Möglichkeiten und Grenzen einer Lehrevaluation.
Eine Analyse am Beispiel der Lehrevaluation an der Philosophischen Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg

Vorläufige Ergebnisdarstellung

Werner Meinefeld

unter Mitarbeit von

Gülsah Adigüzel	Karen Genz	Andreas Göller
Aylin Gülec	Nancy Hollmann	Stefanie Jahl
Monika Karl	Thomas Krombach	Christian Künert
Inga Maubach	Alisa Maul	Robin Pötke
Melanie Riese	Franziska Scharch	Hendriekje Schelten
Kathrin Schmidt	Franziska Schwabl	Anna Sippl
Irmgard Steckdaub-Muller	Michael Stransky	Andreas Tischler
Dagny Tolksdorf	Katharina v. Kracht	MarkusWesthauser

Im Wintersemester 2008/09 nahm auch Juliane Staehler an diesem Forschungsseminar teil.

Der nachfolgende Bericht ist vorläufig und unvollständig: zu mehreren Fragestellungen fehlt noch die Ausarbeitung. Die Ausführungen zu den vorgelegten Gliederungspunkten sind allerdings abgeschlossen.

Erlangen, im Juni 2009

Anmerkung zur Lektüre

Der Bericht ist umfangreicher geworden, als es für eine kurze Information erwünscht ist. Dies schuldet sich nicht zuletzt der Tatsache, daß zum einen die gewonnenen Ergebnisse mittels Tabellen u.ä. zu belegen und diese zu interpretieren waren, zum anderen das methodische Vorgehen zumindest so weit expliziert werden mußte, daß es vom Leser nachvollziehbar und damit in seiner Güte bewertbar ist.

Für eilige Leserinnen und Leser ist die Zusammenfassung der Ergebnisse gedacht (Kapitel 9, S. 65-60). Hier wird auf die Darlegung der Hintergründe der Lehrevaluation, auf ausführliche Belege und Begründungen und methodische Informationen völlig verzichtet.
Verfügen Sie über etwas mehr Zeit, so lesen Sie auch die Zusammenfassung des Vergleichs zwischen der Online-Erhebung und der Erhebung in den Veranstaltungen (S. 11f) sowie die Ausführungen zur multivariaten Analyse der Einflüsse „externer" Variablen auf die Bewertung der Lehrveranstaltungen (S. 40-52).
Wer die Genese der Ergebnisse und der Schlußfolgerungen genauer nachvollziehen möchte, muß sich allerdings der Mühe der Lektüre des gesamten Textes unterziehen.

Danksagung

Den Teilnehmerinnen[1] des Forschungsseminars „Evaluation der Hochschullehre" am Institut für Soziologie im Wintersemester 2008/09 und im Sommersemester 2009 danke ich für ihre Arbeit in diesem Projekt ganz herzlich. Sie haben mit großem Engagement und hoher Kompetenz das Gelingen des Projektes ermöglicht. Für die im folgenden vorgetragenen Ergebnisse und Schlußfolgerungen trage ich allerdings die alleinige Verantwortung.

[1] Den Mehrheitsverhältnissen folgend, werde ich im folgenden sprachlich für die Lehrenden die männliche Form und bei den Studierenden die weibliche Form wählen; sofern es nur um eines der Geschlechter geht, werde ich dies explizit anmerken.

Inhaltsverzeichnis[2]

---

[2] Die kleingedruckten Kapitel sind noch nicht ausgearbeitet.

2. Die Lehrveranstaltungsbewertung an der Philosophischen Fakultät

An der Friedrich-Alexander-Universität Erlangen-Nürnberg wird seit dem Zusammenschluß der ehemaligen Philosophischen Fakultäten I und II mit der Erziehungswissenschaftlichen Fakultät und der Theologischen Fakultät zur Philosophischen Fakultät und Fachbereich Theologie eine lehrveranstaltungsbezogene Befragung der Studierenden aller Lehrveranstaltungen mit einem gemeinsamen Fragebogen durchgeführt. Die Ergebnisse dieser Erhebung werden im gesetzlich geforderten Lehrbericht der Studiendekane berücksichtigt. Diese Befragung soll nicht nur dazu dienen, die Bewertung der Lehrveranstaltungen durch die Studierenden systematisch zu erfassen, sondern sie soll auch die Basis für einen Austausch zwischen den Lehrenden und den Studierenden über die Anlage, die Durchführung und den Lehrerfolg der Vorlesungen und Seminare bilden. Aus diesem Grund wird die Befragung kurz nach der Mitte des Semesters durchgeführt, um die Ergebnisse noch in die laufende Lehrveranstaltung einbringen zu können. Dieses Verfahren ist von Anfang an, z.T. sehr heftig, kritisiert worden, so daß die Studiendekane mit der Ausarbeitung eines neuen Konzeptes für die Lehrevaluation beauftragt wurden.

Als solche kritische Punkte konnten insbesondere folgende Aspekte identifiziert werden:

- die Lehrveranstaltungsbefragung erfolge zu einem zu frühen **Zeitpunkt**, zu dem die Studierenden noch nicht genügend Erfahrungen mit dem Verlauf der Veranstaltung haben sammeln können;
- mehrere Einwände thematisieren die Durchführung der Befragung über einen im **Internet** auszufüllenden Fragenbogen:
  a) dies führe zu einer sehr geringen Beteiligung der Studierenden;
  b) die Verteilung von TAN-Nummern durch die Lehrenden führe ebenfalls zu einem nicht kontrollierbaren Teilnahmeeffekt (dies reicht von der – nicht unbedingt ernst gemeinten, aber als Möglichkeit angesprochenen – Vermutung, Lehrende könnten selbst unter verschiedenen TAN-Nummern an ihrer Bewertung teilnehmen, über den möglichen Verlust dieser kleinen Zettel bis hin zur Möglichkeit der Kumulierung mehrerer Nummern bei einzelnen Studierenden);
  c) es sei ganz ungeklärt, wer sich an einer Online-Erhebung beteilige: die Fans der Lehrenden? die Gruppe derjenigen, die die Gelegenheit zu ungerechtfertigt kritischen Äußerungen nutzen wolle (und dies nur im Schutz der Anonymität des Internets wagten), um sich für Kritik der Lehrenden zu revanchieren? oder beide Gruppen, während die große Mehrheit zufrieden und daher auch nur schwer zu motivieren sei?
- andere Einwände beziehen sich auf den verwendeten **Fragebogen**:
  a) er nehme keine Differenzierung zwischen verschiedenen Veranstaltungstypen vor (Vorlesungen, Seminaren, Sprachkursen o.ä.) und könne daher den Besonderheiten der

einzelnen Typen nicht gerecht werden;

b) auch gebe es keine Berücksichtigung fachspezifischer Anforderungen, so daß manche Fragen für bestimmte Fächer nicht passen, für sie wesentliche Fragen dagegen fehlen;

c) andere Fragen seien unangemessen in ihrem Bezug zum Ziel einer Lehrevaluation bzw. mißverständlich formuliert;

– schließlich seien die **Analyse** und die **Darstellung** der Ergebnisse zu oberflächlich, da sie den Einfluß externer Faktoren (Veranstaltungstyp und -größe, Teilnahmemotivation der Studierenden, Raumsituation etc.) nicht berücksichtigten und damit zu Fehlinterpretationen der Bewertungen der Veranstaltungen führten. Zudem werde mit der Wahl des arithmetischen Mittels ein statistisch ungeeignetes Modell verwendet.

Selten wurden in der Diskussion dagegen über das konkrete Prozedere hinausgehende Fragen angesprochen: Was ist unter einer Evaluation überhaupt zu verstehen? Welche Ziele will und kann man mit einer Lehrevaluation erreichen? Wer ist von ihr betroffen? Wie sollen diese Personengruppen in die Evaluation einbezogen werden? Welche Anforderungen sind an ein solches Projekt zu stellen?: welcher Aufwand ist zu vertreten? welche Methoden sind angemessen? wie sind die Ergebnisse zu verwenden? Wie ist sicherzustellen, daß nicht nur evaluiert wird, sondern daß die Ergebnisse in den Lehralltag umgesetzt werden?

3. Zwischenbemerkung zum Konzept einer „Lehrevaluation"

Im bayerischen Hochschulgesetz ebenso wie im alltäglichen Sprachgebrauch an den Hochschulen ist es üblich, von einer „Lehr*evaluation*" zu sprechen, wenn tatsächlich mehr oder weniger systematische Erhebungen zur Bewertung von Lehr*veranstaltungen* durch die Studierenden gemeint sind. „Lehrevaluation" ist ein griffiger Begriff, und es geht ja auch um eine „Bewertung der Lehre" – die Kriterien für eine professionelle Evaluationsforschung sind allerdings in der Praxis selten bis nie gegeben. In der Regel sind sie nicht einmal bekannt, und die Notwendigkeit ihrer Beachtung wird entweder nicht eingesehen („Das gilt nur für die Forschung, aber nicht für uns.") bzw. man muß bedauernd feststellen, dazu nicht in der Lage zu sein. Der Druck der Verhältnisse („es ist gesetzlich vorgeschrieben") zwingt jedoch dazu, sich – oft genug unter einem enormen Aufwand an Zeit und Engagement – in irgendeiner Weise mit dieser Vorgabe zu arrangieren. Nun ist der alltägliche Sprachgebrauch nicht normiert, insofern ist die Bezeichnung einer Bewertung als Evaluation legitim – wenn es auch befremdlich anmuten mag, daß auch im Kontext der Wissenschaft gedankenlos ein Begriff verwendet wird, dessen Voraussetzungen nicht erfüllt sind, der aber ganz offensichtlich zur Aufwertung einer durchaus als unzureichend empfundenen Praxis dient.

4. Zielsetzung und Anlage der hier vorgelegten Begleitforschung zur Lehrveranstaltungsbewertung

Auf dem Hintergrund der oben erwähnten kontroversen Diskussion um die Lehrveranstaltungsbefragung an der Philosophischen Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg hat der Verfasser im WS 2008/09 und im SS 2009 ein zweisemestriges Forschungsseminar am Institut für Soziologie durchgeführt. Allgemein ist es das Ziel eines solchen Forschungsseminars, die Studierenden anhand eines konkreten Forschungsthemas in die Praxis der empirischen Sozialforschung einzuführen und sie – neben der Vermittlung theoretischer und methodologischer Kenntnisse – mit der praktischen Durchführung eines Forschungsprojektes vertraut zu machen. Hier bot sich zudem die Gelegenheit, eine Forschung durchzuführen, die ein Problem der eigenen Fakultät aufgriff und damit über eine akademische Fragestellung hinausging: von diesem Seminar wurden belastbare Ergebnisse erwartet, die zumindest die eine oder andere kontroverse Diskussion helfen sollten zu entscheiden.

Nach der Auseinandersetzung mit der vorliegenden Literatur zur Evaluationsforschung und speziell mit Arbeiten zur Lehrevaluation haben wir im Forschungsseminar ein Design entwickelt, das die Überprüfung einiger der nachfolgend aufgeführten Kritikpunkte am bisherigen Evaluationsverfahren erlauben sollte. Das Ziel dieses Forschungsseminars ist es somit, zu einer Bewertung der Aussagekraft der an der Philosophischen Fakultät durchgeführten Lehrevaluation zu kommen. Dabei sollen sowohl die von verschiedenen Seiten vorgebrachten Einwände wie auch potentielle weitere Fehlerquellen dieses Vorgehens untersucht werden.

Zum Design der beiden Erhebungen

a) Die Erhebung durch die Studiendekane

Die Erhebung der Veranstaltungsbewertungen durch die Studiendekane (im folgenden als „Online-Erhebung" bezeichnet) erfolgte mittels eines im Internet bereitgestellten Fragebogens. Dieser Fragebogen bestand aus acht Fragen zu verschiedene Dimensionen der Veranstaltung (Aufbau, Verständlichkeit, Betreuung, Gesamtnote ... – vgl. Anhang A), wobei jede Dimension von 1 bis 6 benotet werden konnte. Weitere fünf Fragen bezogen sich auf den Vorbereitungsaufwand, das Anspruchsniveau, das Erleben von Überforderung und Unterforderung u.ä. In einem freien Textfeld konnten zudem Anmerkungen zur Veranstaltung gemacht werden. Zum Abschluß wurden Daten zur Ausübung von Jobs neben dem Studium erhoben.

Die Befragung erfolgte in der dritten und zweiten Woche vor Weihnachten 2008. Die Internet-

adresse für den Fragebogen und TAN-Nummern wurden an alle Lehrenden der Philosophischen Fakultät und Fachbereich Theologie verschickt mit der Bitte, sie an die Teilnehmerinnen ihrer Veranstaltungen zu verteilen. Tatsächlich konnten Bewertungen auch noch nach Ablauf dieser Frist abgegeben werden. Für die Lehrenden bestand aufgrund der gesetzlichen Verpflichtung zur Durchführung der Lehrevaluation eine Pflicht zu Beteiligung an dieser Erhebung. Die Ergebnisse für ihre jeweilige(n) Veranstaltung(en) wurden den Veranstaltungsleitern am 19. Januar 2009 zugeschickt. Zu diesem Zeitpunkt hatten 9.138 Studierende für 919 Veranstaltungen ihre Bewertungen abgegeben.

b) Die Erhebung im Rahmen des Forschungsseminars (im folgenden als „FS-Erhebung" bezeichnet)

Im Forschungsseminar war ein achtseitiger Fragebogen mit 36 Fragen ausgearbeitet worden, der die Überprüfung der im Seminar entwickelten Forschungsfragen erlauben sollte. Alle Lehrenden der beiden früheren Philosophischen Fakultät (also ohne die frühere Erziehungswissenschaftliche und die Theologische Fakultät) waren vom Verfasser angeschrieben und um ihre Zustimmung gebeten worden, in ihren Veranstaltungen eine schriftliche Befragung der Studierenden durchführen zu dürfen. Von den 264 angeschriebenen Lehrenden haben 35 ihre Bereitschaft erklärt, diese Erhebung in 74 Veranstaltungen durchführen zu lassen. Aus verschiedenen Gründen (z.B. wurde die Veranstaltung von mehreren Lehrenden gemeinsam durchgeführt, oder es handelte sich um ein Blockseminar – auch reichte schlicht die Arbeitskapazität des Forschungsseminars nicht aus) wurden nicht alle Veranstaltungen, deren Leiter ihre Bereitschaft bekundet hatten, sondern nur 46 Veranstaltungen von 33 Lehrenden für unsere Befragung ausgewählt.

Die Datenerhebung erfolgte in der Woche vor Weihnachten und in der Woche nach Weihnachten. Damit sollte zum einen (mit dem ersten Termin) eine möglichst große zeitliche Nähe zur Online-Erhebung gewährleistet sein, damit eventuell auftretende Differenzen zwischen den beiden Erhebungen nicht auch auf den Zeitfaktor zurückzuführen sein können. Zum zweiten (der Termin nach Weihnachten) sollte geprüft werden, inwieweit eine zeitliche Konstanz der Bewertungen gegeben ist.[3]

Die Datenerhebung erfolgte in den ersten 15 Minuten der jeweiligen Veranstaltung. Diese

[3] Dabei sind wir uns der Tatsache bewußt, daß es sich hier nicht im strengen Sinne um eine belastbare experimentelle Anordnung handeln kann, da mögliche Unterschiede zwischen den beiden Zeiträumen nicht notwendig dem Zeitfaktor zuzurechnen sein müssen. Sie können z.B. auch den Eigenheiten der in diesen beiden Zeiträumen erfaßten Veranstaltungen geschuldet sein: da wir nicht dieselben Veranstaltungen ein zweites Mal erfassen konnten, sind die Antworten nur en bloc zwischen den Zeiträumen (ev. differenziert nach Veranstaltungstypen) zu vergleichen.

wurde von in der Regel zwei unserer Seminarteilnehmerinnen aufgesucht. Sie stellten unser Projekt (als Begleitung der Online-Befragung) kurz vor, verteilten die Fragebögen und sammelten sie anschließend wieder ein. Damit wurden alle in dieser Sitzung anwesenden Veranstaltungsteilnehmerinnen in die Befragung einbezogen.[4]

Die Veranstaltungen (und damit auch die Lehrenden) wurden anonymisiert, die Daten von den Teilnehmerinnen des Forschungsseminars in SPSS erfaßt und ausgewertet. Der hier vorgelegte Bericht bezieht sich ausschließlich auf die die Einzelveranstaltungen übergreifende Fragestellung nach der Aussagekraft der Lehrveranstaltungsbefragungen – die Ergebnisse für die einzelnen Veranstaltungen wurden den Lehrenden bereits im März zugesandt.[5]

Der Datensatz

Der unserer Analyse zugrundeliegende Datensatz beruht auf den Antworten von 1.528 Studierenden, die in 47 wöchentlich stattfindenden Lehrveranstaltungen von uns befragt wurden. Mit der hier realisierten Auswahl der Lehrveranstaltungen liegt keine repräsentative Stichprobe für die Veranstaltungen der (früheren) Philosophischen Fakultät(en) vor. Aufgrund der Freiwilligkeit der Teilnahme ist eine deutliche Überrepräsentation von Veranstaltungen von Lehrenden zu erkennen, die eine fachliche Nähe zur Sozialforschung aufweisen (sprich: Department Sozialwissenschaften und Philosophie, Pädagogik, Psychologie u.a.), während mit zunehmender Ferne die Teilnahmequote abnahm.[6] Für die von uns angestrebten Analysen ist dies allerdings kaum von Bedeutung, da wir nicht Aussagen z.B. über die Verteilung von Bewertungen an der Fakultät (überwiegen gute oder schlechte Bewertungen u.ä.) vornehmen wollen, sondern analytische Beziehungen zwischen verschiedenen Aspekten solcher Bewertungen untersuchen wollen – diese aber gelten unabhängig von fachlichen Bedingungen und sind damit von Fragen

[4] In 30 Veranstaltungen haben alle Anwesenden die Fragebögen ausgefüllt; in zwölf Veranstaltungen fehlten ein bzw. zwei Fragebögen, in vier Veranstaltungen zwischen drei und sechs Bögen, in einer zwölf. Allerdings waren die beiden Wochen um Weihnachten hinsichtlich der Zahl der Anwesenden nicht immer „erschöpfend": von den Veranstaltungsleitern wurden Fehlzahlen zwischen 0 und 55% berichtet – was einer durchschnittlichen Quote von 17% entspricht.

[5] Hier darf ich mich bei Christian Künert und Dagny Tolksdorf für die von ihnen geleistete Arbeit in der Konzipierung einer benutzerfreundlichen Darstellung der in SPSS berechneten Ergebnisse bedanken.

[6] Vertreter folgender Fächer haben an unserer Untersuchung teilgenommen: *Philosophische Fakultät I:*Pädagogik, Philosophie, Politikwissenschaft, Psychologie, Soziologie, Wirtschaftswissenschaft; *Philosophische Fakultät II:* Amerikanistik/Anglistik, Buchwissenschaft, Germanistik, Romanistik, Theaterwissenschaften.

der Repräsentativität unabhängig.[7]

5. Effekte des Forschungsdesigns: Zur Zuverlässigkeit der Online-Erhebung[8]

Wenden wir uns nun der Frage zu, inwieweit das in der Befragung der Studiendekane gewählte Forschungsdesign: die Durchführung der Befragung über das Internet, zu Verzerrungen in den Bewertungsergebnissen geführt haben kann. Gegen eine Online-Befragung wird oft vorgebracht, daß die Beteiligung zu gering und schon aus diesem Grunde nicht aussagekräftig sei. Zudem wisse man nichts über die Motivation der Antwortenden: handelt es sich um die besonders Engagierten? oder nehmen gerade diejenigen teil, die sich für Kritik oder strenge Beurteilung durch den Lehrenden revanchieren wollen? oder sind diese beiden Gruppen überproportional stark vertreten, während die (schweigende) Mehrheit die Befragung einfach ignoriert?

Der Grad der Beteiligung an der Online-Erhebung

Beteiligen sich weniger als 3 Studierende an der Online-Erhebung, so wird von den Studiendekanen für diese Veranstaltung kein Ergebnisbericht erstellt. Dies traf auf 6 unserer 47 Veranstaltungen zu – für 13% liegen also gar keine Online-Ergebnisse vor. Die nachfolgende Tabelle zeigt, daß die Beteiligung der Studierenden an der Online-Erhebung aber auch in den anderen Veranstaltungen sehr gering ist (in 38 Veranstaltungen lag sie bei durchschnittlich 38%[9]). Rechnet man die erwähnten 6 Veranstaltungen ohne Rücklauf mit den 8 Veranstaltungen mit einer Beteiligung unter 20% zusammen, so kann man sagen, daß die Online-Erhebung für 30% der Veranstaltungen keine zuverlässige Basis für eine Bewertung erbracht hat. Während Online 92% der Veranstaltungen eine Beteiligung *unter* 60% aufweisen, liegen in der FS-Erhebung 95% *über* diesem Wert, 55% weisen sogar eine Beteiligung von mehr als 80% auf.

---

[7] Dies schließt nicht aus, daß z.B. wegen des fast vollständigen Fehlens von Sprachkursen die in diesen Kursen möglicherweise geltenden spezifischen Bedingungen gar nicht erfaßt werden können und somit in unserer Analyse Leerstellen bestehen. Damit könnten sich für den Gesamtbestand aller Veranstaltungen an der Fakultät durchaus z.B. auch etwas andere Zusammenhänge zwischen einzelnen Bewertungen und Einflußfaktoren ergeben, doch schränkt dies nicht die Gültigkeit für die hier berücksichtigten Veranstaltungstypen ein.

[8] Wenn im folgenden die Abkürzung „O" verwendet wird, so steht dies immer für Ergebnisse der Online-Erhebung, „FS" steht entsprechend für die Erhebung, die im Rahmen des Forschungsseminars durchgeführt wurde.

[9] In dieser Zahl von 38% sind die gänzlich fehlenden Veranstaltungen nicht berücksichtigt – die reale Ausschöpfungsquote für *alle* Veranstaltungen ist also noch deutlich niedriger.

Erreichbarkeit der Veranstaltungsteilnehmer in Abhängigkeit von der Erhebungsmethode									
Erhebung	absolute Zahlen		in Prozent[10]		Mittel-wert	Anteilswerte			
	min	max	min	max		≤20%	≤40%	≤ 60%	≤ 70%
Online	3	68	17	69	38%	5%	68%	92%	100%
FS	7	118	45	100	81%	0%	0%	5%	12%

Differenzen in den Bewertungen zwischen Online- und FS-Erhebung

Die hier berichteten Unterschiede belegen allerdings noch nicht, daß die Ergebnisse durch das unterschiedliche Teilnahmeverhalten tatsächlich verzerrt werden: es wäre durchaus möglich, daß die Teilnahme nicht mit dem Bewertungsverhalten zusammenhängt, somit trotz der geringeren Beteiligungsquoten die Online-Bewertungen die Wahrnehmung der Studierenden in den Veranstaltungen in akzeptabler Näherung wiedergeben.

Diese Möglichkeit scheint bestätigt zu werden, wenn wir die in der Online-Erhebung erreichten *Durchschnittswerte* mit den Durchschnittswerten in der FS-Erhebung vergleichen. Um einen solchen Vergleich überhaupt vornehmen zu können, haben wir zentrale Bewertungsfragen der Online-Erhebung in unserer Erhebung im Forschungsseminar wörtlich wiederholt.[11] Zudem haben wir – analog zum Ergebnisbericht der Online-Erhebung – einen als „Globalindikator" bezeichneten Mittelwert aller Antworten auf acht dieser Fragen auch für unsere Daten berechnet.[12] Die nachfolgende Tabelle stellt für jede dieser Fragen (sowie den Globalindikator) die durchschnittliche Bewertung der 41Veranstaltungen der Online-Erhebung, für die Ergebnisse vorliegen, der durchschnittlichen Bewertung aller Veranstaltungen in der FS-Erhebung gegenüber.

[10] In unserer Erhebung hatten wir die Zahl der anwesenden Studierenden festgestellt und die Lehrenden gebeten zu schätzen, wie viele Teilnehmer in etwa fehlen. Die hier berichteten Prozentwerte beziehen sich auf die so ermittelte Gesamtteilnehmerzahl. Vier Lehrende mochten sich bei der Zahl der Fehlenden nicht festlegen, daher fehlen bei diesen Angaben vier Veranstaltungen.

[11] Dies betrifft insbesondere die acht Unterfragen der Frage 1.

[12] Die Berechnung des Globalindikators erfolgte in der Online-Erhebung auf der Basis der Antworten auf alle acht Teilfragen der ersten Frage. Damit wurde in die Bewertung des Dozenten auch die Bewertung des Engagements der Kommilitoninnen einbezogen. Inhaltlich ist dies wenig sinnvoll, doch haben wir diese Berechnung ebenfalls angewendet, da sonst eine Vergleichbarkeit der beiden Bewertungen nicht gegeben gewesen wäre.

Übereinstimmung der Durchschnittswerte für Einzelfragen und Globalindikator zwischen Online-Erhebung und FS-Erhebung		
Frage zu:	**Online-Erhebung**	**FS-Erhebung**
Aufbau	2,26	2,26
Verständlichkeit	2,31	2,24
Qualität der Hilfsmittel	2,31	2,33
Engagement des Dozenten	1,88	1,88
Durchführung der LV	1,64	1,71
Betreuung außerhalb der LV	2,08	2,02
Engagement der Kommilitoninnen	2,78	2,94
Note der LV	2,28	2,27
Globalindikator	2,18	2,20
Zahl der LVen	41	47

Die Übereinstimmung zwischen den Mittelwerten der jeweiligen Dimensionen für *alle* Veranstaltungen ist verblüffend – aber auch irrelevant und irreführend:

- Irrelevant, da es für die Bewertung der Aussagekraft der Befraung nicht auf die Übereinstimmung für den *gesamten* Datensatz ankommt, sondern auf die Übereinstimmung der Ergebnisse für jede *einzelne* Veranstaltung.
- Irreführend, weil die Übereinstimmung eine Zuverlässigkeit der Ergebnisse der Online-Befragung zu belegen scheint, wo tatsächlich eine erhebliche Unzuverlässigkeit vorliegt. Um die *veranstaltungsbezogene* Zuverlässigkeit zu überprüfen (und nur um die kann es ja gehen), ist die Differenz der Ergebnisse der Online-Erhebung und der FS-Erhebung für jede Veranstaltung getrennt zu berechnen.

Die nachfolgende Tabelle zeigt die „Differenz der Bewertungen der einzelnen Dimensionen zwischen der Online- und der FS-Erhebung" auf, wenn wir sie veranstaltungsspezifisch aufeinander beziehen.[13] Dabei wird deutlich:

- in *jeder* der untersuchten Dimensionen („Aufbau der Veranstaltung", „Verständlichkeit" ...)

[13] Berechnet wurde sie als Differenz „Note Online-Erhebung" – „Note FS-Erhebung". Dies bedeutet, daß positive Werte eine schlechtere Bewertung in der Online-Erhebung anzeigen (deren Wert war größer, die Note folglich schlechter), während negative Differenzen für eine bessere Bewertung stehen.

gibt es erhebliche Differenzen; zumeist liegen die Extreme zwischen ±0,6 Bewertungspunkte auseinander, doch gehen die Werte um bis zu +0,89 und −1,4 Notenpunkte auseinander, was erhebliche Ergebnisdifferenzen zwischen den beiden Erhebungen belegt;

- es ist kein Muster in den Differenzen zu erkennen, das es erlauben würde, die Richtung der Verzerrung abzuschätzen (also z.B. zu sagen: „Die Online-Erhebung schätzt die Veranstaltungen *systematisch* besser/schlechter ein als die FS-Erhebung"):
  a) jede Veranstaltung und jede Dimension ist von solchen Differenzen betroffen,
  b) in jeder Dimension gibt es sowohl positive als auch negative Abweichungen, kommt also mal die Online-Erhebung, mal die des Forschungsseminars zu besseren bzw. schlechteren Bewertungen;
- als besonders gravierend erweist sich die Tatsache, daß andererseits einzelne Veranstaltungen in der Online-Erhebung systematisch in (fast) allen Dimensionen besser, andere in (fast) allen Dimensionen schlechter bewertet werden, als dies in der FS-Erhebung der Fall ist:
  a) in fünf Veranstaltungen (20, 29, 31, 37, 41) kommt maximal *eine* positive Abweichung vor[14]: sie werden somit online systematisch „zu gut" bewertet, wenn man die umfassendere Erhebung in der Veranstaltung selbst zum Maßstab nimmt;
  b) neun Veranstaltungen (2, 5, 13, 14, 17, 24, 27, 28, 40) weisen dagegen nur maximal *eine* negative Abweichung auf: sie werden online„zu schlecht" bewertet;
  c) damit ist ein Drittel der Veranstaltungen über- oder unterbewertet worden.

Wir haben es also nicht mit einer systematischen – und damit eventuell rechnerisch auszugleichenden – Über- oder Unterschätzung für die *Gesamtheit* der Veranstaltungen durch die eine oder andere Erhebungsmethode zu tun, wohl aber werden *einzelne* Veranstaltungen systematisch anders bewertet. Diese veranstaltungsspezifischen Verzerrungen lassen sich nicht mit bestimmten Merkmalen der Veranstaltungen in Verbindung bringen: Weder der Typ noch die Größe der Veranstaltung, weder das Geschlecht noch der Status der Lehrenden (Professor vs. Mitarbeiter) hängen mit diesen Verzerrungen zusammen. Es hat den Anschein, als regiere bei der Beteiligung an der Online-Befragung der Zufall – mit entsprechenden Auswirkungen auf die Bewertungen.

(Siehe Tabelle auf der nächsten Seite)

[14] Von acht Meßwerten: der Globalindikator als von den anderen abgeleitete Meßgröße darf hier natürlich nicht berücksichtigt werden.

Unter methodischen Gesichtspunkten wäre es möglich, daß der Zeitpunkt der Datenerhebung bei der Entstehung der beobachteten Differenzen eine Rolle spielt. Um diesen Einfluß auszuschließen, haben wir zwar die FS-Erhebung direkt an die Online-Erhebung anschließen lassen, doch liegt dennoch die Erfahrung von ein bzw. zwei zusätzlichen Sitzungen zwischen den beiden Erhebungsphasen – dies war nicht zu vermeiden. Um zu prüfen, inwieweit sich der Zeitfaktor überhaupt auswirkt, können wir die Durchschnittsbewertungen in der FS-Erhebung vor und nach der Weihnachtspause miteinander vergleichen. Diese Ergebnisse sprechen nicht für einen systematischen Einfluß: über alle Veranstaltungen hinweg waren die Bewertungen in fünf der acht Einzeldimensionen im FS besser als Online, in drei dagegen schlechter – sehr viel gleichmäßiger können die Abweichungen nicht verteilt sein. Wie wir oben sahen, sagt dieser Durchschnittswert für alle Veranstaltungen nichts über die Verzerrungen in den einzelnen Veranstaltungen aus. Wenn allerdings die Bewertungen der einzelnen Veranstaltungen im Abstand von ein oder zwei Sitzungen so stark auseinanderfallen, wie dies die Ergebnisse der Online- und der FS-Erhebung tun, so fehlte den Veranstaltungsbewertungen jegliche zeitliche Stabilität – dies ist unwahrscheinlicher als die Interpretation, daß die selektive Beteiligung an der Online-Erhebung zu unkontrollierbaren Verzerrungen in den Ergebnissen führt.

Differenz der Bewertungen der einzelnen Dimensionen zwischen Online- und FS-Erhebung

	Aufbau	Verständlichkeit	Hilfsmittel	Engagement Dozent	Durchführung	Betreuung	Engagement Kommilitonen	Note für LV	Globalindikator
1	-,04	-,39	,59	,22	,57	,89	,21	,23	,22
2	,41	,30	,23	,19	,26	,22	,11	,20	,25
3	,20	,35	,60	,35	-,05	-,40	-,50	,10	,10
4	-,13	-,21	,15	-,20	-,09	,21	-,44	-,25	-,14
5	,31	,08	,39	,21	,08	,00	,11	,18	,26
6	-,17	-,23	,33	,10	-,07	,05	-,63	-,20	-,12
7	-,27	,07	-,40	-,70	-,67	-,63	,20	-,20	-,31
8	-,31	-,16	,39	-,26	-,15	,49	-,45	-,07	-,13
9	,16	,20	-,30	,12	,28	-,51	-,10	-,15	-,04
10	,00	,06	,18	-,24	-,30	,18	,22	-,07	-,05
11	-,06	,07	,07	-,06	,01	-,07	-,43	-,06	-,04
12	-,23	-,05	,00	,02	-,10	,04	-,46	-,16	-,14
13	,39	,45	,28	,44	,35	,33	-,21	,42	,27
14	,26	,56	,68	,43	-,04	,57	,05	,50	,41
15	,44	,22	-,19	-,07	-,20	-,15	-,14	,00	,07
16	,03	,17	-,15	,08	-,14	,00	,07	-,07	,02
17	,18	,10	,41	,28	,10	,35	,34	,40	,29
18	,02	,09	-,27	-,40	-,62	-,20	-,31	-,15	-,24
19	-,34	,02	-,34	,04	-,14	,27	,18	-,32	-,12
20	-,24	-,25	-,09	-,39	-,21	-,72	-,30	-,42	-,33
21	-,06	-,29	-,27	,17	-,30	,79	,51	,05	,08
22	-,42	,05	-,59	,21	-,35	,10	-1,40	-,32	-,34
23	,37	-,01	,16	,26	-,15	,55	,27	,31	,20
24	,70	,36	,20	,12	,59	-,43	,31	,26	,29
25	-,39	-,22	-,56	-,31	-,14	,00	-,11	-,06	-,21
26	,11	-,28	-,28	-,03	-,13	,23	-,34	-,09	-,10
27	,61	,65	,70	,85	,52	,00	,71	,81	,59
28	,29	,43	,30	,57	,30	,50	-1,14	,14	,18
29	-,15	-,06	-,42	-,20	-,04	,30	-,32	-,19	-,17
30	,04	,24	,05	,01	-,06	,23	-,26	-,06	-,05
31	-,39	-,31	-,28	-,41	-,41	-,56	,41	-,26	-,30
32	,12	-,05	-,23	-,15	,05	-,23	-,32	-,06	-,10
33	-,27	-,20	-,23	,02	,07	,17	-,18	-,08	-,05
34	,05	,52	,06	-,12	,20	-,01	-,07	,31	,10
35	-,24	,51	,00	-,44	-,50	-,48	-,41	,03	-,18
36	-,14	,01	-,32	-,13	,20	,20	,11	,06	-,02
37	-,23	-,45	-,17	-,35	-,49	-,28	,03	-,25	-,27
38	,09	,10	-,07	-,19	,17	-,07	-,13	,23	,07
39	,17	,14	-,31	,26	,19	,25	-,23	,06	,08
40	-,36	,20	,06	,07	,09	,26	,10	,20	,05
41	-,34	-,40	-,18	-,20	-,67	-,36	-,45	-,27	-,37
Insgesamt N	41	41	41	41	41	41	41	41	41

Die Bedeutung dieser Meß-Differenzen für die Veranstaltungsbewertung

Wenn, wie im Fall der vorliegenden Veranstaltungen, 63% Prozent der Veranstaltungen mindestens mit „gut" bewertet werden (und keine einen Durchschnittswert erhält, der schlechter ist als 3,4), bleibt die Frage, was diese Note eigentlich bedeutet: „Ist die Lehre wirklich so gut? Wie kommt es dann zu der kritischen öffentlichen Diskussion um die Lehre?". Spricht man den Studierenden grundsätzlich durchaus die Fähigkeit zu, die Qualität der Lehre zu bewerten, so möchte man an dieser Stelle die Urteile doch nicht für bare Münze nehmen („Im Durchschnitt ist nach Meinung der Studierenden die Lehre gut.") und sucht nach einer Auswertungsart, die eine stärkere Hierarchisierung der Bewertungen ermöglicht. Hier bietet sich die Erstellung einer Rangreihe an: „Wer ist der Beste?". Dazu werden die Veranstaltungen nach dem Durchschnittswert, den sie in den Bewertungen der Studierenden bekommen haben, in eine Rangreihe gebracht.

Dieses Vorgehen ist mit erheblichen methodischen Problemen behaftet. Dies beginnt bereits bei der Wahl des Kriteriums, anhand dessen dieses Urteil erfolgen soll. Nimmt man eine Reihung für jede der vorliegenden Einzeldimensionen vor, so erhielte man ein unübersichtliches Bild mit zu vielen und v.a. uneindeutigen Informationen: eine Veranstaltung, die in bezug auf das Engagement des Dozenten an der Spitze liegt, muß sich nicht gleichzeitig durch einen klaren inhaltlichen Aufbau oder eine gute Verständlichkeit auszeichnen. Es liegt daher nahe, ein summierendes Kriterium zu wählen, wie es z.B. die Gesamtnote für die Lehrveranstaltung darstellt, die von den Studierenden vergeben wurde (Frage 1.8) – hier geben die Befragten selber ein zusammenfassendes Urteil ab, bei dem man allerdings nicht weiß, welche Aspekte für die gewählte Note den Ausschlag geben. Alternativ dazu kann ein Indexwert berechnet werden, der sich aus den Antworten auf die einzelnen Fragen zusammensetzt (wie es beim sogenannten „Globalindikator", wie er von den Studiendekanen berechnet wird, der Fall ist) – bei diesem Vorgehen weiß man, auf welchen Urteilen der Wert beruht, kann sich allerdings nicht sicher sein, daß dies die für die Befragten wesentlichen Dimensionen umfaßt, und auch die dabei unterstellte Gleichgewichtigkeit aller dieser Dimensionen ist sicherlich diskussionsfähig. Eine Überprüfung des Zusammenhangs zwischen der Gesamtnote für die Lehrveranstaltung und dem Globalindikator zeigt allerdings eine sehr hohe Übereinstimmung sowohl in der Online- als auch in der FS-Erhebung. Sie sind (natürlich) nicht völlig identisch, aber die berechnete Korrelation von r = 0,97 in beiden Erhebungen zeigt, daß sie in der Tat dasselbe messen, so daß sich die folgende Betrachtung auf einen der beiden Indikatoren beschränken kann.[15]

[15] Bei „r" handelt es sich um den Produktmomentkorrelationskoeffizienten, der den Wert 0 bei Fehlen eines Zusammenhangs zwischen zwei Variablen und 1 bei einer perfekten Übereinstimmung erreicht.

Auf Sinn oder Unsinn der Bildung von Rangreihen ist an dieser Stelle nicht näher einzugehen. Auch wenn sie nicht explizit aufgestellt werden, liegt der Rezeption der studentischen Bewertungen immer auch ein solch vergleichendes Element zugrunde: „Wo stehe ich im Vergleich zu den Kollegen?". Dabei muß uns allerdings interessieren – und dies soll im folgenden geprüft werden –, inwieweit diese vergleichende Reihung durch die Erhebungsmethode beeinflußt wird. Dazu ist zunächst für jede der beiden Erhebungsmethoden eine Rangliste der Veranstaltungen aufzustellen, die dann in einem zweiten Schritt in bezug auf ihre Übereinstimmung geprüft werden. Die nachfolgende Tabelle zeigt die Differenz der Rangpositionen, die die einzelnen Veranstaltungen in der einen bzw. in der anderen Liste einnehmen.[16] Die nachfolgende Tabelle zeigt, daß die Rangplätze von 41 Veranstaltungen zwischen −14,5 und +19,5 Plätze voneinander abweichen. Um es an einer Veranstaltung plastisch zu zeigen: Die zuletzt erwähnte Veranstaltung mit einer Differenz von 19,5 Plätzen hat online den Rang 36 (mit einer anderen gemeinsam) inne, in der FS-Erhebung steht sie aber auf Platz 17 – für den davon betroffenen Lehrenden liegen zwischen diesen beiden Plazierungen Welten der Bewertung! Eine Abweichung um mehr als fünf Ränge weisen 66% aller Veranstaltungen auf, eine um mehr als zehn Ränge immer noch 17%.

[16] Berechnet wurde die Differenz „Rang in der Gesamtbewertung Online" minus „Rang in der Gesamtbewertung FS" – negative Werte zeigen also an, daß die Veranstaltung Online besser plaziert war, positive Werte stehen für eine schlechtere Online-Plazierung. Für die Online-Erhebung lagen uns nur gerundete Werte vor (1,7 oder 2,.4), so daß Veranstaltungen, die dieselben Durchschnittsnote erhielten, auf einen gemittelten Rangplatz gesetzt werden, während in unserer eigenen Erhebung genauere Zahlen vorliegen und somit nur ganze Rangplätze vergeben werden – dadurch entstehen in der Differenzenbildung halbe Werte.

Differenz der Ränge der Note LV (O vs FS)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	-14,5	2	4,3	4,9	4,9
	-13,5	1	2,1	2,4	7,3
	-12,0	1	2,1	2,4	9,8
	-10,0	1	2,1	2,4	12,2
	-9,5	2	4,3	4,9	17,1
	-8,5	3	6,4	7,3	24,4
	-7,5	1	2,1	2,4	26,8
	-7,0	1	2,1	2,4	29,3
	-6,5	2	4,3	4,9	34,1
	-6,0	3	6,4	7,3	41,5
	-5,5	1	2,1	2,4	43,9
	-5,0	1	2,1	2,4	46,3
	-4,0	2	4,3	4,9	51,2
	-3,5	1	2,1	2,4	53,7
	-3,0	2	4,3	4,9	58,5
	-1,5	3	6,4	7,3	65,9
	-,5	2	4,3	4,9	70,7
	,0	1	2,1	2,4	73,2
	1,0	1	2,1	2,4	75,6
	3,5	1	2,1	2,4	78,0
	5,5	1	2,1	2,4	80,5
	6,0	1	2,1	2,4	82,9
	6,5	1	2,1	2,4	85,4
	7,5	3	6,4	7,3	92,7
	8,0	1	2,1	2,4	95,1
	13,5	1	2,1	2,4	97,6
	19,5	1	2,1	2,4	100,0
	Gesamt	41	87,2	100,0	
Fehlend	System	6	12,8		
Gesamt		47	100,0		

Aufgrund der von vielen Seiten geübten Kritik an einer solchen Erstellung detaillierter Ranglisten (bei einer Massierung von Werten in einem bestimmten Wertebereich führen minimale Meßdifferenzen zu erheblichen Rangdifferenzen) geht man häufig dazu über, statt individueller Rangplätze Ranggruppen zu bilden. Wie die folgende Tabelle zeigt, stimmen aber auch bei dieser wesentlich groberen Zuordnung[17] lediglich 19 der 41 Zuordnungen in der Online- und der FS-Erhebung überein; in zwei Fällen liegt sogar eine Abweichung von zwei bzw. drei Klassen vor. (Selbst wenn wir die Zahl der Ranggruppen auf drei reduzieren: Spitzengruppe – Mittel-

[17] Hier wurden die Ränge in Quintile aufgeteilt und die Zuordnung der Veranstaltungen zu diesen Ranggruppen in den beiden Erhebungen zueinander in Beziehung gesetzt.

gruppe – Schlußgruppe, steigt die Übereinstimmung der Zuordnungen nur von 46% auf 68%, wird also ein Drittel der Veranstaltungen falsch bewertet.)

Ranggruppe nach Note für LV (FS) * Ranggruppe nach Note für LV (O) Kreuztabelle

Anzahl

		Ranggruppe nach Note für LV (O)					
		beste Gruppe	2	3	4	schlechteste Gruppe	Gesamt
Ranggruppe nach Note für LV (FS)	beste Gruppe	6	1	1	0	0	8
	2	4	3	2	0	1	10
	3	0	2	1	3	0	6
	4	0	0	6	3	1	10
	schlechteste Gruppe	0	0	0	1	6	7
Gesamt		10	6	10	7	8	41

Fazit zur Zuverlässigkeit der Ergebnisse der Online-Erhebung

Zu welcher Einschätzung können wir aufgrund der hier vorgelegten Ergebnisse zum Vergleich von Online-Erhebung und FS-Erhebung kommen?

- für ein Drittel der Veranstaltungen weist die Online-Erhebung eine so geringe Beteiligung der Studierenden auf, daß ihre Ergebnisse schon aus diesem Grund unzuverlässig sind bzw. gar keine Bewertung vorliegt;
- der Vergleich der durchschnittlichen Bewertungen in den beiden Bewertungsverfahren belegt, daß die online gewonnenen Ergebnisse nicht zuverlässig sind und somit in die Irre führen: durchgängig weichen sie in allen Einzelfragen und in allen Veranstaltungen von den Ergebnissen der FS-Erhebung ab, bei einem Drittel der Veranstaltungen mit den Effekt einer systematischen Über- bzw. Unterschätzung der tatsächlichen Bewertungslage;
- auch die Verwendung der Ergebnisse in Form einer Rangreihe zeitigt irreführende Ergebnisse der Online-Erhebung, die zudem keine systematische Tendenz erkennen lassen, daher auch nicht in ihrem Ausmaß und ihrer Richtung zu korrigieren sind.

Zum Effekt der unterschiedlichen Ausschöpfungsquote von Online- und FS-Erhebung

Wie wir sahen, realisieren Online-Erhebung und FS-Erhebung ganz unterschiedliche Ausschöpfungsquoten: 38% vs. 81%. Seitens des Forschungsseminars war es beabsichtigt, den Effekt der Ausschöpfungsquote zu kontrollieren, indem wir die Ergebnisse der Online-Erhebung mit den Ergebnissen nur derjenigen Studierenden in der FS-Erhebung verglichen, die an beiden Erhebungen teilgenommen haben. Zu diesem Zweck haben wir eine Frage in den FS-

Fragebogen aufgenommen, in dem die Beteiligung an der Online-Erhebung erfaßt werden sollte. Leider haben sich die dort gemachten Angaben als unzuverlässig erwiesen. Zwar ist eine völlige Übereinstimmung aufgrund möglichen Fehlens in der Veranstaltung zum Zeitpunkt der FS-Erhebung gar nicht zu erwarten – doch sollte die gemeinsame Teilnehmerzahl nicht so weit auseinander gehen, daß mögliche Differenzen in den Ergebnissen auf diese Differenzen in der Teilnahme zurückgehen können. In neun Veranstaltungen gaben aber bei der FS-Befragung zwischen vier und neun Studierende weniger an, online teilgenommen zu haben, als dort tatsächlich dokumentiert sind, und in zwei weiteren Veranstaltungen fehlen sogar 12 bzw. 42 Teilnehmer; außerdem kreuzten in 11 Veranstaltungen zwischen 1 und 3 Personen mehr, als tatsächlich online teilgenommen hatten, ihre Teilnahme an der Online-Erhebung an (was möglicherweise auf Erinnerungsprobleme aufgrund der Teilnahme an mehreren Veranstaltungsbewertungen zurückzuführen ist). Angesichts dieser Divergenzen aber ist die beabsichtigte Prüfung des Effektes der Ausschöpfungsquote leider nicht möglich.

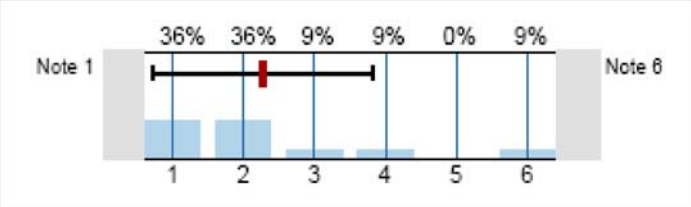
Wer beteiligt sich an der Online-Erhebung?

Einer der eingangs zitierten Einwände gegen eine Online-Erhebung bezog sich auf eine mögliche Selektivität in der Beteiligung der Studierenden: Gibt es eine systematische Verzerrung der Ergebnisse, weil sich bestimmte Gruppen von Studierenden eher als andere der Mühe einer Teilnahme an dieser Befragung unterziehen? Wegen des Fehlens von Informationen über die online teilnehmenden Studierenden können wir keine Aussage darüber machen, ob es sich eher um Hauptfach- oder Nebenfachstudierende handelt, ob ihr Veranstaltungsbesuch eher aus thematischem Interesse oder aufgrund einer Vorschrift der Prüfungsordnung erfolgt, etc. In bezug auf die Frage einer möglichen Verzerrung der Ergebnisse ist aber – und dies dürfte auch die wesentliche interessantere Fragestellung sein – möglich zu prüfen, ob in der Online-Erhebung (im Vergleich zur FS-Erhebung) extreme Bewertungen stärker vertreten sind als mittlere Bewertungen, oder, schärfer formuliert: ob sich eher die Fans oder eher die Kritiker der Lehrenden zur Teilnahme aufgerufen fühlen, oder ob diese beiden extremen Gruppen teilnehmen, nicht aber die „schweigende Mehrheit".

Im ersteren Fall wäre die Bewertung schlechter, als sie der Bewertung durch die Gesamtheit der Teilnehmerinnen entspricht, im zweiten besser; im dritten Fall könnte (bei gleicher Repräsentanz der beiden Extreme) derselbe Mittelwert in beiden Erhebungen herauskommen, allerdings wäre die Streuung der Werte größer (weil die „ausgleichende" Mitte fehlt).

a) Die Extreme: „Fans" und „Revanchisten"

Wenden wir uns als erstes der heiklen Frage zu, inwiefern die Lehrveranstaltungsevaluation von einigen Studierenden dazu genutzt wird, aus – wie auch immer begründeter Frustration – den Lehrenden „eins auszuwischen". Von mehreren Kollegen bin ich darauf angesprochen worden, daß es in den Verteilungen der Antworten in der Online-Befragung auffällige Abweichungen gegeben habe. Während in diesen Veranstaltungen der Schwerpunkt der Verteilung bei den guten Noten lag, gab es in diesen Fällen in (fast) allen Dimensionen zwischen ein und drei Nennungen, die vom Schwerpunkt weit entfernt im negativen Notenfeld lagen, wie es sich im folgenden Beispiel widerspiegelt:



Um diese Frage zu beantworten, haben wir zwei Auswertungen vorgenommen:

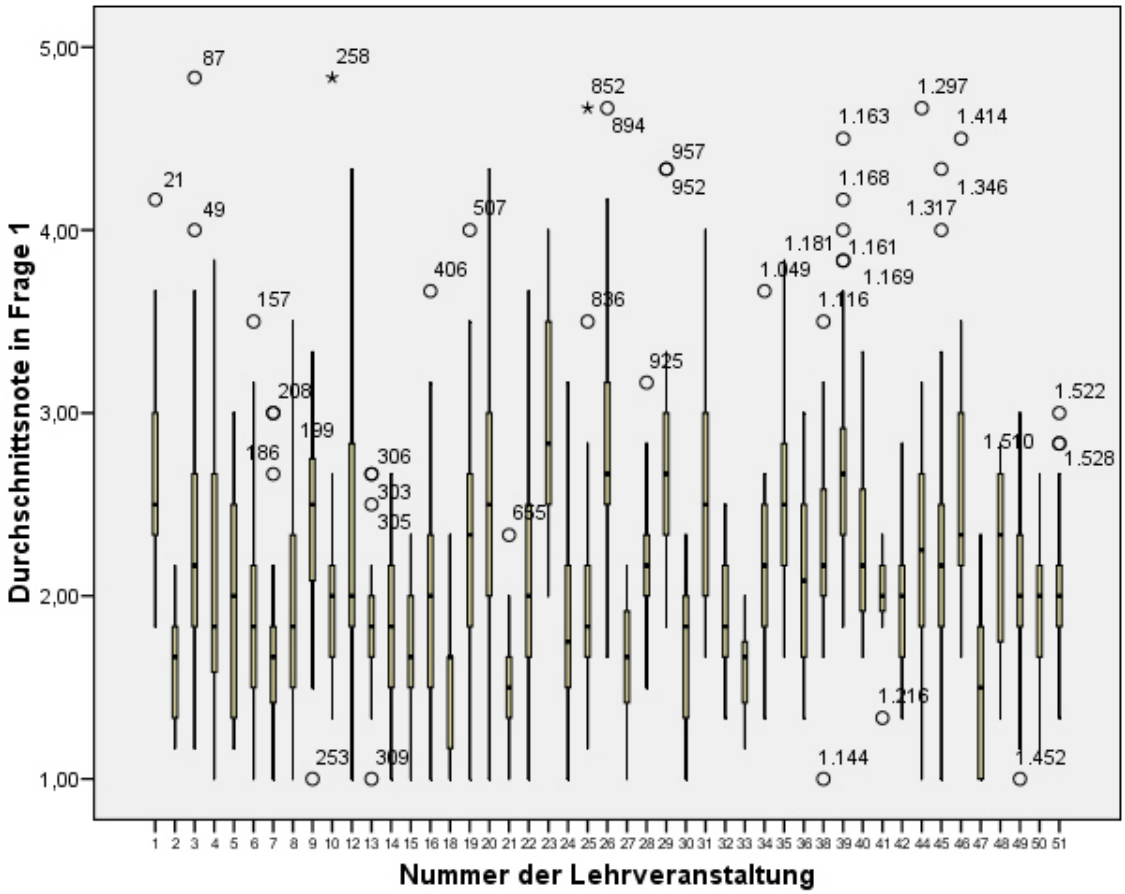
- wir haben die Ergebnisberichte der Studiendekane daraufhin überprüft, ob ein solches Antwortmuster vorliegt (daß also fast alle Bewertungen positiv sind und nur wenige Antworten konstant im extrem negativen Notenbereich liegen);[18] [19]
- anhand der FS-Daten konnten wir über die Daten der Verteilung hinaus prüfen, ob einzelne Veranstaltungsteilnehmer konstant negative Voten abgegeben haben, die von der Bewertung der anderen kategorial abweichen.

[Die Auswertung der Online-Daten steht noch aus.]

[18] Anhand dieser Daten der Online-Erhebung können wir allerdings nicht entscheiden, ob es sich bei diesen abweichenden schlechten Bewertungen immer um dieselben ein bis drei Personen handelt, da uns für den Online-Datensatz die fallbezogenen Daten nicht zur Verfügung stehen. Die uns vorliegenden Informationen beschränken sich auf den Ergebnisbericht, wie er den einzelnen Lehrenden seitens der Studiendekane mitgeteilt wird (und an uns weitergeleitet wurde). Dies umfaßt die Mittelwerte für die einzelnen Fragen, die Zahl der Antworten, die Streuung und die Verteilung auf die einzelnen Antwortkategorien.

[19] In Veranstaltungen, die allgemein ein größeres Bewertungsspektrum aufweisen, sind solche Abweichungen auf der Basis nur der Verteilungsdaten nicht zu identifizieren; auch fallen sie bei der Durchschnittsberechnung kaum ins Gewicht, weil sie weniger stark vom Schwerpunkt der restlichen Bewertungen abweichen.

Anhand der Daten aus der FS-Erhebung konnte fallbezogen die durchschnittliche Bewertung der Dimensionen in Frage 1 bestimmt werden.[20] Das nachfolgend dargestellte Boxplot gibt für die einzelnen Veranstaltungen die Streuung der Bewertungen wieder. Dabei zeigt der mittlere dickere Balken die Verteilung der mittleren 50% (den Median) an, die feinen Linien stehen für die oberen bzw. unteren 25%, und die Kreise bzw. Sterne stehen für „Ausreißer" bzw. „Extremwerte" (die Zahlen geben die Fallnummer wieder).



Die Grafik zeigt sehr schön, daß es in der Tat individuelle Abweichungen vom Schwerpunkt der Bewertungen durch die anderen Veranstaltungsteilnehmerinnen gibt. Dabei überwiegen die Abweichungen in Richtung einer negativen Beurteilung: Nur fünf „Fans" (mit einem abweichenden „sehr gut") stehen 34 Studierende gegenüber, die – statistisch gesehen – abweichend schlecht bewertet haben. Dies umfaßt allerdings auch diejenigen Fälle, die bei einer sehr gut

[20] Diese Berechnung bezieht die Fragen nach dem Engagement der Kommilitoninnen und der Betreuung außerhalb der Veranstaltung nicht ein. Ersteres gehört nicht zu den „Erfolgskriterien" für den Dozenten, letzteres wurde von sehr vielen Befragten nicht beantwortet, weil diese Leistung nicht in Anspruch genommen wurde. (Das Engagement der Kommilitoninnen wurde von den extrem schlechten Bewerterinnen übrigens tendenziell besser bewertet als von den übrigen Veranstaltungsteilnehmerinnen.)

bewerteten Veranstaltung durchschnittlich ein „schlechtes gut" oder ein „befriedigend" gegeben haben (z.B. die Fälle 305, 303, 306, oder 685). Auf mögliche „Revanchegelüste" lassen aber nur Fälle schließen, die in ihrem Durchschnittswert im Notenbereich von 4,5 bis 5 liegen und sich zugleich weit von den anderen Werten absetzen (z.B. die Fälle 87, 258, 852 etc.): dieser Gruppe sind lediglich zwölf Fälle zuzurechnen.

Für sechs Veranstaltungen, die aufgrund der Stärke oder der Zahl der Abweichungen besonders betroffen sind, soll im folgenden geprüft werden, inwieweit diese extremen Bewertungen die Durchschnittsnote beeinflussen.

Die erste Tabelle gibt den Mittelwert wieder, der auf der Basis aller Fragebögen der jeweiligen Veranstaltung berechnet wurde, sowie die Zahl der Studierenden in dieser Veranstaltung und die Standardabweichung.[21]

Durchschnittsnote in Frage 1, alle Fälle

Nummer der Lehrveranstaltung	Mittelwert	N	Standardabweichung
10	2,1444	15	,82102
34	2,2000	15	,58146
39	2,7760	32	,67682
44	2,3043	46	,61494
45	2,1958	80	,59225
46	2,5648	18	,71661
Insgesamt	2,3390	206	,66776

Bei der Berechnung der Werte in der zweiten Tabelle wurden die im Boxplot zu erkennenden Extremfälle nicht berücksichtigt. Ein Vergleich der Mittelwerte zeigt eine Veränderung zwischen 0,05 und 0,2 Notenpunkte; deutlich spiegelt sich der Ausschluß dieser Fälle in einer drastischen Reduktion der Standardabweichungen.

Durchschnittsnote in Frage 1, ohne Extremwerte

Nummer der Lehrveranstaltung	Mittelwert	N	Standardabweichung
10	1,9524	14	,36061
34	2,0952	14	,43222
39	2,5833	28	,45700
44	2,2519	45	,50705
45	2,1453	78	,50564
46	2,4510	17	,54571
Insgesamt	2,2415	196	,51508

Zusammenfassend ist festzuhalten:

- stark abweichende (und damit vermutlich unfaire) Bewertungen kommen durchaus vor;
- in der FS-Erhebung sind sie auf wenige Veranstaltungen mit relativ geringen Auswirkungen auf das Ergebnis beschränkt;
- in der Online-Erhebung tritt dieses Phänomen vermutlich sowohl häufiger als auch mit

[21] Bei der Standardabweichung handelt es sich um ein Maß, das den Grad der Streuung der Einzelwerte um das arithmetische Mittel wiedergibt. Je größer dieser Wert ist, umso unterschiedlicher sind die Bewertungen jeweils ausgefallen.

stärkeren Abweichungen auf, und dort kann es auch zu stärkeren Verzerrungen in den Bewertungen führen als in den FS-Daten, da diese Abweichungen aufgrund der geringeren Fallzahl Online stärker ins Gewicht fallen.[22]

b) Die „schweigende Mehrheit"

Darüber hinaus bestätigt sich die Vermutung, daß sich die Fans und die Kritiker in größerem Ausmaß an der Online-Erhebung beteiligen als diejenigen, die mit ihren Bewertungen eher im Mittelfeld liegen. Diese Vermutung läßt sich mit einem Vergleich der Streuung der Antworten in den beiden Erhebungen belegen. Wenn sich die „Extreme" stärker an einer Befragung beteiligen als die „Mitte", dann führt dies zu einer stärkeren Streuung der Bewertungen, die sich in einem höheren Wert für die Standardabweichung niederschlägt – und genau dies ist hier der Fall. Ziehen wir die Werte der Variationskoeffizienten[23] in der FS-Erhebung von denen der Online-Erhebung ab, so ist die Differenz bei allen Veranstaltungen positiv: die Online-Streuung ist also größer als die FS-Streuung. Allerdings gibt es auch hier keine Systematik in der Abweichung: die Differenzen variieren zwischen 3% und 44%, und es läßt sich – entgegen unserer Erwartung – keine Beziehung zur Größe oder zum Typ der Veranstaltung feststellen: auch hier scheint der Zufall den Grad der Beteiligung an der Online-Befragung zu bestimmen.

6. Einflußfaktoren auf die Bewertung der Lehrveranstaltungen

Wie eingangs dargelegt, wird die Bewertung von Lehrveranstaltungen auch von Faktoren beeinflußt, auf die die Lehrenden keinen oder nur einen sehr begrenzten Einfluß haben. Zu denken ist hier etwa an organisatorische Rahmenbedingungen der Lehre (Räumlichkeit, zeitliche Organisation), an Aspekte der Lehrveranstaltung (Veranstaltungstyp, Teilnehmerzahl, Teilnahmepflicht), an persönliche Charakteristika der Lehrenden (Status, Geschlecht), an Eigenschaften der Teilnehmerinnen (Vorkenntnisse, Arbeitsaufwand) u.ä.

---

[22] „Vermutlich", weil wir eben keine fallbezogenen Daten vorliegen haben. Eine statistische Lösung für die durch diese Extremwerte bedingte Verzerrung könnte in der Berechnung getrimmter Mittelwerte bestehen, bei denen an beiden Seiten einer Verteilung ein gleich großer Prozentsatz der Fälle aus der Berechnung ausgeschlossen wird – angesichts einer relativ geringen Online-Beteiligungsquote würde dies insbesondere bei kleineren Veranstaltungen allerdings zu einer problematischen Reduktion der Fallzahlen führen.

[23] Als Basis für die Berechnung der Streuungsdifferenzen wurde hier der Variationskoeffizient und nicht die Standardabweichung gewählt, weil letztere in ihrer absoluten Höhe vom arithmetischen Mittel abhängt, dieses aber – wie wir sahen – in Online- und FS-Erhebung unterschiedliche Werte annimmt. Der Variationskoeffizient dagegen berechnet sich als Quotient von Standardabweichung und Mittelwert und neutralisiert damit den Effekt unterschiedlicher Mittelwerte: er stellt ein Maß *relativer* Streuung dar.

Diese Faktoren werden in der Literatur zumeist als „externe" Variablen behandelt, d.h. als Faktoren, die einen „verzerrenden" Einfluß auf die Bewertung einer Veranstaltung ausüben. Diese Bezeichnung verführt aber zu einer verkürzten Perspektive auf den Prozeß der Sicherung von Qualität in der Lehre. Nur in einem reinen Modell einer Lehrveranstaltung, in der der Lehrende ungestört von Rahmenbedingungen sein Lehrkonzept verwirklichen könnte (also auch ohne Studierende!), sind diese Variablen „extern". Dieser luftleere Lehrraum, in dem das Ergebnis der Lehre nur abhängig ist vom „Input" des Lehrenden, existiert in der Praxis nicht. Tatsächlich stellt jeder erfahrene Lehrende genau diese Faktoren bei der Konzeption seiner Lehrveranstaltung in Rechnung, muß er sich im Bemühen um die Umsetzung dieser Konzeption mit der durch diese Umstände geschaffenen spezifischen Konstellation auseinandersetzen und daraus resultierende Probleme lösen. Dies aber heißt: für verschiedene Typen, Größen und Zusammensetzungen von Lehrveranstaltungen bestehen unterschiedliche Handlungsbedingungen, die ein spezifisches Lehrkonzept der Lehrenden, damit aber auch ein je spezifisch zugeschnittenes Erhebungsinstrument erfordern – zumindest aber eine Berücksichtigung dieser Faktoren in der Bewertung der Lehrveranstaltungsbewertung. Jedes andere Vorgehen wird gerade denjenigen Veranstaltungen nicht gerecht, die ihre Lehre unter didaktisch erschwerten Bedingungen leisten müssen. Will man den Lehrenden nicht nur eine Rückkoppelung über das Erleben ihrer Veranstaltung durch die Studierenden geben und es dabei ihnen überlassen, in Kenntnis der besonderen Umstände ihrer Veranstaltung eine implizite Kontextuierung dieser studentischen Bewertungen vorzunehmen, sondern den – ja regelmäßig auch praktizierten –Vergleich zwischen den verschiedenen Veranstaltungen auf eine sachlich begründete Basis stellen, so ist eine Berücksichtigung der verschiedenen Einflußfaktoren auf die berichteten Bewertungen unverzichtbar.

Wenn diese Variablen einen Einfluß auf die Bewertung ausüben (also z.B. Veranstaltungen am Freitagnachmittag oder große Vorlesungen schlechter bewertet werden sollten), so läßt sich dies daran erkennen, daß die durchschnittliche Bewertung für diese Veranstaltungen schlechter ist als für Veranstaltungen zu anderen Terminen. Mit diesem „tendenziell" stehen wir angesichts unseres Datensatzes aber vor einem Problem: In ihm sind zwar 1568 Studierende repräsentiert, aber „nur" 47 Veranstaltungen. Haben wir also nur wenige Veranstaltungen an einem Freitag, so ist eine wie auch immer geartete Bewertung nicht zwingend diesem Kriterium „Termin" zuzurechnen: die Terminierung kann ohne Bedeutung sein und die Bewertung tatsächlich nur die Qualität der Lehre wiederspiegeln. Dies bedeutet, daß wir einigermaßen gesicherte Aussagen über den Einfluß solcher Faktoren nur dann vornehmen können, wenn die Besetzung der einzelnen Ausprägungen „groß genug" ist. Wir werden im einzelnen zu diskutieren haben, ob solche Aussagen zu rechtfertigen sind.

An dieser Stelle ist eine auswertungstechnische Anmerkung erforderlich. Wir haben die *Studierenden* befragt, aber wir wollen – auf der Basis ihrer Antworten – *Aussagen über die Veranstaltungen* machen. Da die Bewertungen der einzelnen Veranstaltungen mal nur auf den Antworten von 7, mal auf denen von 118 Teilnehmerinnen beruhen, sind die Aussagen über die Veranstaltungen nicht ohne weitere Bearbeitung aus den Antworten der Studierenden zu gewinnen: die großen Veranstaltungen würden mit ihrer Teilnehmerzahl die kleineren dominieren. Aus diesem Grunde ist es erforderlich, die Aussagen der Befragten mit der Zahl der Teilnehmerinnen an der jeweiligen Veranstaltung zu gewichten: dies neutralisiert die Größe der Veranstaltung. Allerdings ist damit eine etwas ungewohnte Konstellation der Zahlen in den Tabellen verbunden: als absolute Werte tauchen hier je nach Auswertungsmodell keine ganzen Zahlen auf, vielmehr wird die Stärke der Besetzung in den einzelnen Zellen als gewichteter Anteilswert an der Gesamtheit der Befragten ausgedrückt (und dieser addiert sich – von Rundungsungenauigkeiten abgesehen – auf die Zahl der Veranstaltungen: i.d.R. auf 47). Die Prozentzahlen sind wie gewohnt als (gewichteter) Anteil an den jeweiligen (Unter-)Gruppen zu interpretieren.

Der Einfluß organisatorischer Faktoren

Raumnot ist ein vertrautes Phänomen an der Universität. Dies führt zum einen dazu, daß verfügbare Räume nicht groß genug sind, daß die Akustik schlecht ist, oder daß Veranstaltungen in zeitlichen Randlagen stattfinden, die entweder aufgrund der Lebensumstände der Studierenden (abendlicher Nebenerwerb ...) oder aufgrund langer Anfahrtswege nicht sehr beliebt sind. Im FS-Fragebogen waren die Studierenden gebeten worden, die Eignung des Raumes für die Veranstaltung in bezug auf „Ausstattung, Größe, Akustik ..." zu beurteilen (s. Frage 12, Anhang).

Betrachtet man die Bewertung der *Eignung der Veranstaltungsräume*, so äußern sich in 52% der Veranstaltungen die Studierenden positiv oder sehr positiv, in 30% halten sie den Raum für „durchschnittlich geeignet", und in 18% für „eher nicht" bzw. „gar nicht geeignet".[24] Betrachten wir die Bewertungen veranstaltungsbezogen, so zeigt sich eine starke Varianz der Einschätzungen zwischen, aber auch innerhalb der einzelnen Veranstaltungen:

– in vier Veranstaltung liegt der Anteil der sehr guten oder guten Bewertungen unter 10%, in

[24] Diese Bewertung scheint bereits ein Zugeständnis an die Realität der Raumsituation zu beinhalten und spiegelt die Tatsache, daß die Studierenden sich mit den beschränkten Gegebenheiten der universitären Ausstattung arrangieren, denn in den offenen Antworten wird z.T. herbe Kritik an den Räumlichkeiten geübt, und in früheren Befragungen an der Philosophischen Fakultät I schnitten die Räumlichkeiten mit Abstand am schlechtesten ab.

anderen reicht er bis 95%;

– auch ist die Wahrnehmung der Studierenden z.T. sehr unterschiedlich: in 17 Veranstaltungen erstrecken sich die Urteile über vier der fünf möglichen Kategorien, in 16 sogar über alle fünf Kategorien von „sehr gut" bis „gar nicht geeignet".

Betrachten wir die Auswirkungen der so bewerteten Raumsituation auf die Veranstaltungsbewertung, so zeigt sich bei allen Teilfragen von Frage 1 eine Korrelation der Art, daß schlechte Raumbewertungen mit schlechteren Veranstaltungsbewertungen einhergehen.[25]

Durchschnittliche Bewertung der Teilfragen von Frage 1 in Abhängigkeit von der Bewertung der Raumeignung

Eignung des Raums		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommiliton Innen	Note der LV insgesamt (FB der Dekane)
sehr gut geeignet	Mittelwert	2,12	2,09	2,25	1,80	1,59	1,74	2,81	2,07
	N	8	8	8	8	8	6	8	8
gut geeignet	Mittelwert	2,25	2,18	2,44	1,85	1,74	1,98	2,92	2,27
	N	16	16	16	16	16	11	15	16
durchschnittlich geeignet	Mittelwert	2,26	2,18	2,42	1,87	1,65	1,95	2,93	2,23
	N	14	14	14	14	14	9	13	14
eher nicht geeignet	Mittelwert	2,46	2,50	2,75	2,05	1,93	2,14	2,91	2,43
	N	6	6	6	6	6	4	6	6
gar nicht geeignet	Mittelwert	2,74	2,72	3,13	2,40	1,77	2,50	3,27	2,87
	N	2	2	2	2	2	2	2	2
Insgesamt	Mittelwert	2,28	2,23	2,47	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	47	32	44	47

Zur Interpretation der Tabelle:
Die „Mittelwerte" geben jeweils an, wie z.B. der Aufbau der Veranstaltung von denjenigen bewertet wurde, die den Raum für „sehr gut geeignet halten": sie vergaben durchschnittlich die Note 2,12. Wurde der Raum von den Studierenden als „gut geeignet" eingestuft, verschlechtert sich die Note auf 2,25, bis hin zu einer Note von 2,72 von denjenigen Studierenden, die den Raum für „gar nicht geeignet" hielten. Die Zahlen in der Zeile „N" stellen das (gerundete) rechnerische Äquivalent für die Zahl der Studierenden dar, die den Raum in dieser Weise bewerten und aus deren Antworten der Mittelwert bestimmt wurde. (Aufgrund der Gewichtung mit der Veranstaltungsgröße addieren sich alle Antworten jeweils auf die Zahl der in die Rechnung eingegangenen Veranstaltungen).

Zusammenfassend ist also festzustellen, daß die Differenzen der Mittelwerte zwischen den

[25] Die Frage nach der „Betreuung durch die Dozenten außerhalb der Lehrveranstaltung" wurde von sehr vielen Studierenden nicht beantwortet, so daß die Summe der Antworten immer deutlich niedriger ist als in den anderen Variablen.

Untergruppen bei allen Fragen einen – z.T. ausgeprägten – Einfluß der Raumbewertung auf die Veranstaltungsbewertung mit Notendifferenzen bis zu 0,9 Punkten belegen.

Für eine Überprüfung der These über den ungeliebten *Freitag* bzw. *Montag* sind die Besetzungen dieser beiden Tage in unserer Stichprobe zu gering, um zuverlässige Aussagen vornehmen zu können.[26]

Auch die *Tageszeit* scheint für die Bewertung keine Rolle zu spielen: Die Bewertungen aller Veranstaltungen zu den Terminen zwischen 10 Uhr und 16 Uhr (hier liegen jeweils die Daten von mindestens acht Veranstaltungen vor) sind weitgehend gleich (für die Randtermine vor 10 und nach 18 Uhr sind die Besetzungen zu gering für eine zuverlässige Aussage).[27]

Auswirkungen von Charakteristika der Lehrenden

Weitere Faktoren, auf die die Lehrenden (zumindest kurzfristig) keinen Einfluß haben, betreffen ihren akademischen Status und ihr Geschlecht. Dennoch sind sie potentielle Einflußfaktoren für die Bewertung: Professoren könnten die Studierenden qua Status eine höhere Kompetenz und damit „bessere" Lehrveranstaltungen zusprechen / wissenschaftliche Mitarbeiter könnten dies durch eine größere soziale Nähe ausgleichen oder gar überkompensieren; Frauen könnten mit größerer kommunikativer Kompetenz punkten; männliche Studierende eher die Leistung von männlichen Lehrenden honorieren, weibliche Studierende sich von weiblichen Lehrenden eher akzeptiert fühlen.

Betrachten wir zunächst den Status der Lehrenden, so sind Professoren und Privatdozenten einerseits, wissenschaftliche Mitarbeiter andererseits mit 23 bzw. 24 Personen gleich stark vertreten. Die Differenzen zwischen ihren Bewertungen sind nicht sehr groß (zwischen 0,0 Notenpunkten bei der Durchführung der Lehrveranstaltungen und 0,26 in der Gesamtnote für die Veranstaltung), aber sie gehen alle in dieselbe Richtung: Professoren und Privatdozenten werden durchgängig etwas besser bewertet. Dies bedeutet natürlich nicht, daß *alle* Professoren und Privatdozenten bessere Bewertungen bekommen haben als alle wissenschaftlichen Mitarbeiter: die individuellen Bewertungen der Studierenden reichten in beiden Gruppen von sehr gut bis (selten) ungenügend. Ob diese Differenzen nun eine Frage des Images sind oder ob hier

[26] Wären die Daten zuverlässig zu interpretieren, so sprechen die vorliegenden Daten dafür, daß der Freitag der beste Tag ist, um eine gute Bewertung zu bekommen!

[27] Tendenziell am besten haben auch hier die drei Veranstaltungen abgeschnitten, die um 8 Uhr morgens begannen. Mag es an den Dozenten liegen oder an der Auswahl besonders interessierter Studierender zu diesem Zeitpunkt: Die Uhrzeit determiniert jedenfalls nicht die Bewertung.

tatsächlich Leistungsdifferenzen zugrundeliegen, ist an dieser Stelle nicht zu entscheiden.[28]

Ein ähnliches Bild zeigt sich bei der Überprüfung des Einflusses der Variablen „Geschlecht". An der FS-Erhebung nahmen 15 weibliche und 31 männliche Lehrende teil. Die Effekte dieser Variablen sind etwas stärker als die des Status (zwischen 0,1 bei der Betreuung und 0,45 bei der Verständlichkeit), und auch sie weisen in allen Teilfragen in dieselbe Richtung: die weiblichen Lehrenden erhalten durchgängig die besseren Bewertungen.

Durchschnittliche Bewertung der Teilfragen von Frage 1 in Abhängigkeit von Status und Geschlecht der Lehrenden

Status der Lehrenden	Geschlecht der Lehrenden		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der KommilitonInnen	Note der LV insgesamt (FB der Dekane)
Profess. / PD	weiblich	Mittelwert	2,06	1,95	2,31	1,73	1,62	1,93	2,74	2,04
		N	10	10	10	10	10	6	9	10
	männlich	Mittelwert	2,29	2,23	2,45	1,93	1,78	1,91	2,78	2,22
		N	13	13	13	13	13	10	12	13
wiss. Mitarbeiter	weiblich	Mittelwert	2,00	1,84	2,23	1,58	1,65	1,80	2,93	2,01
		N	5	5	5	5	5	3	5	5
	männlich	Mittelwert	2,46	2,45	2,64	2,03	1,73	2,01	3,08	2,47
		N	18	18	18	18	18	12	18	18

Allerdings sind die Variablen Status und Geschlecht nicht unabhängig voneinander: zwei Drittel der weiblichen Lehrenden sind Professorinnen bzw. Privatdozentinnen, während es bei den männlichen nur 42% sind. Die nachfolgende Tabelle zeigt, daß dem Geschlecht der Lehrenden eine größere Bedeutung für die studentischen Bewertungen zukommt als dem Status. Bei Kontrolle des Einflusses des Status bleiben die Differenzen zwischen den Geschlechtern weitgehend erhalten (in der Gruppe der wissenschaftlichen Mitarbeiter werden sie sogar noch größer und steigen bei der Verständlichkeit bis auf 0,6 Notenpunkte[29]), während bei Kontrolle des Geschlechtes die Differenzen zwischen den Statusgruppen geringer werden (ohne Tabelle).[30] Während bei „Raum" und „Zeit" nicht die Gefahr bestand, daß sich „hinter" diesen beiden Variablen tatsächliche Leistungsunterschiede verbergen, ist dies bei den Variablen „Status" und „Geschlecht" durchaus möglich: Professoren und Frauen können möglicherweise tatsächlich bessere Veranstaltungen anbieten als ihr jeweiliges Pendant – auch dies ist hier nicht

[28] Siehe hierzu die multivariate Analyse am Ende dieses Kapitels.

[29] Allerdings ist die Gruppe der wissenschaftlichen Mitarbeiterinnen sehr klein, ist somit stärker von den Werten einzelner Lehrender abhängig; aus diesem Grund ist diese Ergebnis nur begrenzt zu verallgemeinern.

[30] Erwähnt sei, daß diese Differenzen insbesondere bei der Verständlichkeit, beim Aufbau der Veranstaltung, beim Engagement der Dozenten und bei den Hilfsmitteln auftreten.

zu entscheiden.

Auswirkungen von Charakteristika der Lehrveranstaltungen

Unter Lehrenden ist es eine selbstverständliche Erfahrung, daß unterschiedliche *Typen* von Lehrveranstaltungen ganz unterschiedliche Anforderungen an den Lehrenden stellen:

- die Möglichkeiten für die Gestaltung der Lehre in einem Seminar sind andere als in einer Vorlesung;
- eine große Veranstaltung schafft andere Bedingungen als eine kleine;
- eine Pflichtveranstaltung wird von den Studierenden anders angenommen als ein Seminar, das aus Interesse besucht wird.

Prüfen wir im folgenden, ob und ggf. wie sich diese Faktoren auf die Bewertung durch die Studierenden auswirken. Die FS-Erhebung umfaßt 9 Vorlesungen, 13 Proseminare, 16 Hauptseminare, 6 Seminare, 2 Übungen und ein Mittelseminar. Diese Veranstaltungen haben wir nach ihrer Größe und ihrem Typus gruppiert.[31] Wie die nachfolgende Tabelle zeigt, weisen diese Typen deutlich unterschiedliche Teilnehmerzahlen aus.

Zahl der durchschnittlich ausgefüllten Fragebögen

Größe und Typ der Lehrveranstaltung	Mittelwert	N	Standardabweichung
Hauptseminar -25	17,00	10	5,121
Hauptseminar 26+	34,33	6	7,815
Proseminar -30	18,19	16	7,054
Proseminar 30+	34,50	6	3,082
Vorlesung - 60	49,67	3	7,024
Vorlesung 61-80	71,67	3	3,215
Vorlesung 81+	96,67	3	18,583
Insgesamt	32,51	47	23,640

Prüfen wir die durchschnittlichen Bewertungen für diese Typen von Veranstaltungen, so schneiden die Hauptseminare in fast allen Teilfragen besser ab als die Vorlesungen,[32] und diese werden zumeist besser bewertet als die (Pro-)Seminare. Nehmen wir die Größe der Veranstaltungen als weiteres Unterscheidungsmerkmal hinzu, so werden die kleineren Proseminare (mit

[31] Dabei wurden Seminare, Übungen und Sprachkurs mit den Proseminaren zusammengefaßt, da zwischen ihnen strukturell die größten Gemeinsamkeiten bestehen. (Die drei Vorlesungen mit 61 bis 80 Teilnehmerinnen liegen so nahe beieinander, daß ihre Aufteilung auf die kleinere bzw. größere Kategorie nicht zu rechtfertigen war.)

[32] Der Aufbau der Veranstaltung wird bei beiden gleich gut bewertet.

Differenzen bis zu 0,8 Notenpunkten deutlich, die kleineren Vorlesungen nicht immer und nur mit geringen Differenzen besser bewertet als die entsprechenden größeren Veranstaltungen desselbenTyps. Bei den Hauptseminaren wirken sich die Größenunterschiede kaum aus.

Durchschnittliche Bewertung der Teilfragen von Frage 1 in Abhängigkeit vom Veranstaltungstypus

Größe und Typ der Lehrveranstaltung		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitonlnnen	Note der LV insgesamt (FB der Dekane)
Hauptseminar -25	Mittelwert	2,15	2,08	2,20	1,79	1,65	1,82	2,69	2,11
	N	10	10	10	10	10	9	10	10
Hauptseminar 26+	Mittelwert	2,23	2,15	2,27	1,52	1,68	1,71	3,09	2,17
	N	6	6	6	6	6	5	6	6
Prosemina r -30	Mittelwert	2,26	2,10	2,43	1,78	1,57	1,95	2,83	2,20
	N	16	16	16	16	16	11	16	16
Prosemina r 30+	Mittelwert	2,84	2,91	2,98	2,48	1,94	2,25	3,06	2,80
	N	6	6	6	6	6	3	6	6
Vorlesung - 60	Mittelwert	2,06	2,04	2,42	2,03	1,87	2,37	2,76	2,18
	N	3	3	3	3	3	1	2	3
Vorlesung 61-80	Mittelwert	2,19	2,44	2,54	1,99	1,83	2,13	3,22	2,29
	N	3	3	3	3	3	1	2	3
Vorlesung 81+	Mittelwert	2,18	2,27	2,98	2,28	2,03	2,76	3,63	2,45
	N	3	3	3	3	3	1	2	3
Insgesamt	Mittelwert	2,28	2,23	2,47	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	47	32	44	47

Nach aller Erfahrung dürfte es für die Bewertung einer Veranstaltung auch von Bedeutung sein, ob die Studierenden sie besuchen, weil es sich um *Pflichtveranstaltungen* handelt, oder ob sie die Veranstaltung aus *Interesse am Thema* besuchen.[33] Für 45% der Besucherinnen der 47 Veranstaltungen ist die Pflicht das alleinige Besuchsmotiv – 24% sind nur aus Interesse in der Veranstaltung, und 32% weisen eine gemischte Motivation auf. Dabei variieren die Motive mit dem Typ der Veranstaltung: bei den großen Proseminaren und Vorlesungen liegt der Pflichtanteil durchschnittlich bei 64% bzw. 68% (mit einem Anteil der „Nur-Interessierten" von 4%

[33] Frage 23 differenziert zwischen Pflicht, Wahlpflicht, Empfehlung und Interesse als Motiven für einen Veranstaltungsbesuch, wobei diese Motive sich natürlich nicht wechselseitig ausschließen. In der folgenden Auswertung werden drei Typen unterschieden: im ersten Typus werden diejenigen Studierenden zusammengefaßt, die nur aus Pflicht- oder Wahlpflicht-Motiven die Veranstaltung besuchen – dem wird die Gruppe derjenigen gegenübergestellt, die nur aus Interesse für das Thema die Veranstaltung besuchen – und zwischen ihnen steht die Gruppe derjenigen, die mehrere dieser Gründe angekreuzt haben (25 Studierende, die nur auf Empfehlung gekommen sind, werden ebenfalls dieser mittleren Gruppe zugeordnet).

bzw. 6%), während er bei den Hauptsseminaren bei etwa 20%, bei den kleineren Proseminaren und Vorlesungen zwischen 30% und 50% liegt.

Der unterschiedlichen Motivation zum Veranstaltungsbesuch entsprechen in allen Bewertungsdimensionen der Frage 1 unterschiedliche Durchschnittswerte: Interessierten Studierenden erscheint die Veranstaltung z.B. besser strukturiert, verständlicher, und Dozent und Kommilitoninnen erscheinen ihnen engagierter, so daß auch die Durchschnittsnote um eine halbe Note besser ausfällt als die durchschnittliche Bewertung der Studierenden, die die Veranstaltung nur aus Pflichtgründen besuchen. Bedenkt man die sehr unterschiedliche Verteilung des positiv stimmenden Motivs „aus Interesse" (es variiert zwischen 0% und 100%) bei den untersuchten Veranstaltungen, so liegt es nahe anzunehmen, daß Pflichtveranstaltungen tendenziell schlechtere Bewertungen erhalten als solche Veranstaltungen, in denen das Interesse am Thema dominiert – unabhängig von der Lehrleistung des einzelnen Lehrenden.

Durchschnittliche Bewertung der Teilfragen aus Frage 1 in Abhängigkeit vom Pflichtcharakter einer Veranstaltung

Pflichtcharakter der Veranstaltungen		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der KommilitonInnen	Note der LV insgesamt (FB der Dekane)
(Wahl) Pflichtveranstaltung	Mittelwert	2,47	2,60	2,76	2,13	1,84	2,24	3,09	2,59
	N	17	17	17	17	17	11	16	17
(Wahl)Pflicht oder Empfehlung oder Interesse	Mittelwert	2,27	2,14	2,27	1,85	1,68	1,87	2,77	2,15
	N	15	15	15	15	15	10	14	15
Besuch aus Interesse	Mittelwert	2,07	1,88	2,35	1,65	1,60	1,74	2,88	2,02
	N	14	14	14	14	14	9	13	14
Insgesamt	Mittelwert	2,28	2,24	2,48	1,89	1,71	1,97	2,92	2,28
	N	46	46	46	46	46	31	43	46

Auswirkungen von Charakteristika der Studierenden

Drastischer noch fallen die Differenzen zwischen den Extremen aus, wenn wir uns nur auf die inhaltliche Dimension des **Interesses am Thema der Veranstaltung** beziehen[34] (also keinen Bezug zu institutionellen Gründen für den Veranstaltungsbesuch in der Frage herstellen, wie es in Frage 23 der Fall ist). In den Dimensionen Verständlichkeit, Qualität der Hilfsmittel, Betreuung durch den Dozenten wie auch in der Gesamtnote liegt die Differenz bei einem ganzen Notenschritt, z.T. auch darüber!

Durchschnittliche Bewertung der Teilfragen von Frage 1 in Abhängigkeit vom vorgängigen Interesse der Studierenden

Thema hat von Anfang an interessiert		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitonlnnen	Note der LV insgesamt (FB der Dekane)
trifft völlig zu	Mittelwert	2,14	1,96	2,25	1,78	1,59	1,79	2,75	2,02
	N	18	18	18	18	18	12	17	18
2	Mittelwert	2,35	2,29	2,57	1,96	1,81	2,06	2,97	2,38
	N	14	14	14	14	13	9	13	13
3	Mittelwert	2,35	2,40	2,58	1,93	1,79	2,05	3,03	2,38
	N	8	8	8	8	8	5	7	8
4	Mittelwert	2,36	2,41	2,51	1,98	1,70	2,12	3,06	2,40
	N	4	4	4	4	4	3	3	4
5	Mittelwert	2,45	2,69	2,73	2,06	1,69	1,91	3,09	2,64
	N	2	2	2	2	2	1	2	2
trifft überhaupt	Mittelwert	2,62	3,15	3,30	2,32	1,94	2,85	3,18	2,99
	N	1	1	1	1	1	1	1	1
Insgesamt	Mittelwert	2,29	2,24	2,47	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	46	31	43	46

Das vorgängige Interesse am Thema der Veranstaltung beeinflußt also die Bewertung der Veranstaltung in einem erheblichen Ausmaß. Bei der Betrachtung der Auswirkungen dieses Faktors auf die Bewertung soll allerdings ein sehr positives Ergebnis der Frage nach dem Interesse nicht unter den Tisch fallen: 64% der Befragten haben mit den beiden ersten Kategorien ein ausgeprägtes thematisches Interesse bekundet – 28% wählten die beiden mittleren Kategorien – nur 8% zeigten sich völlig desinteressiert. Dies stellt für die Lehrveranstaltungen grundsätzlich eine gute Motivationsbasis dar. Das Problem für eine Evaluation liegt allerdings in der sehr ungleichen Verteilung dieser 8%: während in 27 Veranstaltungen der Anteilswert dieser Desinteressierten unter 5% liegt, überschreitet er in 8 Veranstaltungen den Wert von 15%

[34] Vgl. Frage 8.1: „Das Thema dieser Veranstaltung hat mich vor Beginn der Veranstaltung sehr interessiert" – eine sechsstufige Stellungnahme zwischen „trifft völlig zu" und „trifft überhaupt nicht zu" war möglich.

(mit einem Maximum bei 36%) – für diese Veranstaltungen ist die Ausgangslage nicht gar so gut.

Die Chance auf eine gute Bewertung steigt ebenfalls, wenn an einer Veranstaltung besonders viele Studierende teilnehmen, die das betreffende Fach als ihr **Wunschfach** bezeichnen: sie benoten durchschnittlich um etwa 0,2 bis 0,3 Notenpunkte besser als diejenigen, für die das jeweilige Fach nur eine gute Ergänzung zu ihrem Hauptfach darstellen, die es gar nur als eine Verlegenheitslösung gewählt haben bzw. für die diese Veranstaltung nur eine Pflichtveranstaltung eines anderen Studienfaches darstellt.[35]

Auch ist die Bewertung in starkem Maße abhängig von den **Vorkenntnissen**, die die Studierenden in die Veranstaltung einbringen. In den Dimension Veranstaltungsaufbau, Verständlichkeit, Hilfsmittel und Gesamtnote verbessert sich die Bewertung zwischen einer halben und einer ganzen Note im Vergleich der beiden Gruppen, die der Aussage „Für das Verständnis des Stoffes waren meine Vorkenntnisse eine gute Basis" völlig zustimmen bzw. die ihr überhaupt nicht zustimmen können.[36]

Durchschnittliche Bewertung der Teilfragen von Frage 1 in Abhängigkeit von den Vorkenntnissen der Studierenden

Vorkenntnisse waren gute Basis		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitoninnen	Note der LV insgesamt (FB der Dekane)
trifft völlig zu	Mittelwert	2,01	1,75	2,26	1,73	1,71	1,73	2,83	2,02
	N	7	7	7	7	7	4	6	7
2	Mittelwert	2,15	1,99	2,32	1,80	1,63	1,81	2,91	2,09
	N	12	12	12	12	12	9	11	12
3	Mittelwert	2,33	2,32	2,44	1,87	1,77	2,17	2,82	2,33
	N	12	12	12	12	12	8	11	12
4	Mittelwert	2,54	2,56	2,64	2,00	1,72	2,09	2,92	2,48
	N	6	6	6	6	6	4	6	6
5	Mittelwert	2,37	2,62	2,58	2,08	1,74	1,97	3,07	2,42
	N	5	5	5	5	5	3	5	5
trifft überhaupt nicht zu	Mittelwert	2,60	2,73	2,91	2,17	1,73	2,34	3,10	2,62
	N	3	3	3	3	3	2	3	3
Insgesamt	Mittelwert	2,28	2,23	2,45	1,89	1,71	1,98	2,90	2,27
	N	44	44	44	44	44	31	42	44

[35] Der eigentlich beabsichtigte Vergleich zwischen Hauptfach- und Nebenfachstudierenden macht wenig Sinn, weil es nur wenige Nebenfachstudierende in unserer Stichprobe gab und die Hälfte von ihnen im Bachelorstudiengang eingeschrieben ist, in dem diese Differenz nur noch von geringer Bedeutung ist.

[36] Vgl. Frage 8.6 im Anhang.

Eine systematische Beziehung besteht auch zwischen der Bewertung und der Regelmäßigkeit der studentischen **Veranstaltungsvorbereitung**: in allen Dimensionen verbessert sich die Bewertung mit jeder Zunahme der Vorbereitung, mit Differenzen zwischen 0,2 und 1,0 Notenpunkten. Kritisch muß hier angemerkt werden, daß der Anteil derjenigen, die sich regelmäßig vorbereiten, mit 35% weit unterhalb der für einen Lernerfolg anzusiedelnden Eigenarbeit der Studierenden liegt – rein zahlenmäßig wird die Bewertung einer Veranstaltung aber von den zwei Drittel derjenigen dominiert, die sich nur hin und wieder oder gar nicht vorbereiten.

Durchschnittliche Bewertung der Teilfragen aus Frage 1 in Abhängigkeit von der Häufigkeit der Vorbereitung

Häufigkeit der Vorbereitung		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitoninnen	Note der LV insgesamt (FB der Dekane)
(fast) nie vorbereitet	Mittelwert	2,59	2,47	3,07	2,16	1,82	2,10	3,10	2,55
	N	10	10	10	10	10	6	9	10
hin und wieder vorbereitet	Mittelwert	2,33	2,32	2,52	1,95	1,79	2,04	3,01	2,36
	N	17	17	17	17	17	12	16	17
meistens vorbereitet	Mittelwert	2,21	2,14	2,19	1,74	1,61	1,84	2,83	2,13
	N	11	11	11	11	11	8	10	11
(fast) immer vorbereitet	Mittelwert	1,93	1,92	2,04	1,70	1,57	1,89	2,68	1,97
	N	9	9	9	9	9	6	9	9
Insgesamt	Mittelwert	2,28	2,24	2,47	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	47	32	44	47

Auffällig ist in diesem Zusammenhang,daß fehlende Vorkenntnisse in der Regel nicht zu einer Steigerung der Vorbereitung führen. Die nachfolgende Tabelle zeigt vielmehr, daß gerade diejenigen, die sich selbst mangelnde Vorkenntnisse attestieren, eher noch auf eine regelmäßige Vorbereitung verzichten.[37]

[37] Diese Tabelle ist auf der Basis der Individualdaten berechnet, da eine Gewichtung nach Veranstaltungsgröße hier nicht erforderlich ist: es geht nicht um die Bewertung der Veranstaltungen, sondern um eine von der Veranstaltungsgröße unabhängige Beziehung zwischen diesen Variablen.

Häufigkeit der Vorbereitung in Abhängigkeit von den Vorkenntnissen

			Vorkenntnisse waren gute Basis für Verständnis						
			trifft völlig zu	2	3	4	5	trifft überhaupt nicht zu	Gesamt
Häufigkeit der Vorbereitung	(fast) nie vorbereitet	Anzahl	53	123	92	49	51	38	406
		%	26,9%	32,3%	24,1%	23,6%	33,1%	38,4%	28,6%
	hin und wieder vorbereitet	Anzahl	70	128	143	98	53	30	522
		%	35,5%	33,6%	37,5%	47,1%	34,4%	30,3%	36,8%
	meistens vorbereitet	Anzahl	40	79	80	38	30	21	288
		%	20,3%	20,7%	21,0%	18,3%	19,5%	21,2%	20,3%
	(fast) immer vorbereitet	Anzahl	34	51	66	23	20	10	204
		%	17,3%	13,4%	17,3%	11,1%	13,0%	10,1%	14,4%
Gesamt		Anzahl	197	381	381	208	154	99	1420
		%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Auswirkungen von Veranstaltungscharakteristika, die sich aus der Interaktion von Lehrenden und Studieren ergeben

Nicht alle Aspekte einer Veranstaltung lassen sich eindeutig dem einen oder anderen der am Lernprozeß beteiligten Partner zuordnen. Dazu gehören z.B. die Wahrnehmung der Anforderungen (der Lehrenden) für die Vorbereitung auf die Veranstaltung durch die Studierenden sowie ihr Gefühl der Überforderung bzw. Unterforderung durch diese Leistungsanforderungen. Diese Aspekte entwickeln sich aus den Vorgaben der Lehrenden auf der einen Seite und dem, wie die Studierenden damit umgehen. Alle drei Aspekte stehen aber in einem engen Wechselwirkungsverhältnis mit der Bewertung der Veranstaltungen.

Mit Ausnahme der Verständlichkeit, die – nicht überraschend – von denjenigen am schlechtesten bewertet wird, die die Anforderungen in der Vorbereitung auf die Veranstaltung als hoch oder sehr hoch einschätzen,[38] werden alle Dimensionen der Frage 1 von denjenigen am besten bewertet, die die Anforderungen für angemessen halten – in wechselndem Ausmaß weichen diejenigen, die sie als (sehr) hoch bzw. als (sehr) niedrig erleben, von diesen positiveren Einschätzungen ab (wobei in der einen Dimension die als hoch erlebten Anforderungen, in anderen die als niedrig erlebten Anforderungen mit einer schlechteren Bewertung einhergehen. Es ist also durchaus nicht so, daß als niedrig erlebte Anforderungen durchgängig mit positiven Bewertungen „belohnt" würden (vgl. unten).

[38] Vgl. Frage 6 im Anhang.

Durchschnittliche Bewertung der Teilfragen aus Frage 1 in Abhängigkeit von den Anforderungen an die Vorbereitung

Anforderungen an die Vorbereitung		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der KommilitonInnen	Note der LV insgesamt (FB der Dekane)
sehr hoch	Mittelwert	2,31	2,68	2,28	1,95	1,83	2,24	2,91	2,39
	N	2	2	2	2	2	1	2	2
hoch	Mittelwert	2,29	2,62	2,52	1,97	1,68	2,00	2,93	2,44
	N	9	9	9	9	9	6	9	9
angemessen	Mittelwert	2,22	2,11	2,35	1,83	1,70	1,94	2,88	2,19
	N	30	30	30	30	30	21	28	30
niedrig	Mittelwert	2,65	2,04	2,84	2,04	1,86	1,94	3,05	2,45
	N	3	3	3	3	3	2	3	3
sehr niedrig	Mittelwert	2,36	1,97	3,45	1,92	1,80	2,15	3,26	2,24
	N	1	1	1	1	1	0	1	1
Insgesamt	Mittelwert	2,27	2,22	2,44	1,88	1,71	1,97	2,91	2,26
	N	45	45	45	45	45	31	42	45

Die mit Abstand stärkste Auswirkung einer Variablen auf die Bewertung der Veranstaltungen finden wir im Gefühl, in der Veranstaltung **inhaltlich überfordert** zu werden.[39] Hier liegt in der Dimension der Verständlichkeit eine Differenz von 1,9 Notenschritten zwischen der Gruppe derjenigen, die sich „nie" überfordert fühlen, und denjenigen, die „oft" das Gefühl der Überforderung erleben[40] – und auch in den anderen Dimensionen liegen ungewöhnlich große Differenzen vor. Die Gesamtnote differiert in diesem Bereich um einen Notenschritt.

[39] Vgl. Frage 2.

[40] Die Kategorie „fast immer" wurde nur von 16 Personen angekreuzt – in Relation zur Gesamtheit erscheint in dieser Zeile eine „0" für die gewichtete Anzahl, da diese Personen weniger als ein Veranstaltungsäquivalent ausmachen – vgl. den Interpretationshinweis am Anfang dieses Kapitels zur Tabelle über die Auswirkungen der Eignung des Raumens.

Durchschnittliche Bewertung der Teilfragen aus Frage 1 in Abhängigkeit vom Gefühl der Überforderung

Inhaltliche Überforderung		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitonlnnen	Note der LV insgesamt (FB der Dekane)
nie	Mittelwert	2,12	1,62	2,31	1,75	1,66	1,77	2,86	2,03
	N	10	10	10	10	10	7	9	10
selten	Mittelwert	2,19	2,01	2,44	1,85	1,67	1,88	2,83	2,13
	N	19	19	19	19	19	13	18	19
gelegentlich	Mittelwert	2,41	2,67	2,52	1,98	1,75	2,13	2,95	2,45
	N	14	14	14	14	14	10	13	14
oft	Mittelwert	2,72	3,49	2,82	2,31	1,90	2,42	3,40	3,01
	N	3	3	3	3	3	2	3	3
fast immer	Mittelwert	3,51	4,00	4,26	2,94	2,36	2,61	3,56	3,68
	N	0	0	0	0	0	0	0	0
Insgesamt	Mittelwert	2,29	2,24	2,48	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	47	32	44	47

Aber auch eine Unterforderung zahlt sich für den Dozenten nicht in einer positiven Bewertung aus: mit Ausnahme der Verständlichkeit sind die Bewertungen in den Kategorien „nie” und „selten” durchgängig (z.T. erheblich) besser als in denen, in denen, die eine Unterforderung signalisieren.

Durchschnittliche Bewertung der Teilfragen aus Frage 1 in Abhängigkeit vom Gefühl der Unterforderung

Inhaltliche Unterforderung		Note für Aufbau	Note für Verständlichkeit der LV	Qualität der Hilfsmittel	Engagement des Dozenten	Durchführung der LV	Betreuung durch den Dozenten außerhalb der LV	Engagement der Kommilitonlnnen	Note der LV insgesamt (FB der Dekane)
nie	Mittelwert	2,18	2,31	2,42	1,82	1,60	1,92	2,86	2,25
	N	20	20	20	20	20	13	18	20
selten	Mittelwert	2,25	2,22	2,43	1,91	1,75	2,03	2,86	2,22
	N	18	18	18	18	17	12	16	18
gelegentlich	Mittelwert	2,50	2,05	2,59	1,99	1,82	1,94	3,06	2,33
	N	8	8	8	8	8	6	8	8
oft	Mittelwert	2,83	2,35	3,06	2,54	2,15	2,08	3,74	2,90
	N	1	1	1	1	1	1	1	1
fast immer	Mittelwert	4,20	2,08	4,02	2,20	3,11	2,73	4,33	3,04
	N	0	0	0	0	0	0	0	0
Insgesamt	Mittelwert	2,28	2,24	2,47	1,90	1,71	1,97	2,92	2,27
	N	47	47	47	47	47	32	44	47

Interessanterweise sind Überforderung und Unterforderung keineswegs komplementär zueinander: fast alle Kombinationen der Ausprägungen kommen vor, und mit einem Somers d_{yx} von −0,15 ist die Übereinstimmung zwischen den beiden Antworten erstaunlich gering. Dies bedeutet, daß innerhalb derselben Veranstaltung von denselben Personen sowohl Erfahrungen der Über- wie der Unterforderung gemacht werden.[41]

Kombination der Antworten zur inhaltlichen Überforderung und Unterforderung

Anzahl

		Inhaltliche Überforderung					
		nie	selten	gelegentlich	oft	fast immer	Gesamt
Inhaltliche Unterforderung	nie	114	240	217	68	11	650
	selten	100	249	177	44	3	573
	gelegentlich	67	118	60	5	1	251
	oft	25	14	2	3	0	44
	fast immer	5	0	1	0	0	6
Gesamt		311	621	457	120	15	1524

Einflußfaktoren auf die Bewertung: Eine Zusammenfassung der bivariaten Ergebnisse

Zusammenfassend ist festzuhalten, daß alle hier geprüften Einflußfaktoren – wenn auch in unterschiedlichem Maße – mit der Bewertung der Veranstaltungen korrelieren. Im einzelnen bedeutet dies, daß eine Veranstaltung tendenziell schlechter bewertet wird, wenn

- der Veranstaltungs*raum* als ungeeignet erfahren wird,
- die Veranstaltung von *männlichen* Lehrenden durchgeführt (insbesondere dann, wenn es sich um wissenschaftliche *Mitarbeiter* handelt),
- es sich um ein *größeres Proseminar* oder eine *größere Vorlesung* handelt,
- es sich um eine *Pflicht*veranstaltung handelt,
- es sich um *Nebenfach*studierende handelt,
- *nicht* das *Wunschfach* studiert wird,
- die Teilnehmerinnen vorab *kein Interesse* am Thema der Veranstaltung haben,
- die *Vorkenntnisse* der Teilnehmerinnen keine gute Basis darstellen,
- die Studierenden sich nur bedingt auf die Veranstaltung *vorbereiten*,
- die *Anforderungen* an die Vorbereitung als zu hoch oder als zu niedrig erlebt werden,
- die Studierenden sich *überfordert* bzw. *unterfordert* fühlen.

[41] Auch diese Tabelle wurde auf der Basis der Individualdaten berechnet, da eine Gewichtung nach Veranstaltungsgröße hier nicht erforderlich ist: es geht nicht um eine Veranstaltungsbewertung, sondern um eine von der Veranstaltungsgröße unabhängige Beziehung zwischen diesen Variablen.

Diese Ergebnisse zeigen eines ganz deutlich: die Antworten der Studierenden auf die Fragen in einer Lehrveranstaltungsbewertung sind keineswegs als unverzerrte Information über die Leistung der Lehrenden zu interpretieren. Im Gegenteil sind sie in einem erheblichen Ausmaß von Faktoren abhängig, auf die Lehrenden keine bzw. nur begrenzte Einwirkungsmöglichkeiten haben. Am Beispiel des Veranstaltungstypus etwa ist deutlich zu erkennen, daß gerade diejenigen Veranstaltungen, die unabhängig von der Person des Lehrenden bereits aufgrund ihrer strukturellen Besonderheiten (Teilnehmerzahl, Vorbereitungsaufwand, Anforderungen an die Durchführung ...) erhöhte Anforderungen an die Lehrenden stellen, in der Bewertung durchschnittlich schlechter abschneiden.

Einflußfaktoren auf die Bewertung der Lehrveranstaltungen: Eine multivariate Betrachtung

In der bisherigen Analyse wurde der Einfluß dieser Faktoren auf die Bewertung jeweils isoliert für sich (also in bivariater Analyse) betrachtet – tatsächlich aber bestehen zwischen diesen einzelnen Faktoren vielfältige Wechselwirkungen: so z.B. zwischen Veranstaltungstypus, Pflichtcharakter und Interesse; zwischen Vorkenntnissen, Anforderungen, Vorbereitung, Über- und Unterforderung. Diese Wechselwirkungen führen dazu, daß der Einfluß des einen Faktors in einem anderen bereits enthalten sein kann, aber auch: daß sie sich wechselseitig verstärken können. Um daher ein realistisches Bild der Beziehung zwischen diesen Variablen zu erhalten, ist es notwendig, die bisher nur bivariat betrachteten Beziehungen in einem umfassenderen multivariaten Modell zu analysieren.

Als ein solches Modell bietet sich eine multivariate Regressionsanalyse an. Mit ihrer Hilfe lassen sich drei Aussagen machen:

- Zum einen erlaubt sie es, die Wechselwirkungen zwischen den Einflußvariablen zu neutralisieren, so daß der Gesamteinfluß dieser „externen" Variablen auf die Bewertung der Lehrveranstaltung bestimmt werden kann; damit ist zu erkennen, wie groß der Beitrag dieser Variablen zu dem Entstehen von Unterschieden in der Bewertung der Veranstaltungen ist – und wie groß der Beitrag ist, der auf die uns ja letztlich nur interessierende Lehrleistung des einzelnen Lehrenden zurückzuführen ist.[42]
- Im Vergleich der sogenannten standardisierten Regressionskoeffizienten ist auch die relative Bedeutsamkeit bestimmen, die diesen „externen" Faktoren in ihrem Einfluß auf die Bewertung jeweils zukommt.
- Schließlich ermöglicht die Regressionsanalyse die Berechnung der Bewertungen unabhängig vom Einfluß dieser Einflußfaktoren. Erst auf Basis dieser Werte ist ein gerechter Vergleich

[42] Zur Bezeichnung dieser Variablen als „extern" vergleiche die Ausführungen zu Beginn von Kapitel 6.

der unter z.T. sehr unterschiedlichen Bedingungen erbrachten Lehrleistungen möglich: die „Begünstigung", wie sie im Durchschnitt der Veranstaltungen z.B. mit einem kleinen Hauptseminar verbunden ist, ist hier ebenso neutralisiert wie die „Benachteiligung", die aus einem hohen Anteil von inhaltlich nicht interessierten „Pflichtteilnehmerinnen" resultiert.[43]

Methodische Zwischenbemerkung zur Regressionsanalyse

Für eine solche Regressionsanalyse sind die Daten aufzubereiten. Um die in der bivariaten Analyse betrachteten Variablen in eine Regressionsanalyse einzubeziehen, müssen sie entweder metrisches Meßniveau besitzen, oder sie müssen als Dummy-Variablen kodiert werden. Ersteres kann als gegeben betrachtet werden, letzteres wurde durchgeführt.[44]

Zu klären ist aber zuvor noch, welche Variable denn nun als „Bewertung der Lehrveranstaltung" (also als zu erklärende Variable) gewählt werden soll. In der bivariaten Analyse hatten wir jeweils alle Teilfragen der Frage 1 zum jeweiligen Einflußfaktor in Beziehung gesetzt. Dies würde aber den Analyse-(und Lektüre-)aufwand verachtfachen, zumal nicht zu erwarten ist, daß die Teildimensionen zu wesentlich differenten Ergebnissen führen werden: In der bivariaten Analyse zeigte sich, daß die Tendenz der Bewertung in allen Teilfragen der Frage 1 fast immer in dieselbe Richtung ging. Diese Teilfragen hatten wir unserer Analyse zugrundegelegt, da sie die Basis für die Lehrbewertungen der Studiendekane darstellen – und um die Prüfung der Aussagefähigkeit dieser Lehrbewertung geht es uns ja. Tatsächlich kommt aber nicht allen dieser sieben Teilfragen dasselbe Gewicht in der Bewertung der Lehrqualität zu, und drei von ihnen sind sachlich eher als unangemessen zu bezeichnen.

Letzteres betrifft zum einen die Frage 1.5: „Wie beurteilen Sie die Durchführung der Lehrveranstaltung durch den Dozenten bzw. die Dozentin (z.B. Regelmäßigkeit, Pünktlichkeit, usw.?" Die Frage wurde von Studierenden als unscharf bezeichnet: „Durchführung" ist begrifflich wesentlich weiter als „Regelmäßigkeit" oder „Pünktlichkeit", und was unter „usw" zu fassen ist, bleibt offen. In der Diskussion der Evaluationsergebnisse in einer Veranstaltung, die nur einmal wegen Krankheit ausfiel und immer

---

[43] In einer Institution wie der Universität dürfte sich der Hinweis erübrigen, daß auch diese Werte nur eine Annäherung an die Realität der unverzerrten Bewertung der Lehre darstellen können: Faktoren, die in der Erhebung nicht berücksichtigt wurden/nicht berücksichtigt werden konnten (Aspekte der Persönlichkeit der Lehrenden wie der Studienmotivation der Studierenden, Faktoren wie Sympathie, Sozialkompetenz, fachliche Kompetenz etc.), wurden nicht erhoben, sind in dieser Berechnung folglich nicht berücksichtigt, sind also auch nicht rechnerisch auszuschließen.

[44] „Metrisches Meßniveau" bedeutet, daß die Abstände zwischen den einzelnen Meßpunkten auf der Meßskala gleich groß sind. Dies ist z.B. bei allen Fragen, in denen Antworten der Art „nie" – „selten" – „gelegentlich" tc. anzukreuzen waren, nicht der Fall. Inwieweit dieses Meßniveau bei Notenskalen zu unterstellen ist, wird in der Literatur strittig diskutiert. Nach meinen eigenen Testrechnungen führt die Behandlung der Notenskalen als metrische Variablen nicht zu Verzerrungen in den Ergebnissen (so bleibt die Relation der metrischen und ordinalen Korrelationskoeffizienten für dieselben Variablensätze identisch), so daß ich diese Variablen als metrische Variablen behandeln werde. Die ordinalen Variablen (Überforderung, Raumeignung etc.) wurden von mir in Dummy-Variablen transformiert. In dieser Form können diese Variablen in eine Regressionsanalyse eingeschlossen werden. Vgl. Backhaus u.a., Multivariate Analysemethoden, 1996, 1-55.

pünktlich begann und pünktlich endete, aber als Bewertung nur eine „2" erhielt, gaben Studierende als Erklärung zur Antwort: „Da guckt man gar nicht so genau hin: Man kreuzt den Wert an, den man insgesamt für die Veranstaltung für angemessen hält" – zumindest dieser Frage fehlt also aufgrund ihrer Vagheit die dimensionale Trennschärfe.

Mit der Frage 1.6: „Wie beurteilen Sie die Betreuung durch den Dozenten bzw. die Dozentin außerhalb der Lehrveranstaltung?" wiederum wußte nur ein Teil der befragten Studierenden etwas anzufangen: Während die Verweigerungsrate für die Teilfragen 1.1 bis 1.4 zwischen 0,5% und 5,8% liegt, beträgt sie hier 40%. Die Frage 1.7 schließlich: „Wie beurteilen Sie das Engagement Ihrer Kommilitoninnen und Kommilitonen in der Lehrveranstaltung?" bezieht sich, wenn überhaupt, nur mit einem sehr geringen Teil auf eine Leistung des Dozenten – im Gegenteil scheint sie eher dazu geeignet, eine schwierige Rahmenbedingung für erfolgreiche Lehre zu identifizieren.[45]

Auf diese drei Teilfragen ist also im folgenden zu verzichten, wenn es um ein zusammenfassendes Kriterium für die Lehrbewertung geht. Mit den verbleibenden vier Teildimensionen „Aufbau der Lehrveranstaltung", „Verständlichkeit", „Hilfsmittel", „Engagement des Dozenten" sind zentrale Aspekte einer Lehrveranstaltung erfaßt. Grundsätzlich besteht die Möglichkeit, für jede dieser verbliebenen vier Fragen eine Regressionsrechnung durchzuführen, die den Einfluß der unabhängigen Variablen auf jede einzelne dieser Dimensionen bestimmt. Da aber, wie erwähnt, die Bewertung der Einzelfragen fast immer gleichgerichtet ist, scheint es sinnvoller, einen anderen Weg zu wählen, der die Information der Einzelfragen zu einer einzigen Variablen zusammenfaßt. Dies ist möglich, indem aus den Antworten auf diese vier Teilfragen der Mittelwert für jede Befragte einer jeden Veranstaltung berechnet wird, so daß wir den Einfluß der unabhängigen Variablen auf die durchschnittliche Bewertung dieser vier Dimensionen bestimmen können.[46]

Außerdem haben wir noch die die Einzelbewertungen zusammenfassende Frage 1.8: „Insgesamt beurteile ich die Lehrveranstaltung mit der Note: ...". Sie in die Durchschnittsberechnung einzubeziehen ist nicht sinnvoll, da sie ihrerseits bereits die von den Befragten intuitiv vorgenommene Durchschnitts-

[45] Bei genauerer Betrachtung leistet sie aber auch dies nicht. Es zeigt sich bei allen Auswertungen, daß Studierende, die sich nicht vorbereiten, die kein Interesse am Thema haben etc. das Engagement ihrer Kommilitoninnen positiver einschätzen als die interessierten ... Studierenden. Hier liegt ganz offensichtlich ein aus der Sozialpsychologie bekanntes Phänomen vor, daß die Wahrnehmung der Umwelt durch die Brille der eigenen Einstellung – und insofern systematisch verzerrt – erfolgt. Diese Variable gibt also nicht (nur) eine objektive Bedingung der Lehre wieder, sondern ist (auch) abhängig vom Anteil der Desinteressierten in der Veranstaltung: je höher deren Anteil, desto höher wird – entgegen der Realität – das Engagement der Kommilitoninnen eingeschätzt.

[46] Anzumerken ist, daß diese Art der Mittelwertbildung über verschiedene Dimensionen in der Praxis der Lehrveranstaltungsbewertungen zwar üblich, in der methodischen Literatur dagegen durchaus umstritten ist. Wie wir aber im Vergleich der Teilfrage 1.8 (zusammenfassende Bewertung der Lehrveranstaltung) und dem Globalindikator der Studiendekane eine Korrelation von r = 0,97 beobachtet haben, scheint diese Zusammenfassung der Einzeldimensionen auch nicht völlig in die Irre zu führen. Sie hat den Vorteil, daß sie auf einer überschaubaren Zahl von Dimensionen beruht, deren Gewichtigkeit (vielleicht mit Ausnahme der Frage nach den Hilfsmitteln) als zentral für jede Lehrveranstaltung zu betrachten ist.

bildung darstellt.[47]

Die nachfolgende Regressionsrechnung bezieht sich somit auf den Durchschnitt der Noten, die jede einzelne Studierende in den ersten vier Teilfragen der Frage 1 für die von ihr besuchte Veranstaltung vergeben hat.

Im folgenden werde ich zunächst darauf eingehen, a) wie groß der gemeinsame Einfluß der hier berücksichtigten „externen" Faktoren auf die Bewertung der Veranstaltungen ist, b) werde dann prüfen, welche Bedeutsamkeit jedem einzelnen der vier Faktoren zukommt, um c) abschließend der Frage nachzugehen, inwieweit diese Faktoren zu einer Verzerrung der Veranstaltungsbewertung führen – also einen relevanten Einfluß auf die Bewertung der Lehrveranstaltung ausüben, die eine Interpretation der Bewertungen auf der Basis unkorrigierter Durchschnittswerte verbietet.

a) Der gemeinsame Einfluß der „externen" Faktoren

Wie die nachfolgende Tabelle zeigt, erhalten wir für die Beziehung zwischen dieser summierenden Bewertung der Veranstaltungen und den zwölf in Kapitel 5 behandelten „externen" Faktoren einen Wert von r = 0,53.[48] Das Quadrat des r-Wertes zeigt an, daß 30% der Unterschiede in der Bewertung der Lehrveranstaltungen auf den gemeinsamen Einfluß der hier berücksichtigten Faktoren zurückzuführen sind. Damit wirken in die Bewertung der Lehrleistung in erheblichem Ausmaß Variablen ein, die sich nicht auf die Lehrleistung selber beziehen.

[47] In einer späteren Analyse soll allerdings der Frage nachgegangen werden, inwieweit möglicherweise in einer solchen „globalen" Frage andere Faktoren, die hier nicht berücksichtigt wurden, stärker zum Tragen kommen können, als es in den auf konkrete Dimensionen zielenden Teilfragen der Fall ist. Diese Unsicherheit ist auch der Grund, warum diese Globalfrage nicht zum Bezugspunkt der nachfolgenden Analyse gemacht wurde.

[48] Bei „r" handelt es sich um den Produktmomentkorrelationskoeffizienten, mit dessen Hilfe die Stärke der Beziehung zwischen metrischen Variablen gemessen wird. Auch er bewegt sich zwischen 0 bei fehlendem Zusammenhang und 1, wenn eine perfekte Übereinstimmung zwischen den Variablen vorliegt – wenn also in diesem Fall z.B. die Bewertung der Veranstaltungen völlig aus der Kenntnis der „externen" Faktoren zu prognostizieren wäre.

Erklärungsleistung der Rahmenbedingungen

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,544 [a]	,296	,005	,68572

a. Einflußvariablen : (Konstante), Unterforderung (nein), Geschlecht der Lehrenden (weiblich), Veranstaltungstypus (HS, klein PS u Vorl), Anforderungen (angemessen), Wunschfach (Wunschfach), Eignung des Raumes (geeignet), Vorkenntnisse (gute), Interesse als Motiv (Interesse), Vorbereitung (regelmäßig), Status der Lehrenden (Hochschullehrer), Überforderung (nein), Pflicht als Motiv (keine Pflicht)

b) Die Bedeutsamkeit der einzelnen Faktoren

Über die Stärke des Einflusses der *einzelnen* Faktoren informieren uns die Daten der nachfolgenden Tabelle. Den Zahlen der ersten Spalte (nicht-standardisierte Koeffizienten / B) ist z.B. zu entnehmen, daß Veranstaltungen, deren Räumlichkeit als „gut" oder „sehr gut geeignet" bewertet wurden, durchschnittlich eine um 0,18 Notenpunkte bessere Bewertungen erhielten als diejenigen, deren räumliche Unterbringung schlechter bewertet wurde; weibliche Lehrende wurden durchschnittlich um 0,20 Notenpunkte besser bewertet als männliche, usw.[49] Im Kontext der in dieser Regressionsrechnung berücksichtigten Variablen üben folglich das Gefühl der Überforderung, der Veranstaltungstypus und das Gefühl der Unterforderung den *absolut* stärksten Effekt aus, dicht gefolgt von der Regelmäßigkeit der Vorbereitung, dem Vorliegen ausreichender Vorkenntnisse und dem Geschlecht der Lehrenden.[50]

[49] Die Dummy-Variablen wurden so definiert, daß z.B. zwei Ausprägungen („sehr gut geeignet" und „gut geeignet") zu einer Ausprägung zusammengefaßt und den drei anderen Ausprägungen („durchschnittlich", „eher nicht" bzw. „gar nicht geeignet") gegenübergestellt wurden. Dabei wurden jeweils die zu einer schlechteren Bewertung führenden Ausprägungen als „0" , die sich positiv auswirkenden Ausprägungen mit einer „1" kodiert. Dies führt dazu, daß der Wechsel von einer Ausprägung zur anderen (also von der „0" auf die „1") immer mit einer Verbesserung der Note (also einem negativen Wert in Spalte 1) einhergeht.

[50] In der Betrachtung der bivariaten Beziehungen zeitigte das Interesse der Studierenden am Thema der Veranstaltung die größten Differenzen zwischen den Extremen der Bewertung. Dies schlägt sich in der Regressionsrechnung nicht nieder, weil die beiden Extrempositionen (Kategorien 5 und 6 für „kein Interesse") nur sehr gering besetzt sind und ihre sehr starke Abweichung in der Bewertung bei der Zusammenfassung mit den anderen Kategorien (2 bis 4) neutralisiert wird. Zudem korreliert das Interesse sehr stark mit dem Veranstaltungstypus.

Der Einfluß aller untersuchten "externen" Variablen auf die Bewertung der Veranstaltungen [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	3,304	,431		7,660	,000
	Eignung des Raumes (geeignet)	-,181	,286	-,103	-,634	,531
	Geschlecht der Lehrenden (weiblich)	-,200	,243	-,138	-,823	,417
	Status der Lehrenden (Hochschullehrer)	-,067	,225	-,049	-,298	,768
	Veranstaltungstypus (HS, klein PS u Vorl)	-,246	,259	-,157	-,949	,351
	Pflicht als Motiv (keine Pflicht)	-,099	,264	-,070	-,376	,710
	Interesse als Motiv (Interesse)	-,093	,244	-,066	-,381	,706
	Wunschfach (Wunschfach)	-,043	,230	-,030	-,185	,854
	Vorkenntnisse (gute)	-,201	,228	-,147	-,883	,385
	Anforderungen (angemessen)	-,104	,227	-,073	-,458	,650
	Vorbereitung (regelmäßig)	-,208	,227	-,151	-,918	,366
	Überforderung (nein)	-,275	,245	-,194	-1,123	,271
	Unterforderung (nein)	-,213	,271	-,126	-,786	,438

a. Abhängige Variable: Durchschnittliche Bewertung der Fragen F1 bis F4

Insgesamt addieren sich die Auswirkungen dieser Faktoren auf eine Notendifferenz von 1,93 Notenpunkte. Wenn also ein männlicher Lehrender in einem ungeeigneten Raum ein großes Proseminar abhält, das für die Studierenden verpflichtend ist, an dessen Thema sie kein Interesse haben, für das sie unzureichende Vorkenntnisse mitbringen ..., so wird diese Veranstaltung durchschnittlich um fast zwei Notenstufen schlechter bewertet als eine Veranstaltung, auf die jeweils das Gegenteil zutrifft: ein geeigneter Raum, Professorin, Hauptseminar oder kleine Vorlesung ... Im Einzelfall können damit erhebliche Handicaps kumulieren, die einer guten Bewertung entgegenstehen.

Die gerade diskutierten nicht-standardisierten Koeffizienten informieren uns darüber, um wieviel sich durchschnittlich die Bewertung verändert, wenn wir von einer der Kategorien: z.B. „der Raum ist geeignet", in die andere: „der Raum ist nicht geeignet", wechseln. Für eine deskriptive Aussage über den Einfluß dieser Variablen ist dieser nicht-standardisierte Koeffizient heranzuziehen. Sein Wert hängt allerdings von der Art der Messung der Variablen und ihrer Streuung ab. Wollen wir etwas über die *relative* Bedeutsamkeit der einzelnen Faktoren sagen, so müssen wir uns auf die standardisierten Koeffizienten (die sogenannten Beta-Koeffizienten) beziehen, da in ihnen die methodischen Effekte neutralisiert werden. Auch hier kommt, wie ein Vergleich der entsprechenden Zahlen zeigt, dem Gefühl der Überforderung der relativ größte Einfluß zu, gefolgt vom Veranstaltungstypus, der regelmäßigen Vorbereitung der Studierenden und dem Vorliegen ausreichender Vorkenntnisse (und dem Geschlecht der

Lehrenden und dem Gefühl der Unterforderung).[51] [52]

Die multivariate Analyse zeigt aber nicht nur, welche Faktoren den größten Einfluß haben – sie macht auch deutlich, daß Faktoren, die in der bivariaten Beziehung durchaus als bedeutsam erschienen, nur noch einen geringen Einfluß ausüben, wenn man den Einfluß der anderen Variablen gleichzeitig berücksichtigt. Zu diesen letzteren gehören die Frage nach dem Wunschfach, der Status der Lehrenden, das Interesse am Thema und der Pflichtcharakter der Veranstaltung.[53] Außerdem hatte ich in der bivariaten Analyse darauf hingewiesen, daß einige der in der ersten Regressionsrechnung berücksichtigten Faktoren erst aus der Interaktion zwischen Lehrenden und Studierenden entstehen: so die Wahrnehmung der Anforderungen an die Vorbereitung und das Gefühl von Über- bzw. Unterforderung: die Ausprägung dieser Variablen ist also teilweise auch den Lehrenden selbst zuzurechnen. Da es unser Ziel sein sollte, ein in der Zuordnung der Kausalwirkung eindeutiges, zugleich aber möglichst wenig aufwendiges Erklärungsmodell für die Bewertung zu testen, werde ich in der folgenden Berechnung auf diese wenig erklärungskräftigen bzw. theoretisch uneindeutigen Faktoren verzichten.

Die nachfolgende Regression bezieht also nur noch fünf Rahmenbedingungen der Lehrveranstaltungen in die Berechnung ein. Wie die Tabelle zeigt, ist zwar der gemeinsame Einfluß dieses Modells geringer – allerdings steht der Verringerung der Zahl der erklärenden Variablen um mehr als die Hälfte nur eine Abnahme der Erklärungskraft von 28% auf 22% gegenüber. Eine auf diesem Modell basierende Datenerhebung kommt also bei relativ geringem Erklärungsverlust (bei gleichzeitigem Gewinn an Eindeutigkeit der Zuordnung der Wirkungen zu den beteiligten Personen) mit weniger als der Hälfte der ursprünglich berücksichtigten Fragen aus.

---

[51] Bei der Interpretation dieser Werte ist zu beachten, daß die oben konstatierte Wechselwirkung zwischen den Variablen, die in diese Gleichung aufgenommen wurden, rechnerisch ausgeschaltet ist: es handelt sich jeweils um den vom Einfluß der anderen in dieser Gleichung enthaltenen Variablen „bereinigten" Einfluß der Einzelvariablen auf die Bewertung der Veranstaltung. Der Einfluß, der jetzt noch für die „Überforderung" festgestellt wird, ist folglich unabhängig z.B. vom Umfang der Vorbereitung oder dem Vorhandensein geeigneter Vorkenntnisse.

[52] Auf die Signifikanzen ist hier nicht näher einzugehen, da die Auswahl der Veranstaltungen nicht auf dem Wahrscheinlichkeitsmodell beruhte, somit eine Modellvoraussetzung für die Interpretation von Signifikanzwerten nicht gegeben ist. (Der niedrige Wert der Signifikanzen schuldet sich übrigens der Umrechnung auf die Lehrveranstaltungen – bei einer Berechnung auf der Ebene der Individualdaten sind alle Werte hochsignifikant.)

[53] Interesse am Thema und Pflichtcharakter einer Veranstaltung korrelieren hoch mit dem Veranstaltungstypus – es ist daher nicht erstaunlich, daß diese Variablen bei gemeinsamer Berücksichtigung dieser Variablen an Bedeutung verlieren.

Einfluß von fünf Rahmenbedingungen

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,471 [a]	,221	,117	,64719

a. Einflußvariablen : (Konstante), Vorbereitung (regelmäßig), Eignung des Raumes (geeignet), Vorkenntnisse (gute), Veranstaltungstypus (HS, klein PS u Vorl), Geschlecht der Lehrenden (weiblich)

Die Höhe der nicht-standardisierten Regressionskoeffizienten zeigt, daß die Veranstaltungsbewertungen deutlich von Faktoren beeinflußt werden, die außerhalb des Einwirkungsbereiches der Lehrenden liegen: bei einer Summierung der nicht-standardisierten Regressionskoeffizienten beträgt die durchschnittliche Verbesserung der Noten 1,35 Punkte, wenn die im Variablennamen genannten Ausprägungen vorliegen. Auch dieses Modell belegt somit die Abhängigkeit der Veranstaltungsbewertung von Faktoren, die nicht die dem Lehrenden zuzurechnende Lehrleistung betreffen, sondern Rahmenbedingungen, unter denen er z.T. selber zu leiden hat.

Der Einfluß ausgewählter "externer" Variablen auf die Bewertung der Veranstaltungen [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	2,901	,287		10,099	,000
	Eignung des Raumes (geeignet)	-,227	,259	-,129	-,875	,387
	Geschlecht der Lehrenden (weiblich)	-,264	,218	-,181	-1,209	,234
	Veranstaltungstypus (HS, klein PS u Vorl)	-,328	,231	-,208	-1,418	,164
	Vorkenntnisse (gute)	-,291	,200	-,212	-1,458	,153
	Vorbereitung (regelmäßig)	-,235	,207	-,170	-1,138	,262

a. Abhängige Variable: Durchschnittliche Bewertung der Fragen F1 bis F4

c) Konsequenzen dieser Einflüsse auf die Interpretation der Veranstaltungsbewertungen

Aus den bisher vorlegten Ergebnissen dürfte deutlich geworden sein, daß die Ergebnisse einer Veranstaltungsbewertung – und zwar unabhängig davon, ob die Daten online oder in einer Befragung in der Veranstaltung selbst erhoben wurden – nicht naiv anhand eines Mittelwertes interpretiert werden dürfen, der aus den gegebenen Antworten der Studierenden allein berechnet wird. Wenn, wie gezeigt werden konnte, diese Antworten auch von Faktoren beeinflußt werden, die mit der Lehrleistung nichts zu tun haben, so wird man der Lehrleistung nur gerecht, wenn man den Einfluß dieser Faktoren ausschließt. Im Folgenden möchte ich zeigen, in welcher Weise eine Beschränkung allein auf die Bewertung durch die Studierenden ohne Berücksichtigung dieser Faktoren zu irreführenden Ergebnissen führt.

Bewertungen der Lehre sind – wie jede Messung – relationale Vorgänge: sie setzen ein Objekt (aus der Realität) zu einem anderen Objekt (aus der Meßtheorie) in Beziehung. Während wir in den Naturwissenschaften über objektivierte Vergleichskriterien verfügen (Längen- und Gewichtsmaße z.B.), deren Anwendung standardisiert ist und deren Ausprägung am jeweiligen Gegenstand der Forscher selber feststellt, haben wir es bei sozialwissenschaftlichen Befragungen mit einer Selbstzuordnung der befragten Personen zu tun, die nicht auf einen standardisierten Meßprozeß zurückzuführen sind. Nicht der Sozialforscher entscheidet, ob das Engagement des Dozenten „sehr gut" „gut" oder nur „ausreichend" ist, sondern jede einzelne Studierende nimmt diese Zuordnung vor. Und sie tut es nicht unter Rückgriff auf ein eindeutig definiertes und sozial abgestimmtes Bewertungssystem, sondern:

- unter Rückgriff auf das eigene Verständnis von dem, was sie selbst unter „Engagement des Dozenten" *versteht*;
- unter Rückgriff auf die eigenen *Erwartungen*, die sie an das Engagement von Dozenten unterhält;
- unter Rückgriff auf die *Erfahrungen*, die sie in früheren und parallel laufenden Veranstaltungen mit dem Engagement von Dozenten gemacht hat.

Dennoch handelt es sich hier nicht um willkürliche Anwendungen idiosynkratischer Vorstellungen: das Verständnis dieses Begriffes, die eigenen Erwartungen und die Erfahrungen sind in ihren Grundlagen sozial generiert, sozial vermittelt und werden in ihrer Anwendung sozial kontrolliert – insofern haben sie alle eine gemeinsame Basis. Sie sind aber nicht standardisiert und somit nicht ohne weiteres vergleichbar. In der Praxis der Lehrbewertungen wird ein solcher Vergleich aber immer durchgeführt. Angesichts des Fehlens eines objektiven Maßstabs für die Qualität der Lehre ist dies auch nicht verwunderlich: was bedeutet es denn schon, wenn eine Veranstaltung mit „gut" bewertet wurde?: Erhielten alle anderen ein „sehr gut", so ist diese Bewertung nicht so erfreulich – erhielten die anderen nur ein „befriedigend", dann steht man gut da. Insofern muß sich jede Evaluation von Lehrveranstaltungen auch dem Problem des Vergleichs stellen.

Gerade hier aber setzt – auf der Basis der oben aufgezeigten Einflußfaktoren auf die Lehrbewertung – die Kritik an einer Beschränkung der Interpretation auf die „Rohdaten", sprich: auf die von den Studierenden vergebenen Bewertungen ohne die Berücksichtigung weiterer Einflußfaktoren, an. Wie zu zeigen war, hängen die Bewertungen nicht nur von der individuellen Lehrleistung des Lehrenden, sondern auch von Faktoren ab, auf die er keinen Einfluß hat, die aber seine Veranstaltung mit prägen. Im Folgenden soll geprüft werden, inwieweit der Vergleich der Bewertung der einzelnen Veranstaltungen zu unterschiedlichen Ergebnissen kommt, wenn man ihn einmal auf der Ebene der Rohdaten, zum anderen unter Berücksichtigung

des Einflusses der Rahmenbedingungen durchführt.

In der nachfolgenden Tabelle werden die Lehrveranstaltungen in der Reihenfolge nach ihrem Rang auf der Basis der Rohdaten aufgelistet. Ein Vergleich der Ränge zwischen der dritten und der vierten Spalte zeigt, zu welchen Verschiebungen es im Vergleich der Bewertung der Veranstaltungen kommt, wenn diese Rahmenbedingungen berücksichtigt werden: Veranstaltung 47 fällt von Platz 1 auf Platz 11, Veranstaltung 2 von Platz 2 auf Platz 14 – umgekehrt springt Veranstaltung 49 von Platz 23 auf Platz 1, und die Veranstaltungen 38 und 3 verbessern sich von den Plätzen 32 und 33 auf die Plätze 9 und 10. Selbst wenn man berücksichtigt, daß den unterschiedlichen Rängen nicht immer große Differenzen in den durchschnittlichen Bewertungen zugrundeliegen (was allerdings in der üblichen Interpretation von Ranglisten nur zu häufig unterschlagen wird – da zählt nur: „Wer steht vorne (und hinten)?"), so wird deutlich, daß mit der Berücksichtigung der Rahmenbedingungen die ursprüngliche Reihenfolge auf der Basis der Rohdaten völlig durcheinandergewirbelt wurde.

Gegenüberstellung der Ränge auf der Basis der Rohdaten bzw. der unter Berücksichtigung der Rahmenbedingungen korrigierten Daten

	Nummer der Lehrveranstaltung	Rang auf Basis der Rohdaten	Rang bei Berücksichtigung der Rahmenbedingungen
1	47	1	11
2	2	2	14
3	18	3	17
4	27	4	16
5	21	5	3
6	7	6	19
7	33	7	6
8	15	8	2
9	30	9	31
10	24	10	29
11	6	11	4
12	14	12	34
13	13	13	7
14	50	14	15
15	16	15	12
16	22	16	5
17	5	17	28
18	25	18	32
19	32	19	23
20	42	20	8
21	8	21	39
22	51	22	18
23	49	23	1
24	41	24	20
25	4	25	25
26	36	26	26
27	45	27	21
28	10	28	24
29	48	29	27
30	34	30	33
31	19	31	13
32	38	32	9
33	3	33	10
34	28	34	22
35	44	35	35
36	40	36	30
37	12	37	44
38	9	38	36
39	20	39	40
40	31	40	38
41	46	41	43
42	35	42	.
43	1	43	41
44	29	44	46
45	26	45	42
46	39	46	37
47	23	47	45
Insgesamt N	47	47	46

Wie gering die Übereinstimmung zwischen den beiden Ranglisten ist, zeigt die nachfolgende Tabelle: in 39% der Veranstaltungen beträgt die Differenz 10 und mehr Rangplätze; durchschnittlich liegt die Differenz bei 8 Rangplätzen.

Differenz der Rangplätze der Bewertungen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	-22	2	4,3	4,3	4,3
	-19	1	2,1	2,2	6,5
	-18	1	2,1	2,2	8,7
	-14	2	4,3	4,3	13,0
	-13	1	2,1	2,2	15,2
	-12	2	4,3	4,3	19,6
	-11	1	2,1	2,2	21,7
	-10	1	2,1	2,2	23,9
	-7	1	2,1	2,2	26,1
	-4	1	2,1	2,2	28,3
	-3	1	2,1	2,2	30,4
	-2	2	4,3	4,3	34,8
	-1	2	4,3	4,3	39,1
	0	3	6,4	6,5	45,7
	1	1	2,1	2,2	47,8
	2	6	12,8	13,0	60,9
	3	2	4,3	4,3	65,2
	4	3	6,4	6,5	71,7
	6	4	8,5	8,7	80,4
	7	1	2,1	2,2	82,6
	9	1	2,1	2,2	84,8
	11	1	2,1	2,2	87,0
	12	2	4,3	4,3	91,3
	18	1	2,1	2,2	93,5
	22	1	2,1	2,2	95,7
	23	2	4,3	4,3	100,0
	Gesamt	46	97,9	100,0	
Fehlend	System	1	2,1		
Gesamt		47	100,0		

Selbst wenn man berücksichtigt, daß – wie erwähnt – die Differenzen zwischen den Rangplätzen gering sind, und deswegen zur Bildung von drei Leistungsgruppen übergeht (wie sie z.B. bei den Universitätsrankings in Publikumszeitschriften üblich sind: CHE u.ä.), verschwinden die oben aufgezeigten Defizite einer methodisch unzureichenden Bewertungsmessung nicht. Bei dieser Kategorisierung streuen die Veranstaltungen aus jeder Kategorie der ursprünglichen Rangbildung sogar über alle Kategorien der korrigierten Rangliste: als Mitglied der Spitzengruppe auf der Basis der Rohdaten kann eine Veranstaltung noch in der Schlußgruppe landen, wenn der Einfluß der Rahmenbedingungen ausgeschlossen wird – und umgekehrt.

Übereinstimmung der Rangplätze auf der Basis der Originalwerte der Lehrbewertungen und den drittvariablenkontrollierten Werten (drei Leistungsgruppen)

		Originalwerte der Lehrveranstaltungsbewertung			
		Spitzengruppe	Mittelgruppe	Schlußgruppe	Gesamt
Bewertungen ohne Einfluß der Rahmenbedingungen	Spitzengruppe	10	3	2	15
		62,5%	20,0%	13,3%	32,6%
	Mittelgruppe	5	9	2	16
		31,3%	60,0%	13,3%	34,8%
	Schlußgruppe	1	3	11	15
		6,3%	20,0%	73,3%	32,6%
Gesamt		16	15	15	46
		100,0%	100,0%	100,0%	100,0%

Rangverschiebungen in dieser Größenordnung machen aber einen Vergleich der Lehrleistung auf der Basis unkorrigierter Meßwerte unvertretbar, will man nicht auf die Benachteiligung einiger Lehrenden durch ungünstige Rahmenbedingungen ihrer Lehre noch die ungerechtfertigte (und unnötige) Diskriminierung ihrer Lehrleistung setzen – und dies auf der Basis einer vorgeblich wissenschaftlich abgesicherten Evaluation, denn nichts anderes unterstellt man bei einer Evaluation, die von der Universität in der Universität durchgeführt wird. Angesichts der vorgetragenen Ergebnisse ist es nicht verwunderlich, wenn Lehrende die naive Interpretation der studentischen Bewertungen ihrer Veranstaltung als Spiegel ihrer Lehrleistung für methodisch unreflektiert und sachlich ungerecht empfinden und einer solchen „Evaluation" ablehnend gegenüberstehen.

9. Zur Aussagekraft der „Lehrevaluation" an der Philosophischen Fakultät – eine Zusammenfassung und eine Schlußfolgerung[54]

Zusammenfassung zentraler Ergebnisse der Begleitforschung zur Lehrveranstaltungsbefragung

Die regelmäßige Durchführung einer Lehrevaluation unter Berücksichtigung der Stellungnahme der Studierenden ist an bayerischen Hochschulen gesetzlich vorgeschrieben. Dabei ist der Gesetzestext wesentlich offener als die Praxis an den Hochschulen: Der entsprechende Absatz

[54] Die Daten wurden nur in ausgewählten Veranstaltungen der beiden früheren Philosophischen Fakultäten erhoben. Da aber das dort praktizierte Verfahren in seinen Grundelementen (Online-Erhebung, Mittelwertbildung über verschiedene Dimensionen, fehlende Berücksichtigung von Rahmenbedingungen) auch in anderen Fakultäten und Universitäten angewendet wird, sind die hier gefundenen Ergebnisse über diese Auswahl hinaus von Interesse.

lautet: „Der (jährlich vorzulegende – W.M.) Lehrbericht enthält für den Berichtszeitraum auch Angaben über die Bewertung des Lehrangebots in den einzelnen Studiengängen durch die Studierenden ...".[55] Artikel 10, Absatz (3), Satz 1 konkretisiert dies zwar in der Richtung, daß „Im Rahmen der Bewertung der Lehre ... die Studierenden als Teilnehmer und Teilnehmerinnen von Lehrveranstaltungen anonym über Ablauf sowie Art und Weise der Darbietung des Lehrstoffes befragt" werden „können" (!), doch ist dies keineswegs obligatorisch, und es ist auch nicht die Rede davon, daß diese Befragung der Studierenden in *jedem* Semester erfolgen müsse. Wie an vielen anderen Universitäten und Fakultäten auch wird die Forderung nach einer Beteiligung der Studierenden an der Philosophischen Fakultät jedoch (nur) dadurch eingelöst, daß diese die von ihnen besuchten Veranstaltungen bewerten – andere Aspekte der Lehrsituation wie Studienberatung, Aufbau des Studienganges etc. werden nicht erfaßt. In der Praxis ist also eine Verengung zu beobachten, die weder durch den Gesetzestext vorgegeben noch sachlich geboten ist.

Diese Veranstaltungsbewertung erfolgt an der Philosophischen Fakultät anhand eines Fragebogens, der in seinen Fragen spezifische Aspekte einer Veranstaltung von den Studierenden anhand schulischer Noten (von 1 bis 6) bewerten läßt. Die Datenerfassung wird über einen im Internet von den Studierenden aufzurufenden und auszufüllenden Fragebogen vorgenommen. Unter anderem auf der Basis dieser Ergebnisse erstellt der Studiendekan seinen Lehrbericht, führt ggf. mit einzelnen Lehrenden auch persönliche Gespräche. Die Ergebnisse ihrer eigenen Veranstaltungen werden den Lehrenden mitgeteilt, wobei sie über die Verteilung der Noten zu den einzelnen Dimensionen und deren Mittelwerte und Streuung informiert werden; zudem werden ihre Mittelwerte dimensionsbezogen den Mittelwerten aller Veranstaltungen an der Fakultät gegenübergestellt, so daß jeder Lehrende seine Durchschnittsnote zum Notendurchschnitt aller Veranstaltungen in Beziehung setzen kann.

Unser Vergleich der so gewonnenen Lehrveranstaltungsbewertungen mit den Ergebnissen eines in den Veranstaltungen selbst erhobenen Datensatzes zeigte jedoch, daß die Ergebnisse der OnlineBefragung aufgrund methodischer Begrenzungen in Datenerhebung und Datenanalyse in hohem Maße unzuverlässig sind und eine sachgerechte Beurteilung der Lehrqualität nicht zulassen.

1. Zum einen führt eine Befragung über das Internet zu *unvollständigen* und *falschen* Ergebnissen:
   a) für fast ein Drittel der Veranstaltungen liegt gar keine oder eine so *geringe Beteiligung*

[55] Absatz (3), Satz 2 in Artikel 30 des Bayerischen Hochschulgesetzes in der Fassung vom 23. Mai 2006.

durch die Studierenden vor, daß (zuverlässige) Aussagen für diese Veranstaltungen nicht möglich sind;

b) bei den übrigen Veranstaltungen *divergieren* die Ergebnisse zwischen der Online-Erhebung und der Veranstaltungserhebung so stark, daß beide zu ganz unterschiedlichen Bewertungen für dieselben Veranstaltungen kommen; da die durchschnittliche Beteiligung an der Online-Befragung nur bei 38% lag,[56] die Befragung in den Veranstaltungen aber 81% aller Teilnehmerinnen erreichte, dürften die Ergebnisse der Veranstaltungserhebung die Bewertung der Veranstaltungen durch die Teilnehmerinnen deutlich besser wiedergeben.[57]

2. Im Ergebnisbericht der Studiendekane werden die *unterschiedlichen Rahmenbedingungen* der einzelnen Lehrveranstaltungen bei der Berechnung des arithmetischen Mittels der Studierendenbewertungen nicht berücksichtigt. Der über mehrere Fragen gemittelte Durchschnittswert für die einzelnen Veranstaltungen wird zum Mittelwert aller Veranstaltungen an der Fakultät in Beziehung gesetzt. Dieses Verfahren führt zu einer systematischen Benachteiligung gerade derjenigen Lehrenden, die unter erschwerten Bedingungen ihre Lehre zu erbringen haben: ungeeigneter Raum, große Proseminare und Vorlesungen, nicht ausreichende Vorkenntnisse der Studierenden etc. führen durchschnittlich zu schlechteren Bewertungen der Veranstaltungen und damit der Veranstalter, sind diesen aber nicht als mangelhafte Leistung zuzurechnen.

   Vergleicht man die auf diese Weise gewonnenen unkorrigierten Mittelwerte mit den Mittelwerten, die unter Berücksichtigung der Rahmenbedingungen berechnet wurden, so weichen die Bewertungen so stark voneinander ab, daß dieselbe Veranstaltung in der einen Berechnungsweise in der Spitzengruppe, in der anderen aber sogar in der Schlußgruppe eingestuft sein kann. Diese Unzuverlässigkeit der Messung sollte aber die Verwendung unkorrigierter Mittelwerte verbieten, das sie dem betroffenen Lehrenden wie auch den Rezipienten der Lehrbewertung (Studierenden, Kollegen, Vorgesetzten, der Öffentlichkeit) ein falsches Bild von der tatsächlichen Leistung des einzelnen Lehrenden vermitteln.[58]

---

[56] Dieser Prozentsatz gilt ohne Berücksichtigung derjenigen Veranstaltungen, für die die Online-Befragung mangels Beteiligung der Studierenden gar keine Ergebnisse erbrachte – für *alle* Veranstaltungen liegt der Wert somit noch deutlich niedriger.

[57] Es sei daran erinnert, daß die für diese Analyse herangezogenen Fragen mit den Fragen der Online-Befragung identisch sind – die Divergenz der Ergebnisse kann folglich nicht auf diesen Aspekt der Erhebung zurückgeführt werden.

[58] Angesichts dieses Ergebnisses ist der Vergleich der Bewertung einer Veranstaltung mit dem Mittelwert *aller* Veranstaltungen der Fakultät, wie er im Bericht der Studiendekane vorgenommen wird, informationslos und irreführend: hier werden Veranstaltungen miteinander verglichen, die nicht miteinander verglichen werden können.

Die bisher praktizierte Erhebungs- und die Auswertungsmethode (Online und ohne Berücksichtigung der Rahmenbedingungen) erweisen sich damit als ungeeignet für die Feststellung der Lehrleistung in unterschiedlichen Veranstaltungen. Ihre Ergebnisse unterliegen systematischen Verzerrungen und werden damit den methodischen Anforderungen an eine Messung der Lehrbewertung – und auch den berechtigten Erwartungen der Lehrenden an eine sachbezogene und faire Bewertung ihrer Leistungen – nicht gerecht.

Zur Eignung der Lehrveranstaltungsbewertung als Lehrevaluation

Welche Gründe sprechen für die Wahl der bisher praktizierten Vorgehensweise? Es ist v.a. einer: sie ist mit dem geringsten Arbeitsaufwand verbunden:

- der Aufwand für Datenerhebung und -aufbereitung ist, insbesondere bei wiederkehrender Befragung, minimal;
- die Störung der Lehrveranstaltung hält sich ebenfalls (mit der Verteilung der TAN-Nummern) in engen Grenzen;
- die Datenanalyse stellt nur geringe Anforderungen und produziert exakte Ergebnisse.

Dafür muß man allerdings gewillt sein, die Ungültigkeit der Ergebnisse in Kauf zu nehmen.

Eine Alternative müßte sicherstellen, daß die an einer Veranstaltung teilnehmenden Studierenden so umfassend wie möglich in die Erhebung eingebunden werden, um unkontrollierbare Verzerrungen durch selektive Teilnahme auszuschließen, und sie müßte die wesentlichen Einflußfaktoren auf die Veranstaltungsbewertung mit einbeziehen. Dies bedeutet:

- die Datenerhebung muß im Rahmen der Veranstaltungen durchgeführt werden (dies ist übrigens eine auch von den befragten Studierenden nachdrücklich befürwortete Strategie – das Verteilen der TAN-Nummern auf 3 x 5 cm großen Zetteln stößt bei ihnen auf Spott und Ablehnung);[59] [60]

[59] Dies fordert auch eine Studierendenvertreterin auf einer Tagung der Hochschulrektorenkonferenz im Rahmen des Projekts Qualitätssicherung am 3. und 4. November 2005 im Bonner Wissenschaftszentrum: Weber, in J. Alphei (Hrsg.), Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation, 2006, 55f; sie spricht sich übrigens dezidiert auch gegen eine Beschränkung auf standardisierte Bewertungsmessung aus.

[60] In diesem Zusammenhang ist auch ein gewichtiges Argument für die Gültigkeit der Daten zu erwähnen: durch die Veranstaltungssituation ist die Erinnerung an die gesamte Veranstaltung aktiviert, die Assoziationen beziehen sich auf genau diese Veranstaltung. Werden dagegen (wie es nach Aussagen von Studierenden die Praxis zu sein scheint) am heimischen Computer mehrere Veranstaltungen nacheinander bewertet, so fehlt diese situative Aktualisierung, und es besteht die Gefahr, oberflächlich und in distanzierter Haltung die Fragebögen nacheinander „abzuhandeln": den Antworten fehlt die Spezifizierung auf die konkrete Veranstaltung.

– der Fragebogen ist um Fragen nach diesen Rahmenbedingungen zu erweitern;[61]
– in der Feststellung der Bewertung der Veranstaltungen ist der Einfluß dieser Rahmenbedingungen auszuschalten.

Es liegt auf der Hand, daß eine derart durchgeführte Lehrveranstaltungsbewertung nicht kostenfrei nebenbei zu haben ist: der Aufwand in Datenerhebung und Datenanalyse wäre beträchtlich höher, die Befragung wäre sicherlich nicht von Evaluationslaien im Nebenjob durchzuführen, sie wäre auch nicht für jede Veranstaltung für jedes Semester zu realisieren – und wenn es wirklich um Lehr*evaluation* und nicht nur um Lehr*veranstaltungsbewertung* geht, so müßten andere Schwerpunkte gesetzt und auch die Rahmenbedingungen von Lehre allgemein in die Analyse einbezogen werden.

Eine unverzichtbare Voraussetzung für eine gelingende Evaluation ist aber weder in diesem Bericht noch in der fakultätsinternen Diskussion bisher thematisiert worden: die Frage der zu realisierenden Ziele. Dabei geht es um die Ziele auf zwei unterschiedlichen Ebenen, wobei die letzere die Beantwortung der ersteren voraussetzt: „ Welche Ziele soll eine akademische Lehrveranstaltung erreichen?" und „Welche Ziele soll die Evaluation dieser Lehrveranstaltungen einlösen?"

Die erste Frage ist derartig vielschichtig, daß sie im Kontext dieses Forschungsberichtes gar nicht anzugehen ist. Festzuhalten ist nur: für verschiedene Lehrende in verschiedenen Studienfächern und verschiedenen Veranstaltungen ist diese Frage durchaus unterschiedlich zu beantworten: dies mag von der Wissensvermittlung über die Einübung von Fähigkeiten und Fertigkeiten bis hin zur Konzipierung eigener wissenschaftlicher Fragestellungen unter Beachtung der Regeln wissenschaftlichen Arbeitens gehen. Einer Messung ihrer Realisierung in einer Lehrveranstaltung müßte letztlich eine Erhebung dieser Ziele bei den Lehrenden selbst vorausgehen – eine realitätsferne, sachlich dennoch nicht unberechtigte Forderung.[62] Denn nichts anderes bedeutet „Evaluation": die methodisch abgesicherte Prüfung, inwieweit die Ziele der Handelnden durch ihr Handeln tatsächlich erreicht worden sind. Selbst wenn man die Evaluation (und auch dies mit gutem Grund) nicht auf die Ziele der individuellen Lehrenden abstellen will, würde auch die Identifikation unterschiedlicher Ziele z.B. für verschiedene Veranstaltungstypen

[61] Über die Eignung der im bisherigen Verfahren gestellten Fragen (an denen sich ursprünglich die Kritik der Kolleginnen und Kollegen an der Fakultät entzündete) ist an dieser Stelle noch gar nicht gesprochen worden. Angesichts der aufgezeigten Verzerrungsmechanismen – denen auch „bessere" Fragen unterliegen würden – erscheint die Neuformulierung zwar nicht unwichtig, eher doch nachrangig.

[62] Vgl. hierzu die Forderungen von Günter Endruweit, Programmevaluation als Laienspiel, in: Soziologie, Heft 2, 1992, 107-115.

oder Studienfächer die Komplexität einer Lehrevaluation bis zur Unpraktikabilität steigern. So ist es kein Zufall, daß diese Frage ausgeklammert wurde und man sich mit der Annahme zufriedengab, daß es allgemeine Dimensionen gebe, an denen die Qualität einer Lehrveranstaltung zu beurteilen sei (Struktur, Verständlichkeit etc. – diese Dimensionen finden sich in fast allen einschlägigen Fragebögen).

Aber auch auf der zweiten Ebene: der Frage nach den Zielen einer Lehrevaluation, ist die Diskussion nicht zu einem klaren Ergebnis geführt worden. Die Antwort auf diese Frage hat aber entscheidende Konsequenzen für die Art der Durchführung der Lehrevaluation:

- Geht es um die Feststellung von *Mißständen* in der Lehre, so benötigt man den Aufwand einer methodisch differenzierten Bewertung aller Veranstaltungen nicht: hier genügt die Einrichtung eines Verfahrens, das die Benennung solcher Mißstände im Einzelfall ermöglicht (ggf. auch durch ein sehr rudimentäres Befragungsverfahren in den Veranstaltungen – hier mag auch eine Online-Befragung das angemessene Mittel sein, sichert sie doch maximale Anonymität).

- Geht es um die *Rückkoppelung* der Erfahrungen der Studierenden an die Lehrenden, so sind standardisierte Bewertungsverfahren eher ungeeignet. Für diesen Zweck empfehlen sich offene Fragen, in denen die Studierenden ihre Erfahrung mit der Veranstaltung darlegen und begründen können – erfahrungsgemäß lernen Lehrende aus diesen Anmerkungen mehr für eine Verbesserung ihrer Veranstaltungen als aus Mittelwerten. Inwieweit dies eine zentralisierte Organisation der Befragung erfordert, ist nicht zuletzt davon abhängig, wie man die Kooperationsbereitschaft der Lehrenden bewertet. Ob allerdings unwillige Lehrende in irgendeiner Weise zum Lernen aus diesen Rückmeldungen der Studierenden gebracht werden können, erscheint in jedem Fall fraglich – dies wirft also die Frage nach einer externen Kontrolle auf.

- Geht es um eine *Bewertung* der Lehrleistungen der einzelnen Lehrenden, die über die Rückmeldung der studentischen Erfahrungen an ihn hinausgeht und Teil einer offiziellen und möglicherweise (siehe Transparenz) auch öffentlichen Leistungsbewertung werden soll, so kann kein Weg daran vorbeigehen sicherzustellen, daß in diese Leistungsbewertung nur die Aspekte einfließen, auf die der einzelne Lehrende Einfluß hat, die er zu verantworten hat – es ist also im oben beschriebenen Sinne eine methodisch reflektierte und abgesicherte Erhebung und Analyse der Daten erforderlich.

- Geht es um die *Transparenz* der Lehrqualität, um z.B. Studierende für das Studium an einer

Hochschule zu gewinnen, dann ist, wie im Fall der Bewertung, ebenfalls eine differenzierte Erfassung relevanter Aspekte der Lehre erforderlich.[63]

– Geht es um die *Sicherung und Steigerung der Lehrqualität* durch den Lehrenden, so sind offene Rückmeldungen sinnvoll, die es erlauben, Schwachstellen zu identifizieren – aus Mittelwerten standardisierter Antworten sind keine klaren Handlungsanregungen zu entnehmen. Eine Standardisierung dieser Rückmeldungen ist allerdings dann erforderlich, wenn nicht – z.B. in einem System kollegialen Austausches – die Rückmeldungen veranstaltungsbezogen inhaltlich besprochen werden sollen, sondern übergeordneten Kontrollinstanzen ein schneller Überblick über potentielle Schwachstellen in der Institution ermöglicht werden soll.

Für diesen Zweck aber ist es – wie im Fall der offiziellen Bewertung und der Transparenz – unverzichtbar, daß diese standardisierte Erfassung dem Stand der Evaluationsforschung entspricht. Und dies nicht nur aus Gründen der Gerechtigkeit gegenüber den betroffenen Lehrenden, sondern auch im Interesse eines Erfolges im Bemühen um die Sicherung der Lehrqualität. Vermittelt das Evaluationsverfahren den Eindruck einer unfairen und sachlich unzutreffenden Bewertung, so provoziert es nicht nur Widerstände hinsichtlich der Lehrevaluation, es wird auch negative Konsequenzen auf die Ausübung der Lehre selbst haben, wenn der Eindruck entsteht, erbrachte Leistungen würden systematisch nicht gewürdigt. Lehrende werden sich strategisch auf die Bewertungskriterien wie auf die Bewertungsrahmenbedingungen einstellen, so daß es z.B. schwieriger werden wird, für „undankbare" Veranstaltungen überhaupt Lehrende zu finden.

Die oben genannten Ziele gilt es zu unterscheiden, will man eine sachgerechte Entscheidung über das erforderliche Vorgehen in der Lehrevaluation treffen. Die bisherige Diskussion zeichnet sich allerdings durch einen ausgeprägten Zielwirrwarr aus: in unterschiedlichen Kontexten werden unterschiedliche Ziele benannt. Dabei ziehen sich Befürworter der bisherigen Lehrevaluation in kontroversen Diskussionen gerne auf die Funktionen der Aufdeckung von Mißständen und auf die Rückmeldung an die Lehrenden zurück – während die Konzeption des

[63] Angesichts der zunehmenden Konkurrenz zwischen Hochschulen und Studiengängen sollte sich die Öffentlichkeit aber keine zu große Hoffnung auf interessenunabhängige Ergebnisdarstellungen machen. In solchen Fragen agieren Hochschulen nicht anders als andere Handlungsträger: sie werden damit Reklame machen (wollen und müssen), und die kritische Öffentlichkeit ist mittlerweile genügend über soziale Prozesse aufgeklärt, um zu wissen, daß Reklame nicht den Status einer unverfälschten Tatsachenbeschreibung genießt. Wie die kontroversen – und keineswegs schon entschiedenen – Diskussionen über die Qualität der Informationen von Ranglisten von Studiengängen und Hochschulen bereits zur Genüge gezeigt hat, wird man es also lernen müssen, Informationen über Lehrqualität richtig zu bewerten. Skepsis ist angebracht, daß dies in der breiten Öffentlichkeit in der erforderlichen Weise geleistet werden kann – die Verselbständigung dieser Prozesse wird in den Hochschulen selbst zumindest beklagt.

Evaluationsverfahren tatsächlich auf den Vergleich der Lehrleistung abgestellt ist. In programmatischen Reden und Entwürfen wird denn auch explizit ausgesprochen, daß das Ziel der Einführung von Lehrevaluationen die Etablierung eines finanziellen Belohnungs- und Bestrafungssystems für Lehrende und Institute ist – natürlich mit dem Ziel, auf diese Weise eine Verbesserung der Lehrqualität zu erreichen.[64]

Will man sich aber tatsächlich um die Sicherung und Steigerung der Lehrqualität bemühen, so wird man die Rahmenbedingungen der Lehre nicht weiter ausklammern können. Betrachten wir die Bewertungen der Lehrveranstaltungen durch die Studierenden, so zeigt sich in dem den obigen Berechnungen zugrundegelegten Bewertungsindex, daß die Studierenden einerseits das Notenspektrum durchaus ausschöpfen (sie vergeben Noten zwischen sehr gut und mangelhaft, in Einzelfällen auch ungenügend – sie differenzieren also zwischen den Veranstaltungen). Andererseits bewerten sie 66% der Veranstaltungen mit einer Note, die besser ist als 2,5, und die durchschnittliche Benotung der zugrundegelegten Fragen liegt bei 2,2. Billigt man also den Studierenden die Fähigkeit zur Bewertung der von ihnen besuchten Lehrveranstaltungen zu, so ist zur Kenntnis zu nehmen, daß sie ihren Lehrenden durchschnittlich eine gute Lehrleistung attestieren. Will man sich also um eine Verbesserung der akademischen Lehre bemühen, so wäre zu fragen, ob in den Rahmenbedingungen (Größe der Veranstaltungen, räumliche Ausstattung, Belastung der Lehrenden mit lehr- und forschungsfernen Aufgaben etc.) nicht ein wesentlich größeres Potential zur Verbesserung der Lehre bestünde. Hier stellt sich für die betroffenen Lehrenden die Frage, ob die Verengung des Konzeptes der Lehrevaluation auf eine Lehr*veranstaltungsbewertung* nicht v.a. die Funktion einer Problemverschiebung hat: indem die einzelne Lehrveranstaltung ins Zentrum der Betrachtung gerückt wird, wird die Aufmerksamkeit von anderen Faktoren, denen tatsächlich eine wesentlich größere Bedeutung für die Lehrqualität zukommt, abgelenkt – sei es, weil man dort keine Veränderungsmöglichkeit sieht, sei es, weil es für Hochschulpolitiker opportun erscheinen mag, die Verantwortung für Mängel einer strukturell seit Jahren vernachlässigten Hochschulausbildung den dort Lehrenden anzulasten, statt die hochschulpolitischen Versäumnisse zu thematisieren und zu beheben.[65]

Im Lichte der oben berichteten Ergebnisse erscheint es erforderlich, bevor über Details wie die

---

[64] Siehe z.B. Künzel und Weber in J. Alphei (Hrsg.), Qualitätsentwicklung an Hochschulen ..., 2006, 26, 27, 55. (Vgl. die Fußnote zu Beginn dieses Abschnittes.)

[65] Dabei soll nicht verschwiegen werden, daß die Hochschulen selbst in den vergangenen Jahrzehnten durch die Blockade einer umfassenden Hochschulreform die Chance verpaßt haben, die Struktur der Universität an die sich verändernden gesellschaftlichen Rahmenbedingungen anzupassen. Dies ändert aber nichts an der Tatsache, daß die jetzige Beschränkung auf die Qualität der Lehrveranstaltungen wiederum von den zentralen Einflußfaktoren ablenkt und eine weitere Chance verpaßt wird, eine angemessene akademische Ausbildung für zunehmend mehr Studierende sicherzustellen.

Formulierung einzelner Fragen in einem Bewertungsbogen für Lehrveranstaltungen zu diskutieren ist, zu klären, welche Ziele mit einer Lehrevaluation überhaupt angestrebt werden sollen. In Abhängigkeit von dieser Entscheidung ließe sich dann die Frage nach der erforderlichen – und praktisch möglichen – Vorgehensweise für eine Evaluierung der Qualität der Lehre beantworten. Um zu verhindern, daß diese Evaluierung als Beschäftigungsprogramm für die Hochschulmitglieder endet und um sicherzustellen, daß die Ergebnisse einer Lehrevaluierung tatsächlich zur Verbesserung der Lehre beitragen, wäre es hilfreich, die in der bisherigen Evaluationsforschung gesammelten Erfahrungen über Erfolgsbedingungen von Evaluierungen zu nutzen. Hier besteht ein erheblicher Nachholbedarf, der von der Zieldefinition bis hin zur Sicherstellung der Umsetzung erforderlicher Konsequenzen aus den Evaluationsergebnissen reicht.[66]

[66] Vgl. die Erfahrungen mit Evaluationen im Bildungsbereich in der der Schweiz in M. Stamm, Evaluation im Spiegel ihrer Nutzung: Grand idée oder grande illusion des 21. Jahrhunderts?, in: Zeitschrift für Evaluation, 2003, 183-200.

Anhang

Der im Forschungsseminar entwickelte Fragebogen liegt dieser Arbeit entweder bei, oder er ist im Internet unter der folgenden Adresse aufzurufen:

http://www.soziologie.phil.uni-erlangen.de/files/lehre/Fragebogen%20Endfassung.pdf