bias“ zutreffend zu beschreiben, wonach Fragebogen zur Lehrevaluation sich erwiesen haben als „relatively unaffected by a variety of variables hypothesized as potential biases“ (Marsh & Roche, 1997, S. 1187).

Hinsichtlich der Publikationen im deutschen Sprachraum fällt es schwer, den allgemeinen Stand der Biasforschung zu benennen, da „fast in jeder Publikation zur Lehrevaluation der Zusammenhang mit Verzerrungsvariablen untersucht wird“ (Rindermann, 2001, S. 183). Spiel und Gössler (2000) fanden lediglich „Interesse“ als einzige Variable, die über verschiedene Stichproben hinweg systematisch mit dem studentischen Urteil korrelierte[12]. Dies wird in der aktuellsten Literatursicht von Spiel (2001) erneut angeführt. Sie konnte dies auch in ihren eigenen Untersuchungen wiederholt bestätigen. Weiterhin nennt Spiel noch einige Variablen, deren Einfluss aber wesentlich geringer sei. Darunter fallen der Besuchsgrund, der Veranstaltungstyp und Enthusiasmus des bzw. der Dozierenden.

Konsequenz-Validität

[12] Gelegentlich wird darin aber keine Störvariable gesehen, da nicht zu klären ist, ob das Interesse nicht erst durch die Veranstaltung hervorgerufen bzw. moderiert wird (Marsh, 1987; Spiel, 2001).

Hinsichtlich der Konsequenz-Validität berichtet Marsh in seinem Literatur-Review von 1984 für den anglo-amerikanischen Sprachraum, dass die Erforschung dieses Validitätsaspektes bisher wenig systematisch betrieben wurde. Insgesamt gäbe es bisher auch nur wenige Versuche zu bestimmen, ob, wie und in welchem Ausmaß Daten aus studentischen Veranstaltungsbeurteilungen zur systematischen Optimierung der Lehre verwendet werden. Seit den 1990ern avancierte die Frage nach der Konsequenz-Validität zum zentralen Validitätsaspekt. McKeachies Artikel "Student Ratings. The Validity of Use" schließt eine Serie von Artikeln zu den verschiedenen Validitätsaspekten ab, die 1997 im American Psychologist publiziert wurden. Er fasst den Forschungsstand zu den verschiedenen Aspekten der Validität von Fragebogenverfahren zur Lehrevaluation wie folgt zusammen: Fragebogenverfahren seien valide und praktische Quellen zur Datengewinnung in Fragen der Lehreffektivität. Das grundlegendere Validitätsproblem sei hingegen vielmehr die Art und Weise der Verwendung dieser Daten in Komitees und Verwaltungen. Dort geschehe Missbrauch aufgrund mangelnden Wissens, wie diese Daten zu interpretieren sind (siehe auch Braskamp & Ory, 1994). Bereits 1987 mahnte McKeachie deshalb an, dass Forschung sich nicht nur auf die (kriterienbezogene) Validität der Ratings von Studierenden beziehen sollte. Vielmehr müssten auch Fragen der Glaubwürdigkeit von Ratings sowie ihrer Rolle in Entscheidungsprozessen stärker untersucht werden. Theall und Franklin (1990b), die sich insbesondere den Problemen der praktischen Optimierung von Lehre unter Verwendung von Fragebogenverfahren zur Lehrevaluation widmen, charakterisieren die Situation ähnlich. Nach Meinung der Autoren ist es vor allem die Beziehung zwischen der Gewinnung der Daten und ihrer Verwendung - die Synergie von Theorie und Praxis - die erhöhter Aufmerksamkeit bedarf.

Erste Anmerkungen zum Thema Konsequenz-Validität finden sich auch bei deutschen Autoren. Beispielsweise ist für Rindermann (2001) „weniger die Validität des studentischen Urteils ein Problem, als die Art der Nutzung oder nicht Nutzung der Ergebnisse durch die Universität“ (S. 205). Nähere Ausführungen hierzu fehlen leider. Ähnlich bezeichnete bereits el Hage (1996) die Validitätsdebatte als „Nebenkriegsschauplatz“.

6.4 Fazit

Marsh (1987) hat für die *anglo-amerikanischen Forschungsaktivitäten* in seinem umfassenden Literaturreview den viel zitierten Satz geprägt, dass Fragebogen zur Lehrevaluation multidimensional, reliabel, stabil und

primär eine Funktion des/der Dozierenden einer Veranstaltung sind; dass sie valide sind gegenüber anderen Indikatoren von Lehreffektivität und relativ wenig durch mögliche Bias-Variablen beeinflusst werden (Marsh, 1987, S. 707). Bei anderen Autoren sind abschließende Einschätzungen dieser Art in ähnlicher Weise zu finden (vgl. Cashin, 1995; Cohen, 1990; McKeachie, 1997).

Insbesondere die Validität der Verfahren war oftmals Gegenstand der wissenschaftlichen Auseinandersetzung. Der Rückgang während der 1990er Jahre zu Fragen der Validität zeigt bereits an, dass bisher gewonnene Forschungsergebnisse die Hauptfragen befriedigend beantworten. Fragebogen zur Lehrevaluation gelten als gründlichst erforscht und empirisch abgesicherte Instrumente zur Personal-Evaluation (Marsh & Bailey, 1993). Nach amerikanischer Auffassung hat die Schwäche der mittels studentischer Lehrevaluation gewonnen Daten, vergleichsweise wenig mit der (kriterienbezogenen) Validität oder Reliabilität der Fragebogen zu tun. Wie zuvor erwähnt, treten seit den 1990er Jahren insbesondere Fragen des angemessenen Einsatzes von Fragebogen und Nutzungsaspekte der Daten in den Vordergrund.

Für *Deutschland* sieht die Gesamteinschätzung der Gütekriterien von Fragebogen zur Lehrevaluation heterogener aus. Hier kennzeichnen noch sehr grundlegende Debatten beispielsweise um Urteilskompetenz der Studierenden (z. B. die Konsistenz der Urteile) und die Validität der Verfahren den Stand der Forschung. Trotz vielfacher Aktivitäten und Publikationen scheint sich das Forschungsfeld noch nicht voll etabliert zu haben. So berichtet el Hage (1996), dass nicht bei allen anzutreffenden Verfahren mit „einem gewissen Allgemeinheitsanspruch“ die entsprechenden Gütekriterien ermittelt wurden.

Vor diesem Hintergrund wundert es wenig, dass in den bisherigen deutschen Veröffentlichungen die Haltungen und Positionen teilweise weit auseinander liegen. Dies gilt um so mehr angesichts der verschiedenen möglichen Verwendungszwecke der Evaluationsdaten oder auch Evaluationsmodelle (vgl. Kapitel 5). Während die Beurteilung des Nutzen von Lehrevaluationsfragebogen für formative Zwecke auch von Kritikern wie Kromrey (1995) oder Rosemann und Schweer (1996a,b) nicht abgesprochen wird, sieht das Bild vor allem in Bezug auf die Verwendung der Daten im Rahmen von Steuerungsmodellen z. B. zum Zwecke der Personalbeurteilung sehr viel kontroverser aus.

Zur Frage, ob die Urteile der Studierenden als valides Maß von Lehrqualität oder Lehreffektivität gelten können, vertritt Rindermann die posi-

tivste Einstellung: Seine Position basiert auf der intensiven Rezeption der anglo-amerikanischen Literatur sowie auf eigenen Untersuchungen mittels des Verfahrens HILVE:

> Es „kann begründet von studentischen Lehrevaluationen als ein Maß universitärer Lehrqualität (und nicht bloß der Messung von Akzeptanz) gesprochen werden. Allerdings sollten hierzu nicht Ergebnisse einzelner Veranstaltungen oder gar einzelner Studierender herangezogen werden“ (Rindermann 1997a, S. 38).

Für die ablehnende Haltung gegenüber der summativen Verwendung von studentischen Lehrevaluationsurteilen sei die Position von Diehl angeführt:

> „Erzwungene Teilnahme mit öffentlichem Vergleich der Ergebnisse und öffentlichen Urteil über die »Qualität« der Veranstaltungen und Lehrenden, womöglich verknüpft mit daraus abgeleiteten Belohnungen und ‚Bestrafungen‘. Für eine derartige Anwendung sind VBVOR und VBREF eindeutig nicht konzipiert worden. (...) Ihr primärer Wert liegt in der von den Lehrenden individuell und privat durchgeführten Ergebnisanalyse“ (Diehl, in Druck, S. 23).

Vergleichbar ist dazu auch die Haltung von Rosemann und Schweer (1996b):

> „Dies bedeutet nicht, gegen (...) Verbesserung (...) argumentieren zu wollen. Nur: Von einer Evaluation der Lehre durch Studierende sollte man sich nicht erhoffen, was auch weitaus komplexere Instrumente der Personalbeurteilung im Bereich von Wirtschaft und Verwaltung bis heute nicht zu leisten imstande sind; nämlich eine einigermaßen objektive und valide Beurteilung der individuellen Arbeitsleistung, hier der Lehrleistung von Dozenten“ (Rosemann & Schweer, 1996b, S. 99).

Die heterogenen Einschätzungen aus Deutschland, aber auch die Diskussionen in der anglo-amerikanischen Literatur, die trotz mehrheitlich positiver Einschätzung der Gütekriterien noch nicht abgeschlossen sind (vgl. Artikelserie im American Psychologist von 1997), deuten darauf hin, dass die Qualität der Instrumente bisher nicht völlig geklärt ist. Zwei Problembereiche aus dem Spannungsfeld von Inhalts- bzw. Konstruktvalidität kristallisieren sich bei der Sichtung der Literatur heraus: (1) die fehlende theoretische Fundierung sowie Definitionsprobleme und (2) die Verlage-

rung der Debatte um die Konstruktvalidität auf die Dimensionalität von Fragebogen.

Beide Problemfelder stehen im Zusammenhang mit der „angemessenen“ Repräsentativität des Konstruktes 'gute Lehre' in Fragebogenverfahren. Ihnen ist das folgende Kapitel gewidmet, das auch der Hinleitung auf die empirische Fragestellung dieser Arbeit dient.

7. Problemfelder bei der Erfassung des Konstruktes ‚gute Lehre‘ in Fragebogenverfahren

Die Frage nach der Erfassung oder auch angemessenen Repräsentativität des Konstruktes 'gute Lehre' in Fragebogenverfahren kann in zwei Problemfelder gegliedert werden.

Das erste Problemfeld bezieht sich auf die theoretische Fundierung der Fragebogen zur Lehrveranstaltungsevaluation: Inwieweit wurden bei der Konzeption der Instrumente Theorien über das Konstrukt 'gute Lehre' herangezogen? Dies ist eng mit der Frage nach der Definition von ‚guter Lehre‘ verknüpft. Im Vergleich zu den zahlreichen Veröffentlichungen zu möglichen Definitionen von Lehre thematisieren nur wenige Autoren die theoretische Fundierung von Fragebogen zur Lehrveranstaltungsevaluation (Ausnahmen sind beispielsweise Marsh, 1984; Rindermann, 1998a,b, 1999a).

Das zweite Problemfeld behandelt Aspekte der Dimensionalität des Konstruktes. Ist 'gute Lehre' ein unidimensionales oder multidimensionales Konstrukt? Was bedeutet die Dimensionalität des Konstruktes für die Konzeption bzw. die Dimensionalität eines Fragebogens, der dieses Konstrukt erfassen soll? Um die erste der beiden Fragen ist in den USA eine heftige Diskussion entbrannt (vgl. verschiedene Veröffentlichungen von Abrami und Kollegen sowie Marsh und anderen, s.u.).

Beide Problemfelder werden nun ausführlicher dargestellt, um dann die daran unmittelbar anschließende Frage zu betrachten, welche Dimensionen in verschiedenen Fragebogen zur Lehrveranstaltungsevaluation repräsentiert sind und ob es sich dabei um jeweils die gleichen Dimensionen handelt.

7.1 Fehlende theoretische Fundierung und Definitionsprobleme

> „Any instrument designed to assess teaching must be based on some notion of what constitutes good teaching“ (Greenwood, Bridges, Ware, & McLean, 1973).

Diesem Zitat von Greenwood et al. aus dem Jahre 1973 kann man entnehmen, dass eine theoretische Fundierung von Instrumenten zur Lehrevaluation nicht vernachlässigt werden sollte. Das gilt für die Entwicklung von Peer-Ratings oder Leitlinien zur Verhaltensbeobachtung wie auch für Fragebogen zur studentischen Lehrevaluation. Im Anschluss an dieses Zitat ließe sich argumentieren, dass jedes zur Evaluation der Lehre

eingesetzte Instrument auf Qualitätskonzeptionen basieren sollte, die Auskunft darüber geben, was unter ‚guter Lehre' oder ‚Qualität in der Lehre' im Einzelnen zu verstehen sei. Für die Ableitung von Qualitätskonzepten sind die Arbeiten von Michal Scriven hilfreich. Scriven (1980) unterscheidet vier verschiedene Phasen im Evaluationsprozess.

Die ersten beiden Schritte eines jeden Evaluationsvorhabens beinhalten nach Scriven (1980) die Festlegung von Kriterien und Standards, die die Grundlage der Evaluation (der Analyse, des Evaluationsinstrumentes) sein sollen. Kriterien definieren, welche Leistungsaspekte evaluiert werden sollen. Standards definieren das erwünschte Mindestmaß an Leistung, das es zu erzielen gilt. Die Schritte drei und vier umfassen dann die eigentliche Analyse sowie die Integration aller Ergebnisse zu einem Werturteil (Synthese). Auf die Entwicklung eines Qualitätskonzepts 'gute Lehre' bezogen heißt dies, dass vor der Konstruktion von Instrumenten und ihrem Einsatz

1. definiert werden müsste, aus welchen Komponenten das Konstrukt ‚Lehre' besteht (Vereinbarung von Kriterien)
2. ein Leistungsmaß definiert werden müsste, das festlegt, was ‚gute' Lehre ist (Vereinbarung von Standards)

Solche Kriterien und Standards ließen sich aus wissenschaftlichen Theorien zum Konstrukt 'gute Lehre' ableiten. Fragebogen zur Lehrevaluation wären damit theoretisch fundiert, sofern sie sich explizit auf Qualitätskonzepte des Konstruktes ‚gute Lehre' bezögen. Ziel wäre also eine enge Verknüpfung zwischen der Konzeption der Lehrevaluationsinstrumente und dessen, was die wissenschaftliche Forschung an Konzepten und Theorien bereitstellt. Wie auch immer diese Ableitungen oder Verknüpfungen im Einzelfall aussehen mögen, der Kerngedanke des Konzeptes von Scriven ist die Transparenz und die Explikation der Grundlagen für ein Instrument oder einen Evaluationsprozess.

Finden sich ähnliche Gedanken und Ableitungen auch in der Literatur zur Konzeption von Fragebogen zur Lehrevaluation? Werden sie umgesetzt?

Leitgedanken wie die von Scriven (1980) zur Konzeption von Lehrevaluationsinstrumenten werden von einigen Autoren der anglo-amerikanischen Literatur angeführt (vgl. Arreola, 1995, 2000; Braskamp & Ory, 1994; Cashin 1996). Beispielsweise sehen Braskamp et al. (1984) in der Definition dessen, was unter ‚guter Lehre' zu verstehen sei, die oberste Voraussetzung von Lehrevaluation. Jeder Evaluationsprozess beginne mit dem Schritt der Definition. Dies entspräche dem Vorgehen Scrivens, mit der Definition des Kriteriums zu beginnen. Auch in Deutschland wird die

Verzahnung von theoretischen Annahmen über das Konstrukt ‚gute Lehre' mit der Konstruktion des Fragebogens erkannt. Beispielsweise findet sich bei Richter (1994) in einem Leitfaden für die Entwicklung von Fragebogen zur Lehrveranstaltungsevaluation der explizite Hinweis, zunächst mit der Entwicklung eines Lehrmodells zu beginnen. Und auch Engel (2001, S. 7) eröffnet seinen Beitrag mit den Worten „Lehre und Studium kann ohne Anlegung geeigneter Qualitätskriterien nicht beurteilt werden."

Als Konsequenz solcher Empfehlungen zur theoretischen Fundierung müsste sich die Diskussion um diese Kriterien und dann die darauf basierende theoretisch fundierte Konzeption der Fragebogen zur Lehrevaluation anschließen. Es sei aber darauf hingewiesen, dass die alleinige Verwendung von Theorien oder Konzepten für eine ausgewogenen Konzeption nicht ausreichend oder anderen Konstruktionsmethoden überlegen ist (Rindermann, 2001). Andere Methoden zur Itemgenerierung sollten bei der Konstruktion eines soliden Verfahrens hinzukommen (s. auch Kapitel 5). Dennoch: Bei der Entwicklung der Items auf eine Einbeziehung wissenschaftlicher Befunde über das interessierende Konstrukt zu verzichten, könnte mit Blick auf die Validität des Verfahrens problematisch sein. „Der Geltungsbereich eines Fragebogens kann nicht weiter oder enger, besser oder schlechter sein als die aufgrund vorhandener oder nur schwach vorhandener theoretischer Überlegungen bestimmte Itemsammlung." (Mummendey, 1995, S. 61-62). Weiterhin muss nach Meinung Mummendeys gewährleistet sein, dass sich jedes Item des Fragebogens vernünftig auf das theoretische Konstrukt beziehen lässt (Mummendey, 1995). Dies könnte durch die Einbeziehung von Theorien und Befunden zum interessierenden Konstrukt erreicht werden.

Für die *Praxis* in den USA verweist Marsh (1984) darauf, dass im Gegensatz zu den überwiegend empirischen Ansätzen ein theoriebasierter Ansatz noch nicht beschritten wurde: "An alternative approach based on a theory of teaching or learning could be used to posit evaluation dimensions, though such an approach does not seem to have been used in student evaluation research" (Marsh, 1984, S. 709).

In Deutschland haben zwar einige Autoren ihre Instrumente explizit unter Zuhilfenahme wissenschaftlicher Theorien und Konstrukte entwickelt (z. B. Westermann et al., 1998; Rindermann & Amelang, 1994), doch bilden sie eher die Ausnahme. Gelegentlich herrscht sogar Unklarheit darüber, aus welchen Quellen die Items generiert wurden. Weiterhin kritisiert Webler (1993, S. 417): „In jede Konstruktion eines Evaluationsfragebogens gehen regelmäßig Annahmen [didaktische Hypothesen] über Merk-

male guter Lehre mit ein. Sie werden aber meist nicht explizit formuliert, bestimmen aber die Auswahl der Variablen."

Die Umsetzung der Verknüpfung von theoretischen Konzepten zum Konstrukt ‚gute Lehre' mit einer explizit auf diesen Konzepten basierten Entwicklung eines Instrumentes zur Erfassung des Konstruktes ist gering. „In der Regel orientierten sich neue Verfahren mehr an Anforderungen der Praxis und weniger an theoretischen Modellen" (Rindermann, 2001, S. 55). Insgesamt können nur wenige Verfahren in diesem Sinne als theoretisch fundiert bezeichnet werden. In den meisten Fällen der Evaluation der Lehre wird direkt in der dritten Phase des Evaluationsprozesses (Scriven, 1980) eingestiegen und „es werden all jene heftig kritisiert, die auf die ersten beiden Phasen hinweisen, ohne die Evaluation bloß zu einer bürokratischen Übung verkommt" (Stangl, 2000, S. 3).

Auf den zweiten Blick verwundert die karge theoretische Fundierung der Verfahren jedoch nicht, da der Schritt der Festlegung von Kriterien, hier der Komponenten des Konstruktes ‚gute Lehre', schwierig ist. Dazu benötigte man als erstes eine Definition von ‚guter Lehre'.

Cashin (1989) untersucht bisherige Ansätze zur *Definition von Lehre* und schlussfolgert, dass diese Ansätze und damit die Forschung zur Evaluation von Lehre auf einer höchst unvollständigen Definition von teaching beruht. Sein Beitrag zeigt die Definitionsvielfalt zwischen verschiedenen Ansätzen auf, die versuchen, Kriterien oder Komponenten und Definitionen 'guter Lehre' zu explizieren. Diese Vielfalt sei anhand einiger ausgewählter Beispiele illustriert:

- Cashin (1989) präsentiert eine "erweiterte Definition" von Lehre, die sieben Bereiche umfasst: (1) subject matter mastery, (2) curriculum development, (3) course design, (4) delivery of instruction, (5) assessment of instruction, (6) availability to students, and (7) administrative requirements.
- Centra (1993) führt sechs Kriterien an, die effektive Lehre charakterisieren: (1) good organization of subject matter and course, (2) effective communication, (3) knowledge of and enthusiasm for the subject matter and teaching, (4) positive attitude toward students, (5) fairness in examination and grading, (6) flexibility in approaches to teaching.
- Arreola (2000) nennt vier Komponenten von Lehre: (1) instructional delivery skills, (2) instructional design skills, (3) content expertise, and (4) course management.

Obwohl diese Definitionen deutliche Überschneidungen haben, besteht dennoch das Problem einer fehlenden allgemein anerkannten Definition von guter Lehre (Cohen, 1981). Auch in Deutschland wurden zahlreiche Taxonomien und Definitionen zur „Lehre" publiziert:

- Für Rindermann (2001) besteht Lehrkompetenz aus den drei „Fähigkeitsbündeln": (1) Strukturierung sowie didaktische Methodenvielfalt und -sicherheit unter verschiedenen Unterrichtsbedingungen, (2) soziale Kompetenzen und (3) Persönlichkeitsmerkmale (Freundlichkeit, Offenheit, Engagement).
- Kramis (1990) orientiert sich an der Unterrichtsforschung und postuliert die drei folgenden „grundlegenden Gütekriterien für Unterricht": (1) Bedeutsamkeit der gewählten Unterrichtsinhalte und Ziele, (2) Effizienz der gewählten Lernorganisation, Lernaktivitäten und Medien sowie (3) gutes Lernklima.

Es könnten viele weitere Beispiele genannt werden, aber insgesamt gilt auch hier: Man ist weit davon entfernt, eine allgemein anerkannte Definition von ‚guter Lehre' zu erzielen. Schweer (2001, S. 160) resümiert im Handwörterbuch der Pädagogischen Psychologie: Es „ist noch völlig offen, was unter dem Globalbegriff ‚Qualität der Lehre' überhaupt zu verstehen ist. (...) Eindeutige theoretische Konzepte hierzu liegen bislang *nicht* vor." Kromrey (2001) hält die Definition von Kriterien und Standards im Bereich der Lehrevaluation für unmöglich, bzw. lehnt ein Anknüpfen an diesen Leitgedanken aus der Literatur zu Qualitätskonzepten ab. „Es wurde des weiteren festgestellt, auch das ‚Messen' von Qualität als Aufgabe von Evaluation sei nicht einlösbar" (Kromrey, 2001, S. 34).

Die Gründe für die fehlende Einigkeit oder die Unmöglichkeit einer allgemeinen Definition sind vielfältig. An erster Stelle steht vor allem die Komplexität des Lehrgeschehens. Zur Illustration sei hier nur genannt, dass es nach Abrami et al. (1996) allein drei verschiedene Perspektiven gibt, aus denen „lehren" bzw. „Lehre" definiert werden könne: eine Produkt-orientierte Definition, eine Prozess-orientierte Definition und eine Prozess-Produkt-orientierte Definition. Noch grundlegender wird das Problem bei el Hage (1996, S. 90) genannt, wonach bisher eine „Theorie des Unterrichts an Hochschulen" fehlt bzw. bei Rindermann (1999b, S. 3), wonach „grundlegende *theoretische* Konzeptionen zum Hochschulunterricht (...) vergleichsweise rar" seien.

Bezüglich eines Aspektes des Konstruktes ‚gute Lehre' besteht jedoch Einigkeit. Nach der allgemeinen Auffassung der Evaluationsforscher ist *das Konstrukt ‚Lehre' eindeutig multidimensional* (Abrami, 1989; Marsh,

1991; Marsh & Hocevar, 1991). Ein multidimensionales Konstrukt weist verschiedene Komponenten auf. In der Konsequenz bedeutet dies in methodischer Hinsicht, dass das Konstrukt nicht durch einen allgemein gültigen Indikator repräsentiert werden kann (McKeachie, 1990). Aus praktischer Sicht bedeutet es, dass ein Dozent gute Lehre leisten kann hinsichtlich einer Dimension (z. B. Enthusiasmus), aber unzureichende Lehre hinsichtlich einer anderen (z. B. Veranstaltungsorganisation).

Wie die Definitionsvielfalt von Lehre bereits vermuten ließ, besteht zwischen den Experten hinsichtlich der *Komponenten des Konstruktes und ihrer Beziehung zueinander keine Einigung*. Das Konstrukt ‚gute Lehre' gilt also als multidimensional, wobei unklar ist, aus welchen Komponenten es sich zusammensetzt.

7.2 Verlagerung der Debatte auf die Dimensionalität von Fragebogen

In engem Zusammenhang mit der Frage nach der Dimensionalität des Konstruktes ‚gute Lehre' steht die Frage nach der Dimensionalität von Fragebogen zur Lehrevaluation. Die Debatte um die Komponenten des Konstruktes ‚gute Lehre' wurde und wird in der Literatur zur Lehrevaluationsforschung gar nicht explizit geführt. Sie verlagerte sich auf die Diskussion der Dimensionen in Fragebogen zur Lehrevaluation, was an folgender Auseinandersetzung zu entnehmen ist:

Seit den 1980er Jahren wird von einigen Autoren eine Auseinandersetzung geführt über die Frage, ob Fragebogen zur Lehrveranstaltungsevaluation multidimensional oder unidimensional konzipiert werden sollten. Zumindest hinsichtlich der summativen Verwendung der Daten spricht sich Abrami vielfach dafür aus, dass es trotz der Multidimensionalität des Konstruktes sinnvoller sei, gute Lehre unidimensional zu erfassen (vgl. Abrami, 1985, 1989, 2001). Marsh und Kollegen widersprechen dieser Position vehement und fordern multidimensionale Instrumente (vgl. Marsh, 1984, 1991; Marsh & Bailey, 1993).

In Deutschland dagegen gibt es diese Diskussion nicht, vermutlich weil die sinnvolle Verwendung von Daten studentischer Lehrevaluationen für Personalentscheidungen und ähnliche Zwecke noch grundsätzlich angezweifelt wird (vgl. Kap. 6). Nichts desto weniger gehen hierzulande Autoren davon aus, dass „die Multidimensionalität der Erfassungsmethode der wohl wichtigste Grundstein für eine möglichste hohe Güte des Verfahrens und die Gültigkeit der Befunde" ist (el Hage, 1996, S. 86).

Selbst bei Einigkeit in der Meinung, dass Fragebogen multiple Dimensionen enthalten sollten, bliebe immer noch zu klären, *welche Dimensionen* in Fragebogen zur Lehrevaluation aufgenommen werden sollten (el Hage, 1996, S. 87). Um die Dimensionen eines Fragebogens zu bestimmten, lassen sich nach Sichtung der Literatur neben der Orientierung an Theorien oder Konzepten zu den Komponenten des Konstruktes zwei weitere Vorgehensweisen unterscheiden:

1) *Empfehlungen der Literatur (prospektive Entscheidung):* Bei dieser Vorgehensweise werden die Dimensionen des Fragebogens im Vorfeld festgelegt, indem man sich auf Empfehlungen der gängigen Literatur beruft. Einige Autoren proklamieren in Handbüchern und Ratgebern, dass ein Set typischer Dimensionen für Fragebogen zur Lehrevaluation existiere:

   Centra (1993), ebenso Braskamp und Ory (1994) führen sechs solcher Faktoren an: (1) Organization, planning or structure, (2) Teacher student interaction or rapport, (3) Clarity, communication skill, (4) Work load, course difficulty, (5) Grading and examinations, assignments, (6) Student learning, student self ratings of accomplishments or progress. Die Autoren gelangten zu dieser Einschätzung durch Literaturstudium. Im deutschen Sprachraum hat bisher lediglich el Hage Dimensionen zusammengestellt, die „als typisch gelten dürfen" (el Hage, 1996, S. 93). Sie nennt acht ebenfalls durch Literaturstudium extrahierte Faktoren: (1) Zuwendung, (2) Fairness von Prüfungen und Benotungen, (3) Kommunikationsfähigkeit, (4) Kurs- bzw. Stofforganisation, (5) Stimulierung, (6) Variabilität vs. Monotonie, (7) Enthusiasmus, (8) Kurswert.

2) *Empirisch (retrospektive Entscheidung)*: Die Dimensionen des Fragebogens können auch im Nachhinein über empirische Methoden ermittelt werden. An erster Stelle stehen dazu faktorenanalytische Verfahren zur Verfügung. Viele Veröffentlichungen beider Kulturräume beschäftigten sich mit faktorenanalytischen Befunden und Fragestellungen der Instrumente zur Lehrveranstaltungsevaluation (z. B. Banz & Rodgers, 1985; Burdsal & Bardo, 1986; Feldman, 1989). Für Deutschland bleibt darüber hinaus festzuhalten, dass auf diese Form der Bestimmung der Dimensionalität eines Instrumentes vielfach verzichtet wurde: „Häufig kann nur indirekt aus den zur Bewertung vorgelegten Items erschlossen werden, welche Faktoren für eine gute Lehre aus Sicht der Initiatoren ausschlaggebend sind (Bülow-Schramm & Reissert, 1993, S. 401; ähnlich Feuerstein,

1997). Eine Zusammenstellung faktorenanalytischer Befunde verschiedener Instrumente bietet Rindermann (2001).

In der Praxis sind vielfach Mischformen beider Vorgehensweisen anzutreffen (Rindermann, 2001). Publiziert werden aber vor allem die Befunde der Faktorenanalyse. Ein allgemeines Problem dieser faktorenanalytischen Ergebnisse ist, dass die Faktorenstrukturen nicht gut repliziert werden konnten (z. B. Hofmann, 1988; Astleitner & Krumm, 1996). Eine Ausnahme stellt der SEEQ von Marsh dar, dessen 9-Faktoren-Lösung in mehr als 30 explorativen Faktorenanalysen über verschiedene Kontexte hinweg Unterstützung fand (vgl. Marsh 1982, 1984, 1987, 1991; Marsh & Bailey, 1993; Marsh & Hocevar, 1984, 1991). Aber auch eindrückliche Befunde zur Replizierbarkeit der Faktorenstruktur eines Instrumentes über Kontexte hinweg (z. B. Marsh's SEEQ) garantiert nicht die Replizierbarkeit einer Faktorenstruktur über verschiedene Instrumente (Abrami & d'Apollonia, 1990).

Die *entscheidende Frage* ist also die, *ob sich gleiche oder zumindest ähnliche Faktoren über verschiedene Instrumente hinweg finden lassen*. Oder wie Abrami et al. (1996, S. 226) formulieren: „Another type of evidence concerning the validity of rating forms comes from comparisons of items on different rating forms". Sollte sich also so etwas wie ein „typisches Faktorenbündel" über verschiedene Instrumente hinweg empirisch auffinden lassen, so könnte dies für die Klärung der Erfassung des Konstruktes ‚gute Lehre' hilfreich sein. In einem solchen Falle käme man zumindest einer operational einheitlich gefassten Definition der Komponenten des Konstruktes ‚gute Lehre' näher: Gute Lehre bestünde dann aus den Dimensionen, die typischerweise in Fragebogen zur Lehrveranstaltungsevaluation repräsentiert sind.

Bei der Suche nach einem allgemeingültigen oder typischen Bündel relevanter Dimensionen hilft ein einfacher Vergleich der Ergebnisse von Faktorenanalysen nicht weiter. Die in der Literatur zu Lehrevaluationsfragebogen beschrittenen faktorenanalytischen Vorgehensweisen sind zu verschieden und Ergebnisse nur begrenzt vergleichbar. Auch sagt die Interpretation der erhaltenen Faktoren (ihre inhaltliche Benennung) nichts über die den Faktoren zugeordneten Items aus. ‚Transparenz' im Fragebogen A kann völlig anders definiert sein als in Fragebogen B. Die Ähnlichkeit oder Unähnlichkeit von Faktoren aus Fragebogen über verschiedene Instrumente hinweg ist nur auf Itemebene abschätzbar (vgl. el Hage, 1996; Diehl, in Druck).

Der Frage nach der Invarianz oder Spezifität der Dimensionen über verschiedene Fragebogen hinweg hat sich in empirischer Form bisher nur die Forschergruppe um Philip C. Abrami zugewandt. Bislang publizierten sie dazu zwei Studien, die in diese Thematik passen (Abrami & d'Apollonia, 1990; Abrami et al., 1996). Sie werden wegen ihres Stellenwertes für die hier aufgeworfene Fragestellung im nächsten Abschnitt beschrieben.

7.3 Vergleichende Studien zur Dimensionalität von Fragebogen

Im Folgenden werden zwei für die Forschung bedeutsame Studien vorgestellt, die sich mit Aspekten der Frage der Invarianz oder Homogenität von Dimensionen über verschiedene Lehrevaluationsinstrumenten beschäftigt haben. Die erste Studie von Abrami und d'Apollonia (1990) untersucht die Häufigkeiten und die Uniformität von Dimensionen verschiedener Fragebogen zur Lehrveranstaltungsevaluation. In der zweiten Studie versuchen Abrami, d'Apollonia und Rosenfield (1996) ein Set von Faktoren zu ermitteln, das über verschiedene Fragebogen zur Lehrveranstaltungsevaluation hinweg invariant, also in allen Verfahren anzutreffen ist.

Die Studie von Abrami und d'Apollonia (1990)

Um die Uniformität der einzelnen Dimensionen in verschiedenen Fragebogen der Lehrevaluation zu untersuchen, wurden die Items aus verschiedenen Instrumenten herangezogen. In die Untersuchung gingen Items aus Fragebogen ein, die in verfügbaren und methodischen Ansprüchen genügenden Validitätsstudien mit Multisection-Design (vgl. Kap. 6) vorkamen und für die Korrelationen mit einem Kriterium für Lehrerfolg (z. B. Klausurnote) vorlagen[13]. Bei den in den Studien berichteten Validitätskoeffizienten handelt es sich jeweils um Korrelationen zwischen Rating (Veranstaltungsmittelwerte) und studentischer Leistung (Abschlussklausur, Mittelwerte). Insgesamt wurden 44 Studien aufgenommen, was zu einer Berücksichtigung von 752 Validitätskoeffizienten aus 154 berichteten Einzelergebnissen führte.

Die Autoren konzipierten ein Maß für die Uniformität, den Uniformitätsindex (UI). Dieser versteht sich als ein Maß für die Dimensionalität der Koeffizienten einer Dimension über verschiedene Fragebogen hinweg. Ein hoher Index bedeutet, dass die berichteten Validitätskoeffizienten

[13] Nähere Informationen zur Auswahl der Studien und den methodischen Anforderungen sind bei Abrami et al. (1988) zu finden.

eindimensional sind, d.h. dass diese Koeffizienten über die verschiedenen Fragebogen hinweg *eine Dimension* repräsentieren.

Um diesen Index ermitteln zu können, wurden die Items von 2 unabhängigen Urteilern 24 Kategorien möglicher Dimensionen von Fragebogen zur Lehrveranstaltungsevaluation zugeordnet. Das dabei verwendete Kodierschema basiert auf einem von Feldman (1977) entwickelten Schema mit ursprünglich 21 Kategorien. Das Feldman-Schema wurde in dieser Studie noch um die drei Kategorien 'globales Urteil Dozent/Dozentin', 'globales Urteil Veranstaltung' und 'Misceallaneaous' erweitert. Die Interrater-Reliabilität betrug .93 (Cohen's kappa).

Die Untersuchung ergab, dass alle Kategorien - jedoch mit unterschiedlicher Häufigkeit - in den Studien repräsentiert sind. Die am häufigsten vorkommenden Kategorien waren 'Klarheit und Verständlichkeit' (N = 112)[14] sowie 'globales Urteil Veranstaltung' (N = 109). Die am geringsten repräsentierten Kategorien waren 'Enthusiasmus' (N = 30) und 'zusätzliches Material' (N = 26).

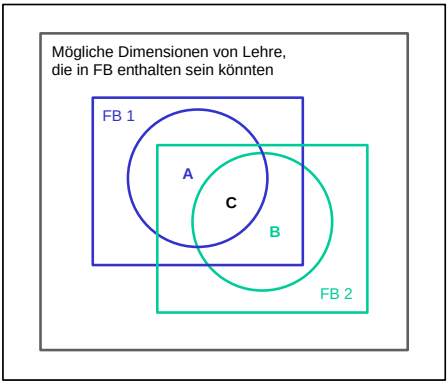
Mit Ausnahme der globalen Kategorien ergaben sich jedoch durchweg niedrige Uniformitätskoeffizienten (bei zusätzlich hohen Standardabweichungen). Die Autoren schließen daraus, dass die berichteten Validitätskoeffizienten multidimensional seien. Demnach variieren die Items - wie auch die Faktoren, die sie repräsentieren - in den verschiedenen Validitätsstudien. Inhaltlich bedeutet das, dass die Kategorie 'Veranstaltungsorganisation' in Fragebogen A etwas anderes erfasst als die gleichnamige Dimension in Fragebogen B. Schließlich zeigen auch die hohen Standardabweichungen der Uniformitätsindizes, dass die Dimensionen in den verschiedenen Fragebogen wenig konsistent sind.

Die Studie von Abrami, d'Apollonia und Rosenfield (1996)

In einer neueren Untersuchung von 1996 greifen Abrami und Kollegen die Frage nach der Replizierbarkeit von Faktoren über verschiedene Fragebogen hinweg auf. Sie versuchten mittels Faktorenanalyse ein sogenanntes "common core" von Faktoren aus verschiedenen Fragebogenverfahren zu extrahieren. Ein auf diesem Wege definiertes „common core" entspräche dem Bereich C in der nachfolgenden Abbildung.

[14] Das maximale N, das pro Kategorie erzielt werden konnte, entsprach der Anzahl der Einzelergebnisse aus den Studien, also $N_{(max)} = 154$.

Abb. 6. Die Suche nach dem „common core" über Faktorenanalysen



Erläuterung:

Großes Rechteck: alle Dimensionen, die evtl. in Fragebogen zur Lehrevaluation enthalten sein können.

Kleines Rechteck: Dimensionen eines Fragebogens [FB 1 und FB 2]

Kreis: Anteil, der durch Faktorenanalyse des jeweiligen FB aufgeklärt werden kann [A und B]

Schnittmenge der Kreise: Anteil, der durch eine Faktorenanalyse über beide FB aufgeklärt werden kann [**C**]

Die Abbildung beschränkt sich dabei zur besseren Anschaulichkeit auf die Darstellung der Verhältnisse nur für zwei Fragebogen, wobei in die Analyse aber mehrere Instrumente eingegangen sind. Aus 17 Studien, in denen „die meisten üblichen Fragebogen repräsentiert sind"[15], wurden 225 Items extrahiert und in 40 verschiedene Kategorien kodiert. Für die Kodierung verwendeten die Autoren eine erneut variierte Version des Feldman-Schemas (Feldman, 1977). Fünf Kategorien wurden nicht besetzt. Das Kernziel der Studie war, eine Faktorenanalyse über alle in den Studien enthaltenen Interitemkorrelation aus den verschiedenen Fragebogen durchzuführen. Nach verschiedenen Schritten der Datenanalyse (u.a. Aggregation für die Kategorien) wurde eine Faktorenanalyse über eine 35x35 Faktoren-Interkorrelationsmatrix gerechnet. In die Matrix gingen 6788 Interitemkorrelationen der bereits genannten 225 Items ein. Die Autoren extrahierten 4 interkorrelierte Faktoren (vgl. Tab. 10).

[15] Es gibt keine weiteren Hinweise, um welche Verfahren es sich dabei handelte.

Tab. 10. Faktorenanalytische Ergebnisse der Studie von Abrami et al. (1996)

Faktor	Eigenwert	Erklärte Varianz	Kategorien [N*]
Instructor viewed in an instructional role	8.0	62.8%	13
Instructor viewed as a person	5.6	4.2%	16
Instructor viewed as regulator	2.5	3.7%	2
wenig einheitlich, nicht interpretierbar	1.8	2.9%	4
		Σ 73,6%	Σ 35

Anmerkung. *N gibt die Anzahl der Kategorien an, die in diesem Faktor repräsentiert sind.

Auf den ersten dieser vier Faktoren laden 13 Kategorien, die zusammen genommen den Veranstaltungsleiter in seiner Rolle als *Lehrenden* beschreiben (z. B. Enthusiasmus zu lehren, Stimulation des Interesses, Klarheit in der Lehre). Dieser Faktor wird von den Autoren auch als „g-Faktor" für die „allgemeine Fähigkeit zu lehren" (general instructional skill) interpretiert, da dieser Faktor allein etwa 63% Varianz erklärt. Der Lehrende in seiner Rolle *als Person* ist im zweiten Faktor repräsentiert. Auf diesen Faktor laden 16 Kategorien (u.a. Beantwortung von Fragen, Respekt für andere, Enthusiasmus für Studierende). Der dritte Faktor subsumiert die Kategorien Feedback und Evaluation und beschreibt den Lehrenden als *Regulator*. Der vierte Faktor ist nach Meinung der Autoren schwierig zu interpretieren, da wenig einheitliche Kategorien hoch auf ihn laden (z. B. Supervision, Expertenwissen, Wahl des Veranstaltungsmaterials etc.).

Aus diesen Ergebnissen schließt die Autorengruppe, dass das „common set" durch die 4-Faktoren Lösung abgesteckt sei: „Our factor analysis across the multiple rating forms indicates that there is a ‚common' structure to instructional effectiveness" (Abrami et al., 1996, S. 251). Die Autoren führen fort, dass die ersten beiden Faktoren aufgrund deutlich höherer Eigenwerte bedeutsamer seien als die Faktoren drei und vier. Dies gelte insbesondere für den ersten Faktor, einen g-Faktor, der die allgemeine Fähigkeit zu lehren beschreibe. Die Autoren sind ferner der Meinung, dass „Lehre" multidimensional sei, die Dimensionen von Lehre aber inkonsistent über die verschiedenen Fragebogen hinweg seien.

„We also believe that effective teaching is multidimensional but that there are differences across rating forms concerning the specific dimensions, that underlie effective instruction" (Abrami et al., 1996, S. 252).

7.4 Zusammenfassung und Fazit

Infolge der Auseinandersetzung mit dem Forschungsstand von Fragebogen zur studentischen Lehrveranstaltungsevaluation wurden zwei Problemfelder herausgearbeit: (1) die fehlende theoretische Fundierung sowie Definitionsprobleme und (2) die Verlagerung der Debatte um die Konstruktvalidität auf die Dimensionalität von Fragebogen. Aus der Diskussion dieser beiden Problemfelder bleibt folgendes festzuhalten:

Die anglo-amerikanische und die deutschsprachige Lehrevaluationsforschung sind gekennzeichnet durch fehlende oder allenfalls geringe theoretische Fundierung der Fragebogen zur Lehrveranstaltungsevaluation. Dies steht in engem Zusammenhang mit der fehlenden Einigkeit bezüglich einer allgemein anerkannten Definitionen des Konstruktes ‚Lehre' bzw. ‚guter Lehre'. Sind sich alle Beteiligten noch darüber einig, dass das Konstrukt multidimensional ist, herrscht Uneinigkeit darüber, welche Komponenten das Konstrukt umfasse und in welcher Beziehung die Komponenten zueinander stünden.

Es wurden weiterhin festgestellt, dass sich die Debatte von der Ebene der Komponenten des Konstruktes ‚gute Lehre' auf die Ebene der Diskussion der Dimensionen von Fragebogen verlagert hat. Diesbezüglich kam die Frage auf, welche Dimensionen in Fragebogen anzutreffen sind. In der Erörterung dieser Frage wurde insbesondere auf faktorenanalytische Befunde aus Einzelstudien und Literaturübersichten eingegangen. Als Hinweis auf eine bisher wenig diskutierte Form der Validität von Fragebogen zur Lehrevaluation – und somit auch als Hinweis auf das dahinter liegende Konstrukt - könnte die Extraktion eines gemeinsamen Bündels von Dimensionen sein, das über verschiedene Instrumente hinweg invariant ist.

Zwei Studien wurden in diesem Zusammenhang vorgestellt. Die Ergebnisse der Studien zeigten, dass

- Dimensionen über verschiedene anglo-amerikanische Fragebogen hinweg inkonsistent definiert sind (niedrige Uniformitätsindizes bei Abrami & d'Apollonia, 1990).

- eine 4-Faktorenlösung mit einem starkem g-Faktor „allgemeine Lehrfähigkeit" aus verschiedenen anglo-amerikanischen Instrumenten extrahiert werden konnte (Abrami et al., 1996).
- nach wie vor „Inkonsistenz besteht hinsichtlich der Dimensionen von Lehre, insbesondere über verschieden Fragebogen hinweg" (Abrami et al., 1996, S. 251).

Bei der Darstellung der beiden Problemfelder wurden beinahe ausschließlich anglo-amerikanische Autoren herangezogen. Dies galt insbesondere für die beiden aus dem amerikanischen Raum stammenden Studien. Folglich beziehen sich die dargestellten Ergebnisse nur auf anglo-amerikanische Instrumente. Weiterhin ist der Generalisierungsgrad der beiden berichteten Studien eingeschränkt, da die Berücksichtigung von Fragebogen bzw. Items an Publikationen mit besonderen methodischen Anforderungen gebunden war.

8. Empirischer Teil

8.1 Gesamtrahmen der Fragestellung: Die Dimensionalitätsdebatte

Diese Arbeit möchte einen Beitrag zur Transparenz der Aktivitäten und Verfahren im Bereich der Lehrevaluation aus wissenschaftlicher Sicht leisten. Dies erforderte die kritische Auseinandersetzung mit dem aktuellen wissenschaftlichen Forschungstand bzw. der deutschsprachigen und anglo-amerikanischen Literatur.

Als Ergebnis dieser Auseinandersetzung wurden in Kapitel 7 bereits einige problematische Aspekte hinsichtlich der Erfassung des Konstruktes ‚gute Lehre' in Fragebogenverfahren diskutiert. Unter den dort genannten erscheint insbesondere die Frage nach der *Dimensionalität des Konstruktes* ‚gute Lehre' von zentraler Bedeutung. Wie gezeigt wurde, ist mit der Frage nach der Dimensionalität des Konstruktes der Aspekt der *Dimensionalität von Fragebogen* eng verknüpft. Hinsichtlich der Frage der Homogenität bzw. Heterogenität von Dimensionen über verschiedene Instrumente wurden bereits einige Ergebnisse vorgestellt. Sie stammen aus dem amerikanischen Raum und bilden den Ansatzpunkt für die empirische Fragestellung dieser Arbeit. Die darin thematisierten Fragen können aber auch in Bezug auf deutschsprachige Fragebogen geführt werden:

Welche Faktoren sind in deutschsprachigen Instrumenten zu finden? In welchem Ausmaß sind verschiedene Faktoren repräsentiert? Lässt sich ein Bündel typischer Faktoren aus deutschsprachigen Verfahren extrahieren?

Als Hinweis auf eine Antwort können aus dem deutschsprachigen Raum bisher nur rein tabellarische Auflistungen der Faktoren verschiedener deutschsprachiger Instrumente (z. B. Rindermann, 2001; Diehl, in Druck) herangezogen werden. Empirische Analysen im Sinne von Abrami und d'Apollonia (1990) fehlen. Diese Situation bildet den Hintergrund für die empirische Fragestellung dieser Arbeit.

8.2 Wahl der Untersuchungsart

Gemäß der oben genannten Fragestellung sollte eine Methode zur Analyse der Instrumente angewendet werden, die die einzelnen Verfahren einem Vergleich zugänglich macht. Zugleich erschien eine umfangreiche und aufwendige Datenerhebung mit verschiedenen, parallel eingesetzten Verfahren in ausreichend großen Lehrveranstaltungen nicht realisierbar, da dies den Zeitrahmen und Umfang einer Diplomarbeit sprengen würde.

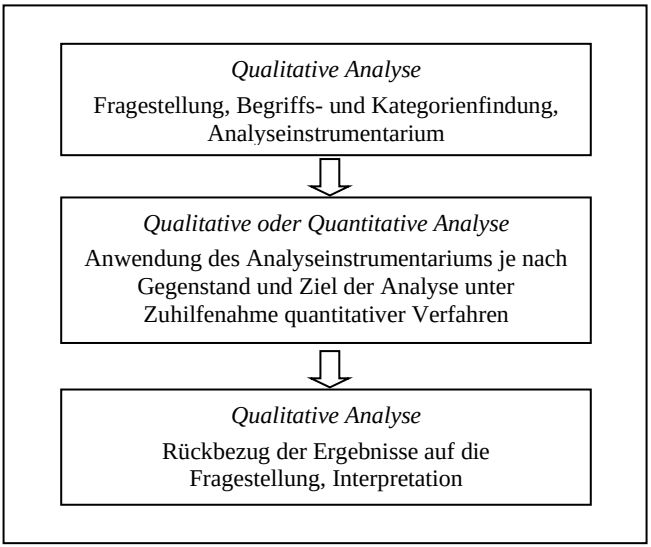
Als Datenmaterial für die Untersuchung sollten daher die Items aus bereits existierenden Fragebogen herangezogen selbst werden. Eine klassische Methodengruppe zur Auswertung verbaler Daten sind inhaltsanalytische Verfahren, wobei im allgemeinen zwischen qualitativen und quantitativen Inhaltsanalysen unterschieden wird.

Die *quantitative Inhaltsanalyse* „strebt eine Zuordnung der einzelnen Teile eines Textes zu ausgewählten, übergreifenden Bedeutungseinheiten (Kategorien) an“ (Bortz & Döring, 1995, S. 138). Ziel ist es, das Wortmaterial hinsichtlich bestimmter Aspekte zu quantifizieren. Demgegenüber werden bei *qualitativen Inhaltsanalysen* die zugeordneten Textteile interpretiert (Bortz & Döring, 1995).

In der Geschichte der Inhaltsanalyse gab es heftige Diskussionen, ausgelöst und getragen durch jeweils einseitige Forderungen nach einer rein quantitativen bzw. rein qualitativen Inhaltsanalyse (ausführlicher Merten, 1995).

Mayring (2000) dagegen plädiert für eine Überwindung des Gegensatzes qualitativ-quantitativ, da im gesamten Verlauf des Forschungsprozesses eine Abfolge quantitativer und qualitativer Schritte auftreten und in ein gemeinsames Ablaufmodell integriert werden können (vgl. Abb. 7).

Abb.7. Phasenmodell zum Verhältnis quantitativer und qualitativer Analyse (Mayring, 2000, S. 20)



Vor der Wahl einer bestimmten Form der Inhaltsanalyse sind die verschiedenen möglichen Ziele oder Grundformen zu bedenken (vgl. Mayring, 2000):

Zusammenfassung	Ziel: Datenreduktion, wobei die wesentlichen Inhalte erhalten bleiben sollen, aber durch Abstraktion auf einen überschaubaren Korpus verdichtet werden, der jedoch noch immer ein Abbild des Gesamtmaterials ist.
Explikation	Ziel: Erweiterung des Verständnisses, wozu zu einzelnen fraglichen Textteilen zusätzliches Material herangetragen wird, das die Textstelle erläutert, erklärt, ausdeutet.
Strukturierung	Ziel: Bestimmte Aspekte aus dem Material herauszufiltern, unter vorher festgelegten Ordnungskriterien einen Querschnitt durch das Material zu legen oder das Material aufgrund bestimmter Kriterien einzuschätzen.

Die Stoßrichtung der vorliegenden Arbeit fällt in die Grundformen Strukturierung und Zusammenfassung. Inhaltsanalysen mit dieser Zielrichtung werden bei Lamnek (1993) unter dem Stichwort „reduktive, und damit eher quantitative Inhaltsanalyse“ (S. 191ff) betrachtet. Merten (1995) behandelt Inhaltsanalysen mit der Zielrichtung Strukturierung bzw. Ordnen und Zusammenfassung bzw. Verdichtung ebenfalls unter dem Stichwort quantifizierende Verfahren:

> „[Die] *Quantifzierende* Vorgehensweise hat ordnende Funktion, indem sie Vergleichbarkeit erzeugt, Informationen verdichtet und vor allem die Verwendung von *Ziffern* gestattet, also durch Abstraktion die semiotische Dimension hinter sich läßt“ (Merten, 1995, S. 50).

Ähnlich argumentieren auch Bortz und Döring (1995), wenn sie vorschlagen, dass „quantitative Inhaltsanalysen immer dann indiziert [sind], wenn es darum geht, ausgewählte Einzelaspekte von Texten oder eng umrissene Fragestellungen systematisch und u.U. auch hypothesengeleitet zu untersuchen“ (Bortz & Döring, 1995, S. 140). Denn „indem man qualitativ erhobene Daten später quantifiziert und quantitativ weiterverarbeitet, vollzieht man den Übergang vom qualitativen zum quantitativen Ansatz.“ (Bortz & Döring, 1995, S. 273).

Der Zielsetzung und der eng formulierten Fragestellung der vorliegenden Arbeit entsprechend, wurde ein *quantitativer Ansatz* gewählt. Doch auch innerhalb der quantitativ ausgerichteten Inhaltsanalyse gibt es verschiedene Vorgehensmöglichkeiten. Mayring (2000) unterscheidet drei klassische Grundtechniken: *Häufigkeitsanalysen* als Auszählung der Ausprägungen pro Kategorie; *Kontingenzanalysen*, die in Kontingenztafeln münden und über das gemeinsame Auftreten von mindestens zwei Merkmalen informieren; sowie *Valenz- und Intensitätsanalysen*, bei denen Textbestandteile durch Schätzurteile quantifiziert werden. Die verschiedenen Formen und ihre Zielorientierungen sind in Tabelle 11 zusammengefasst.

Tab. 11. Analyseformen de Quantitativen Inhaltsanalyse (aus Mayring 2000, S. 57)

Analyse	Charakterisierung	Grundform
Häufigkeits-analysen	Herausfiltern bestimmter Textbestandteile durch Kategoriensystem; Aussagen über relatives Gewicht dieser Textbestandteile per Häufigkeit	Strukturierung Zusammenfassung
Kontingenz-analyse	Herausfiltern bestimmter Textbestandteile durch Kategoriensystem; Herausarbeiten einer Struktur durch häufige Kontingenzen; Erklärung einzelner Textbestandteile durch Kontingenzen	Strukturierung Zusammenfassung Explikation
Valenz- und Intensitäts-analyse	Herausfiltern bestimmter Textbestandteile durch Kategoriensystem; Einschätzung (Skalierung) aufgrund des Kontextes; Zusammenfassung der Einschätzungen;	Strukturierung Zusammenfassung

Als Ansatz für diese Untersuchung schieden Valenz- bzw. Intensitäts- und Kontingenzanalyse aus, da weder eine skalierende Einschätzung der Fragebogenitems auf Skalen (Intensitätsschätzung) noch die Frage nach typischen Kombinationen (Kontingenzen) von Fragebogenmerkmalen einen Beitrag zum oben skizzierten Dimensionalitätsproblem liefern würde. Geeignet dagegen schien die Häufigkeitsanalyse; denn es interessierten primär, *welche* Dimensionen in Fragebogen repräsentiert sind (Kategorie x vorhanden versus nicht vorhanden) bzw. *in welchem Ausmaß* sie vorhanden sind (Häufigkeit von Kategorie x). Tabelle 12 gibt die Analyseschritte der Häufigkeitsanalyse wieder, die auch das methodische Vorgehen der vorliegenden Untersuchung leiteten.

Tab. 12. Analyseschritte einer Häufigkeitsanalyse (nach Mayring, 2000, S. 14)

Vorgehen bei Häufigkeitsanalysen
• Formulierung der Fragestellung
• Bestimmung der Materialstichprobe
• Aufstellen eines Kategoriensystems
• Definition der Kategorien
• Bestimmung der Analyseeinheiten
• Kodierung
• Verrechnung, d.h. Feststellen und Vergleichen der Häufigkeiten
• Darstellung und Interpretation der Ergebnisse

Wenn in dieser Arbeit vom quantitativen Ansatz gesprochen wird, schließt dies explizit ein, dass einige Analyseschritte genuin qualitativer Natur sind (z. B. die Konkretisierung der Fragestellung, die Wahl eines Kategoriensystems, die Bestimmung der Analyseeinheiten).

8.3 Methodisches Vorgehen

8.3.1 Konkretisierung der Fragestellung

Aufgrund der bereits dargestellten Diskussion (vgl. Kap. 7) entstand der Eindruck, dass Fragebogen zur Lehrevaluation sehr heterogen sind hinsichtlich der in ihnen berücksichtigten Dimensionen des Konstruktes „Lehre“. Der Diskussion der Literatur folgend erschien es deshalb lohnend, den Versuch zu unternehmen, zu vergleichenden Aussagen hinsichtlich der Dimensionalität von Fragebogen zur Lehrevaluation zu gelangen. Die vorliegende Untersuchung zielte daher darauf hin, ein Ergebnis zu erhalten, das ordnende Funktion übernehmen kann und Informationen verdichtet (Strukturierung, Zusammenfassung). Es wurde angestrebt, die Dimensionen verschiedener Verfahren herauszuarbeiten und miteinander zu vergleichen. Aus methodischer Sicht verfolgte die Untersuchung das Ziel, einen Beitrag zur Konstruktvalidität von Fragebogen zur Lehrevaluation zu leisten.

Die konkreten Fragen, die die vorliegende Untersuchung zu beantworten versucht, lauten:

- Welche Faktoren sind in den verschiedenen Instrumenten zur Lehrveranstaltungsevaluation zu finden - und in welchem Ausmaß? Gibt es ein Bündel typischer Faktoren, das in deutschen Fragebogen zur Lehrveranstaltungsevaluation repräsentiert ist?
- Mit Blick auf die einflussreichen anglo-amerikanischen Forschung interessierte insbesondere auch, ob sich Verfahren aus dem deutschen Sprachraum von amerikanischen Verfahren hinsichtlich der Dimensionalität unterscheiden. Die vorliegende Untersuchung blieb daher nicht auf deutschsprachige Instrumente beschränkt. Weiterhin konnte dann gefragt werden, ob sich ein für beide Kulturräume gemeinsames Bündel typischer Faktoren extrahieren lässt.
- Eng verwoben mit der Frage nach typischen Dimensionen ist die nach der Frage Verteilung der Dimensionen in den Verfahren. Gestalteten sich Fragebogen zur Lehrevaluation homogen, so wären in ihnen nicht nur die gleichen Dimensionen anzutreffen, sondern die Dimensionen würden auch mit gleicher Häufigkeit auftreten (Abrami et al., 1996).

8.3.2 Bestimmung der Materialstichprobe

Die theoretisch angestrebte Grundgesamtheit (vgl. Kromrey, 1998), auf die sich diese Untersuchung beziehen will, sind alle entwickelten und aktuell eingesetzten Fragebogen zur Evaluation von Lehre in Deutschland bzw. derselben in den USA. Da die Grundgesamtheit niemals vollständig und korrekt zu erfassen ist, musste eine Erhebungsgesamtheit gebildet werden, aus der die Fälle später ausgewählt wurden (s. Kromrey, 1998).

Die Sammlung der Verfahren für die Erhebungsgesamtheit folgte einer zu diesem Zweck entstandenen Checkliste, die bei Arreola (2000) zu finden ist. Die Sammlung vollzog sich in zwei Schritten: Der Checkliste folgend sollten als erstes mögliche Verfahren gesichtet werden. Dazu wurden Literaturdatenbanken des Feldes (z. B. Eric, Psyclit, Psyndex) durchsucht; ausgewählte Zeitschriften (z. B. Instructional Education, Journal of Educational Measurement, Zeitschrift für Pädagogische Psychologie) gezielt durchgeschaut und den Literaturhinweisen aus Artikeln gezielt nachgegangen. Von den dort genannten Verfahren wurde in einem zweiten Schritt versucht, möglichst viele Verfahren zu beschaffen. Dazu wurden die in Artikeln oder im Verlag veröffentlichten Fragebogen bestellt; Autoren und Universitäten angeschrieben und um die Zusendung des Verfahrens und ggf. zugehöriger Unterlagen (Manual etc.) gebeten sowie im Internet zugängliche Verfahren heruntergeladen. Auf diese Weise ent-

standen die als Anhänge A und B aufgenommenen Listen verfügbarer deutscher und amerikanischer Verfahren. Sie bildeten die beiden Erhebungsgesamtheiten der Untersuchung, woraus anhand der folgenden Kriterien jeweils eine nicht willkürliche Auswahl (vgl. Kromrey, 1998) von Fällen getroffen wurde[16]:

Länge des Verfahrens: Die ausgewählten Verfahren sollten nicht zu viele Items umfassen, um nicht die Konzentration der Kodierer zu überfordern. Es wurden 60 Items pro Verfahren als obere Grenze festgelegt.

Anzahl: Es sollten insgesamt mindestens 10 Verfahren in die Stichprobe aufgenommen werden, wobei eine gleiche Anzahl deutscher und amerikanischer Verfahren in die Stichprobe gelangen sollte (mindestens je 5).

Verfahrenstyp: Aus dem zuvor genannten Aspekt der Länge des Verfahrens ergibt sich nahezu zwingend, dass nur Standardformen oder einzelne Verfahren aus multiplen Standardformen in Frage kamen[17], nicht dagegen Itempools oder Cafeteria-Systeme. Keine Rolle für die Auswahl spielte dagegen, ob das jeweilige Verfahren zu summativen oder formativen Zwecken entwickelt wurde.

Quelle: Es kamen nur Verfahren in Frage, die entweder als Originalbogen vorlagen oder im Original in Büchern oder Zeitschriften von den Erstautoren veröffentlicht wurden.

Aktualität: Es sollten nur Verfahren aufgenommen werden, die gegenwärtig eingesetzt werden oder solche, die sich explizit auf aktuelle Kurzverfahren beziehen[18].

Nach diesen Kriterien konnten schließlich je 7 Verfahren aus den in den Anhängen A und B genannten deutschen und amerikanischen Verfahren in die Untersuchung eingehen. Die Kurzbezeichnungen der Verfahren und die Anzahl der aufgenommenen Items sind in Tabelle 13 wiedergegeben (vgl. auch die Anhänge A und B).

Tab. 13. Stichproben dieser Untersuchung: Ausgewählte Verfahren (und Anzahl der Items)

[16] Für ein zufallsgesteurtes Auswahlverfahren war die Erhebungsgesamtheit zu klein und zu heterogen, da sie Standardformen, Itempools, Multiple Standardformen und Cafeteria Systeme umfasste.

[17] Lagen verschiedene Varianten für jeweils unterschiedliche Veranstaltungsformen vor (Seminar versus Vorlesung), so wurde grundsätzlich die Variante „Vorlesung“ in die Stichprobe aufgenommen.

[18] Beispielsweise versteht sich der VBVOR (Diehl, 1998) explizit als Kurzform des VBPYSCH (Diehl & Kohr, 1977).

Stichprobe deutscher Verfahren		**Stichprobe amerikanischer Verfahren**	
Name	Items (N)	Name	Items (N)
HILVE	37 [43]*	IDEA	47
BEVA	40	SIR II	40
FELL-V	23	SIRS	21
MFAL	17	SEEQ	31
VBPSYCH	40	CIEQ	21
FB-LV	55	ICE	55
FEVOR	20	SPTE	39
Gesamt	Σ 232	Gesamt	Σ 254
Mittel (gerundet)	**33**	Mittel (gerundet)	36

Anmerkung. *Von den 43 Items des HILVE gingen nur diejenige Items in die Analyse ein, die auch von Rindermann in die Faktorenstruktur einbezogen werden (N = 37). Bei den „fehlenden" 6 Items handelt es sich um 5 vom Veranstaltungsleiter frei formulierbare Items und um das Item Nr. 43, das nach dem Grund für den Besuch einer Veranstaltung fragt.

Alle Items gingen im originalen Wortlaut und in originaler Schreibweise ein. Einzige Ausnahme bildeten einige Items des BEVA und des FB-LV. Zur Erhöhung der Lesbarkeit wurden die Bezeichnungen „Seminar" im BEVA und „Vorlesung" im FB-LV den anderen Verfahren angeglichen und durch den allgemeineren Begriff „Veranstaltung" ersetzt. Nicht beachtet wurden die in einigen Verfahren anzutreffenden zusätzlichen Erhebungen wie z. B. demographische Variablen.

8.3.3 Das Kategoriensystem

Das Kategoriensystem stellt das zentrale Instrument einer quantitativen Inhaltsanalyse dar (Breakwell, Hammond & Fife-Schaw, 1995). Die Anwendung eines wohl definierten Kategoriensystems bildet den entscheiden Punkt für einen Vergleich bzw. die Intersubjektivität der Ergebnisse und die Abschätzung der Reliabilität der Analyse (Mayring, 2000). Im gleichen Sinne ist die Aussage Berelsons zu verstehen: „Content Analysis stands or falls by its categories." (Berelson, 1952, zitiert nach Merten, 1995, S. 147). Aus diesen Gründen werden hohe methodische Anforderungen an die Systeme zur Kategorisierung qualitativer Merkmale gestellt. Nach Bortz (1999) müssen die folgenden Kriterien für Kategoriensysteme erfüllt sein:

Genauigkeits-Kriterium:	Die Kategorien müssen exakt definiert sein.
Exklusivitäts-Kriterium:	Die Kategorien müssen sich gegenseitig ausschließen.
Exhaustivitäts-Kriterium:	Die Kategorien müssen das Merkmal erschöpfend beschreiben.

Es können induktiv oder deduktiv erstellte Kategoriensysteme verwendet werden. Im ersten Fall wird das Kategoriensystem aus dem ungeordneten, zu untersuchenden Textmaterial gebildet. Im zweiten Fall hingegen wird ein bereits ausgearbeitetes – oft theoriegeleitetes – Kategoriensystem für die Kategorisierung verwendet. Da die Entwicklung eines Kategoriensystems aufwendig ist und die methodischen Ansprüche hoch sind, lag die Wahl eines bereits vorhandenen und erprobten Systems nahe.

Für die vorliegende Untersuchung wurde daher das Kategoriensystem von Feldman (1989, vgl. Anhang D) ausgewählt, das aus 28 sogenannten "instruction characteristics" besteht. Für das System sprach, dass es die folgenden positiven Eigenschaften aufweist:

- Es wurde 1976 von Feldman theoriegeleitet und speziell zur Kodierung von Items aus Fragebogen zur Lehrevaluation (Standardformen) entwickelt.
- Es wurde seither mehrfach überarbeitet und ergänzt (Feldman 1977, 1988, 1989). Inzwischen liegen 28 definierte Kategorien mit Ankerbeispielen vor, womit Exklusivitäts- und Genauigkeits-Kriterium erfüllt sein dürften.
- Es gilt als die ausführlichste Sammlung möglicher Dimensionen von Fragebogen zur Lehrevaluation (Marsh, 1991). Damit erfüllt es im Vergleich zu weniger umfangreichen Kategoriensystemen am ehesten das Exhaustivitäts-Kiterium.
- Feldmans Kategoriensystem wurde bereits in mehreren Analysen verschiedener Autoren verwendet. Das hat den Vorteil, dass Ergebnisse dieser Untersuchung in Bezug zu anderen Ergebnissen, die mit dem Kategoriensystem gewonnen wurden (z. B. Abrami & d'Apollonia, 1990), diskutiert werden können.

Obwohl Abrami & d'Apollonia (1990) bzw. Abrami et al. (1996) das Schema in ihren Analysen aufgrund von Problemen modifizierten, wurde das Schema von Feldman in der Version von 1989 für die Analyse der vorliegenden Arbeit verwendet. Der ausschlaggebende Grund war, dass die modifizierte Variante von Abrami et al. (1996) mit 40 Kategorien

noch mal um einiges umfangreicher ist und weniger Ankerbeispiele für die Definition der Kategorien aufweist.

Zwei Varianten des Feldman Schemas (1989)

In die Untersuchung gingen zwei Varianten des Feldman Schemas ein. Für die Kategorisierung der Items aus deutschsprachigen Verfahren sollte ein deutschsprachiges Schema verwendet werden. Dazu wurde das Kategoriensystem von Feldman (1989) durch die Verfasserin dieser Arbeit aus dem Englischen ins Deutsche übersetzt (vgl. Anhang E). Die Übersetzung wurde von einer Anglistin überprüft. Die englischsprachige Version des Schemas wurde nicht verändert. Beide Systeme wurden allerdings um die Kategorien ‚Gesamturteil Veranstaltung' (overall course evaluation) und ‚Gesamturteil Dozent' (overall instructer evaluation) ergänzt. Weiterhin wurde eine Kategorie „00" eingeführt, um wirklich jedem Item eine Kategorie zuweisen zu können. Die Kategorie steht für Items, die weder den Kategorien Feldmans noch den globalen Kategorien zugeordnet werden konnten. Es war denkbar, dass in den letzten 12 Jahren neue Aspekte der Lehre in die Fragebogen aufgenommen wurden, an die 1989 weniger zu denken war (z. B. im Zusammenhang mit dem Internet).

8.3.4 Bestimmung der Analyseeinheiten

In diesem Schritt der Inhaltsanalyse sind nach Mayring die Kodier-, die Kontext- sowie die Auswertungseinheit zu definieren (vgl. Mayring, 2000). In der vorliegenden Untersuchung bildete das *Item* die zentrale Ebene der Betrachtung. Ein Item war zugleich Kodier- und Kontexteinheit. Es definierte die Kodiereinheit, da ein Item als kleinster Materialbestandteil in die Auswertung eingehen sollte (also keine Zuordnung einzelner Wörter). Ein Item stellte zugleich die Kontexteinheit dar, da das Item ebenfalls als größter auszuwertender Materialbestandteil angesehen wurde (keine Zuordnung ganzer Skalen oder Skalengruppen). Die Auswertungseinheit, die festlegt, welche Textbestandteile nacheinander kodiert werden, war ebenfalls durch ein Item definiert. Davon zu unterscheiden ist die Auswahleinheit (vgl. Merten, 1995), welche durch die Ebene eines Fragebogens festgelegt wurde.

8.3.5 Kodierung

Die Durchführung der Kodierung leisteten vier unabhängige Kodierer. Zwei Kodierer wurden im Zusammenhang mit den deutschen Verfahren

eingesetzt; zwei weitere (bilinguale) Kodierer mit Englisch als Muttersprache und sehr guten Deutschkenntnissen kodierten die Items aus den amerikanischen Verfahren. Das Kodiersystem, d.h. die Benennung der Kategorien und die Sprache der Ankerbeispiele, korrespondierte jeweils mit der Sprache der zu kategoriesierenden Items im Übungsbogen bzw. im Kodierbogen. Die Instruktionen wurde der Vergleichbarkeit wegen schriftlich und für alle Kodierer in deutscher Sprache verfasst. Der praktische Ablauf gliederte sich in drei Phasen.

1. In der *Lesephase* sollten sich die Kodierer mit dem Kategoriensystem und den dazugehörigen Ankerbeispielen, vertraut machen (vgl. die Anhänge C bis E).

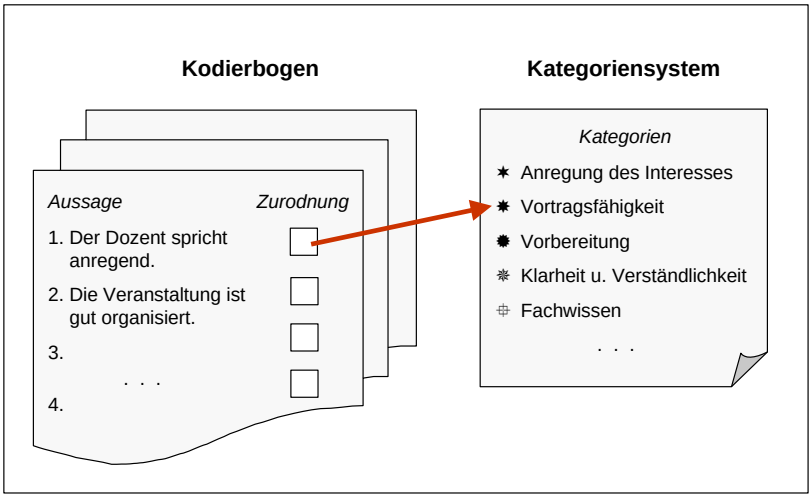
2. Darauf folgte die *Trainingsphase*, in der die Kodierer einen Übungsbogen erhielten. Dieser umfasste jeweils 20 Items, die für die deutsche Version aus einer Veröffentlichung von Stangl (2000) und für die amerikanische Version aus dem SIR zufällig ausgewählt wurden. Die Instruktion und die beiden Übungsbogen sind in den Anhängen F bis H wiedergegeben. Die Kodierer wurden gebeten, die einzelnen Aussagen des Übungsbogens einer der 30 Kategorien zuzuordnen oder die Kategorie „00“ (sonstige) zu wählen, wenn die Aussage ihrer Meinung nach keiner Kategorie zugeordnet werden könne. Es musste jede Aussage bearbeitet und exakt einer Kategorie zugewiesen werden. Die Trainingsphase diente dazu, sich mit dem Kategoriensystem weiter vertraut zu machen und sich an das Arbeiten mit dem Kategoriensystem sowie der Gestaltung des Kodierbogens zu gewöhnen.

3. In der dritten Phase, der *Kodierphase*, erhielten die Kodierer den für die Auswertung relevanten Kodierbogen. Der Bogen gibt die Items der 7 Verfahren in aneinandergereihter Form wieder. Innerhalb eines Verfahrens traten die Items in ihrer originalen Reihenfolge auf. Dies führte zu 232 zu kodierenden Aussagen in der deutschen und zu 255 zu kodierenden Aussagen in der amerikanischen Fassung des Kodierbogens. Beide Fassungen sowie eine Übersicht der Numerierungen sind in den Anhängen I bis K zu finden. Instruktion und Kodierungsablauf waren identisch zur Übungsphase. Abschließend konnten die Kodierer noch Anmerkungen auf einem freien Blatt Papier notieren.

In keiner Phase gab es eine zeitliche Begrenzung oder einen Ansporn zur zügigen Bearbeitung, so dass die Kodierer jeweils ihr eigenes Arbeits-

tempo wählen konnten. Die folgende Abbildung veranschaulicht den Kodiervorgang.

Abb.8. Der Kodiervorgang



8.4 Auswertung

Für alle statistischen Auswertungen wurde das Softwarepaket SPSS 10.0 verwendet. Die untersuchten Aspekte lassen sich inhaltlich in die drei Teile Kodierung, Häufigkeiten und Tests gliedern. Bei allen drei Bereichen wurden zwei Analyseebenen unterschieden: die Ebene des einzelnen Verfahrens (Fragebogen X) und die Ebene des ‚Kulturraumes', also der Gruppe der deutschsprachigen versus der amerikanischen Verfahren.

1. *Kodierung:* Hinsichtlich des Prozesses der Kodierung interessierte das Ausmaß der Beurteilerübereinstimmung.
2. *Häufigkeiten:* Eine deskriptiv-qualitative Analyse der Häufigkeiten diente der Beantwortung folgender Fragen: Wieviele der 31 möglichen Dimensionen sind in deutschen und amerikanischen Instrumenten repräsentiert? Wurden Kategorien nicht besetzt? Wenn dies der Fall war, um welche Kategorien bzw. Dimensionen handelt es sich? In welchem Ausmaß sind die einzelnen Dimensionen in den Verfahren repräsentiert?
3. *Tests*: Da „die deduktive Strategie der Kategorienvorgabe gut mit einem hypothesenprüfenden Vorgehen zu verbinden ist, indem man

die Hypothesen über die Art der Zellenbesetzung im Kategoriensystem formuliert“(Bortz & Döring, 1995, S. 143), wurden im Vorfeld auch für die vorliegende Untersuchung Hypothesen aufgestellt. Alle Hypothesen betreffen die erwarteten Häufigkeiten für die einzelnen Dimensionen und lauten:

H_{01}: Die Anteile der Dimensionen sind in den einzelnen Verfahren gleichverteilt.

H_{02}: Die Anteile der Dimensionen sind innerhalb eines Kulturraumes über die einzelnen Verfahren gleichverteilt.

H_{03}: Die Anteile der Dimensionen sind über beide Kulturräume gleichverteilt.

Tabelle 14 gibt einen Überblick über die Leitfragen der drei Auswertungsaspekte Kodierung, Dimensionen und Häufigkeiten sowie über die zur Analyse gewählten Auswertungsmethoden.

Tab. 14. Übersicht Auswertung und Fragestellungen

Inhaltliche Ebene	**Auswertungsebene und -methodik**	
	Einzelne Verfahren	Verfahrensgruppe Kulturraum
Kodierung	Wie hoch ist die Beurteilerüber-einstimmung? *Cohen's kappa*	Wie hoch ist die Beurteilerüber-einstimmung *Cohen's kappa*
Häufigkeiten	Wieviele Dimensionen sind im Fragebogen X repräsentiert? *Bzw.* Mit welchen Häufigkeiten sind die Dimensionen des Fragebogen X repräsentiert? *Deskriptiv / Häufigkeitstabellen*	Wieviele Dimensionen sind in den Kulturräumen repräsentiert? *bzw.* Mit welchen Häufigkeiten sind die Dimensionen in den Kulturräumen repräsentiert? *deskriptiv / Häufigkeitstabellen*
Tests	Sind die Dimensionen im Frage-bogen X gleichmäßig verteilt? *Chi-Quadrat-Test (H_{01})*	Sind die Dimensionen inner-halb des Kulturraumes über die Verfah-ren gleichmäßig verteilt? *Chi-Quadrat-Test (H_{02})* Sind die Dimensionen über beide Kulturräume gleich-mäßig verteilt? *Chi-Quadrat-Test (H_{03})*

Datenaggregierung: Die Kategorie 31

Einige Kodierer gaben an, dass es ihnen besonders schwer fiel, zwischen den Kategorien 23 und 24 (Schwierigkeit und Arbeitsaufwand – deskriptiv; Schwierigkeit und Arbeitsaufwand – bewertend) zu unterscheiden. Durch die Ankerbeispiele war nicht eindeutig ersichtlich, wann eine Aussage nur beschreibenden oder wertenden Charakter hat. Dieses Problem zeigte sich deshalb konsequenter Weise in den vorgenommenen Zuordnungen. Die Kodierer benutzten nicht beide, sondern meist durchgängig nur eine der beiden Kategorien (entweder Kategorie 23 oder 24). Aus diesem Grunde wurden in einer separaten Analyse die Kategorien 23 und 24 zu einer neuen Kategorie mit der Nummer 31 zusammengefasst (vgl. auch Problemaspekte bei Abrami und d'Apollonia, 1990). Ergebnisse dieser Analyse sind jedes Mal explizit ausgewiesen.

8.4.1 Beurteilerübereinstimmung

Um ein Maß für die Güte der Kategorisierung zu erhalten, wurde Cohen‘s kappa (Cohen, 1960) als Index gewählt. Der kappa-Koeffizient beschreibt das Ausmaß der Beurteilerübereinstimmung (auch Interrater- oder Intercoderreliabilität) und ist weit verbreitet. Er hat gegenüber anderen Assoziationskoeffizienten zwei Vorteile. Kappa bereinigt die Zahl der übereinstimmenden Urteile um die Zahl der zufällig zu erwartenden Übereinstimmung (Krauth, 1995). Weiterhin ist kappa im Gegensatz zu anderen Assoziationskoeffizienten nach Carletta (1995) eine interpretierbare Größe. Krippendorff (1980, S. 147) spricht bei kappa Werten von $\kappa > .80$ als gute Übereinstimmung, wobei er bei Übereinstimmungen im Bereich von $.67 < \kappa < 0.80$ bereits vorsichtige Schlussfolgerungen für möglich hält. Nach Bortz und Döring (1995, S. 254) sind dagegen schon kappa-Werte von $\kappa > 0.70$ als gut zu bezeichnen.

In Bezug auf diese Grenzwerte für akzeptable und gute Beurteilerübereinstimmungen, können die hier erzielten Werte für beide Stichproben als nicht befriedigend bezeichnet werden (vgl. Tab. 15). Werden jedoch, wie bereits erwähnt, die Kategorien 23 und 24 zu einer Kategorie (31) zusammengefasst, erreicht das Ausmaß der Übereinstimmung in beiden Stichproben den von Krippendorff (1980) genannten Bereich, der bereits vorsichtige Schlussfolgerungen zulässt. Sie liegen dann nur noch knapp unter der bei Bortz und Döring (1995) genannten Grenze von $\kappa = .70$.

Tab. 15. Das Ausmaß der Beurteilerübereinstimmung über alle Verfahren eines Kulturraumes

Deutsche Verfahren (gesamt)		Amerikanische Verfahren (gesamt)	
κ	κ mit Kategorie 31	κ	κ mit Kategorie 31
.630	.679	.655	.684

Die Übereinstimmung der Urteile auf der Ebene der einzelnen Verfahren stellt sich sehr heterogen dar (vgl. Tab. 16).

Tab. 16. Das Ausmaß der Beurteilerübereinstimmung für die einzelnen Verfahren

Deutsche Verfahren (einzeln)			Amerikanische Verfahren (einzeln)		
Name	κ	κ mit Kat. 31	Name	κ	κ mit Kat. 31
HILVE	.493	.566	IDEA	.436	.485
BEVA	.562	.591	SIR II	.596	.678
FELL-V	.511	.564	SIRS	.665	.665
MAFAL	.525	.525	SEEQ	.816	.816
VBPSYCH	.645	.769	CIEQ	.583	.583
FB-LV	.728	.782	ICE	.788	.788
FEVOR	.669	.775	SPTE	.658	.688

Wie der Tabelle 16 zu entnehmen ist, haben die erreichten Übereinstimmungen eine Spannweite von $\kappa = .493$ bis $\kappa = .728$ in der deutschen bzw. von $\kappa = .436$ bis $\kappa = .816$ in der amerikanischen Stichprobe. Die Zusammenfassung der Kategorien 23 und 24 führt in Einzelfällen zu einer deutlichen Erhöhung der Beurteilerübereinstimmung. Jedoch wirkt sich dies in beiden Stichproben nur gering aus, da die Grenze von $\kappa = .67$ für akzeptable Übereinstimmungen jeweils nur in einem einzigen Fall überschritten wurde (VBPSYCH und SIR II). Auffällig sind die besonders geringen Übereinstimmungen für das deutsche Verfahren HILVE und das amerikanische Verfahren IDEA.

Die folgende Analyse der Häufigkeiten und der Hypothesentests muss also vor dem Hintergrund der teilweise sehr geringen Beurteilerübereinstimmung gesehen werden. Um ein möglichst zutreffendes Bild wiederzugeben, gingen die im Folgenden genannten Häufigkeiten nicht über beide Kodierer gemittelt in die Analysen ein, sondern jeweils für beide Kodierer getrennt.

8.4.2 Deskriptive Analyse der Häufigkeiten

Die vollständigen Häufigkeitstabellen der Dimensionen für die beiden Kulturräume sind im Anhang L aufgeführt. Im Folgenden werden daraus einige besonders markante Ergebnisse vorgestellt.

Die erste Frage, die anhand der Häufigkeiten untersucht wurde, galt der Anzahl der benutzten Kategorien. Wieviele Kategorien verwendeten die Kodierer?

Für die Zuordnung standen den Kodierern insgesamt 31 Kategorien zur Verfügung. Davon entfielen 28 Kategorien auf das Schema von Feldman (1989); zwei globale Kategorien (29, 30) sowie eine Kategorie „sonstige“ (00) wurden in Anlehnung an die Studie von Abrami und d’Apollonia (1990) aufgenommen. Insgesamt verwendeten die Kodierer zwischen 24 und 30 Kategorien der 31 möglichen Kategorien; für die Kategorisierung der deutschen Items im Mittel 26,5 und für die amerikanischen im Mittel 29,5 (vgl. Tab. 17).

Tab. 17. Übersicht zur Anzahl der repräsentierten Dimensionen in den beiden Kulturräumen

Anzahl der Dimensionen* (in 7 deutschen Verfahren, N = 232 Items)			**Anzahl der Dimensionen*** (in 7 amerikanischen Verfahren, N = 254 Items)		
Kodierer A	Kodierer B	Mittel	Kodierer A	Kodierer B	Mittel
29	24	26,5	30	29	29,5

Anmerkung. *Die maximale Anzahl beträgt 31

Bezüglich der Analyseebene der einzelnen Verfahren wurden wesentlich weniger Kategorien verwendet; im Gesamtmittel über beide Kodierer bei 14,79 Kategorien für deutsche Instrumente und bei 17,5 Kategorien für amerikanische (vgl. Tab. 18). Dieser Unterschied wurde bei einer Prüfung mit einem 2-seitigen T-Test nicht signifikant (t (26) = -1,723, p = .097).

Tab. 18. Übersicht zur Anzahl der repräsentierten Dimensionen in verschiedenen Fragebogen

Deutsche	**Dimensionen [N]**		**Items**	**Amerik.**	**Anzahl Dimensionen**		**Items**
Verfahren	**Kod. A**	**Kod. B**	**[N]**	**Verfahren**	**Kod. A**	**Kod. B**	**[N]**
HILVE	16	16	37	IDEA	20	18	47
BEVA	16	16	40	SIR II	22	19	40
FELL	13	15	23	SIRS	13	9	21
MFAL	13	9	17	SEEQ	20	19	31
VBPSYCH	12	13	40	CIEQ	10	12	21
FB-LV	24	19	55	ICE	19	18	55
FEVOR	13	12	20	SPTE	22	24	39
Mittel	15,29	14,29	33,14	Mittel	18,00	17,00	36,29

Zur Beantwortung der Frage, in welchem Ausmaß die Dimensionen repräsentiert sind bzw. welche Kategorien am häufigsten vorkommen, wurden die Häufigkeitstabellen auf hohe und niedrige relative Häufigkeiten untersucht.

Nicht und gering besetzte Kategorien

Da in der Gesamtgruppe der Verfahren eines Kulturraumes nicht alle Kategorien besetzt wurden, ist zu fragen, um welche Kategorien es sich dabei handelt. Auch ist interessant, ob die gleichen Kategorien in beiden Kulturräumen betroffen sind. In Tabelle 19 sind die Kategorien zusammengetragen, die bei *beiden* Kodierern eines Kulturraumes übereinstimmend nicht besetzt oder nur selten besetzt wurden.

Hinsichtlich der deutschen Stichprobe wurden von beiden Kodierern übereinstimmend keine Items den Kategorien „Erreichbarkeit und Hilfsbereitschaft des Dozenten“ (22) bzw. „Produktivität des Dozenten hinsichtlich seiner forschungsbezogenen Tätigkeiten“ (19) zugeordnet. Neben den nicht besetzten Kategorien gab es zwei weitere Kategorien, die aufgrund geringer Häufigkeit bei beiden Kodierern auffielen. In die Dimension 13 „Fairness und Unvoreingenommenheit bei der Bewertung“ fiel einmal keines, im anderen Falle nur ein Item (0.4%). Auch war der Aspekt „Vortragsfähigkeiten des Dozenten“ sehr schwach repräsentiert (übereinstimmend 0.4%). Darüber hinaus gab es bei beiden Kodierern abweichend, einzelne Kategorien, die entweder nicht besetzt oder gering besetzt sind (vgl. Anhang L).

In Bezug auf die amerikanischen Instrumente wurde übereinstimmend lediglich die Kategorie 25 „Classroom Management“ nicht besetzt. Weiterhin traten die Kategorien 22 und 26 bei beiden Kodierern gar nicht oder mit sehr geringen Häufigkeiten (0.4%) auf. In allen übrigen Kategorien lagen die Häufigkeiten bei beiden Kodierern übereinstimmend höher.

Mit Ausnahme der Kategorie 22 sind es also unterschiedliche Kategorien, die in den beiden Kulturräumen gar nicht oder gering repräsentiert sind.

Tab. 19. Nicht besetzte und wenig besetzte Kategorien bei beiden Kodierern für deutsche und amerikanische Verfahren

Übereinstimmend nicht besetzte oder seltene Kategorien			
	In deutschen Verfahren		**In amerikanischen Verfahren**
7	Vortragsfähigkeiten des Dozenten	22	Teacher's Productivity in Research-Related Areas
13	Fairness, Unvoreingenommenheit bei der Bewertung durch den Dozenten; Qualität der Leistungsüberprüfung	25	Classroom Management
19	Erreichbarkeit und Hilfsbereitschaft des Dozenten	26	Pleasantness of Classroom Atmosphere
22	Produktivität des Dozenten hinsichtlich seiner forschungsbezogenen Tätigkeiten		

Auf einzelne Verfahren bezogen, ist aus der Übersicht zu der Anzahl der verwendeten Kategorien pro Verfahren (vgl. Tab. 18) zu erschließen, wie viele Kategorien jeweils nicht besetzt wurden. Da die deutschen Verfahren im Schnitt 33, die amerikanischen im Schnitt 36 Items umfassen, war zu erwarten, dass bei den einzelnen Verfahren relativ viele Kategorien unbesetzt bleiben. Eine vertiefende Analyse der unbesetzten Kategorien erscheint daher nicht sinnvoll und wurde nicht vorgenommen.

Häufige Kategorien

Bei der Durchsicht der prozentualen Häufigkeiten der Kategorien in beiden Kulturräumen fielen einige Kategorien mit besonders hohen Anteilen auf (vgl. Anhang L).

Für die Gruppe der deutschen Verfahren traten bei beiden Kodierern nur sechs Kategorien mit prozentualen Häufigkeiten größer als 5% auf. Diese sechs Kategorien sind zur besseren Übersicht in Tabelle 20 aufgeführt. Sie sind für beide Kodierer identisch, sofern man nicht zwischen den ohnehin schwer zu differenzierenden Kategorien 23 und 24 unterscheidet. An vorderster Stelle steht übereinstimmend die Kategorie „Klarheit und Verständlichkeit" mit 12,5% bzw. 12,9%. Zusammengenommen sind in den 6 Kategorien mehr als 50% der Items repräsentiert.

Tab. 20. Die häufigsten Kategorien in deutschen Verfahren

Deutsch	**Kodierer A**			**Kodierer B**		
	Kategorie*	f	> 5%	Kategorie*	F	> 5%
	6	30	12.9	6	29	12.5
	0	24	10.3	0	29	12.5
	12	23	9.9	10	26	11.2
	16	18	7.8	12	21	9.1
	10	17	7.3	23	16	6.9
	24	13	5.6	16	13	5.6
Summe	N = 6	125	53.8	N = 6	134	57.8
[max.]	[31]	[232]	[100]	[31]	[232]	[100]

Anmerkung. *(6) Klarheit und Verständlichkeit, (10) Art und Wert des Stoffes (12) Wahrgenommenes Ergebnis oder Wirkung, (16) Ermutigung des Dozenten, Fragen zu stellen; seine Offenheit gegenüber der Meinung anderer, (23, 24) Schwierigkeit der Veranstaltung (und Aufwand) – beschreibend, bewertend

In amerikanischen Verfahren lagen bei beiden Kodierern jeweils fünf Dimensionen über einer prozentualen Häufigkeit von 6%. Gemeinsam wurden dadurch jeweils etwa 40% der Items repräsentiert. Von diesen Kategorien sind vier bei beiden Kodierern identisch, wobei es sich um die folgenden Dimensionen handelt: „Wahrgenommenes Ergebnis oder Wirkung der Veranstaltung“, „Art und Nützlichkeit des zusätzlichen Materials und der Veranstaltungshilfen“, „Ermutigung des Dozenten, Fragen zu stellen, zu diskutieren; seine Offenheit gegenüber der Meinung anderer“ und die Kategorie „Sonstige“. Auffällig ist, dass die Kategorie 30, bei Kodierer A an erster Stelle steht, bei Kodierer B jedoch nicht unter den 5 häufigsten Kategorien ist. Sie erreicht bei Kodierer B aber immer noch 5,1% (vgl. Anhang L). In umgekehrter Weise gilt dies auch für die Kategorie 6 „Klarheit und Verständlichkeit“.

Tabelle 21 gibt die häufigen und seltenen Kategorien in amerikanischen Verfahren sowie die originalen Bezeichnungen der Kategorien wieder.

Tab. 21. Die häufigsten Kategorien in amerikanischen Verfahren

Amerik.	**Kodierer A**			**Kodierer B**		
	Kategorie*	f	> 6%	Kategorie*	f	> 6%
	30	22	8.7	12	33	13.0
	0	22	8.7	0	19	7.5
	12	19	7.5	11	18	7.1
	11	17	6.7	16	18	7.1
	16	17	6.7	6	17	6.7
Summe	N = 5	97	38.3	N = 5	105	41,4
[max.]	[31]	[254]	[100]	[31]	[254]	[100]

Anmerkung. *(6) Clarity and Understandablenes, (11) Nature and Usefulness of Supplementary Materials and Teaching Aids, (12) Perceived Outcome or Impact of Instruction, (16) Teacher's Encouragement of Questions and Discussion, and Openess to Opinions of Others

Für die Ebene der einzelnen Verfahren bleibt festzuhalten, dass in vielen Verfahren eine oder zwei Dimensionen von auffallend hoher Häufigkeit existieren, die in der Regel bei beiden Kodierern übereinstimmen. Da der Fokus dieser Arbeit nicht auf der Analyse der einzelnen Fragebogen liegt, wurde diese Analyse nicht weiter vertieft. Eine Übersicht über häufig besetzte Kategorien einzelner Verfahren befindet sich in Anhang M.

8.4.3 Chi-Quadrat-Tests

Wie im Vorfeld bereits vermutet und durch die Inspektion der Häufigkeitstabelle unterstützt, wurde insbesondere die Gleichverteilung der Dimensionen bzw. Kategorien in Frage gestellt. Insgesamt lagen vier spezifische Hypothesen vor, die mit Chi-Quadrat-Verfahren getestet wurden.

H_{01}: Die Anteile der Dimensionen sind in den einzelnen Verfahren gleichverteilt.

Um die Gleichverteilungshypothesen für die einzelnen Verfahren zu überprüfen, wurden Chi-Quadrat-Anpasssungstests durchgeführt. Dabei wurde jedes Verfahren einzeln betrachtet und geprüft, ob sich die Dimensionen über das Merkmal Fragebogen X gleich verteilen. Zur Prüfung wurde der exakte Test (Fischer) eingesetzt, da die erwarteten Häufigkei-

ten aufgrund der hohen Anzahl zur Verfügung stehender Dimensionen in allen Zellen kleiner 5 waren.

Die Tests ergaben, dass die Häufigkeiten der Dimensionen bei drei von sieben deutschen Verfahren nicht gleichverteilt sind (BEVA, VBPSYCH und FB-LV). Bei den amerikanischen Verfahren wurde die Gleichverteilungshypothese nur in einem Fall verworfen (ICE). In allen genannten Fällen lag die Irrtumswahrscheinlichkeit unter $\alpha < .01$. Für die Verfahren HILVE und IDEA liegen keine übereinstimmenden Ergebnisse vor. Die detaillierten Ergebnisse sind Tabelle 22 aufgeführt.

Tab. 22. Ergebnisse der Chi-Quadrat-Anpassungstest für die einzelnen Verfahren

Deutsche Verfahren (einzeln)					**Amerikanische Verfahren (einzeln)**				
Verfahren [Items N]	K*	X^2	df	Exakte Signifikanz (2-seitig)	Verfahren [Items N]	K*	X^2	df	Exakte Signifikanz (2-seitig)
HILVE	A	9,270	15	.895	IDEA	A	17,255	19	.602
37	B	32,622	15	.007	47	B	50,660	17	.000
BEVA	A	43,200	15	.000	SIR II	A	18,300	21	.668
40	B	33,500	14	.003	40	B	26,500	18	.095
FELL-V	A	9,217	12	.744	SIRS	A	5,619	12	.968
23	B	10,000	10	.485	21	B	6,857	8	.610
MAFAL	A	2,118	12	1.000	SEEQ	A	8,355	19	.968
17	B	1,529	8	.999	31	B	8,839	18	.610
VBPSYCH	A	29,600	11	.002	CIEQ	A	7,095	9	.686
40	B	22,450	12	.001	21	B	8,143	11	.766
FB-LV	A	62,382	23	.000	ICE	A	47,600	18	.000
55	B	53,818	18	.000	55	B	36,964	17	.004
FEVOR	A	9,900	12	.694	SPTE	A	15,718	21	.824
20	B	7,600	11	.818	39	B	12,077	23	.963

Anmerkung. * K steht für Kodierer. Es traten in allen Zellen erwartete Häufigkeiten kleiner 5 auf. Die unterschiedliche Anzahl der Freiheitsgerade auch innerhalb einer Verfahrensgruppe geht auf die unterschiedliche Anzahl der verwendeten Kategorien zurück.

Fasst man die verschiedenen Verfahren als r unabhängige Stichproben auf, so stehen für die Prüfung von Häufigkeitsunterschieden über die c Kategorien Chi-Quadrat-Verfahren für r*c-Tafeln zur Verfügung (Chi-

Quadrat-Unabhängigkeitstest). Diese Verfahren überprüfen, ob die Anteile der c Kategorien (hier die Häufigkeiten der 31 möglichen Kategorien) in allen r unabhängigen Stichproben (hier den jeweils 7 einzelnen Verfahren) die gleichen sind. Ein signifikanter r*c-Chi-Quadrat-Wert gibt dementsprechend an, dass sich die prozentuale Verteilung der Kategorien in den Verfahren unterscheidet (vgl. Bortz, 1999, S. 167). Dieses Verfahren ist damit geeignet die folgende Hypothese zu testen:

H_{02}: Die Anteile der Dimensionen sind innerhalb eines Kulturraumes über die einzelnen Verfahren gleichverteilt.

Da mehr als 20% der Zellen erwartete Häufigkeiten kleiner 5 aufwiesen, wurde neben der asymptotischen Signifikanz auch die Signifikanz über die Option „Monte-Carlo-Studien“ geschätzt[19]. Die Nullhypothese konnte bei drei Kodierern für die Verteilung der Dimensionen über die deutschen wie auch über die amerikanischen Verfahren mit einer Irrtumswahrscheinlichkeit von $\alpha < .001$ zurückgewiesen werden (vgl. Tab. 23).

Tab. 23. Ergebnisse der Chi-Quadrat-Unabhängigkeitstests der Variablen Verfahren und Kategorie

Deutsche Verfahren [N=7]					**Amerikanische Verfahren [N=7]**				
Kod.	X^2	df	Signifikanz (2-seitig)		Kod.	X^2	df	Signifikanz (2-seitig)	
			Asymptotisch*[1]	Monte Carlo*[2]				Asymptotisch*[1]	Monte Carlo*[2]
A	211,967	138	.000	.000	A	214,768	174	.019	.012
B	244,643	168	.000	.000	B	245,454	168	.000	.000

Anmerkung. *[1] Es traten in jeweils mehr als 90% der Fälle erwartete Zellhäufigkeiten kleiner 5 auf. *[2] Basierend auf 10000 Stichprobentabellen mit dem Startwert 957002199 (deutsch) bzw. bei 79654295 (amerikanisch).
Die unterschiedliche Anzahl der Freiheitsgerade auch innerhalb einer Verfahrensgruppe geht auf die unterschiedliche Anzahl der verwendeten Kategorien zurück.

[19] Für die Durchführung des exakten Tests nach Fischer konnten die hohen Anforderungen an die Systemtechnik nicht erfüllt werden.

Eine ähnliche Fragestellung, die Hypothese H_{03}, wurde ebenfalls mit einem Chi-Quadrat-Unabhängigkeitstest überprüft. Die Hypothese lautete:

H_{03}: Die Anteile der Dimensionen verteilen sich über die beiden Kulturräume gleich.

Diesbezügliche Hypothesentests lassen insgesamt den Schluss zu, dass sich die Verteilung der Dimensionen des einen Kulturraumes von der Verteilung der Dimensionen des anderen signifikant unterscheidet ($\alpha < .001$). Inhaltlich bedeuten dies, dass die Dimensionen in deutschen Verfahren in anderen Anteilen repräsentiert sind als in amerikanischen Verfahren. Die einzelnen Ergebnisse sind in Tabelle 24 (S. 84) zusammengestellt.

Tab. 24. Ergebnisse des Chi-Quadrat-Unabhängigkeitstests der Variablen Kulturraum (deutsch vs. amerikanisch) und Kategorie (max. 31)

Kulturraum [N = 486 Items]				
Kodierer	X^2	df	Signifikanz (2-seitig)	
			Asymptotisch*[1]	Monte Carlo*[2]
A beider Stichproben	96,709	29	.000	.000
B beider Stichproben	115,587	29	.000	.000

Anmerkung. *[1]In 25 Zellen (41,7% der Fälle) traten erwartete Zellhäufigkeiten kleiner 5 auf. *[2] Basierend auf 10000 Stichprobentabellen mit einem Startwert von 2000000.

8.5 Zusammenfassung und Diskussion der Ergebnisse

Anlage der Untersuchung und Beurteilerübereinstimmung

Die erzielten Übereinstimmungen der Kodierer können nur bezüglich einzelner Verfahren als gut bezeichnet werden. Für die Gruppe der untersuchten deutschen und amerikanischen Fragebogen insgesamt erreichen sie nach Aggregierung der Kategorien 23 und 24 akzeptable Werte, so dass „vorsichtige Schlussfolgerungen“ im Sinne Krippendorffs (1980) möglich sind.

Die relativ niedrigen Übereinstimmungen sind vor allem mit Blick auf die große Anzahl der Kategorien plausibel, denn „mit wachsendem Umfang des Kategoriensystems leidet die Zuverlässigkeit der Kodierung, da bei den Kodierern Grenzen der Gedächtnisleistung und Aufmerksamkeit erreicht werden“ (Bortz & Döring, 1995, S. 142). Eine Reduzierung der Anzahl der Kategorien wäre eine theoretisch sinnvolle Maßnahme (s. Merten, 1995), scheint für die vorliegende Fragestellung jedoch weniger angemessen. Wie aus der Diskussion der Literatur hervorging, hatte das verwendete Schema gerade den Vorteil, in sehr differenzierter Weise mögliche Dimensionen der Lehrevaluation in Fragebogen zu repräsentieren. Dies war insbesondere wichtig, da die Experten über ein reduziertes Set möglicher Dimensionen bzw. Kategorien bisher keine Einigkeit erzielen konnten. Weiterhin könnte mit der Reduzierung der Anzahl der Kategorien die Gültigkeit sinken, da in einem solchen Falle ein „differenziertes Bild sozialer Wirklichkeit nun auf ein tendenziell simples Raster vergröbert wird“ (Merten, 1995, S. 308). Es scheint insgesamt von Vorteil, ein umfangreiches Schema verwendet zu haben, zumal sich in manchen Fällen die Aussagekraft einer Untersuchung durch die Anwendung komplexer Schemata steigern lässt, trotz der Schwierigkeit, zuverlässige Resultate zu erzielen (vgl. Ritsert, 1972 in Mayring, 2000).

Es gibt zwei weitere Punkte, an denen die vorliegende Untersuchung methodisch verbessert werden könnte, was auch eine positive Auswirkung auf die Resultate der Beurteilerübereinstimmung erwarten ließe.

Der erste Aspekt betrifft die *Stichprobenziehung* und die Frage der *Repräsentativität* der vorliegenden Stichproben. Beide Materialstichproben wurden jeweils aus einer heterogenen Erhebungsgesamtheit nach bestimmten Kriterien ausgewählt (vgl. Abschnitt 8.3.2). Die Heterogenität hinsichtlich der Entstehung der Verfahren, der Präferenzen der Autoren und somit der Items ist sicherlich wünschenswert, erhöht dies doch die Generalisierbarkeit der Auswertungen. Hinsichtlich der Heterogenität der

Verfahrenstypen wurden bereits Extreme vermieden, indem beispielsweise Itempools aus der Analyse ausgeschlossen wurden. Eine relative Heterogenität in den Verfahrens*typen* blieb jedoch bestehen. Zum Beispiel befinden sich mit den Verfahren VEFOR oder FELL-V zwei Fragebogen in der Stichprobe, die für die spezifische Veranstaltungsform der Vorlesung konzipiert wurden, während andere Verfahren einen allgemeineren Anspruch vertreten (z. B. HILVE). Diese Form der Heterogenität sollte zukünftig besser vermieden werden, wozu allerdings größere Erhebungsgesamtheiten notwendig wären. Diese waren in der vorliegenden Untersuchung nicht zu erzielen. Lägen größere und homogenere Erhebungsgesamtheiten vor, könnten auch zufallsgesteuerte Methoden der Stichprobenziehung eingesetzt werden (vgl. Lamnek, 1993).

In Folgeuntersuchungen könnte weiterhin das *Training* der Kodierer intensiviert werden. Krippendorff (1980) berichtet, dass in manchen Studien Kodierer mehrere Wochen oder Monate lang vorbereitet werden. Im Vergleich dazu muss das Training dieser Untersuchung – ein einzelner Übungsbogen mit 20 Items – wohl als gering bewertet werden.

Häufigkeiten von Dimensionen in Lehrevaluationsfragebogen, Hypothesentests

Bei der Diskussion der Ergebnisse ist zu bedenken, dass sie aufgrund der vergleichsweise niedrigen Beurteilerübereinstimmung unter Vorbehalt und mit Vorsicht zu betrachten sind. Interessant ist aber, dass sich die nachfolgend genannten Ergebnisse zu häufig und selten repräsentierten Dimensionen relativ klar ergaben.

Vergleicht man die Resultate beider Kulturräume bzw. beider Stichproben, so fällt auf, dass

- in beiden Kulturräumen nicht alle möglichen Kategorien Feldmans (1989) repräsentiert sind.
- sich Häufigkeiten der einzelnen Kategorien in der Regel nicht gleichverteilen (s.u.).
- jeweils übereinstimmend relativ wenige Kategorien nicht besetzt wurden, diese sich aber uneinheitlich darstellen. Eine Ausnahme bildet die Kategorie 22 „Forschungsproduktivität“, die in beiden Stichproben gar nicht oder mit einer sehr geringen Häufigkeit (0,4%) repräsentiert ist.
- jeweils übereinstimmend relativ viele Items nicht zugeordnet werden konnten (zwischen 7,5% und 12,5%) und die Kategorie „Sonstige“

bei allen Kodieren unter den häufig besetzten Kategorien zu finden ist.

- jeweils übereinstimmend relativ wenige Kategorien (5 oder 6 von 31) mit hohen Häufigkeitswerten auftraten, diese sich aber relativ homogen darstellen und einen Großteil der Items auf sich vereinigen (zwischen 40% und 60% Prozent).
- neben der Kategorie „Sonstige“ zwei der fünf bzw. sechs häufigsten Kategorien in beiden Kulturräumen identisch sind und die Kategorie 6 bei drei der vier Kodierer unter den häufigsten Kategorien genannt wird. Zusammen vereinigen sie bei allen Kodierern etwa 30% der Items und können als kulturübergreifende „Spitzenreiter“ gelten. Es sind dies die Kategorien:
  - 6) Klarheit und Verständlichkeit,
  - 12) Wahrgenommenes Ergebnisse oder Wirkung der Veranstaltung und
  - 16) Ermutigung des Dozenten, Fragen zu stellen, zu diskutieren; seine Offenheit gegenüber der Meinung anderer.

Die Hypothesen zur Gleichverteilung der Dimensionen konnten in den meisten Fällen verworfen werden. Die folgende Tabelle fasst die Ergebnisse der Chi-Quadrat-Tests abschließend zusammen.

Tab. 25. Überblick über die Ergebnisse der Hypothesentests

Ergebnisse der Hypothesentests	
Einzelne Verfahren	**Verfahrensgruppe Kulturraum**
Die Dimensionen verteilen sich **nicht** gleichmäßig in den deutschen Verfahren BEVA, VBPSYCH und FB-LV sowie im amerikanischen Verfahren ICE.	Die Dimensionen innerhalb eines Kulturraumes sind über die Verfahren **nicht** gleichmäßig verteilt *(Zurückweisung von H_{02}).*
Die Nullhypothese der Gleichverteilung konnte nicht verworfen werden für die deutschen Verfahren HILVE, FELL-V, MFAL und FEVOR sowie für alle anderen amerikanischen Verfahren.	Die Dimensionen verteilen sich **nicht** gleichmäßig über die beiden Kulturräume bzw. die Gruppe der amerikanischen und deutschen Verfahren
(Insgesamt ein uneinheitliches Bild bzgl. H_{01})	*(Zurückweisung von H_{03}).*

Die empirische Fragestellung dieser Arbeit wurde angeregt durch die in Kapitel 7 diskutierten Problemfelder bei der Erfassung des Konstruktes ‚gute Lehre' in Fragebogenverfahren. Aus diesen Problemfeldern wurde die Dimensionalitätsdebatte bzw. die Frage nach der Homogenität oder Heterogenität von Dimensionen in Lehrevaluationsfragebogen herausgegriffen. Sie bildete den Ansatzpunkt für den empirischen Teil der Arbeit.

Was bedeuten die vorangehend beschriebenen Ergebnisse für die Problematik der Homogenität von Dimensionen in Fragebogen zur Lehrveranstaltungsevaluation?

Die vorliegende Untersuchung konnte zu dieser Problematik zwei Hinweise liefern:

Der erste betrifft die Verteilung von Dimensionen (hier Verteilung der Häufigkeiten der Kategorien 1-30). Auf der Ebene des einzelnen Fragebogens konnte die Hypothese der Gleichverteilung der Anteile der Dimensionen für viele Verfahren nicht verworfen werden, was für eine ausgewogene Konstruktion dieser Verfahren spricht (H_{01}). Das Bild verändert sich bei der Betrachtung der Verfahren als Gruppe. Auf der Ebene der Verfahrensgruppe sind die möglichen Lehrdimensionen im Sinne Feldmans (1989) nicht gleich häufig anzutreffen (Zurückweisung von H_{02}). Ebenso wenig verteilen sich die Dimensionen gleichmäßig über beide Kulturräume (Zurückweisung von H_{03}).

Für diese Ergebnisstruktur bieten sich zwei Erklärungen an: Einmal kann es daran liegen, dass diejenigen Fragebogen die Effekte in der Gruppe der Verfahren hervorgerufen haben, deren Dimensionen sich nicht gleichverteilten. Gegen eine solche Interpretation spricht aber, dass sich die Effekte auch innerhalb der Gruppe der amerikanischen Verfahren zeigten, obwohl diese lediglich nur ein nicht ausgewogenes Verfahren umfasst (ICE). Eine andere Erklärung wäre, dass sich die einzelnen ausgewogen konzipierten Fragebogen untereinander in den berücksichtigten Dimensionen unterscheiden. In einem solchen Falle wären die Fragebogen A und B zwar ausgewogen konzipiert und enthielten jeweils eine Anzahl gleichmäßig repräsentierter Dimensionen, unterschieden sich aber in der Zusammenstellung der berücksichtigten Dimensionen. Würden dann einige Dimensionen häufiger berücksichtigt als andere, führte dies zu einer unausgewogenen Verteilung der Dimensionen in der Verfahrensgruppe.

Ergebnis 1: Die Befunde über die Verteilung der Häufigkeiten der Dimensionen legen nahe, dass in sich überwiegend ausgewogen konstruierte Fragebogen existieren. Aus der Unausgewogenheit der Anteile der Dimensionen auf der Ebene der Verfahrensgruppe (deutsche wie amerikanische) wird geschlossen, dass sich die einzelnen Fragebogen darin unterscheiden, dass sie *verschiedene Dimension* berücksichtigten, wobei einige Dimensionen häufiger aufgenommen wurden als andere. Dies spricht eher für die Heterogenität der untersuchten Verfahren.

Als Konsequenz der ungleichen Verteilung der Dimensionen in der Verfahrensgruppe könnten als zweiter oder weiterführender Hinweis zur Frage der Homogenität oder Heterogenität die Ergebnisse der deskriptiv-qualitative Analyse der Häufigkeitstabellen herangezogen werden. Interpretiert man hohe Häufigkeiten im Sinne einer faktischen Relevanz, so lässt sich aus der vorliegenden Untersuchung schließen, dass es insgesamt wichtigere und unwichtigere Aspekte von Lehre gibt. Zumindest zeigte sich in beiden Kulturräumen, dass typische oder häufige und weniger typische bzw. selten berücksichtigte Aspekte existieren.

In beiden Stichproben trat jeweils nur eine geringe Anzahl an häufig genannten Kategorien auf (5 bzw. 6 Kategorien). Darunter wurden nur zwei Dimensionen identifiziert, die sich bei allen Kodierern unter den häufig anzutreffenden Dimensionen befanden. Hierbei handelt es sich um die Dimensionen „Ermutigung des Dozenten, Fragen zu stellen, zu diskutieren; seine Offenheit gegenüber der Meinung anderer“ und „Wahrgenommenes Ergebnis oder Wirkung der Veranstaltung“. Bei drei der vier

Kodierer ist übereinstimmend auch die Kategorie 6 „Klarheit und Verständlichkeit“ unter den häufigen Dimensionen aufgeführt.

Für eine Einschätzung dieser Ergebnisse in Bezug auf die „Spitzenreiter“-Dimensionen bieten sich zwei Vergleiche an. Als erstes können dazu die Ergebnisse von Abrami und d’Apollonia (1990) herangezogen werden. Darüber hinaus ist ein Vergleich mit den in Abschnitt 7.2 vorgestellten typischen Dimensionen anderer (z. B. Centra, 1993) möglich.

a) Vergleich mit den Ergebnissen von Abrami und d’Apollonia (1990)

Dieser Vergleich lag nahe, da das von den Autoren verwendete Kategorienschema ebenso wie das in der vorliegenden Untersuchung verwendete Schema auf Feldman (1976 bzw. 1989) zurückgehen. Die bei Abrami und d’Apollonia berücksichtigten 24 Kategorien wurden anhand ihrer absoluten Häufigkeiten in eine Rangfolge gebracht und in nachfolgender Tabelle zusammengestellt (vgl. Tab. 26, Spalte 2). Weiterhin zeigt die Tabelle in der dritten Spalte die in der vorliegenden Untersuchung identifizierten „Spitzenreiter“-Kategorien.

Tab. 26. Die 10 häufigsten Kategorien bei Abrami und d'Apollonia (1990) im Vergleich zu den „Spitzenreiter"-Dimensionen der vorliegenden Untersuchung

Dimension	Ränge bei Abrami und d'Apollonia (1990)	„Spitzenreiter", vorliegende Untersuchung (2001)
Clarity and understandableness	1	X (Kategorie 6)
Overall instructor	2	
Overall course	3	
Encouragement of class discussion	4	X (Kategorie 16)
Preparation and organization	5	
Stimulation of Interest	6	
Workload	7	
Classroom management	8	
Class level and progress	9	
Perceived Outcome	10	X (Kategorie 12)
Concern and respect for students	10	

Wie aus dem Vergleich der Spalten zwei und drei der Tabelle 26 zu entnehmen ist, entfallen zwei der drei „Spitzenreiter"-Dimensionen auf die ersten vier Ränge bei Abrami und d'Apollonia (1990); die dritte Dimension fällt dagegen auf einen mittleren Rang (Rang 10 von 24). Dies kann als stützender Hinweis für die Ergebnisse der vorliegenden Untersuchung angesehen werden.

b) Vergleich mit Listen „typischer" Dimensionen verschiedener Autoren.

Vergleicht man die drei „Spitzenreiter"-Dimensionen der vorliegenden Untersuchung mit Listen „typischer Dimensionen" anderer Autoren (vgl. Abschnitt 7.2), so lassen sich inhaltliche Ähnlichkeiten feststellen. Zum Beispiel weist die Kategorie „Student learning, student self ratings of accomplishments or progress" von Centra (1993) deutliche Ähnlichkeit mit der Kategorie 12 der vorliegenden Untersuchung auf („Perceived Outcome or Impact of Instruction"). Diese Ähnlichkeiten zu Centra (1993) lassen sich ebenfalls für die beiden anderen Dimensionen aufzei-

gen. Gleiches gilt auch für die bei el Hage (1996) genannten Bezeichnungen. Die entsprechenden Dimensionen sind gegenüberstellend in Tabelle 27 (S. 90) zusammengetragen.

Tab. 27. Ähnlichkeiten der als „Spitzenreiter" identifizierten Dimensionen der vorliegenden Untersuchung mit den als „typisch" identifizierten Dimensionen andere Autoren

Amerikanische Verfahren	
Centra (1993) sowie Braskamp und Ory (1994)	***Vorliegende Untersuchung (2001)***
Clarity, communication skill	Clarity and Understandableness (6)
Student learning, student self ratings of accomplishments or progress.	Perceived Outcome or Impact of Instruction (12)
Teacher student interaction or rapport	Teacher's Encouragement of Questions and Diskussion, and Openess to Opinions of others (16)
Deutsche Verfahren	
El Hage (1996)	***Vorliegende Untersuchung (2001)***
Kurs- bzw. Stofforganisation	Klarheit und Verständlichkeit (6)
Zuwendung Kommunikationsfähigkeit	Ermutigung des Dozenten, Fragen zu stellen, zu diskutieren; seine Offenheit gegenüber der Meinung anderer (16)
Kurswert	Wahrgenommenes Ergebnis oder Wirkung der Veranstaltung (12)

Einschränkend ist zu beachten, dass es sich bei diesem Vergleich um Ähnlichkeiten handelt, die durch die Benennungen der Dimensionen nahegelegt werden und nicht durch Vergleiche auf Itemebene. Ein Vergleich von Items ist nicht möglich, da die anderen Autoren keine Ankerbeispiele oder Items für die Beschreibung der Dimensionen anführen. Auch sei an die Studie von Abrami und d'Apollonia (1990) erinnert, die darauf aufmerksam machte, dass die gleichnamige Dimensionen in einem Fragebogen A etwas anderes bedeuten kann als im Fragebogen B.

Ergebnis 2: Die deskriptiv-qualitative Analyse der Häufigkeiten der einzelnen Kategorien zeigte, dass in beiden Kulturräumen eine Anzahl von 5-6 typischer oder häufig auftretender Dimensionen existiert. Drei dieser Dimensionen wurden für beide Kulturräumen als gemeinsame „Spitzenreiter“ identifiziert (Kategorie 6, 12 und 16). Sie repräsentieren zusammen etwa 30% aller Items. Weiterhin weisen sie relative Ähnlichkeiten zu den häufigen Dimensionen anderer empirischer Analysen und tabellarischer Zusammenstellungen typischer Dimensionen auf.

Insgesamt können aber auch diese Ergebnisse nicht als Beleg für die Homogenität der Dimensionen von Fragebogen zur Lehrveranstaltungsevaluation gewertet werden, da neben den „Spitzenreitern“ ein nicht zu vernachlässigender Anteil nicht übereinstimmender Dimensionen bleibt.

9. Diskussion

9.1 Konsequenzen für die Erfassung des Konstruktes ‚gute Lehre': Rückbezug zur Dimensionalitätsdebatte

Die Kategorisierung von 232 Items aus sieben deutschen und von 254 Items aus sieben anglo-amerikanischen Fragebogen zur Lehrevaluation ergab, dass sie sich auf 30 Kategorien möglicher Dimensionen von Lehre ungleichmäßig verteilen. Über eine Häufigkeitsanalyse konnten drei Dimensionen als gemeinsame „Spitzenreiter" für beide Kulturräume identifiziert werden. In der Gruppe der deutschen wie der anglo-amerikanischen Verfahren vereinigen sie jeweils etwa 30% der Items. Die drei Dimensionen lauten in deutscher Bezeichnung:

- Klarheit und Verständlichkeit
- Wahrgenommenes Ergebnis oder Wirkung der Veranstaltung
- Ermutigung des Dozenten, Fragen zu stellen; seine Offenheit gegenüber der Meinung anderer.

Es wurde weiterhin festgestellt, dass diese drei Dimensionen relative Ähnlichkeiten zu als häufig oder typisch deklarierten Dimensionen anderer Autoren aufweisen; seien es die auf empirischem Weg ermittelten Dimensionen von Abrami und d'Apollonia (1990) oder die durch Literaturstudium identifizierten Dimensionen von Centra (1993) und el Hage (1996). Trotz dieser Ähnlichkeiten wurde aber auch betont, dass sich ein respektabler Anteil von Items auf die übrigen Dimensionen verteilt (etwa 70%). Insgesamt müssen die untersuchten Fragebogenverfahren zur Lehrveranstaltungsevaluation hinsichtlich der in ihnen berücksichtigten Dimensionen und ihrer Verteilung eher als heterogen denn als homogen bezeichnet werden.

Was bedeuten diese Ergebnisse für die Dimensionalitätsdebatte?

Ausgangspunkt für die empirische Fragestellung der Arbeit waren einige als Problemfelder bezeichneten Themenkomplexe; darunter insbesondere die Dimensionalitätsdebatte bzw. die Frage nach der Homogenität von Faktoren in Fragebogen zur Lehrevaluation. Diese Thematik steht in engem Zusammenhang mit einer fehlenden anerkannten Definition von Lehre bzw. der mangelnden Einigkeit der Experten bezüglich der Komponenten von Lehre. Empirische Analysen der Dimensionen von Fragebogen entsprangen der Hoffnung, über die Dimensionalität der Fragebogen, die das Konstrukt ‚gute Lehre' zu erfassen beabsichtigen, einen „common core" als ein gemeinsames Set typischer Dimensionen von

Lehre zu extrahieren (vgl. Abrami et al., 1996). Auch im Rahmen der empirischen Fragestellung dieser Arbeit wurde angeführt, dass man auf diesem Wege unter Umständen zumindest eine einheitliche *operationale Definition* des Konstruktes und seiner Komponenten erzielen könnte. Gute Lehre bestünde damit aus den Komponenten, die in Fragebogen zur Lehrevaluation gemeinsam anzutreffen sind.

Überträgt man diesen Gedanken auf die Ergebnisse dieser Untersuchung, umfasste gute Lehre aus einer operationalen Sicht die Komponenten klar und verständlich zu sein (Dimension 6), für die Meinung anderer offen zu sein und zur Diskussion anzuregen (Dimension 12) sowie – nicht näher bestimmte – Ergebnisse zu erzielen, die von den Studierenden wahrgenommen werden (16). Diese Komponenten beschreiben zumindest das, was viele der untersuchten Fragebogen zur Evaluation von Lehre abfragen.

Es wäre vermutlich aber vermessen, dieser operationalen Definition von Lehre einen allgemeinen Anspruch zuschreiben zu wollen. Eine Reihe von methodischen Einwänden und inhaltlichen Kritikpunkten stehen dem entgegen. Zwei Aspekte beziehen sich sehr spezifisch auf die vorliegende Untersuchung. Zwei weitere treffen auch für jede andere Untersuchung zu, die mit strukturierenden Verfahren wie inhaltsanalytischen Kategorisierungen oder Faktorenanalyse arbeiten würde.

1) *Datenbasis*

   Um den Ergebnissen dieser Untersuchung das entsprechende Gewicht beimessen zu können, das für die Ableitung einer *allgemeinen* operationalen Definition notwendig gewesen wäre, hätten die Ergebnisse auf einer höheren Beurteilerübereinstimmung und einer repräsentativeren Stichprobe basieren müssen (vgl. die Diskussion dieser Aspekte in Abschnitt 8.5).

2) *Anzahl der Dimensionen*

   Die oben angeführte Definition wäre zwar multidimensional, bestünde aber lediglich aus drei Komponenten (den Dimensionen 6, 12 und 16). In Anbetracht der Komplexität des Lehrgeschehens und der Vielfältigkeit der bisherigen Ansätze von guter Lehre (vgl. Abschnitt 7.1) ist es fraglich, ob nicht weitere Dimensionen für eine adäquate Beschreibung des Lehrgeschehens notwendig sind. Vor diesem Hintergrund erscheint die oben angeführte Definition unvollständig.

3) Grenzen der Methodik: Unvollständigkeit

Weiterhin können die methodischen Einwände, die gegenüber faktorenanalytischen Ergebnissen vorgebracht werden, auch gegen diese Untersuchung gerichtet werden. Die inhaltsanalytische Vorgehensweise kann wie die Faktorenanalyse nur solche Aspekte hervorbringen, die im Datenmaterial in manifester (oder latenter Weise) enthalten sind. Infolgedessen bliebe jede Definition, die auf Basis einer solchen Vorgehensweise gewonnen würde, mit der Unsicherheit behaftet, möglicherweise unvollständig zu sein. Dies ist um so eher der Fall, je weniger repräsentativ das Ausgangsmaterial für die interessierende Grundgesamtheit ist.

4) Grenzen der Methodik: Relevanz

Schließlich bleibt zu bedenken, dass die in der vorliegenden Untersuchung als „Spitzenreiter" identifizierten Dimensionen nicht zugleich relevante Aspekte von Lehre verkörpern müssen. Die Gleichsetzung von Häufigkeit und Relevanz ist nur insofern zutreffend, als dass sich argumentieren ließe, diese Dimensionen seien aus der Sicht der Fragebogenautoren relevant. Rindermann (2001) spricht in einem ähnlichen Zusammenhang auch von der „definitorischen Macht" der Fragebogen(-autoren). Nichtsdestoweniger mögen andere Interessengruppen (Studierende, Dozierende, Politiker) andere Vorstellungen davon haben, was ein Fragebogen zur Lehrevaluation erfassen sollte. Unabhängig von einer durch Personengruppen definierten Relevanz gilt: Ob die in Fragebogen gefundenen Dimensionen für das Konstrukt ‚gute Lehre' relevant sind, kann ebenso wenig wie bei der Faktorenanalyse (Marsh, 1984) über eine Häufigkeitsanalyse entschieden werden.

Eine gewisse Homogenität oder Übereinstimmung von Faktoren zwischen verschiedenen Fragebogen konnte mehrfach aufgezeigt werden – sei es über Literatursichten (vgl. Centra, 1993; Braskamp & Ory, 1994, el Hage, 1996) oder auf empirischen Weg (vgl. Abrami & d'Apollonia, 1990; Abrami et al., 1996). In schwächerem Maße konnte dies auch die vorliegende Untersuchung zeigen. Unklar ist aber, worauf die geringe – aber vorhandene – Ähnlichkeit in den Instrumenten zurückgeführt werden kann. Bisher konnte nicht entschieden werden, ob darin ein gemeinsamer Kern des Konstruktes ‚gute Lehre' repräsentiert ist, ober ob es sich um die gemeinsamen Vorstellungen der Fragebogenautoren oder gar um ein Methodenartefakt handelt. Letzteres könnte vermutet werden, da für die Konstruktion neuer Verfahren immer wieder auf Itempools bereits existierender Verfahren zurückgegriffen wurde (vgl. Kapitel 5). Weiterhin

wird von vielen Autoren immer wieder auf die verbleibende Heterogenität hingewiesen. Diese Heterogenität blieb bisher in allen Studien bestehen (vgl. Abschnitt 7.3) und scheint nicht auflösbar.

Ob darin allerdings ein Mangel liegt bzw. ob die Suche nach einer allgemein gültigen Faktorenstruktur überhaupt sinnvoll ist, ist eine Frage der Perspektive: Die Perspektive von Invarianz oder Situationsspezifität von Lehre (Abrami et al., 1996).

Die Perspektive von Invarianz versus Situationsspezifität von Lehre

Zwei mögliche Perspektiven, die der Invarianz versus die der Situationsspezifität von Lehre, werfen ein jeweils unterschiedliches Licht auf die Frage, ob in der Heterogenität von Dimensionen in Fragebogen zur Lehrevaluation ein Makel zu sehen ist.

Die Heterogenität hinsichtlich der berücksichtigten Dimension von Lehre wäre als Mangel zu betrachten, wenn es allgemein anerkannte Kriterien für die Qualität von Lehre gäbe und ein Fragebogen genau diese im Vorfeld definierten Kriterien erfassen sollte. Hinter einer solchen Idee verbirgt sich der Gedanke einer *Invarianz* der relevanten Aspekte von Lehre. Von diesem Standpunkt aus gilt gute Lehre als unabhängig von verschiedenen Kontexten, verschiedenen Hochschulen, verschiedenen Fächern, verschiedenen Dozierenden (Marsh & Hocevar, 1984). In einem solchen Falle wären die Faktoren multidimensionaler Fragebogen gleich und somit über verschiedene Fragebogen hinweg replizierbar. „Moreover, the relative type and proportion of items representing the factors would also not vary across forms – that is, are the factors from evaluation instrument A similar to those from instrument B?“ (Abrami & d'Apollonia, 1990, S. 99-100).

Abrami und d'Apollonia führen fort: Es sei nicht logisch anzunehmen, dass die Faktoren für gute Lehre bekannt und invariant seien, wenn die Faktoren nicht replizierbar sind und überdies in unterschiedlichen Häufigkeiten auftreten. Die Ergebnisse ihrer eigenen Untersuchungen (vgl. Abschnitt 7.3) sprächen gegen die Invarianzannahme. „The inconsistency suggests that any one of the existing multidimensional rating form may not represent teaching for all instructors, courses, and settings“ (Abrami et al., 1996, S. 251).

Abrami und d'Apollonia (1990) weisen aber darauf hin, dass eine *alternative Sicht* zur Invarianzannahme existiert. Dieser Alternative, der Annahme der *Situationsspezifität* von Fragebogen zur Lehrevaluation, sei bisher zu wenig Aufmerksamkeit geschenkt worden. Abrami et al. (1996)

vertreten die Perspektive der Situationsspezifität und fordern, dass künftige Forschung situationsspezifische Einflüsse stärker berücksichtigen solle und mehr Augenmerk auf multiple Ansätze für die Definition von Lehre gerichtet werden müsse. Aus der Perspektive der Situationsspezifität ist die Heterogenität von Fragebogenverfahren eine logische Konsequenz. Sie würde aus der Perspektive der Situationsspezifität nicht als Mangel betrachtet. Sie bietet Raum für vielfältige Definitionen von Lehre und Instrumenten zu ihrer Erfassung.

Weitere Autoren des anglo-amerikanischen Sprachraumes argumentieren gegen die Invarianzannahme von Lehre und verdeutlichen die Konsequenzen für die Praxis. In einer sehr anschaulichen Weise ist das bei McKeachie (1997) zu sehen, der auf die Erfahrungen von Greenwald (1997) Bezug nimmt. Greenwald (1997) berichtet von der für ihn überraschenden Erfahrung, in einem Jahr die besten Evaluationsergebnisse für eine Veranstaltung erhalten zu haben und im darauffolgenden Jahr davon stark abweichende, negativere Bewertungen zu bekommen – für die in Inhalt, Didaktik und Aufbau nicht veränderte Veranstaltung. McKeachie (1997, S. 1224) rekurriert eindeutig auf die Situationsspezifität von Lehre, wenn er entgegnet: „Veranstaltungen sind unterschiedlich. Gute Dozenten bauen Brücken zwischen dem eigenen Wissen und dem der Studierenden. Was für den einen Studenten oder die eine Veranstaltung funktioniert, muss nicht für andere funktionieren."

In *Deutschland* wird die Frage nach der Heterogenität oder Homogenität von Faktoren über verschiedene Instrumente hinweg nicht explizit geführt (vgl. Kapitel 7). Dennoch finden sich auch hierzulande Äußerungen gegen die Annahme der Invarianz von Lehre. Diese beziehen sich allerdings nicht auf die Fragebogen zur Lehrevaluation, sondern direkt auf die Definition bzw. das Konstrukt ‚gute Lehre'.

Nach Bülow-Schramm und Reissert (1993) ist die Qualität der Lehre ein multipler Begriff, der nicht eindeutig definiert werden könne und der in Abhängigkeit von den Interessen der jeweiligen Evaluatoren differiere. Infolgedessen differierten dann auch die Qualitätsmaßstäbe. Ebenso ist Lehre in der Auffassung von Kromrey (2001) zielgruppenorientiert; Lehrqualität könne sinnvoll immer nur „relational" definiert werden „als Angemessenheit des Angebots (der Lehrenden) für definierte „Kunden" (Studierende)" (Kromrey, 2001, S. 36). Insgesamt scheint von vielen deutschsprachigen Autoren eine Haltung vertreten zu werden, wie sie bei Gold (1996, S. 149) formuliert wird: „Erfolgreiche Lehre ist notwendigerweise immer adaptiv."

9.2 Konsequenzen für die Praxis: Bedingungen für den Nutzen von Fragebogen zur Lehrevaluation

Ausgangspunkt der Diskussion um die Lehrevaluation in Deutschland war die „Qualitätssorge in der Lehre“ (Richter, 1994) und damit verbunden das Interesse, die Qualität der Lehre zu messen. Die Diskussion verengte sich rasch auf die Frage, ob diesbezüglich Fragebogen zur Lehrevaluation geeignete oder ungeeignete Instrumente darstellten. Innerhalb dieser Diskussion ist die Tendenz zu beobachten, den Einsatz von Fragebogen mit Evaluation gleichzusetzen (vgl. Kap. 4). Lehrevaluation wird reduziert auf Umfrageforschung (Kromrey, 1995). Diese Fokussierung auf die Befragung von Studierenden mittels Fragebogen, verbunden mit der Absicht, damit die Qualität der Lehre zu messen, wurde zu Recht von einigen Autoren kritisiert.

In diesem Sinne versteht sich die Position Rosemanns (1999) als fundamentale Kritik der o. g. Praxis: „Soll die studentische Veranstaltungsbeurteilung genutzt werden im Sinne einer validen Messung und Beurteilung der Qualität der Lehre bzw. der Lehrkompetenz des Dozenten bzw. zu Vergleichen zwischen Dozenten, dann ist die Titelfrage [„Qualität der Lehre durch Befragung?“] negativ zu beantworten“ (Rosemann, 1999, S. 50). In Konsequenz dessen sieht der Autor den Nutzen von Fragebogen zur Lehrevaluation nur für formative Zwecke gegeben – im Sinne eines „Anstoßes für Innovationen in der Lehre“ (Rosemann, 1999, S. 51). Aber selbst Befürworter wie Rindermann (2001) weisen darauf hin, dass Ergebnisse der studentischen Lehrevaluation nur unter ganz bestimmten Randbedingungen verwendbar sind (für aggregierte Daten, bei mindestens 5 Veranstaltungen etc.).

Wie ist der Nutzen von Fragebogen zur Lehrevaluation nun abschließend zu beurteilen? Die Antwort darauf scheint eine Frage der Perspektive zu sein – nicht nur die Frage der Perspektive, ob studentische Lehrevaluationen zu summativen oder formativen Zwecken eingesetzt werden können, sondern die einer grundsätzlich veränderten Perspektive vor dem Hintergrund des multiplen Evaluationsansatzes.

Die Rolle von Fragebogenverfahren zur Lehrevaluation: Das Evaluationssetting

Dem multiplen Ansatz der Lehrevaluation folgend sollte Evaluation auf einen Datenpool zurückgreifen, der aus der Anwendung verschiedener Methoden aus verschiedenen Quellen erzeugt wurde. Die konsequente Umsetzung des multiplen Ansatzes in der Lehrevaluation schlägt sich in

der Entwicklung eines ganzen Evaluationssystems nieder. Vor dem Hintergrund des multiplen Evaluationsansatzes ergibt sich eine veränderte Perspektive und es gilt zu fragen:

Können Fragebogen zur Lehrevaluation einen sinnvollen Beitrag leisten in einem umfassenden System zur Evaluation von Lehre?

Es interessiert also die Bewertung der Rolle und des Beitrags von Fragebogen zur Lehrevaluation als Baustein in einem komplexen Evaluationsmodell (vgl. auch Preißer, 1992). Können Daten, die ein Fragebogen bereitstellt, sinnvoll genutzt werden? Sind sie sinnvoll im Rahmen des Datenpools, der die Grundlage darstellt für ein durch den Bewertungsprozess der Experten entstehendes (Ab-)Bild der Qualität der Lehrveranstaltungen / der Lehre? Zugespitzt könnte man formulieren: Es geht nicht um Evaluation der Lehre **durch** Lehrevaluationsfragebogen, sondern umfassende Evaluation der Lehre mittels eines komplexen Evaluationssystems, das **auch** Fragebogen zur Lehrevaluation zur Datengewinnung einsetzt.

Insgesamt scheint es notwendig, das Gesamtsystem in den Blick zu nehmen. Die Diskussion von Fragebogen ohne diesen Kontext muss dagegen „steril“ und praxisfern wirken. Die Rolle von Fragebogen im Rahmen solcher Modelle zu diskutieren, sollte die Aufgabe zukünftiger Forschung sein. Für die Entwicklung solcher Evaluationssysteme und Modelle sei auf die Erfahrungen des amerikanischen Raumes (vgl. Abrami, 2001; Arreola, 2000; Theall & Franklin, 1990b) sowie auf erste Ansätze aus Deutschland verwiesen (überwiegend als Modelle der Evaluationsverbünde; z. B. Reissert & Konnerth, 2001; Webler, 1999; Winter, 2000).

Dabei ist zu bedenken, dass der Nutzen von Fragebogenverfahren nicht allein an teststatistische Gütekriterien gebunden ist. Er ist vielmehr an eine ganze Reihe weiterer Bedingungen geknüpft. Doch ehe einige Bedingungen vorgestellt werden, sei auf ein weiteres Phänomen verwiesen: Die Verflechtung des Themas mit politischen und normativen Überzeugungen.

Exkurs: Die Verflechtung mit politischen und normativen Überzeugungen

Viele Entscheidungen im Evaluationsprozess entziehen sich einer wissenschaftlichen Klärung. Beispielsweise kann Forschung „prinzipiell nicht bestimmen, welche Ziele durch das Evaluationsvorhaben anzustreben sind“ (Wottawa, 2001, S. 154). Ziele müssen anderweitig bestimmt werden. Das bietet Raum für normative Einflüsse und ist zugleich die Grundlage für Widerstände, emotionale Ablehnung etc.

Auf die Verflechtung der Diskussion um die Evaluation von Lehre mit politischen und normativen Dimensionen weisen eine Reihe von Autoren hin. Die folgende Auswahl an Umschreibungen für diesen Sachverhalt mag dies exemplarisch verdeutlichen:

- Für Preißer (1993) ist die Evaluation von Lehre nicht viel mehr als eine „symbolische Politik", die der Ablenkung von tieferen Problemen im Hochschulbereich (z. B. extreme asymmetrische Kommunikationsstruktur, fehlendes Managementwissen) dient.
- Ähnlich sieht Schick (1992) in der Evaluation der Lehre ein „Ablenkungsmanöver" von einem Kernproblem der Hochschulmisere, „nämlich der Nivellierung durch zu viele und nicht geeignete (...) Abiturienten" (S. 369).
- Die Debatte um Gütekriterien bezeichnet el Hage (1996) als „Nebenkriegsschauplatz" von Abwehr und Verweigerung (S. 84).
- Rinderman (2001) konstatiert eine „selektiven Taubheit" der Forschenden, da die bereits vorhandenen Antworten auf die Fragen der Gütekriterien „immer noch und immer wieder überhört werden" (S. 206).
- Unter dem Stichwort „Selbstobjektivierungsproblem im akademischen Milieu" diskutiert Rindermann (2000a,b) Aspekte, die aus der Tatsache resultieren, dass Forschende von den Ergebnissen der Lehrevaluationsforschung selbst betroffen sind, da sie quasi selbst Gegenstand der Forschung sind.

Widerstände, normative Grundsatzfragen und politischer Wille sind Determinanten, die aus der Diskussion um die Evaluation von Lehre nicht wegzudenken sind. Gerade deswegen scheint es wichtig aufzuzeigen, an welcher Stelle die Wissenschaft Beiträge für die Diskussion liefern kann, wo ihre Grenzen sind und an welche Bedingungen der Einsatz von Fragebogen in einem Evaluationsmodell gebunden ist.

Bedingungen für den Einsatz von Lehrevaluationsfragebogen in komplexen Modellen zur Lehrevaluation

Im Folgenden seien einige der Bedingungen genannt für den Einsatz von Lehrevaluationsfragebogen in umfassenden Modellen zur Evaluation. Die ersten beiden Aspekte sind grundsätzlicher Natur. Denn erst wenn der Evaluationsgedanke als solcher akzeptiert ist, können konkrete Ziele geklärt und angestrebt werden (Wottawa & Thierau, 1998). Tabelle 28 führt diese und andere Bedingungen auf. Des Weiteren wird versucht, mögliche

Beiträge der wissenschaftlichen Forschung aufzuzeigen (vgl. ebenfalls Tab. 28, S. 101).

Tab. 28. Bedingungen für den Einsatz von Lehrevaluationsfragebogen im Rahmen komplexer Evaluationsmodelle

a)	*Die normative Grundsatzfrage: Ist umfassende Evaluation gewollt?* Diese Frage scheint in ihrer Grundsätzlichkeit kaum öffentlich diskuiert zu werden. Gleichwohl aber finden sich immer wieder Hinweise, dass Evaluation ohne die Akzeptanz und den Willen zur selben wirkungslos bleibt, unter Umständen sogar kontraproduktive Effekte erzeugt (vgl. bereits Spitzer, 1976). • Möglicher Beitrag der Wissenschaft: Bereitstellung von Information
b)	*Akzeptanz der Veränderbarkeit und Veränderungsbedürftigkeit* Akzeptanz ist notwendig, wenn Evaluation Wirkungen erzielen soll. Dieser Aspekt ist auch an individuelle Voraussetzungen geknüpft. So erfordert die Akzeptanz von Evaluation unter anderem die Bereitschaft, Gewohntes aufzugeben, sich der Furcht vor Misserfolg auszusetzen und den Glauben an einen erzielbaren Fortschritt (Wottawa & Thierau, 1998). • Möglicher Beitrag der Wissenschaft: Bereitstellung von Informationen, Methoden der Einstellungsänderung, Moderation der Prozesse
c)	*Evaluationsziele und -kriterien* Die Relevanz von Zielen und Kriterien arbeitet sehr deutlich Gold heraus: „Der Validitätsdiskussion müsste also eine Kriterienklärung vorangehen. Was würde man von guter Lehre erwarten: Daß sie den relevanten Stoff in möglichst kurzer Zeit korrekt und umfassend vermittelt? Daß sie objektivierbare Lernerfolge begünstigt? Daß sie Lernende und Lehrende zufriedenstellt? Selbst der studentische Lernerfolg, der im Zuge der konvergenten Validität häufig als besonders ‚hartes' und erwünschtes Kriterium der Lehrqualität bezeichnet, ist hierzu offenbar nur bedingt verwendungsfähig. (...) Wenn aber inhaltlich unklar bleibt, welche Kriterien Lehrqualität indizieren, müssen Validierungsversuche [bzw. das Evaluationsvorhaben insgesamt, eigene Anmerkung M.H.] unvollständig bleiben" (Gold, 1996, S. 149). • Möglicher Beitrag der Wissenschaft: Aufzeigen bisheriger Problemfelder in dieser Diskussion, Methoden zur Konsensfindung, Moderation der Prozesse

d) Institutionelle Rahmenbedingungen

„Erfolgsfaktoren“ für Evaluation liegen nach Reissert und Konnerth (2001) auch in den hochschulpolitischen Rahmenbedingungen; zum Beispiel ausreichende Autonomie sowie das Vorhandensein von Kooperation und Vertrauen.

- Möglicher Beitrag der Wissenschaft: Erforschung der Wirkungszusammenhänge, Modellentwicklung (Qualitätssicherung, Organisationsentwicklung)

e) Administration

Kernaufgabe eines Evaluationsmodells ist, auch die Aspekte der Administration des Evaluationsvohabens festzulegen. Darunter fällt die Beantwortung der Fragen: Wer, wann, wie und was evaluieren soll. Zum Beispiel wurden Einflüsse des Durchführungszeitpunktes der Evaluation nachgewiesen (Centra, 1993). Empfehlungen und Hinweise zu diesen Aspekten finden sich bei Arreola (1995, 2000) sowie bei Thierau-Brunner, Stangel-Meseke & Wottawa (1998).

- Möglicher Beitrag der Wissenschaft: Erforschung der Wirkungszusammenhänge, Modellentwicklung

f) Entscheidungsfindung und Verwendung der Daten

Nachdem die Daten erhoben sind, wird die Frage wichtig, wie diese ausgewertet werden sollen, um zu einer Entscheidungsfindung zu kommen. Beispiele für die Handhabung und Interpretation von Daten aus Lehrevaluationsbogen zum Zwecke der Personalentscheidung oder der Vergabe von „Benefits“ finden sich bei Abrami (2001) und Huberty (2000).

- Möglicher Beitrag der Wissenschaft: Modelle der Entscheidungsfindung, Nutzen von Teaching Portfolios (vgl. Centra, 1993).

g) Mögliche Konsequenzen

Hierbei handelt es sich um Aspekte, die bereits unter dem Stichwort „Konsequenz-Validität“ erörtert wurden (vgl. Greenwald, 1997). Wie können studentische Veranstaltungsbeurteilungen zur systematischen Optimierung der Lehre verwendet werden?. Die Art und Weise der Verwendung der Daten in Komitees und Verwaltungen stehen diesbezüglich im Mittelpunkt des Interesses: z. B. mangelndes Wissen, wie diese Daten zu interpretieren sind.

- Möglicher Beitrag der Wissenschaft: Untersuchung verschiedener Anreizsysteme, Training der Bewertenden, Wirkungsanalysen

Während el Hage (1996, S. 121) in Bezug auf die Gütekriterien von Fragebogen noch anführte „Diese Art der Qualitätsbestimmung wird auch erst dann wirklich relevant, wenn es beim Vergleich zwischen Lehrenden um tatsächliche Auswirkungen gehen wird.“, dann ist dies spätestens mit

dem neuen Besoldungsgesetz für Professor der Fall. Die Relevanz dieser Bestimmung und – so könnte ergänzt werden – der Auseinandersetzung mit allen anderen genannten Bedingungen scheint dringend notwendig: sieht das neue Gesetz doch die Bewertung von Leistung in Lehre und Forschung vor.

Die genannten Rahmenbedingungen klären, wer welche Rolle im Evaluationsprozess übernimmt, was mit den Daten geschieht, warum und wozu diese erhoben werden etc. Die grundlegende Funktion eines komplexen Evaluationsmodells scheint darin zu liegen, dass es alle Prozesse und Bedingungen expliziert: Gemeinsamer Nenner all dieser Bedingungen wären damit die Aspekte *Transparenz, Offenheit und Information*.

Diese könnte man auch als Charakterisierung für den Dialog heranziehen, der notwendig ist, um die genannten Bedingungen auszuhandeln. Besonders essentiell im Evaluationsprozess erscheint die Ziel- und Kriteriendiskussion (vgl. Thierau-Brunner et al., 1998). Der Dialog kann nicht allein Ziel in sich sein, sondern muss vor allem als Mittel für Kriterienfindung fungieren (Bülow-Schramm, 1995). Dem Prozess der Kriterienfindung und der Explikation kommt vor allem auch deswegen so hohe Bedeutung zu, da Evaluation in ihrem Wesen zielgerichtet und zweckorientiert ist (vgl. Kap. 2). Zwecke und Ziele müssen ausgehandelt werden, letztlich auch, um Evaluation einer späteren Wirkungsanalyse zugänglich zu machen. Denn schließlich sind Evaluationen

> „auf die Wirklichkeit bezogene Aktivitäten. Was am Ende zählt, ist (...) das Ausmaß, in dem die Evaluationen zu Veränderungen in Richtlinien, Programmen oder Praktiken führen – Veränderungen, die auf kurze oder lange Sicht die Bedingungen des menschlichen Lebens verbessern“ (Rossi & Freeman, 1993, S.403).

10. Zusammenfassung und Ausblick

Anlass für diese Arbeit waren die in Deutschland seit den 1990er Jahren immer wieder geführten Auseinandersetzungen um die Evaluation von Lehre. Die wissenschaftliche Diskussion in Deutschland, aber auch in den USA, ist geprägt von Kontroversen und polarisierenden Stellungnahmen (Braskamp & Ory, 1994). Gesamtziel der vorliegenden Arbeit war daher, einen Beitrag zur Transparenz der Aktivitäten, Verfahren und Positionen im Bereich der Lehrevaluation aus wissenschaftlicher Sicht zu leisten.

Dazu wurden im zweiten Kapitel zunächst einige grundlegende Begriffe geklärt und die Lehrevaluation in das weite Feld der Evaluation im Anwendungsbereich Hochschule eingeordnet. Es wurde deutlich, dass die Lehrevaluation nur einen Teilausschnitt möglicher Evaluationsfelder in der Hochschule repräsentiert. Das dritte Kapitel gab einen Überblick über die historischen Entwicklungen der Evaluation der Lehre in den USA ebenso wie in Deutschland, wonach sich die Situation in Deutschland gegenüber den USA als heterogener, weniger etabliert und umstrittener darstellte.

Die in Kapitel 4 aufgeworfenen Fragen nach Studierenden als gewinnbringende Informationsquelle und nach Fragebogen als sinnvoller Methodik der studentischen Lehrevaluation leiteten in den Kernbereich dieser Arbeit über, der Lehrveranstaltungsevaluation. Das Kapitel stellte heraus, dass in Deutschland eine Fokussierung auf Evaluation im Sinne der Befragung von Studierenden via Fragebogen konstatiert werden muss. Die Auseinandersetzung mit den Instrumenten, den verschiedenen Fragebogenverfahren, die in diesem Zusammenhang eingesetzt werden, stand im Zentrum der vorliegenden Arbeit. In Kapitel 5 erfolgte daher ein Überblick über Aufbau, Form und mögliche Zielsetzungen solcher Verfahren. Dieser Überblick wies auf die Heterogenität der Instrumente hin. Die Diskussion der Qualität der Fragebogenverfahren in Form einer intensiven Auseinandersetzung mit dem aktuellen wissenschaftlichen Forschungstand der deutschsprachigen und anglo-amerikanischen Literatur wurde im sechsten Kapitel vertieft. Ergebnis dieser Auseinandersetzung war, dass sich Deutschland und die USA im Stand um die Debatte der Gütekriterien unterscheiden. Während in Deutschland noch vielerorts Grundsatzdiskussionen geführt werden, rücken im amerikanischen Sprachraum vor allem Aspekte der Konsequenz-Validität und des richtigen Einsatzes von Fragebogen in den Vordergrund. Jedoch ist die Diskussion um die Qualität der Verfahren auch in den USA nicht endgültig abgeschlossen – wie aktuelle Veröffentlichungen zeigen. So wurden als weiteres Ergebnis der Auseinandersetzung mit der Fachliteratur in Kapitel

7 zwei Problemfelder für die Erfassung des Konstruktes ‚gute Lehre' abgeleitet: (1) Die fehlende theoretische Fundierung und Definitionsprobleme sowie (2) die Verlagerung der Debatte von den Komponenten des Konstruktes ‚gute Lehre' auf die Diskussion der Dimensionalität von Fragebogen.

Kapitel 8 greift schließlich die Frage der Dimensionalität von Fragebogenverfahren als möglichen Hinweis auf die Dimensionalität des Konstruktes auf und macht sie zum Gegenstand des empirischen Teils dieser Arbeit. Konkret wurden 7 amerikanische und 7 deutsche Fragebogen zur Lehrevaluation hinsichtlich der in ihnen enthaltenen Dimensionen untersucht. Die Häufigkeitsanalyse der Kategorisierung der insgesamt 486 Items zu 31 möglichen Dimensionen von Lehre zeigte, dass drei Dimensionen in den Verfahren beider Kulturräume besonders häufig anzutreffen sind. Ferner verteilten sich die Dimensionen über die meisten einzelnen Verfahren gleichmäßig, jedoch nicht innerhalb eines Kulturraumes und auch nicht über beide Kulturräume. In der Gruppe der amerikanischen Verfahren traten signifikant andere Anteile der Dimensionen auf als in der Gruppe der deutschen Verfahren. Die Verfahren müssen im Vergleich zueinander insgesamt als heterogen charakterisiert werden. Eine Ableitung allgemein verbindlicher oder allgemein anzutreffender Komponenten von Lehre scheint nicht möglich. Diese Ergebnisse stützen die erste bislang wenig vorhandenen Ergebnisse anderer Autoren (z. B. Abrami et al., 1996).

Ob in der Heterogenität der in den Verfahren berücksichtigten Dimensionen ein Mangel zu sehen ist, wurde in der Diskussion des neunten Kapitels als eine Frage der Perspektive beantwortet. Für die Annahme der Invarianz von Lehre musste dies bejaht werden; nicht dagegen aus Sicht der Situationsspezifität von Lehre, wonach Lehre notwendiger Weise immer adaptiv ist (Gold, 1997).

Die Diskussion zeigte weiterhin auf, dass die Frage nach dem Nutzen von Fragebogenverfahren ebenfalls eine Frage der Perspektive ist. Aus der Sicht des multiplen Evaluationsansatzes wäre die Rolle der Fragebogen im Zusammenhang mit komplexen Evaluationsmodellen zu betrachten. Der Nutzen von Fragebogen in einem solchen Modell sollte daher künftig weiter erforscht werden, wobei die vielfältigen Bedingungen für den Einsatz von Fragebogen zur Lehrevaluation beachtet werden müssen. Die Diskussion versuchte auch, auf die bei einigen Autoren genannte Verflechtung des Themas mit politischen und normativen Überzeugungen hinzuweisen. Zum Abschluss der Diskussion wurden einige Bedingungen

für den Nutzen von Fragebogen im Rahmen komplexer Evaluationssysteme genannt und mögliche Beiträge der Wissenschaft skizziert.

Den in dieser Arbeit diskutierten Argumentationen und empirischen Ergebnissen folgend, ließe sich der Schluss ziehen, dass Generalkonzepte im Sinne eines allgemein guten und anerkannten Fragebogens zur Lehrevaluation oder einer allgemein verbindlichen Definition von Lehre unter Ausgrenzung situationsspezifischer Einflüsse nicht möglich sind. Insgesamt scheint es lohnend, in einem durch Offenheit und Transparenz charakterisierten Dialog aller Beteiligten individuelle oder lokale, aber komplexe Modelle zur Evaluation von Lehre zu entwickeln.

Aufgaben zukünftiger Forschung lägen damit weniger in der methodischen Optimierung eines einzelnen Instrumentes (z. B. Fragebogen) als eher darin, die Gestaltung und das Wirken der verschiedenen Komponenten der jeweiligen Evaluationsmodelle zu analysieren.

11. Literatur

Abrami, P. C. (1985). Dimensions of effective college instruction. *The Review of Higher Education, 8 (3)*, 211-228.

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30 (2)*, 221-227.

Abrami, P. C. (2001). Improving judgments about teaching effectivness using teacher rating forms. In *New Directions for Institutional Research Nr. 109* (pp. 59-87). San Fransico: Jossey-Bass Publishers.

Abrami, P. C. & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning, No. 43, pp. 97-11). San Fransisco: Jossey-Bass Publishers.

Abrami, P. C., d'Apollonia, S. & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82 (2)*, 219-231.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Ed.), *Higher Education: Handbook of theory and research* (Vol. 11, pp. 213-264). New York: Agathon Press.

Arreola, R. A. (1995). *Developing a comprehensive faculty evaluation system:* Bolton, MA: Anker Publishing.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system:* (2. Aufl.). Bolton, MA: Anker Publishing.

Arreola, R. A., & Aleamoni, L. M. (1990). Practical decisions in developing and operating a faculty evaluation system. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning, No. 43, pp. 37-55). San Fransisco, CA: Jossey-Bass Publishers.

Astleitner, H. & Krumm, V. (1996). Dimensionen von Lehrverhalten: Faktorenstrukturen 1. und 2. Ordnung mit Kreuzvalidierung. *Empirische Pädagogik, 10 (1)*, 7-26.

Banz, M. L. & Rodgers, J. L. (1985). Dimensions underlying student ratings of instruction: a multidimensional scaling analysis. *American Educational Research Journal*, *22 (2)*, 267-272.

Basler, H.-D. (1978). Ein 20-Item Fragebogen zur Evaluation des Unterrichts. *Medizinische Psychologie*, *3*, 203-204.

Basler, H.-D., Bolm, G., Dickescheid, T. & Herda, C. (1995). Marburger Fragebogen zur Akzeptanz der Lehre. *Diagnostica*, *41 (1)*, 62-79.

Beiträge zur Hochschulforschung (1993). Bayrisches Staatsinstitut für Hochschulforschung und Hochschulplanung (Hrsg), *Band 4*.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5. Aufl.). Berlin: Springer.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer.

Brandenburg, G. C. & Remmers, H. H. (1927). A rating scale for instructors. *Educational Administration and Supervision*, *13*, 399-406.

Braskamp, L. A., Brandenburg, D. C. & Ory, J. C. (1984). *Evaluating teaching effectiveness: A practical guide*. Beverly Hills, CA: Sage.

Braskamp, L. A. & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance.* San Fransisco: Jossey-Bass Publishers.

Breakwell, G. M., Hammond, S. & Fife-Schaw, C. (1995). *Research Methods in Psychology*. London: Sage.

Bülow, M. (1977). Evaluation I: Verfahren - Methoden - Erfahrungen zur Überprüfung universitärer Ausbildung. In *Hochschuldidaktische Arbeitspapiere* (Bd. 8). IZHD/Hamburg: **???.**

Bülow-Schramm, M. (1995). Wer hat Angst vor den Evalutoren? In Raabe - Redaktion (Hrsg.), *Handbuch Hochschullehre Highlights* (Bde. 1, Evaluation der Lehre: Ziele - Akzeptanz - Methoden, S. 1-19, D 1.6). Stuttgart: Raabe.

Bülow-Schramm, M. & Reissert, R. (1993). Evaluation der Lehre in Deutschland - eine hochschulpolitische Aufgabe mit Vergangenheit und Zukunft. *Beiträge zur Hochschulforschung, 4*, 393-405.

Bundesministerium für Bildung und Forschung; Bundesministerium des Inneren. (2001, 30.05.). Kabinett beschließt neues Dienstrecht für Professoren [Online]. In *Pressemitteilung 80/2001*. available online: http://www.bmbf.de/.

Burdsal, C. A. & Bardo, J. W. (1986). Measuring student's perceptions of teaching: dimensions of evaluation. *Educational and Psychological Measurement, 46*, 63-79.

Carletta, J. (1995). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics, 30 (11)*, 1-6.

Carstensen, D. & Reissert, R. (1997). *Qualitätsförderung an Hochschulen. Das Verfahren der internen und externen Evaluation* [On-line]. available: http://bawue.gew.de/fundusho/evaluation.html.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research.* IDEA Paper No. 20. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1989). *Defining and evaluating college teaching.* IDEA Paper No. 21. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1990). *Student ratings of teaching: Recommendations for use.* IDEA Paper No. 22. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited.* IDEA Paper No. 32. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E. (1996). *Developing an effective faculty evaluation system.* IDEA Paper No. 33. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement.* San Francisco, CA: Jossey-Bass Publishers.

Centra, J. A. (1993). *Reflective faculty evaluation: enhancing teaching and determining faculty effectiveness.* San Francisco, CA: Jossey-Bass Publishers.

Cohen, P. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20 (1)*, 37-46.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13 (4)*, 321-341.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51 (3)*, 281-309.

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education, 54 (4)*, 448-458.

Cohen, P. A. (1990). Bringing research into practice. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning, No. 43, pp. 123-132). San Fransisco, CA: Jossey-Bass Publishers.

d'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52 (11)*, 1198-1209.

DER SPIEGEL. (1989). Welche Uni ist die beste? *DER SPIEGEL, 50*.

DER SPIEGEL. (1993, 19. April). Welche Uni ist die beste? Spiegel-Rangliste der deutschen Hochschulen. *DER SPIEGEL, 16*, 80-101.

DER STERN. (1993, 15. April). Deutschlands beste Universitäten. *DER STERN, 16*.

Detchen, L. (1940). Shall the student rate the professor? *Journal of Higher Education, 11 (3)*, 146-154.

Dickenberger, D., Tracy, R. & Nystroem, A. (2000). Modularisierte Evaluation der Lehre [Online]. In *ProfiLe, Arbeitsbericht II*. available: http://www.

Diehl, J. M. (1996). Studentische Evaluation von Hochschulveranstaltungen. Ein Kommentar zu Kromrey. *Zeitschrift für Pädagogische Psychologie, 10 (3/4)*, 167-170.

Diehl, J. M. (1998). *Fragebögen zur studentischen Evaluation von Hochschulveranstaltungen (Version 5.0).* Gießen: Fachbereich 06 Psychologie (Abteilung Methodik).

Diehl, J. M. (in Druck). Studentische Lehrevaluation in den Sozialwissenschaften: Fragebögen, Normen, Probleme. In E. Keiner (Hrsg.), *Evaluation* (in) der Erziehungswissenschaft. Weinheim: Deutscher Studienvorlag.

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. & Reuschling, H. (1998). *VBOR 5.0, VBREF 5.0 - Dokumentation und Auswertungsprogramme* [3,5" Diskette]. Justus-Liebig-Unversität Gießen: Fachbereich 06 Psychologie.

Diehl, J. M. & Staufenbiel, T. (1997). *Evaluation von Lehre. Normen für VBVOR und VBREF.* Gießen: Fachbereich 06 Psychologie (Abteilung Methodik).

Diehl, P. F. (1989). *Understanding student evaluations of teaching effectiveness: A handbook for administrators and faculty* (2. Aufl.). Athens, Georgia: University of Georgia, The Office of Instructional Development.

Dowell, D. A. & Neal, J. A. (1982). A selective review of the validty of student ratings of teaching. *Journal of Higher Education, 53 (1)*, 51-62.

Doyle, K. O. (1975). *Student evaluation of instruction.* Lexington, MA: D.C. Heath & Co.

Eble, K. E. (1984). New directions in faculty evaluation. In P. Seldin (Hrsg.), *Changing practices in faculty evaluation. A critical assessment and recommendations for improvement* (pp. 96-100). San Fransisco, CA: Jossey-Bass Publishers.

Ellis, R. (Ed.) (1993). *Quality assurance for university teaching.* Bristol, PA: Open University Press.

Enders, J. & Teichler, U. (1995). *Der Hochschullehrerberuf. Aktuelle Studien und ihre hochschulpolitische Diskussion.* Neuwied: Luchterhand.

Engel, U. (Hrsg.). (2001). *Hochschul-Ranking. Zur Qualitätssicherung von Studium und Lehre.* Frankfurt a. M.: Campus.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243-288.

Feldman, K. A. (1977). Consistency and variability among student college students in rating their teachers and courses: A review and analysis. *Research in Higher Education, 6*, 223-274.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education, 28*, 291-344.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30 (6)*, 583-645.

Feuerstein, H. J. (1997). *"Prüf den Prof". Evaluation der Lehre an der Fachhochschule Kehl. Evaluation: Ein heißes Eisen der bundesdeutschen Hochschulpolitik* [On-line]. available: http://www.fh-kehl.de/zheaf/eva_5.htm.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik* (2. Aufl.). Göttingen: Hogrefe.

FOCUS MAGAZIN VERLAG. (1993, 27. September). Die besten deutschen Universitäten. *FOCUS, 39*.

Franklin, J. & Theall, M. (1990). Communicating student ratings to decision makers: design for good practice. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning, No. 43, pp. 75-93). San Fransisco, CA: Jossey-Bass Publishers.

Friedeburg, L. von. (1996). Einleitung: Qualität von Lehr und Studium. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 16 (2)*, 115-118.

Gold, A. (1996). Können Studierende die Qualität der Lehre beurteilen? *Zeitschrift für Pädagogische Psychologie, 10 (3/4)*, 147-150.

Gräf, L. (1991). Fragwürdige Experten - Sekundäranalyse der SPIEGEL-Untersuchung zur Qualität westdeutscher Universitäten. *Soziologie. Mitteilungsblatt der DGfS, 1*, 69-85.

Gralki, H. O. & Hecht, H. (1992). Hochschuldidaktische Aspekte der Beurteilung von Lehrveranstaltungen durch Studenten. In D. Grühn & H. Gattwinkel (Hrsg.), *Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 99-113). Berlin: FU-Dokumentationsreihe.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52 (11)*, 1182-1186.

Greenwood, G. E., Bridges, C. M. Jr., Ware, W. B. & McLean, J. E. (1973). Student evaluation of college teaching behaviors instrument: a factor analysis. *Journal of Higher Education, 44 (8)*, 596-604.

Habel, E. (1995). Hochschulen zum Rapport**???** In Raabe - Redaktion (Hrsg.), *Handbuch Hochschullehre Highlights* (Bde. 1, Evaluation der Lehre: Ziele - Akzeptanz - Methoden, S. 1-28, D 1.5). Stuttgart: Raabe.

Hage, N. el. (1996). *Lehrevaluation und studentische Veranstaltungskritik. Projekte, Instrumente und Grundlagen.* Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Hofmann, J. M. (1988). Studienmotivation und Veranstaltungsbeurteilung. Eine Korrelationsstudie am Beispiel der Beurteilung psychologischer Lehrveranstaltungen. *Psychologie in Erziehung und Unterricht, 35*, 119-126.

Hornbostel, S. & Daniel, H.-D. (1995). Das SPIEGEL-Ranking. Mediensensation oder ein Beitrag zur hochschulvergleichenden Lehrevaluation? In P. Ph. Mohler (Hrsg.), *Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung* (2. Aufl., S. 29-44). Münster: Waxmann.

Hoyt, D. P. & Pallett, W. H. (1999). *Appraising teaching effectiveness: Beyond student ratings.* IDEA Paper No. 36. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Huberty, C. (2000, April). An Approach to Annual Assessment and Evaluation of University Faculty. *Paper presented at annual meeting of the American Educational Research Association, New Orleans*.

Klein, S. (1993). Studentische Lehrevaluation. Entwicklung, Methoden und Ergebnisse am Beispiel der Fakultät für Betriebswirtschaftslehre an der Universität Mannheim. *Beiträge zur Hochschulforschung*, *4*, 429-446.

Kleine, D. & Merkens, H. (1979). Überprüfung eines Fragebogens zur Beurteilung von Lehrveranstaltungen. *Psychologie in Erziehung und Unterricht*, *26*, 149-153.

Kramis, J. (1990). Bedeutsamkeit, Effizienz, Lernklima. Grundlegende Gütekriterien für Unterricht und Didaktische Prinzipien. *Beiträge zur Lehrerbildung*, *8*, 279-296.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Psychologie Verlags Union.

Krippendorff, K. (1980). *Content Analysis: An introduction to its methodology*. London: Sage.

Kriz, J. (1995). Die Wirklichkeit von (Vor-)Urteilen. Über die inhaltlichen und methodischen Hintergründe der STERN-Image-Analyse. In P. Ph. Mohler (Hrsg.), *Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung* (2. Aufl., S. 11-28). Münster: Waxmann.

Kromrey, H. (1994). Wie erkannt man "gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen. *Empirische Pädagogik*, *8 (2)*, 153-168.

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

Kromrey, H. (1996a). <<Gute>> oder <<schlechte>> Sozialforschung? *Zeitschrift für Pädagogische Psychologie*, *10 (3/4)*, 171-173.

Kromrey, H. (1996b). Qualitätsverbesserung in Lehre und Studium statt sogenannter Lehrevaluation. Ein Plädoyer für gute Lehre gegen schlechte Sozialforschung. *Zeitschrift für Pädagogische Psychologie*, *10 (3/4)*, 153-166.

Kromrey, H. (1998). Empirische Sozialforschung. In *UTB für Wissenschaft 1040* (8. Aufl.). Opalden: Leske+Budrich.

Kromrey, H. (2001). Studierendenbefragungen als Evaluation der Lehre? Anforderungen an Methodik und Design. In U. Engel (Hrsg.), *Hochschul-Ranking* (S. 11-47). Frankfurt a. M.: Campus.

Lamnek, S. (1990). Zur Professionalität der Studie "Welche Uni ist die beste?". *Soziologie. Mitteilungsblatt der DGfS, 2*, 91-99.

Lamnek, S. (1993). *Qualitative Sozialforschung* (Bd. 2, Methoden und Techniken, 3. Aufl.). Weinheim: Beltz.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Psychologie Verlags Union.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of educational Psychology*, *52*, 77-95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76 (5)*, 707-754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *Journal of Educational Research, 11*, 253-388.

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83 (2)*, 285-296.

Marsh, H. W. & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64 (1)*, 1-18.

Marsh, H. W. & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21 (2)*, 341-366.

Marsh, H. W. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9-18.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52 (11)*, 1187-1197.

Mayring, P. (2000). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (7. Aufl.). Weinheim: Beltz: Psychologie Verlags Union.

McKeachie, W. J. (1987). Instructional evaluation: current issues and possible improvements. *Journal of Higher Education, 58 (3)*, 344-350.

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82 (2)*, 189-200.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52 (11)*, 1218-1125.

Merten, K. (1995). *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis* (2. Aufl.). Opladen: Westdeutscher Verlag.

Meyer-Althoff, M. (1978). Evaluation II: Versuche und Erfahrungen mit Selbstevaluation. *Hochschuldidaktische Arbeitspapiere, 11*.

Millman, J. (1981). *Handbook of teacher evaluation.* Beverly Hills, CA: Sage.

Moosbrugger, H., Hartig, J. & Naumann, J. (1997). Fragebogen zur Evaluation des Lehr- und Lernverhalten (FELL). In *Arbeiten aus dem Institut für Psychologie* (Bd. 3/1997). Frankfurt a. M.: Institut für Psychologie der Universität Frankfurt.

Müller-Wolf, H.-M. (1977). *Lehrverhalten an der Hochschule.* München: Verlag Dokumentation.

Mummendey, H.-D. (1995). *Die Fragebogenmethode* (2. Aufl.). Göttingen: Hogrefe.

Neidhardt, F. (1990). Lob und Tadel sind befangen. Über den Umgang mit dem SPIEGEL-Ranking. *SPIEGEL-Spezial, 1*, 118-125.

Neidhardt, F. (1991). Die Ranking-Debatte: Evaluationsversuche im Lehrbereich der Hochschulen. In W.-D. Webler & H.-U. Otto (Hrsg.), *Der Ort der Lehre in der Hochschule. Lehrleistungen, Prestige und Hochschulwettbewerb* (S. 283-294). Weinheim: Deutscher Studienverlag.

Ory, J. C. (1990). Student ratings of instruction: Ehtics and practice. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning No. 43, pp. 63-74). San Fransisco, CA: Jossey-Bass Publishers.

Preißer, R. (1992). Schlußfolgerungen aus den bisherigen Erfahrungen mit Lehrveranstaltungskritikern für eine Evaluation der Lehre. In D. Grühn & H. Gattwinkel (Hrsg.), *Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 197-217). Berlin: FU-Dokumentationsreihe.

Preißer, R. (1993). Evaluation der Lehre als symbolische Politik? *Beiträge zur Hochschulforschung, 4*, 495-524.

Professionalisierung individueller Lehre (ProfiLe), Universität Mannheim. (2000, Mai). *Projekt Evaluation der Lehre. Modulares Instrumentarium zur Lehrevaluation. Version II. Manual* [On-line]. Universität Mannheim: http://www.uni-mannheim.de/profile.profile_evaluation_modulseite.html.

Reischmann, J. (1995). *Bamberger Seminar-Evaluierungsbögen BEVA. Beiheft, Kopiervorlagen. Auswertungsprogramm.* Bamberg: hektoraphiertes Manuskript.

Reissert, R. (1992). Evaluation der Lehre - Aktuelle Aktivitäten an deutschen Hochschulen. *Hochschul Informations System* (HIS).

Reissert, R. (1994). Evaluation der Lehre - interne Selbstevaluation und externe Begutachtung durch Peers. *Hochschul Informations System* (HIS).

Reissert, R. & Konnerth, T. (2001). Evaluation von Studium und Lehre - Ein wirksames Instrument zur Qualitätsverbesserung? In U. Engel (Ed.), *Hochschul-Ranking. Zur Qualitätssicherung von Studium und Lehre* (pp. 177-194). Frankfurt a. M.: Campus.

Remmers, H. H. (1928). The relationship between students' marks and student attitudes towards instructors. *School and Society, 28*, 759-760.

Remmers, H. H. (1930). To what extent do grades influence student ratings of instructors? *Journal of Educational Research, 21*, 314-316.

Remmers, H. H. (1931). The equivalence of judgements to test items in the sense of the Spearman-Brown formula. *Journal of Educational Psychology*, *22*, 66-71.

Remmers, H. H. (1934). Reliability and halo effect on high school and college students' judgements of their teachers. *Journal of Applied Psychology*, *18*, 619-630.

Remmers, H. H. & Brandenburg, G. C. (1927). Experimental data on the Purdue Rating Scale for instructors. *Educational and Administrational Supervision*, *13*, 519-527.

Richter, R. (1994). *Qualitätssorge in der Lehre. Leitfaden für die studentische Lehrevaluation.* Neuwied: Luchterhand.

Rindermann, H. (1996a). Qualitative Urteile über die Qualität der Lehre durch Studierende? *Zeitschrift für Pädagogische Psychologie*, *10 (3/4)*, 151-166.

Rindermann, H. (1996b). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen.* Landau: Empirische Pädagogik.

Rindermann, H. (1996c). Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf die Kritik an der Lehrevaluation. *Zeitschrift für Pädagogische Psychologie*, *10 (3/4)*, 129-145.

Rindermann, H. (1997a). Die studentische Beurteilung von Lehrveranstaltungen: Forschungsstand und Implikationen für den Einsatz von Lehrevaluation. In R. S. Jäger, G. Trost & R. H. Lehmann (Hrsg.), *Tests und Trends. Jahrbuch der Pädagogischen Diagnostik* (Bd. 11, S. 12-53). Weinheim: Beltz.

Rindermann, H. (1997b). Generalisierbarkeit studentischer Veranstaltungsbeurteilungen. Sind Lehrevaluationsresultate generalisierbar auf andere Veranstaltungen eines Dozenten oder auf inhaltsgleiche Parallelveranstaltungen verschiedener Dozenten? *Psychologie in Erziehung und Unterricht*, *44 (3)*, 216-234.

Rindermann, H. (1998a). Das Münchener multifaktorielle Modell der Lehrveranstaltungsqualität: Entwicklung, Begründung und Überprüfung. *Beiträge zur Hochschulforschung*, *3*, 189-224.

Rindermann, H. (1998b). Skalen der Lehrevaluation: Welche Aspekte sollen in universitären Lehrveranstaltungen beurteilt werden? In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation* (S. 295-316). Bonn: Deutscher Psychologen Verlag.

Rindermann, H. (1998c). Übereinstimmung und Divergenz bei der studentischen Beurteilung von Lehrveranstaltungen: Methoden zu ihrer Berechnung und Konsequenzen für die Lehrevaluation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 19 (2)*, 73-92.

Rindermann, H. (1999a). Bedingungs- und Effektvariablen in der Lehrevaluationsforschung: Konzeption und Prüfung des Münchener multifaktoriellen Modells der Lehrveranstaltungsqualität. *Unterrichtswissenschaft, 27 (4)*, 357-380.

Rindermann, H. (1999b). *Die studentische Beurteilung von Lehrveranstaltungen - Forschungsstand und Implikationen.* Graz: Eingeladener Gastvortrag auf dem Symposium Evaluierung an der Universität - zwischen Qualitätsmanagement und Selbstzweck, 2.12.99.

Rindermann, H. (2000a). Das Selbstobjektivierungsproblem im akademischen Milieu I. *Das Hochschulwesen, 48 (3)*, 74-82.

Rindermann, H. (2000b). Das Selbstobjektivierungsproblem im akademischen Milieu II. *Das Hochschulwesen, 48 (4)*, 117-123.

Rindermann, H. (2001). *Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen.* Landau: Verlag Empirische Pädagogik.

Rindermann, H. & Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation.* Heidelberg: Asanger.

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

Rosemann, B. (1999). Qualität der Lehre durch Befragung? Eine Frage der Perspektive. *Armis et litteris, 4/1999*, 37-52.

Rosemann, B. & Schweer, M. K. W. (1996a). Evaluation universitärer Lehre und der Wissenszuwachs bei den Studierenden. *Zeitschrift für Pädagogische Psychologie, 10 (3/4)*, 175-180.

Rosemann, B. & Schweer, M. K. W. (1996b). Sisyphos in der Hochschule - Von der Fiktion, es allen recht machen zu können. In G. Brinek & A. Schirlbauer (Hrsg.), *Vom Sinn und Unsinn der Hochschuldidaktik* (S. 77-99). Wien: Wiener-Universitäts-Verlag.

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation. A systematic approach* (5. Aufl.). Newbury Park: Sage.

Schick, M. (1992). Nur ein Ablenkungsmanöver: Zur Problematik von Lehrevaluationen. *Mitteilungen des Hoschulverbandes*, *40*, 368-369.

Schweer, M. (2001). Evaluation der Lehre. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 159-164). Weinheim: Beltz. Psychologie Verlags Union.

Schweer, M. K. W. & Rosemann, B. (1995). Qualität der Lehre: Bedingungsvariablen des studentischen Urteils. *Zeitschrift für Pädagogische Psychologie*, *9 (3/4)*, 189-196.

Scriven, M. (1980). *The logic of evaluation.* Inverness: Edgepress.

Seldin, P. (1984). *Changing practice in faculty evaluation. A critical assessment and recommendations for improvement.* San Francisco, CA: Jossey-Bass Publishers.

Seldin, P. (Ed.) (1995). *Improving college teaching.* Bolton, MA: Anker Publishing.

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

Spiel, C. (2001). Der differentielle Einfluß von Biasvariablen auf studentische Lehrveranstaltungsbewertung. In U. Engel (Hrsg.), *Hochschul-Ranking. Zur Qualitätssicherung von Studium und Lehre* (S. 61-82). Frankfurt a. M.: Campus.

Spiel, C. & Gössler, P. M. (2000). Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. *Zeitschrift für Pädagogische Psychologie*, *14 (1)*, 38-47.

Spitzer, D. R. (1976). The importance of faculty attitudes in the planning for instructional development. *Research in Higher Education*, *5*, 97-111.

Sprung, L. & Sprung, H. (2000). Methodenlehre der Psychologie: System und Geschichte. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *21 (1)*, 31-48.

Stangl, W. (2000). *Die Evaluation universitärer Lehrveranstaltungen. Version 2.6* [On-line]. available: http://paedpsych.jk.uni-linz.ac.at/PAEDPSYCH/EVALUATION.

Staufenbiel, Th. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica*, *46 (4)*, 169-181.

Süllwold, F. (1992). Universitäre Lehre: Welche Realität wird bei der Beurteilung von Hochschullehrern durch Studierende erfaßt? *Mitteilungen des Hochschulverbandes*, *40 (1)*, 34-35.

Theall, M. & Franklin, J. (1990a). Student ratings in the context of complex evaluation systems. In M. Theall & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (Bd. New directions for teaching and learning, No. 43, pp. 17-34). San Fransisco, CA: Jossey-Bass Publishers.

Theall, M. & Franklin, J. (Eds.) (1990b). Student ratings of instruction: Issues for improving practice. In *New directions for teaching and learning* (Bd. No. 43). San Fransisco, CA: Jossey-Bass Publishers.

Thierau-Brunner, H., Stangel-Meseke, M. & Wottawa, H. (1998). Evaluation von Personalentwicklungsmaßnahmen. In K. Sonntag (Hrsg.), *Personalentwicklung in Organisationen* (2. Aufl., S. 261-286). Göttingen: Hogrefe.

Todt, E. & Götz, Ch. (1997). *Veranstaltungsrückmeldung (VR). Programm zur Evaluation von Veranstaltungen an der Universität.* Gießen: Fachbereich 06 Psychologie.

Webler, W.-D. (1992). Qualität der Lehre - Zwischenbilanz einer unübersichtlichen Entwicklung. *Das Hochschulwesen*, *40 (4)*, 153-161, 169-176.

Webler, W.-D. (1993). Evaluation der Lehre: Praxiserfahrungen und Methodenhinweise. *Beiträge zur Hochschulforschung*, *4*, 407-428.

Webler, W.-D. (1996). Qualitätssicherung in Lehre und Studium an deutschen Hochschulen. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 16 (2),* 119-148.

Webler, W.-D. (1999). Evaluation und dauerhafte Qualitätssicherung - eine Alternative zum Peer Review. *Das Hochschulwesen, 5/99,* 149-153.

Webler, W.-D. & Otto, H.-U. (1991). Der Ort der Lehre in der Hochschule. Lehrleistungen, Prestige und Hochschulwettbewerb. In *Blickpunkt Hochschuldidaktik* (Bd. 90). Weinheim: Deutscher Studien Verlag.

Weiss, R. (1991). Ziele und Probleme einer Lehrveranstaltungskritik. *Zeitschrift für Hochschuldidaktik, 15 (1-2),* 35-42.

Westermann, R., Spies, K., Heise, E. & Wollburg-Claar, S. (1998). Bewertung von Lehrveranstaltungen und Studienbedingungen durch Studierende: Theorieorientierte Entwicklung von Fragebögen. *Empirische Pädagogik, 12 (2),* 133-166.

Whitley, J. S. (1984). Are student evaluations constructive criticism? *Community and College Journal, 54 (7),* 41-42.

Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and cnsultants. *Journal of Higher Education, 57 (March/April),* 196-211.

Winter, M. (2000). Evaluation und Qualitätsentwicklung von Studium und Lehre. *Das Hochschulwesen, 48 (6),* 185-191.

Wottawa, H. (1993). *Psychologische Methodenlehre. Eine orientierende Einführung* (2. Aufl.). Weinheim: Juventa.

Wottawa, H. (2001). Evaluation. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 152-158). Weinheim: Beltz. Psychologie Verlags Union.

Wottawa, H. & Thierau, H. (1998). *Lehrbuch Evaluation* (2. Aufl.). Bern: Hans Huber.

Anhang

Anhang A: Liste verfügbarer amerikanischer Verfahren

Aleamoni Course/Instructor Evaluation Questionnaire (CIEQ)
Comprehensive Data Evaluation Services, Tucson

Instructor and Course Evaluation (ICE)
Instructional Evaluation Office, Southern Illinois University at Carbondale

Students' Evaluation of Educational Quality (SEEQ)
University of Western Sydney

Student Instructional Report (SIR)
Educational Testing System, Princeton

Student Instructional Ratings System (SIRS)
Scoring Office, Computer Center, Michigan State University

Instructional Assessment System (IAS)
Educational Assessment Center, Office of Educational Assessment, University of Washington (Seattle)

Instructional Development and Effectiveness Assessment System (IDEA)
Center for Faculty Evaluation and Development
Kansas State University

Purdue Cafeteria System (Cafeteria)
Center for Instructional Services, Purdue University

Instructor and Course Evaluation Form (ICES)
Division of Measurement and Research, University of Illinois (Urbana)

Instructor Designed Questionnaire (IDQ)
Center for Research on Learning and Teaching, University of Michigan (Ann Arbor)

Student Perceptions of Teaching (SPOT)
Evaluation and Examination Service, University of Iowa

Student Perceptions of Teaching Effectiveness (SPTE)
The Social Science Research Laboratory, Wichita State University

Anhang B: Liste verfügbarer deutschsprachiger Verfahren

Bamberger Seminar-Evaluierungsbögen (BEVA)

Reischmann, J. (1995). *Bamberger Seminar-Evaluierungsbögen BEVA. Beiheft, Kopiervorlagen. Auswertungsprogramm.* Bamberg: hektoraphiertes Manuskript.

Fragebogen zur Beurteilung einer Lehrveranstaltung (FB-LV) und Fragebogen zur Beurteilung von Studienbedingungen (FB-ST)

Westermann, R., Spies, K., Heise, E. & Wollburg-Claar, S. (1998). Bewertung von Lehrveranstaltungen und Studienbedingungen durch Studierende: Theorieorientierte Entwicklung von Fragebögen. *Empirische Pädagogik, 12 (2)*, 133-166.

FEVOR, FEREF, VEPRA

Staufenbiel, Th. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica, 46 (4)*, 169-181.

Fragebogen zur Evaluation des Lehr- und Lernverhaltens (FELL)

Moosbrugger, H., Hartig, J. & Naumann, J. (1997). Fragebogen zur Evaluation des Lehr- und Lernverhalten (FELL). In *Arbeiten aus dem Institut für Psychologie* (Vol. 3/1997). Frankfurt a. M.: Institut für Psychologie der Universität Frankfurt.

Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE)

Rindermann, H. & Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation.* Heidelberg: Asanger.

Modulares Instrumentarium zur Lehrevaluation

Professionalisierung individueller Lehre (ProfiLe), Universität Mannheim. (2000, Mai). *Projekt Evaluation der Lehre. Modulares Instrumentarium zur Lehrevaluation. Version II. Manual* [On-line]. Universität Mannheim: Available online:

http://www.uni-mannheim.de/profile.profile_evaluation_modulseite.html

Marburger Fragebogen zur Akzeptant der Lehre

Basler, H.-D., Bolm, G., Dickescheid, T. & Herda, C. (1995). Marburger Fragebogen zur Akzeptanz der Lehre. *Diagnostica, 41 (1)*, 62-79.

Studentische Bewertng von Lehrveranstaltungen Itempool

Winter, M. (2000, 2. Februar). *Itempool. Studentische Bewertung von Lehrveranstaltungen* [On-line]. available online:

http://www.verwaltung.uni-halle.de/prorstu/eval/itempool.tex.

VBVOR 5.0 und VBREF 5.0

Diehl, J. M. (1998). *Fragebögen zur studentischen Evaluation von Hochschulveranstaltungen (Version 5.0).* Gießen: Fachbereich 06 Psychologie (Abteilung Methodik).

Diehl, J. M. & Staufenbiel, T. (1997). *Evaluation von Lehre. Normen für VBVOR und VBREF*. Gießen: Fachbereich 06 Psychologie (Abteilung Methodik).

Diehl, J. M. & Reuschling, H. (1998). *VBOR 5.0, VBREF 5.0 - Dokumentation und Auswertungsprogramme* [3,5" Diskette]. Justus-Liebig-Unversität Gießen: Fachbereich 06 Psychologie.

VBPSYCH

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

Veranstaltungs-Rückmeldung (VR)

Todt, E. & Götz, Ch. (1997). *Veranstaltungs-Rückmeldung (VR). Programm zur Evaluation von Veranstaltungen an der Universität.* Gießen: Fachbereich 06 Psychologie.

Aus Österreich

Lehrveranstaltungsbegleitende Evaluation (LBVE, online)

Stangl, W. (2000). *Die Evaluation universitärer Lehrveranstaltungen. Version 2.6* [On-line]. Universität Linz, Österreich: available: http://paedpsych.jk.uni-linz.ac.at/PAEDPSYCH/EVALUATION.

Lehrveranstaltungsrückmeldung für die Studierenden (Juni 2000, Testphase)

Patry, O. (2000, Juni). *Lehrveranstaltungs-Rückmeldung für die Studierenden* [Online]. Universität Salzburg, Österreich: available online: http://www.sbg.ac.at/erz/evaluation/uebersicht.html.

Anhang C: Instruktion Kategoriensystem

Im Folgenden finden Sie ein **Kategoriensystem**, bestehend aus den Bezeichnungen für 30 verschiedene Kategorien. Die Kategorien beziehen sich auf Aspekte des Lehrgeschehens an Hochschulen. Sie dienen dazu, Aussagen aus dem Bereich der Hochschullehre zu ordnen. Zu jeder Kategorie finden Sie daher einige beispielhafte Aussagen, die für die jeweilige Kategorie typisch sind.

Bitte machen Sie sich nun mit dem Kategoriensystem vertraut, indem Sie die Bezeichnungen der Kategorien und alle dazugehörigen beispielhaften Aussagen sorgfältig lesen.

Nehmen Sie sich so viel Zeit, wie Sie möchten

Anhang D: Kategoriensystem (englische Fassung)

Die 28 Kategorien und Ankerbeispiele von Feldman (1989)

	Bezeichnung der Kategorie	Ankerbeispiele
1.	Teacher's **Stimulation of Interest** in the Course and its Subject Matter	„The teacher stimulated intellectual curiosity", "It was easy to remain attentive"
2.	Teacher's **Enthusiasm** (for Subject or for Teaching)	The instructor shows energy and excitement", "The instructor seems to enjoy teaching"
3.	Instructor's **Knowledge** of Subject Matter	"The instructor has a good knowledge about or beyond the textbook", "The Teacher keeps the lecture material updated"
4.	Teacher's **Intellectual Expansiveness** (and Intelligence)	"The teacher is well informed in all related fields", "The teacher has respect for other subject areas and indicates their relationship to his or her own subject of presentation"
5.	Teacher's **Preparation** (Organization of the Course)	"The presentation of the material is well organized", "The instructor was well prepared for each day's lecture"
6.	**Clarity** and **Understandableness**	"The instructor gave clear explanation", "The teacher effectively synthesizes and summarizes the material"
7.	Instructor's **Elocutionary Skills**	"The teacher has the ability to speak distinctly and be clearly heard", "The teacher changed pitch, volume or quality of speech"
8.	Teacher's **Sensitivity** to, and Concern with, Class Level and Progress	"The teacher takes an active personal interest in the progress of the class and show a desire for students to learn" "The teacher was skilled in observing student reactions"
9.	**Clarity of Course Objectives** and Requirements	"The instructor gave a clear idea of the student requirements", "The teacher clearly defined student responsibilities in the course"
10.	Nature and Value of the **Course Material** (Including Usefulness and Relevance)	"The teacher has the ability to apply material to real life", "There is worthwhile and informative material in lectures that doesn't duplicate the text"
11.	Nature and Usefulness of **Supplementary Materials** and Teaching Aids	"The instructor provided a variety of activities in class and used a variety of media (slides, films, projections, drawings) and outside resource persons", "The homework assignments and supplementary readings were helpful in understanding the course"
12.	Perceived **Outcome** or Impact of Instruction	"The instructor has given me tools for attacking problems", "The course has

	increased my general knowledge"
13. Instructor's **Fairness**; Impartiality of Evaluation of Students; Quality of Examinations	"Test questions were clear", "Grading on the course was fair"
14. **Personality** Characteristics („Personality„) of the Teacher	"The teacher has a good sense of humor", "The instructor was not autocratic and does not try to force us to accept his ideas and interpretations"
15. Nature, Quality, and Frequency **of Feedback** from the Teachers to Students	"The teacher told the students when they had done a good job", "The teacher is prompt in returning tests and assignments"
16. Teacher's **Encouragement of Questions** and Discussion, and **Openness to Opinions** of others	"Students felt free to ask questions and express opinions", "The instructor invited criticisms of his or her own ideas"
17. **Intellectual Challenge** and Encouragement of **Independent Thought** (by the Teacher and the Course)	"This course challenged students intellectually" "The teacher attempts to stimulate creativity"
18. Teacher's Concern and **Respect** for Students; **Friendliness** of the Teacher	"The teacher took the students seriously", "The teacher was friendly towards all students"
19. Teacher's **Availability** and Helpfulness	"The instructor is willing to help students having difficulty" "The teacher was accessible to students outside of class"
20. Teacher **Motivates** Students **to Do Their Best**: High Standard of Performance Required	"The instructor motivates students to do their best work", "The instructor sets high standards of achievement for students"
21. Teacher's **Encouragement of Self-Initiated Learning**	"Students are encouraged to work independently", "Students assume much responsibility for their learning"
22. **Teacher's Productivity in Research**-Related Activities	"The instructor publishes material related to his subject field", "The teacher talks about his own research"
23. **Difficulty** of the Course (and Workload) – **Description**	"The workload and pace of the course was difficult" "I spent many hours studying for this course"
24. **Difficulty** of the Course (and Workload) – **Evaluation**	"The instructor often asked for more than students could get done", "The teacher's lectures and oral presentations are 'over my head' "
25. **Classroom Management**	"The instructor controls class discussion to prevent rambling and confusion", "Students are allowed to participate in deciding the course content"
26. Pleasantness of **Classroom**	"I felt comfortable in this class", "The class

Atmosphere	does not make me nervous", "The teacher is always criticizing and arguing with students"
27. **Individualization of Teaching**	"Instead of expecting every student to do the same thing, the instructor provides different activities for different students", "My grade depends primarily on my improvement over my past performance"
28. Instructor Pursued and/or **Met Course Objectives**	"There was a close agreement between the announced objectives of the course and what was actually taught"
29. **Overall Instructor***	"Rate the overall teacher's effectiveness", "How would you rate your instructor with respect to general (all-around) teaching abilities?"
30. **Overall Course***	"How would you rate the overall value of this course?", "Have you enjoyed taking this course?"

* Diese beiden Kategorien wurden in Anlehnung an Abrami und d'Apollonia (1990) ergänzend hinzugefügt.

Anhang E: Kategoriensystem (deutsche Fassung)

Die 28 Kategorien und Ankerbeispiele von Feldman (1989, eigene Übersetzung)

	Bezeichnung der Kategorie	Ankerbeispiele
1.	**Anregung** des Interesses für das Thema der Veranstaltung durch den Dozenten	„Der Dozent regte zu intellektueller Neugier an", „Es fiel leicht, aufmerksam zu bleiben"
2.	**Begeisterungsfähigkeit** des Dozenten (für das Thema oder im Lehrverhalten)	„Der Dozent zeigte Aktivität und Freude", „Der Dozent scheint gern zu lehren"
3.	**Fachwissen** des Dozenten	„Der Dozent hat fundiertes Wissen bzgl. der Begleitungsliteratur und darüber hinaus" "Der Dozent hält sein Vorlesungsmaterial aktuell / auf aktuellem Stand"
4.	Kognitive **Aufgeschlossenheit** des Dozenten gegenüber anderen Feldern	„Der Dozent ist gut über angrenzende Bereiche informiert", „Der Dozent respektiert andere Themenbereiche und weist auf deren Bezüge zur eigenen Veranstaltung hin"
5.	**Vorbereitung** des Dozenten (Organisation der Veranstaltung)	„Die Präsentation des Lehrstoffs war gut organisiert", „Der Dozent war auf seine Veranstaltung stets gut vorbereitet"
6.	Klarheit und Verständlichkeit	„Der Dozent gab verständliche Erklärungen", „Der Dozent fasste das Material effektiv zusammen"
7.	Vortragsfähigkeiten des Dozenten	„Der Dozent hat die Fähigkeit, deutlich und gut hörbar zu sprechen", „Der Dozent variierte Tonlage und Lautstärke seiner Sprechweise"
8.	Die **Sensibilität** des Dozenten und seine empfundene Verantwortung/sein Verantwortungsgefühl **für** das **Leistungsniveau** und den **Fortschritt** in der Veranstaltung	„Der Dozent interessierte sich für den Fortschritt der Veranstaltungsteilnehmer und dafür, dass die Studierenden etwas lernen", „Der Dozent hatte die Fähigkeit, die Reaktionen der Studierenden zu beachten"
9.	Klarheit der Veranstaltungsziele und –anforderungen	„Der Dozent vermittelte eine klare Vorstellung von den Anforderungen, die er an die Studierenden stellt", „Der Dozent definierte die Verantwortlichkeiten der Studierenden für diese Veranstaltung klar"
10.	Art und Wert des **Stoffes** der Veranstaltung (beinhaltet auch Nützlichkeit und Relevanz)	„Der Dozent hat die Fähigkeit, den Stoff auf das 'wirkliche Leben' zu übertragen", „Der Stoff der Veranstaltung war wertvoll und informativ und nicht eine Wiederholung der Begleitliteratur"
11.	Art und Nützlichkeit des **zusätzlichen**	„Der Dozent gestaltete die Veranstaltung mit

	Materials und der Veranstaltungshilfen	verschiedenen Aktivitäten und benutzte vielfältige Medien (Folien, Filme, Diaprojektionen, Zeichnungen)“, „Die Hausaufgaben und die zusätzliche Lektüre waren für das Verständnis der Veranstaltung hilfreich“
12.	Wahrgenommenes Ergebnis oder Wirkung der Veranstaltung	„Der Dozent gab mir Hilfen an die Hand, um Probleme zu lösen“, „Durch die Veranstaltung hat sich mein Allgemeinwissen verbessert“
13.	**Fairness, Unvoreingenommenheit** bei der **Bewertung** durch den Dozenten; Qualität der Leistungsüberprüfung	„Die Testfragen waren klar zu verstehen“, “Die Bewertung der Studierenden durch den Dozenten in diesem Kurs war fair“
14.	**Persönliche Charakteristika** des Dozenten („Persönlichkeit“)	„Der Dozent hat einen guten Sinn für Humor“, „Der Dozent war nicht autoritär und versuchte nicht, uns seine Ideen und Interpretationen aufzuzwängen“
15.	Art, Qualität und Häufigkeit des **Feedback**s an die Studierenden durch den Dozenten	„Der Dozent sagte den Studierenden, wenn sie etwas gut gemacht hatten“, „Der Dozent gab Tests und eingereichte Aufgaben zügig zurück“
16.	**Ermutigung** des Dozenten, **Fragen** zu stellen, zu **diskutieren**; **seine Offenheit** gegenüber der Meinung anderer	„Die Studierenden fühlten sich frei, Fragen zu stellen und ihre Meinung zu äußern“, „Der Dozent forderte zur Kritik seiner eigenen Ideen auf“
17.	**Intellektuelle Herausforderung** und Ermutigung zu **unabhängigem Denken** (durch den Dozenten und die Veranstaltung)	“Diese Veranstaltung forderte die Studierenden intellektuell heraus“, Der Dozent versucht, die Kreativität zu stimulieren“
18.	Empfundene Verantwortung und **Respekt** gegenüber Studierenden; **Freundlichkeit** des Dozenten	„Der Dozent nahm die Studierenden ernst“, „Der Dozent verhielt sich gegenüber allen Studierenden freundlich“
19.	Erreichbarkeit und Hilfsbereitschaft des Dozenten	„Der Dozent war für die Studierenden auch außerhalb der Veranstaltung erreichbar“, „Der Dozent war bereit, Studierenden bei Problemen zu helfen“
20.	Motivation durch den Dozenten, sein Bestes zu geben; hohe Leistungsanforderungen	„Der Dozent motivierte die Studierenden, ihr Bestes zu geben“, „Der Dozent setzte hohe Leistungsstandards“
21.	Ermutigung zu **selbstinitiiertem Lernen** durch den Dozenten	„Die Studierenden wurden ermutigt, unabhängig zu arbeiten“, Die Studierenden übernehmen viel Verantwortung für ihre Lernfortschritte“
22.	Produktivität des Dozenten hinsichtlich seiner **forschungsbezogenen Aktivitäten**	„Der Dozent publiziert auf seinem Forschungsfeld“, „Der Dozent berichtet von seiner Forschungsarbeit“
23.	**Schwierigkeit** der Veranstaltung (und	„Der Arbeitsaufwand und das Lerntempo der

Aufwand) – **beschreibend**	Veranstaltung waren schwierig“, „Ich habe viele Stunden damit verbracht, für diese Veranstaltung zu lernen“
24. **Schwierigkeit** der Veranstaltung (und Aufwand) – **bewertend**	„Der Dozent erwartete von den Studierenden mehr als sie im Stande zu leisten sind“, „Die Vorlesung und Präsentationen des Dozenten waren zu abgehoben“
25. Veranstaltungsmanagement	„Der Dozent kontrollierte die Diskussionen, um Abschweifungen und Verwirrungen zu vermeiden“
26. Veranstaltungsklima	„Ich fühlte mich wohl in der Veranstaltung“, „Die Veranstaltung macht mich nicht nervös“, „Der Dozent kritisiert ständig und diskutiert mit den Studierenden“
27. **Anpassung der Lehre** an die individuellen Bedürfnisse der Studierenden	„Anstatt zu verlangen, dass alle Studierenden das Gleiche tun, bot der Dozent verschiedenen Aktivitäten für verschiedene Studierende an“, „Meine Bewertung hängt vorwiegend von der Verbesserung zu meinen bisherigen Leistungen ab.“
28. Der Dozent **erreichte die Ziele** der Veranstaltung	„Die angekündigten Veranstaltungsziele stimmten überein mit dem, was tatsächlich gelehrt wurde“
29. Gesamturteil Dozent*	„Bewerten Sie die Effektivität des Dozenten insgesamt“, “Wie würden Sie Ihren Dozenten hinsichtlich seiner allgemeinen Lehrkompetenz insgesamt bewerten? ”
30. Gesamturteil Veranstaltung*	„Wie würden Sie den Wert der Veranstaltung insgesamt bewerten? “Hat Ihnen die Veranstaltung gefallen?“

* Diese beiden Kategorien wurden in Anlehnung an Abrami und d’Apollonia (1990) ergänzend hinzugefügt.

Anhang F: Instruktion Übungsbogen / Kodierbogen

Sie haben sich bereits mit dem **Kategoriensystem** bestehend aus 30 Kategorien vertraut gemacht.

Im folgenden finden Sie eine Reihe von Aussagen, die sich auf Lehrveranstaltungen an Hochschulen beziehen. Gehen Sie bitte alle Aussagen durch und entscheiden Sie jeweils, zu welcher Kategorie die vorliegende Aussage ihrer Meinung nach **am besten passt**. Es gibt also keine falschen und richten Zuordnungen.

Tragen Sie bitte in das leere Kästchen hinter jeder Aussage die Nummer der Kategorie ein, zu der Sie die Aussage zuordnen würden (1-30). Sollten Sie der Meinung sein, dass eine Aussage keinesfalls zu einer der vorliegenden 30 Kategorien passt, so tragen Sie bitte die Nummer 00 ein.

Für die Bearbeitung haben Sie so viel Zeit wie Sie möchten.

Vielen Dank für Ihre Mitarbeit!

Beispiel

Sie lesen:

Der Dozent begegnete Studierenden mit Achtung.

Wenn Sie der Meinung sind, dass die Aussage zur **Kategorie 18: Respekt gegenüber Studierenden** passt, dann tragen Sie die Nummer **18** in das leere Kästchen ein.

Der Dozent begegnete Studierenden mit Achtung.

Wenn Sie der Meinung sind, dass die Aussage zur **keiner** der Kategorien passt, dann tragen Sie die Nummer 00 in das leere Kästchen.

Der Dozent begegnete Studierenden mit Achtung.

Anhang G: Übungsbogen (englische Fassung)

	Aussage	*Zuordnung*
1.	The instructor's objectives for the course have been made clear.	
2.	There was considerable agreement between the announced objectives of the course and what was actually taught.	
3.	The instructor used class time well.	
4.	The instructor was readily available for consultation with students.	
5.	The instructor seemed to know when students didn't understand the material.	
6.	Lectures were too repetitive of what was in the textbook(s).	
7.	The instructor encouraged students to think for themselves.	
8.	The instructor seemed genuinely concerned with students' progress and was actively helpful.	
9.	The instructor made helpful comments on papers or exams.	
10.	The instructor raised challenging questions or problems for discussion.	
11.	In this class I felt free to ask questions or express my opinions.	
12.	The instructor was well prepared for each class.	
13.	The instructor told students how they would be evaluated in the course.	
14.	The instructor summarized or emphasized major points in his lectures or discussions.	
15.	My interest in the subject area has been stimulated by this course.	
16.	The scope of the course has been too limited; not enough material has been covered.	
17.	Examinations reflected the important aspects of the course.	
18.	I have been putting a good deal of effort into this course.	
19.	The instructor was open to other viewpoints.	
20.	In my opinion, the instructor has accomplished (is accomplishing) his or her objectives for the course.	

Anhang H: Übungsbogen (deutsche Fassung)

	Aussage	*Zuordnung*
1.	Wortmeldungen von Teilnehmern wurden berücksichtigt.	
2.	Die Inhalte der Lehrveranstaltung wurden unter verschiedenen Aspekten erörtert.	
3.	Die Erklärungen des Lehrveranstaltungs-Leiters waren verständlich.	
4.	Die gestellten Aufgaben waren den Teilnehmern angepaßt.	
5.	Die Eigenverantwortlichkeit der Teilnehmer wurde gefördert.	
6.	Die Arbeitsatmosphäre wurde vom Lehrveranstaltungs-Leiter gefördert.	
7.	Die Arbeitsformen in der Lehrveranstaltung wurden vom Lehrveranstaltungs-Leiter variiert.	
8.	Audiovisuelle Hilfsmittel wurden vom Lehrveranstaltungs-Leiter benutzt.	
9.	Schwierige Sachverhalte wurden durch Beispiel erläutert.	
10.	Eine Kritik an den Inhalten der Lehrveranstaltung war möglich.	
11.	Auf Fragen der Teilnehmer wurde eingegangen.	
12.	Die Inhalte sind für das weitere Studium brauchbar.	
13.	Die Fragestellungen der Lehrveranstaltung waren schwierig.	
14.	Der Praxisbezug der Lehrveranstaltung wurde hergestellt.	
15.	Die Informationsmenge in der Lehrveranstaltung war angemessen.	
16.	Arbeitsunterlagen wurden ausreichend zur Verfügung gestellt.	
17.	Die Lehrveranstaltung hat das Interesse an den Inhalten geweckt.	
18.	Der Lehrveranstaltungs-Leiter hat sich für den Erfolg der Lehrveranstaltung engagiert.	
19.	Die Bedingungen für einen positiven Abschluß wurden offengelegt.	
20.	Der Raum und seine Ausstattung waren für Lehrveranstaltung geeignet.	

Anhang I: Kodierbogen (englische Fassung)

	Aussage	*Zuordnung*
1.	The Instructor displayed a personal interest in students and their learning.	
2.	The Instructor found ways to help students answer their own questions.	
3.	The instructor scheduled course work (class activities, tests, projects) in ways which encouraged students to stay up-to date in their work.	
4.	The Instructor demonstrated the importance and significance of the subject matter.	
5.	The instructor formed "teams" or "discussion groups" to facilitate learning.	
6.	The Instructor made it clear how each topic fit into the course.	
7.	The Instructor explained the reasons for criticisms of students' academic performance.	
8.	The Instructor stimulated students to intellectual effort beyond that required by most courses.	
9.	The Instructor encouraged students to use multiple resources (e.g. data banks, library holdings, outside experts) to improve understanding.	
10.	The Instructor explained course material clearly and concisely.	
11.	The Instructor related course material to real life situations.	
12.	The Instructor gave tests, projects, etc. that covered the most important points of the course.	
13.	The Instructor introduced stimulating ideas about the subject.	
14.	The Instructor involved students in "hands on" projects such as research, case studies, or "real life" activities.	
15.	The Instructor inspired students to set and achieve goals which really challenged them.	
16.	The Instructor asked students to share ideas and experiences with others whose backgrounds and viewpoints differ from their own.	
17.	The Instructor provided timely and frequent feedback on tests, reports, projects, etc. to help students improve.	
18.	The Instructor asked students to help each other understand ideas and concepts.	
19.	The Instructor gave projects, tests or assignments that required original or creative thinking.	
20.	The Instructor encouraged student-faculty interaction outside of class (office visits, phone calls, e-mail, etc.).	
21.	Progress on gaining factual knowledge (terminology, classifications, methods, trends).	

22.	In this course, my progress on learning fundamental principles, generalizations, or theories.	
23.	In this course, my progress on learning to apply course material (to improve rational thinking, problem-solving and decisions).	
24.	In this course, my progress on developing specific skills, competencies and points of view needed by professionals in the field most closely related to this course.	
25.	In this course, my progress on acquiring skill in working with others as a member of a team.	
26.	In this course, my progress on developing creative capacities (writing, inventing, designing, performing in art, music, drama etc.).	
27.	In this course, my progress on gaining a broader understanding and appreciation of intellectual-cultural activity (music, science, literature, etc.).	
28.	In this course, my progress on developing skill in expressing myself orally or in writing.	
29.	In this course, my progress on learning how to find and use resources for answering questions or solving problems.	
30.	In this course, my progress on developing a clearer understanding of, and commitment to, personal values.	
31.	In this course, my progress on learning to analyze and critically evaluate ideas, arguments, and points of view.	
32.	In this course, my progress on acquiring an interest in learning more by asking my own questions and seeking answers.	
33.	The Course: Amount of reading.	
34.	The Course: Amount of work in other (non-reading) assignments.	
35.	The Course: Difficulty of subject matter.	
36.	I had a strong desire to take this course.	
37.	I worked harder on this course than on most courses I have taken.	
38.	I really wanted to take a course from this instructor.	
39.	I really wanted to take this course regardless of who taught it.	
40.	As a result of this course, I have more positive feelings toward this field of study.	
41.	Overall, I rate this instructor an excellent teacher.	
42.	Overall, I rate this course as excellent.	
43.	As a rule, I put forth more effort than other students on academic work.	
44.	The Instructor used a variety of methods - not only tests - to evaluate student progress on course objectives.	
45.	The Instructor expected students to take their share of responsibility for learning.	
46.	The Instructor had high achievement standards in this class.	

47.	The Instructor used educational technology (e.g., internet, e-mail, computer exercises, multi-media presentation, etc.) to promote learning.	
48.	The instructor's explanation of course requirements	
49.	The instructor's preparation for each class period	
50.	The instructor's command of the subject matter	
51.	The instructor's use of class time	
52.	The instructor's way of summarizing or emphasizing important points in class	
53.	The instructor's ability to make clear and understandable presentations	
54.	The instructor's command of spoken English (or the language used in the course)	
55.	The instructor's use of examples or illustrations to clarify course material	
56.	The instructor's use of challenging questions or problems	
57.	The instructor's enthusiasm for the course material	
58.	The instructor's helpfulness and responsiveness to students	
59.	The instructor's respect for students	
60.	The instructor's concern for students progress	
61.	The availability of extra help for this class (taking into account the size of the class)	
62.	The instructor's willingness to listen to students questions and opinions	
63.	The information given to students about how they would be graded	
64.	The clarity of exam questions	
65.	The exam's coverage of important aspects of the course	
66.	The instructor's comments on assignments and exams	
67.	The overall quality of the textbook(s)	
68.	The helpfulness of assignments in understanding course material	
69.	Problems or questions presented by the instructor for small group discussions contributed to learning ...	
70.	Term paper(s) or project(s) contributed to learning ...	
71.	Laboratory exercises for understanding important course concepts contributed to learning ...	
72.	Assigned projects in which students worked together contributed to learning ...	
73.	Case studies, simulations, or role playing contributed to learning ...	
74.	Course journals or logs required of students contributed to learning ...	
75.	Instructor's use of computers as aids in instruction contributed to learning ...	

76.	My learning increased in this course	
77.	I made progress toward achieving course objectives	
78.	My interest in the subject areas has increased	
79.	This course helped me to think independently about the subject matter	
80.	This course actively involved me in what I was learning	
81.	I studied and put effort into the course	
82.	I was prepared for each class (writing and reading assignments)	
83.	I was challenged by this course	
84.	For my preparation and ability, the level of difficulty of this course was ...	
85.	The work load of this course in relation to other courses of equal credit was ...	
86.	For me, the pace at which the instructor covered the material during the term was ...	
87.	Rate the quality of instruction in this course as it contributed to your learning (try to set aside your feeling about the course content)	
88.	The instructor's enthusiasm when presenting material	
89.	The instructor's interest in teaching	
90.	The instructor's use of examples or personal experiences to help get points across in class .	
91.	The instructor's concern with whether the students learned the material.	
92.	Your interest in learning the course material.	
93.	Your general attentiveness in class.	
94.	The course as an intellectual challenge.	
95.	Improvement in your competence in this area due to this course.	
96.	The instructor's encouragement to students to express opinions.	
97.	The instructor's receptiveness to new ideas and others' viewpoints.	
98.	The students' opportunity to ask questions.	
99.	The instructor's stimulation of class discussion.	
100.	The appropriateness of the amount of material the instructor attempted to cover.	
101.	The appropriateness of the amount of pace at which the instructor attempted to cover the material.	
102.	The contribution of homework assignments to your understanding of the course material relative to the amount of time required.	
103.	The appropriateness of the difficulty of assigned reading topics.	
104.	The instructor's ability to relate the course concepts in a systematic manner.	
105.	The course organization.	
106.	The ease of taking notes on the instructor's presentation.	

107.	The adequacy of outlined direction of the course.	
108.	Your general enjoyment of the course.	
109.	You found the class intellectually challenging and stimulating.	
110.	You have learned something which you considered valuable.	
111.	Your interest in the subject has increased as a consequence of this class.	
112.	You have learned and understood the subject materials in this class.	
113.	Lecturer was enthusiastic about teaching the class.	
114.	Lecturer was dynamic and energetic in conducting the class.	
115.	Lecturer enhanced presentations with the use of humor.	
116.	Lecturer's style of presentation held your interest during class.	
117.	Lecturer's explanations were clear.	
118.	Class materials were well prepared and carefully explained.	
119.	Proposed objectives agreed with those actually taught so you knew where the class was going.	
120.	Lecturer gave presentations that facilitated taking notes.	
121.	Students were encouraged to participate in class discussion	
122.	Students were invited to share their ideas and knowledge.	
123.	Students were encouraged to ask questions and were given meaningful answers.	
124.	Students were encouraged to express their own ideas and/or question the lecturer.	
125.	Lecturer was friendly towards individual students.	
126.	Lecturer had a genuine interest in individual students.	
127.	Lecturer made students feel welcome in seeking help/advice in or outside of class	
128.	Lecturer was adequately accessible to students during office hours after class.	
129.	Lecturer contrasted the implications of various theories.	
130.	Lecturer presented the background or origin of ideas/concepts developed in class.	
131.	Lecturer presented points of view other than his/her own when appropriate.	
132.	Lecturer adequately discussed current developments in the field.	
133.	Feedback on examination/graded material was valuable.	
134.	Methods of evaluating student work were fair and appropriate	
135.	Examinations/graded materials tested class content as emphasised by the lecturer.	
136.	Required readings/texts were valuable.	
137.	Readings, homework, etc. contributed to appreciation and understanding of the subject.	

138.	Overall, how does this class compare with other classes at this institution?	
139.	Overall, how does the lecturer compare with other lecturers at this institution?	
140.	It was a very worthwhile course.	
141.	I would rather take another course that was taught this way.	
142.	The instructor seemed to be interested in students as individuals.	
143.	The course material was too difficult.	
144.	It was easy to remain attentive.	
145.	NOT much was gained by taking this course.	
146.	I would have preferred another method of teaching in this course.	
147.	The course material seemed worthwhile.	
148.	The instructor did NOT synthesize, integrate or summarize effectively.	
149.	The course was quite interesting.	
150.	The instructor encouraged development of new viewpoints and appreciations.	
151.	I learn more when other teaching methods are used.	
152.	Some things were NOT explained well.	
153.	The instructor demonstrated a thorough knowledge of the subject matter.	
154.	This was one of my poorest courses.	
155.	The course content was excellent.	
156.	Some days I was NOT very interested in this course.	
157.	I think that the course was taught quite well.	
158.	The course was quite boring.	
159.	The instructor seemed to consider teaching as a chore routine activity.	
160.	Overall, the course was good.	
161.	The instructor prepared for class.	
162.	The instructor made clear assignments.	
163.	The instructor set clear standards for grading.	
164.	The instructor graded fairly.	
165.	The instructor knew if students understood her/him.	
166.	The instructor spoke understandably.	
167.	The instructor answered impromptu questions satisfactorily.	
168.	The instructor showed an interest in the course.	
169.	The instructor accepted criticism and suggestions.	
170.	The instructor gave several examples to explain complex ideas.	
171.	The instructor increased your appreciation for the subject.	
172.	The instructor organized and presented subject matter well.	
173.	The instructor specified objectives of the course.	

174.	The instructor achieved the specified objectives of the course.	
175.	The instructor explained the subject clearly.	
176.	The instructor showed an interest in students.	
177.	The instructor was enthusiastic about the subject.	
178.	The instructor was available outside of class.	
179.	The instructor encouraged student participation.	
180.	The instructor in general, taught the class effectively.	
181.	This course was a good learning experience.	
182.	The content of this course was good.	
183.	The course was well organized.	
184	I had trouble paying attention in class.	
185.	There should be additional prerequisites.	
186.	There should be fewer prerequisites.	
187.	This course was very interesting.	
188.	The amount of required work was appropriate.	
189.	This course was one of the best I have taken.	
190.	The tests covered the course material well.	
191.	This course was a waste of time.	
192.	The textbook was good.	
193.	Audio-visuals could be used more effectively.	
194.	This course should be taught in some other way.	
195.	I covered this material in other courses.	
196.	The course material was too difficult.	
197.	This course should continue to be offered.	
198.	The reading assignments were hard to understand.	
199.	I was often confused.	
200.	Generally, the course was good.	
201.	In this course I made progress on gaining factual knowledge (terminology, classifications, methods, trends).	
202.	In this course I made progress on learning fundamental principles, generalizations, or theories.	
203.	In this course I made progress on learning to apply course material to improve rational thinking.	
204.	In this course I made progress on developing specific competencies needed by professionals in the field.	
205.	In this course I made progress on learning how professionals in this field gain knowledge.	
206.	In this course I made progress on developing creative capacities.	
207.	In this course I made progress on developing a sense of personal responsibility (self-reliance, self-discipline).	
208.	In this course I made progress on gaining a broader appreciation of intellectual-cultural activity.	

209.	In this course I made progress on developing skill in expressing myself orally or in writing.	
210.	In this course I made progress on developing the implications of the course material for understanding myself.	
211.	My reason for taking this course: strong interest in the material.	
212.	My reason for taking this course: a strong interest in the discipline.	
213.	My reason for taking this course: to obtain a good grade.	
214.	My reason for taking this course: to satisfy a requirement for my major.	
215.	My reason for taking this course: to fulfill requirements for electives.	
216.	Considering the nature of the subject, the entire course (was extremely well organized).	
217.	As an aid to learning, the number and difficulty of assignments were (excessive).	
218.	Considering the nature of the course in terms of subject and class size, the method of presentation of material , i.e., lecture, discussion groups, etc.) was most (appropriate).	
219.	The general objectives of the course were (clearly understood).	
220.	The degree to which the material covered in this course was interrelated and consistent with subject area was (very high).	
221.	The learning resources, including the text and all other required sources of content beyond the classroom presentation were (of great value).	
222.	The method of assigning grades was (clearly understood and consistent).	
223.	Considering other courses of similar credit and level, the work load for this course was very (heavy).	
224.	Considering the level of the course, class composition, prerequisites, etc., the level of the material presented was (very high).	
225.	Considering the nature of the course and subject material, the rate of coverage of the material was (too slow).	
226.	The number and type of evaluations, i.e., exams, assignments, papers, etc., used in determining the grade were (sufficient to reflect achievement).	
227.	The instructor's use of, or directions of students to, nonrequired references or resources was (appropriate).	
228.	The instructor's classroom presentation was (well prepared at all times).	
229.	Based upon your own experience the instructors attitude towards student was (respectful).	
230.	Judging only on the basis of your own experience, the instructor's knowledge of the subject matter of the course appeared to be (exceptional).	

231.	By raising challenging questions or problems for discussion, the instructor stimulated students to think for themselves (never).	
232.	As reflected in the classroom and in the presentation of course material, the instructor was (very enthusiastic).	
233.	With respect to your progress in this course, the instructor was (concerned and actively helpful).	
234.	In terms of voice level, rate of speaking, appearance, and mannerisms, the personal characteristics of the instructor were judged to be (conductive to learning).	
235.	The ability of the instructor in handling questions and answering them to the student's satisfaction was (quite satisfactory).	
236.	With respect to students' freedom to express opinions and ask questions in the classroom, the instructor (freely permitted comments).	
237.	The instructor helped the class achieve the objectives set forth in the course (very satisfactorily).	
238.	For the purpose of explanation and clarification outside of the formal class period, the instructor was (readily available).	
239.	The examination questions, or other evaluative methods used by this instructor, seemed (very clear and fair).	
240.	As a result of this course my interest in pursuing additional knowledge in this area has been (stimulated).	
241.	As a result of this course my knowledge level in the area has (greatly increased).	
242.	With respect to my ability and prior preparation, the level of difficulty of this course was (very high).	
243.	My feeling is that the instructor graded (very hard).	
244.	I have usually gone to the meetings of this class with (eager anticipation).	
245.	From my own experience, the instructor came across as a person as well as a teacher.	
246.	Compared with other courses I have taken, I feel my retention of material concepts, applications etc., in this course will be (high).	
247.	In terms of what I have gained from the course, the grade I expect to obtain is (a good reflection).	
248.	With respect to the goals of the course the amount of material presented was (insufficient).	
249.	I like the instructor very much.	
250.	In one way or another (whether in relationship to my major, other courses, or just life in general) this course has been (very useful and worthwhile).	
251.	On the basis of the subject material, would you recommend this course to another student (highly)?	
252.	In conveying the concepts of this course in a clear, meaningful, and appropriate way, the instructor's ability was (very evident).	

253.	On the basis of performance in this class, would you recommend this instructor to another student?	
254.	Considering all the instructors I have had, on the basis of general quality of instruction, I would rank this instructor as (among the best).	

Anhang J: Kodierbogen (deutsche Fassung)

	Aussage	*Zuordnung*
1.	Der inhaltliche Aufbau der Veranstaltung ist logisch/nachvollziehbar.	
2.	Die Veranstaltung ist gut organisiert.	
3.	Der Stoff wird anhand von Beispielen veranschaulicht.	
4.	Die Bedeutung/Relevanz/Nutzen der behandelten Themen wird nahegelegt.	
5.	Ein Bezug zwischen Theorie und Praxis wird hergestellt.	
6.	Ich werde zum Mitdenken angeregt.	
7.	Zur kritischen Auseinandersetzung mit den behandelten Themen wird angeregt.	
8.	Der Dozent kann Kompliziertes verständlich machen.	
9.	Der Dozent wirkt gut vorbereitet.	
10.	Der Dozent spricht anregend.	
11.	Der Dozent zeigt Engagement in seiner Lehrtätigkeit und versucht Begeisterung zu vermitteln.	
12.	Der Dozent nimmt die Lehre wichtig.	
13.	Dem Dozenten ist es wichtig, daß die Studierenden etwas in der Veranstaltung lernen können.	
14.	Der Dozent ist im Umgang mit den Studierenden freundlich und aufgeschlossen.	
15.	Der Dozent ist kooperativ.	
16.	Die Veranstaltung ist interessant.	
17.	Die Veranstaltung zieht sich schleppend dahin.	
18.	Die Stoffmenge kann ich noch verkraften.	
19.	Das Tempo der Veranstaltung ist zu schnell.	
20.	Ich verstehe alles.	
21.	Höhe der Anforderungen (angemessen).	
22.	Ich lerne viel in der Veranstaltung.	
23.	Ich lerne etwas Sinnvolles und Wichtiges.	
24.	Das Thema der Veranstaltung interessiert mich.	
25.	Die Referate der Studierenden sind interessant.	
26.	Die Referate sind strukturiert und verständlich.	
27.	Die Referate sind nützlich und wertvoll.	
28.	Die Referenten werden durch den Dozenten adäquat ergänzt.	
29.	Ich bereite mich auf die Veranstaltung vor oder bereite sie nach (z.B. durch Lesen der Literatur).	
30.	Mein Arbeitsaufwand für die Veranstaltung ist verglichen mit anderen Veranstaltungen hoch.	

31.	Ich beteilige mich mit Wortbeiträgen.	
32.	Beim Einbringen eigener Beiträge fühle ich mich frei und äußerungsfähig.	
33.	Es finden ausreichend Diskussionen statt.	
34.	Die Diskussionen in der Veranstaltung sind produktiv.	
35.	Der Besuch der Veranstaltung lohnt sich.	
36.	Die Veranstaltung fördert mein Interesse am Studienfach.	
37.	Wenn man alles in einer Note zusammenfassen könnte, würde ich der Veranstaltung die folgende Note geben: (Notenskala von 1-6).	
38.	Ich beherrsche jetzt die Inhalte dieser Veranstaltung.	
39.	Das Lernen fiel mir öfters schwer, weil die Darstellung wenig anschaulich war.	
40.	Die Veranstaltung hat mich in die Lage versetzt, mir den Stoff selbständig weiter zu erarbeiten.	
41.	Es war ein gemeinsames Bemühen spürbar, sich gegenseitig zu helfen.	
42.	Es fällt mir jedoch schwer, die Bedeutung des Gelernten für meine spätere Praxis zu sehen.	
43.	Alle TeilnehmerInnen wurden zur aktiven Mitarbeit motiviert.	
44.	Durch die Veranstaltung angeregt, habe ich mich über die Pflichtaufgaben hinaus noch zusätzlich mit dem Thema beschäftigt.	
45.	Ich kann jetzt endlich mehr TeilnehmerInnen mit ihrem Namen ansprechen als am Anfang.	
46.	Diese Veranstaltung hat mir geholfen, mich selbst besser zu verstehen.	
47.	Um mitzukommen, mußte man sich immer sehr konzentrieren.	
48.	Ich habe durch diese Veranstaltung mehr Spaß an meinem Studium bekommen.	
49.	Wir StudentInnen konnten an der Gestaltung der Veranstaltung mitwirken.	
50.	Für mich hat sich dieses Seminar gelohnt.	
51.	Das Seminar war so angelegt, daß erfolgreiches Lernen leicht fiel.	
52.	Ich habe hier (Veranstaltungs)Material, Bücher oder andere Hilfsmittel kennengelernt, mit denen ich mir bei Bedarf selbst weiter helfen kann.	
53.	In der Gruppe gibt es jetzt ein herzliches Gefühl der Zusammengehörigkeit.	
54.	Im durchgenommenen Lernstoff habe ich noch Lücken.	
55.	Der Veranstaltungsverlauf war interessant und lebendig gestaltet.	
56.	Diese Veranstaltung hat mein Selbstvertrauen gestärkt, daß ich erfolgreich lernen kann.	
57.	In den Sitzungen wurde immer wieder miteinander über die Veranstaltung (Ablauf, Inhalte, Bewertung) gesprochen.	

58.	Ich bekam genügend Kenntnisse vermittelt.	
59.	Die StudentInnen wurden ermutigt, eigene Fragen und Erfahrungen einzubringen.	
60.	Ich habe in dieser Veranstaltung Ratschläge erhalten, die mein Lernen in Zukunft leichter machen.	
61.	Das Klima in der Veranstaltung war eher unangenehm.	
62.	Ich fühle mich jetzt im behandelten Stoff sicher.	
63.	Der Veranstaltungsleiter konnte zuhören und auf die Argumente der StudentInnen eingehen.	
64.	Zu einer solchen Veranstaltung würde ich gerne wieder gehen.	
65.	Ich habe die anderen TeilnehmerInnen auch als Mensch kennengelernt.	
66.	Wenn am Ende der Veranstaltung eine Prüfung wäre, würde ich sicher eine gute Note bekommen.	
67.	Der Nutzen des Stoffs wurde immer wieder an Beispielen aus der Sicht der StudentInnen verdeutlicht.	
68.	Ich werde anderen unbedingt empfehlen, an dieser Veranstaltung teilzunehmen, wenn sie wiederholt wird.	
69.	Es gab genügend Anlässe, miteinander herzlich zu lachen.	
70.	Die Veranstaltung hat mir persönlich für mein Studium viel gebracht.	
71.	Manche Inhalte hätten besser erklärt werden müssen.	
72.	Durch diese Veranstaltung habe ich ein besseres Verständnis für die Probleme unserer Gesellschaft gewonnen.	
73.	Es fiel schwer, sich in den Sitzungen zu Wort zu melden.	
74.	Ich könnte die Inhalte der Veranstaltung jetzt auch jemand anderem beibringen.	
75.	Jede Sitzung war klar und übersichtlich gegliedert.	
76.	Der Umgang miteinander war immer partnerschaftlich.	
77.	Diese Veranstaltung hat meinen Horizont auch über das Fachliche hinaus erweitert.	
78.	Aktuelle Literatur und neue Forschungsergebnisse werden in die Veranstaltung integriert.	
79.	Der in der Veranstaltung vermittelte Stoff ist in der zur Verfügung stehenden Zeit gut zu bewältigen.	
80.	Der Veranstalter geht auf die Fragen der TeilnehmerInnen angemessen ein.	
81.	Die Bedeutung der Veranstaltungsinhalte im Kontext des gesamten Studienganges wird deutlich gemacht.	
82.	Die Veranstaltung vermittelt Zusammenhänge und nicht nur Einzelfakten.	
83.	In der Veranstaltung werden auch schwierige Inhalte verständlich erklärt.	

84.	In der Veranstaltung werden ausreichend Hilfsmittel zu Aneignung des Lehrstoffes (Skripts, Literaturlisten, u.ä.) zur Verfügung gestellt.	
85.	Die Veranstaltung weckt bzw. verstärkt das Interesse an dem behandelten Stoff.	
86.	Die Vermittlung des Lehrstoffes wird durch den Einsatz von Medien wie Tafel, Overheadfolien oder Video angemessen unterstützt.	
87.	Die TeilnehmerInnen erscheinen zu der Veranstaltung pünktlich.	
88.	Die TeilnehmerInnen zeigen Interesse an den Veranstaltungsinhalten.	
89.	Ein inhaltlicher "roter Faden" ist während des Verlaufs der Veranstaltung immer erkennbar.	
90.	Die Veranstaltung vermittelt Kenntnisse, die die eigenständige Weiterbeschäftigung mit dem behandelten Stoff ermöglichen.	
91.	In der Veranstaltung werden auch tiefergehende Kenntnisse über die behandelten Inhalte vermittelt.	
92.	Für die Veranstaltung stehen ausreichend technische Hilfsmittel (Tafel, Overheadprojektor, Video u.a.) zur Verfügung.	
93.	Die verwendeten Folien, Tafelbilder etc. sind gut verständlich.	
94.	Der Besuch der Veranstaltung führt zu einem spürbaren Wissenszuwachs.	
95.	Der in der Veranstaltung vermittelte Stoff ist gut strukturiert.	
96.	Eine selbständige und aktive Auseinandersetzung mit den Veranstaltungsinhalten wird in der Veranstaltung gefördert.	
97.	Für die Veranstaltung stehen geeignete Räumlichkeiten zur Verfügung.	
98.	In der Veranstaltung werden Bezüge der Veranstaltungsinhalte zur Berufspraxis hergestellt.	
99.	In der Veranstaltung wird den zentralen Themen im Vergleich zu nebensächlicheren Inhalten entsprechend ihres Stellenwertes Platz eingeräumt.	
100.	In der Veranstaltung wird ein guter Überblick über das behandelte Stoffgebiet vermittelt.	
101.	Ich wurde motiviert, mich weiterhin mit dem Stoff zu beschäftigen.	
102.	Der Unterrichtende ermutigte die Teilnehmer zu eigener Aktivität.	
103.	Es gelang dem Unterrichtenden, Zusammenhänge verständlich zu machen.	
104.	Der Lehrende konnte die Studierenden dazu motivieren, gemeinsam Problemstellungen zu erörtern.	
105.	Die Art, wie der Unterricht gestaltet wurde, hat wesentlich zu meinem Lernerfolg beigetragen.	
106.	Die Veranstaltung motivierte mich, fachlichen Austausch mit Kommilitonen zu führen.	

107.	Der Lehrinhalt wurde gut veranschaulicht (durch Beispiele, Dias, Folien, Tafel, Skripte, etc.).	
108.	Der Unterrichtende scheint viel Ahnung von seinem Fach zu haben.	
109.	Die Organisation des Unterrichts war gut geplant.	
110.	Der Unterrichtende hatten den Ablauf gut im Griff.	
111.	Der Unterrichtende förderte meine Bereitschaft zum fachlich übergreifenden Lernen.	
112.	Die Art der Unterrichtsgestaltung führte zu guten Lernerfolgen bei den Teilnehmern.	
113.	Die Veranstaltung hat mir für die spätere Berufspraxis viel gebracht.	
114.	Die fachliche Autorität des Unterrichtenden hat mich überzeugt.	
115.	Der Unterrichtende zeigte Interesse an den Wünschen der Teilnehmer.	
116.	Der Unterricht hat mich insgesamt weitergebracht.	
117.	Ich konnte einen Zusammenhang zwischen Thema und zukünftigen Beruf finden.	
118.	Man konnte dem Stoff der Veranstaltung leicht folgen.	
119.	Der Dozent war einer kritischen Haltung dem Inhalt seiner Veranstaltung gegenüber aufgeschlossen.	
120.	Das Verhalten des Dozenten hat die Veranstaltung aufgelockert.	
121.	Der Dozent war am Lernerfolg der Teilnehmer nicht sonderlich interessiert.	
122.	Der dargebotene Stoff war zu umfangreich.	
123.	Das Verhalten des Dozenten den Veranstaltungsteilnehmern gegenüber wirkte kühl und unpersönlich.	
124.	Einige Sachverhalte wurden nicht genügend gut erklärt.	
125.	Der Dozent hat das Stoffgebiet zu unkritisch dargestellt.	
126.	Die Veranstaltung war häufig verwirrend, weil keine Gliederung mehr zu erkennen war und einem so der Überblick verloren ging.	
127.	Der Dozent ging auf Verständnisschwierigkeiten bei den Teilnehmern nicht genügend ein.	
128.	Das in der Veranstaltung erworbene Wissen habe ich gut im übrigen Studium anwenden können.	
129.	Der Dozent schien Lehre als reine Pflichtübung und Routinetätigkeit zu betrachten.	
130.	Das Thema dieser Veranstaltung hatte mit Psychologie nicht viel zu tun.	
131.	In der Veranstaltung wurde zu viel Stoff behandelt.	
132.	Die in dieser Veranstaltung erworbenen Kenntnisse haben mir in anderen Veranstaltungen sehr geholfen.	
133.	Der Stoff wurde durch Beispiele gut verdeutlicht.	

134.	Das Thema dieser Veranstaltung halte ich im Rahmen des Psychologiestudiums für sehr sinnvoll.	
135.	Die Relevanz des Stoffes wurde durch genügend Beispiele verdeutlicht.	
136.	Der Dozent ging auf die Vorstellungen und Vorschläge der Studenten ein.	
137.	Der Dozent wirkte arrogant.	
138.	Es wurde zu schnell vorgegangen.	
139.	Die Veranstaltung machte mich gut mit psychologischen Fragestellungen vertraut.	
140.	Durch die Behandlung zu vieler Einzelheiten ging der Gesamtzusammenhang verloren.	
141.	Der Stoff war übersichtlich gegliedert.	
142.	Unter psychologischen Inhalten stelle ich mir etwas anderes vor.	
143.	Der Dozent wiederholte den Stoff nicht oft genug und auch nicht so, daß die Studenten erkennen konnten, worin ihre Verständnisschwierigkeiten bestanden.	
144.	Der Stoff der Veranstaltung war zu schwierig.	
145.	Der Dozent vermittelte den Stoff anschaulich und verständlich.	
146.	Die in der Veranstaltung erworbenen Kenntnisse kann man gut in der späteren Berufspraxis gebrauchen.	
147.	Der Dozent hat die Veranstaltung didaktisch gut aufgebaut und durchgeführt.	
148.	Dem Dozenten ging es offensichtlich nur um das Durchziehen des Stoffes	
149.	Der dargebotene Stoff kam einen oft ziemlich unwichtig vor.	
150.	Grundgedanken und Begriffe wurden vom Dozenten zu schnell entwickelt.	
151.	Der durchgenommene Stoff gewann durch die verwendeten Beispiele an Anschaulichkeit und Praxisnähe.	
152.	Der Dozent drückte sich meist verständlich aus.	
153.	Um der Veranschaulichung folgen zu können, war zu viel zusätzliche Arbeit notwendig.	
154.	Der Dozent berücksichtigte weitgehend die Interessen der Teilnehmer.	
155.	Die Zusammenhänge zwischen dem vermittelten Stoff und der späteren beruflichen Praxis wurden gut aufgewiesen.	
156.	Die Veranstaltung verlief nach einer klaren Gliederung.	
157.	Die Veranstaltung ermöglichte ein besseres Verstehen anderer Veranstaltungen.	
158.	Ich besuche diese Lehrveranstaltung um für meinen angestrebten Beruf Kenntnisse oder Fertigkeiten zu erwerben.	
159.	Ich besuche diese Lehrveranstaltung um Vorgehensweisen und Ergebnisse der Wissenschaft kennenzulernen.	

160.	Ich besuche diese Lehrveranstaltung um mich persönlich weiterzubilden.	
161.	Ich besuche diese Lehrveranstaltung um mir prüfungsrelevanten Stoff anzueignen.	
162.	Ich besuche diese Lehrveranstaltung um einen Schein zu erwerben.	
163.	Ich besuche diese Lehrveranstaltung um einen tieferen Einblick in die behandelte Thematik zu bekommen.	
164.	Ich besuche diese Lehrveranstaltung um ein Verständnis für Probleme und Zusammenhänge zu bekommen.	
165.	Insgesamt habe ich in dieser Lehrveranstaltung meine Ziele erreicht.	
166.	Ich habe in dieser Lehrveranstaltung viel gelernt.	
167.	Diese Lehrveranstaltung ist langweilig und einschläfernd.	
168.	Insgesamt bin ich mit der Lehrveranstaltung zufrieden.	
169.	Unabhängig von der Art der Vermittlung finde ich das Thema der Lehrveranstaltung interessant.	
170.	Der Lehrende erläutert schwierige Sachverhalte verständlich.	
171.	Der Lehrende ist gut vorbereitet.	
172.	Der Lehrende verfügt über ein gutes Fachwissen.	
173.	Der Lehrende faßt die behandelten Inhalte gut zusammen.	
174.	Der Lehrende setzt mögliche Darbietungshilfen (Tafel, Folien, Filme, Skripte etc.) zu wenig ein.	
175.	Der Lehrende gestaltet Tafelbild oder Folien leserlich und übersichtlich.	
176.	Der Lehrende benützt öfter Beispiele, die zum Verständnis der Lehrinhalte beitragen.	
177.	Der Lehrende knüpft mit neuen Inhalten an Bekanntes an.	
178.	Der Lehrende nimmt die Lehrtätigkeit ernst.	
179.	Der Lehrende legt Wert darauf, daß die Studierenden etwas in der Lehrveranstaltung lernen können.	
180.	Der Lehrende zeigt persönliches Interesse am Stoff.	
181.	Der Lehrende gestaltet diese Lehrveranstaltung lebendig und engagiert.	
182.	Der Lehrende hat diese Lehrveranstaltung übersichtlich gegliedert.	
183.	Der Lehrende versteht es, die Studierenden zur Mitarbeit zu motivieren.	
184	Der Lehrende vermittelt mir neue Einsichten.	
185.	Der Lehrende regt zur kritischen Auseinandersetzung mit den behandelten Themen an.	
186.	Der Lehrende versteht es, die Studierenden für den Stoff der Lehrveranstaltung zu interessieren.	
187.	Der Lehrende betont sehr deutlich wichtige Aspekte des Stoffes.	

188.	Der Lehrende ist offen für andere Auffassungen.	
189.	Der Lehrende respektiert die Studierenden als Persönlichkeiten.	
190.	Der Lehrende äußert Kritik konstruktiv.	
191.	Der Lehrende ist im Umgang mit den Studierenden freundlich und aufgeschlossen.	
192.	Der Lehrende geht während der Lehrveranstaltung auf Fragen, Anregungen und Einwände sorgfältig ein.	
193.	Der Lehrende ist an der Meinung der Studierenden zur Lehrveranstaltung interessiert.	
194.	Der Lehrende sorgt für eine angenehme Atmosphäre in der Lehrveranstaltung.	
195.	Der Lehrende benachteiligt oder diskriminiert bestimmte Studierende.	
196.	Der Lehrende gibt den Studierenden zu wenig Rückmeldung.	
197.	Die Lernziele der Lehrveranstaltung sind klar definiert.	
198.	Die Bedeutung der angebotenen Lehrinhalte für das Studium ist unklar.	
199.	Einige Verhaltensweisen des Lehrenden gegenüber Studierenden finde ich sehr störend.	
200.	Man wird dazu angeregt, über die praktische Anwendbarkeit theoretischen Wissens nachzudenken.	
201.	In dieser Lehrveranstaltung herrscht eine Arbeitsatmosphäre, die dazu ermutigt, sich zu beteiligen.	
202.	Durch dieses Lehrveranstaltung wird man zu wenig auf die Prüfung vorbereitet.	
203.	Die Stoffmenge in dieser Lehrveranstaltung ist zu umfangreich.	
204.	Die Schwierigkeit der Veranstaltungsinhalte ist zu hoch.	
205.	Der verlangte Arbeitsaufwand für diese Lehrveranstaltung ist zu hoch.	
206.	Ich arbeite regelmäßig für diese Lehrveranstaltung (Vor- oder Nachbereitung).	
207.	Ich besuche in diesem Semester diese Veranstaltung regelmäßig.	
208.	Die Beiträge der Studierenden in der Lehrveranstaltung sind für mich interessant.	
209.	Die Studierenden zeigen großes Engagement für diese Lehrveranstaltung.	
210.	Das Klima unter den Studierenden ist kooperativ.	
211.	Diese Lehrveranstaltung ist überfüllt.	
212.	Während dieser Lehrveranstaltung fühle ich mich durch äußere Bedingungen (z.B. schlechte Luft oder Akustik) beeinträchtigt.	
213.	Die Veranstaltung verläuft nach einer klaren Gliederung.	
214.	Der Dozent kommt häufig vom Thema ab.	
215.	Der Dozent verdeutlicht Zusammenhänge zu wenig.	
216.	Der Dozent drückt sich klar und verständlich aus.	

217.	Die Veranstaltung gibt einen guten Überblick über das Themengebiet.	
218.	Die Art, wie die Veranstaltung gestaltet ist, trägt zum Verständnis des Stoffes bei.	
219.	Die Hilfsmittel zur Unterstützung des Lernens (z.B. Literatur, Skript, Folien) sind ausreichend und in guter Qualität.	
220.	Dem Dozenten scheint der Lernerfolg der Studierenden gleichgültig zu sein.	
221.	Der Dozent verhält sich den Studierenden gegenüber freundlich und respektvoll.	
222.	Der Dozent geht auf Fragen und Anregungen der Studierenden ausreichend ein.	
223.	Der Dozent gestaltet die Veranstaltung interessant.	
224.	Die Veranstaltung ist vermutlich für die spätere Berufspraxis sehr nützlich.	
225.	Der Dozent verdeutlicht zu wenig die Verwendbarkeit und den Nutzen des behandelten Stoffes.	
226.	Der Dozent fördert mein Interesse am Themenbereich.	
227.	Der Schwierigkeitsgrad der Veranstaltung ist (viel zu niedrig/gering - viel zu hoch/groß).	
228.	Der Stoffumfang der Veranstaltung ist (viel zu niedrig/gering - viel zu hoch/groß).	
229.	Das Tempo der Veranstaltung ist (viel zu niedrig/gering - viel zu hoch/groß).	
230.	Welche "Schulnote" (1 bis 6) würden Sie der Veranstaltung insgesamt geben?	
231.	Welche "Schulnote" (1 bis 6) würden Sie dem Dozenten als Veranstaltungsleiter geben?	
232.	Ich habe in der Veranstaltung viel gelernt.	

Anhang K: Übersicht Itemnumerierung

Numerierung im englischen Übungsbogen

Quelle	Item Nr. im Original	Item Nr. Kodierbogen	Anzahl Items
SIR	1-20 [von 23]	1-20	20

Numerierung im deutschen Übungsbogen

Quelle	Item Nr. im Original	Item Nr. Kodierbogen	Anzahl Items
Stangl (2000)	Nicht numeriert	1-20	20

Numerierung im englischen Kodierbogen

	Name	Item Nr. im Original	Item Nr. Kodierbogen	Anzahl Items
1)	IDEA	1-47	1-47	47
2)	SIR II	1-40	48-87	40
3)	SIRS	1-21	88-108	21
4)	SEEQ	1-32	109-139	31
5)	CIEQ	1-21	140-160	21
6)	ICE	1-56	161-216	56
7)	SPTE	1-39 (-1)	217-254	38
				254 gesamt

Numerierung im deutschen Kodierbogen

	Name	Item Nr. im Original	Item Nr. Kodierbogen	Anzahl Items
1)	HILVE	1-34, 40-42 [35-39 freie, 43]	1-37	37 [43]
2)	BEVA	1-40	38-77	40
3)	FELL-V	1-23	78-100	23
4)	MFZAL	1-17	101-117	17
5)	VBPSYCH	1-40	118-157	40
6)	FB-LV	1-55	158-212	55
7)	FEVOR	1-20	213-232	20
				232 gesamt

Anhang L: Häufigkeitstabelle der Dimensionen für beide Kulturräume

Kat.	Kulturraum							
	deutsch				amerikanisch			
	Kodierer B		Kodierer A		Kodierer B		Kodierer A	
	Anzahl	%	Anzahl	%	Anzahl	%	Anzahl	%
0	29	12,5%	24	1,03%	19	7,5%	22	8,7%
1	8	3,4%	8	3,4%	13	5,1%	9	3,5%
2	6	2,6%	7	3,0%	9	3,5%	8	3,1%
3	4	1,7%	4	1,7%	4	1,6%	7	2,8%
4			6	2,6%	4	1,6%	2	0,8%
5	8	3,4%	3	1,3%	5	2,0%	10	3,9%
6	29	12,5%	30	12,9%	17	6,7%	11	4,3%
7	1	0,4%	1	0,4%	3	1,2%	3	1,2%
8	7	3,0%	6	2,6%	8	3,1%	8	3,1%
9	2	0,9%	2	0,9%	7	2,8%	7	2,8%
10	26	11,2%	17	7,3%	7	2,8%	11	4,3%
11	5	2,2%	9	3,9%	18	7,1%	17	6,7%
12	21	9,1%	23	9,9%	33	13,0%	19	7,5%
13			1	0,4%	12	4,7%	12	4,7%
14	4	1,7%	1	0,4%	3	1,2%	3	1,2%
15	2	0,9%	2	0,9%	4	1,6%	4	1,6%
16	13	5,6%	18	7,8%	18	7,1%	17	6,7%
17	6	2,6%	6	2,6%	7	2,8%	7	2,8%
18	7	3,0%	8	3,4%	3	1,2%	3	1,2%
19					7	2,8%	6	2,4%
20			4	1,7%	5	2,0%	3	1,2%
21	8	3,4%	5	2,2%	3	1,2%	4	1,6%
22							1	0,4%
23	16	6,9%	1	0,4%	9	3,5%	5	2,0%
24	2	0,9%	14	6,0%	11	4,3%	12	4,7%
25	7	3,0%	3	1,3%				
26	10	4,3%	10	4,3%	1	0,4%	1	0,4%
27			5	2,2%	2	0,8%	2	0,8%
28			3	1,3%	2	0,8%	7	2,8%
29	1	0,4%	3	1,3%	7	2,8%	11	4,3%
30	10	4,3%	8	3,4%	13	5,1%	22	8,7%

Anhang M: Übersichten über häufige Kategorien in einzelnen Verfahren

Tab. 01: Übersicht über häufige Kategorien in den einzelnen deutschen Verfahren

Name	Items	Kodierer A			Kodierer B		
	N	Kategorie	F	p	Kategorie	f	p
HILVE	37	16	5	13,5%	0	10	27,0%
		10	4	10,8%			
		0	4	10,8%			
BEVA	40	12	11	27,5%	12	6	15,0%
		26	6	15,0%	26	4	10,0%
		16	4	10,0%	16	4	10,0%
FELL-V	23	11	4	17,4%	10	5	21,7%
		4	4	17,4%	6	4	17,4%
MFAL	17	Alle	≤ 2	≤ 11,8	21	3	17,6%
PSYCH	40	6	11	27,5%	6	10	25,0%
FB-LV	55	0	13	23,6%	0	14	25,5%
		16	5	9,1%	6	5	9,1%
		6	4	7,3%	16	4	7,3%
FEVOR	20	6	5	25,0	6	4	20,0

Tab. 02: Übersicht über häufige Kategorien in den einzelnen amerikanischen Verfahren

Name	Items	Kodierer A			Kodierer B		
	N	Kategorie	f	P	Kategorie	f	p
IDEA	47	12	13	27,7%	0	5	10,6%
					11	5	10,6%
					16	5	10,6%
SIR II	40	11	8	20,0%	11	6	15,0%
		23	5	12,5%			
SIRS	21	0	4	12,9%	16	4	19,0%
		16	4	9,7%			
SEEQ	31	16	4	12,9%	16	4	12,9%
CIEQ	21	1	4	19,0%	30	5	23,8%
		30	4	19,0%			
ICE	55	12	11	20,0%	12	10	18,2%
		0	8	14,5%	30	10	18,2%
					0	6	10,9%
SPTE	39	13	4	10,3%	13	4	10,3%
		24	4	10,3%	28	4	10,3%
					29	4	10,3%