Nutzung von Lehrevaluationsdaten für die Qualitätssicherung der Evaluationsinstrumente am Beispiel der Universität Zürich

Rüdiger Mutz, Hans-Dieter Daniel

Die Kernintention dieses Beitrags ist es, am Beispiel eines Pilotprojekts der Universität Zürich sowohl konzeptionell als auch empirisch aufzuzeigen, wie Daten studentischer Lehrevaluationsbeurteilungen für die Qualitätssicherung der Lehrevaluationsinstrumente genutzt werden können, insbesondere für die Prüfung der drei Testgütekriterien der Reliabilität, Validität und Fairness. Eine solche Sicherung der Qualität des Instrumentariums ist dabei nicht Selbstzweck, sondern im Rahmen von „Quality Audits" oder Systemakkreditierungen auch explizite Aufgabe von Universitäten.

1 Einleitung

Lehrveranstaltungsbeurteilungen an Universitäten waren in den 1990er Jahren in Deutschland heftig umstritten, sowohl in der Hochschulpraxis als auch in der wissenschaftlichen Diskussion. Beispielsweise wurden schwerwiegende rechtliche Probleme wie die Beeinträchtigung der Freiheit der Lehre (Art. 5 GG) geltend gemacht *(z. B. Hufen 1995; Mußgnug 1992)* oder die Zuverlässigkeit (Reliabilität), Gültigkeit (Validität) und Fairness von studentischen Urteilen generell in Frage gestellt *(z. B. Kromrey 1994; Süllwold 1992)*. Mittlerweile verebbt an den Hochschulen mit der breiten Einführung von auf studentischen Befragungen basierenden Lehrevaluationen die Diskussion. Den in Universitäten eingesetzten Instrumentarien wird mit Bezug auf die einschlägige, meist angloamerikanische Forschungsliteratur der 1980er und 1990er Jahre *(z. B. Daniel 1996; 2000; Marsh 1984; Marsh/Roche 1997)* unterstellt, sie seien im Allgemeinen reliabel, valide und fair. Auch in der Forschungsliteratur zur Lehrevaluation finden sich kaum noch aktuelle Beiträge zu Fragen der Reliabilität, Validität und Fairness von Lehrevaluationsverfahren *(*Ausnahmen z. B. *Greimel-Fuhrmann/Geyer 2005; Klein/Rosar 2006; Mutz 2000; 2003)*.

Während die Datenlage in den 1990er Jahren im deutschsprachigen Bereich noch sehr spärlich war *(Rindermann 1995)*, liegen mittlerweile durch die EDV-gestützten, meist zentral organisierten Lehrveranstaltungsevaluationen (z. B. mit dem Software-System EVA-SYS, www.electricpaper.de) schon an einer einzelnen Hochschule riesige Datenmengen

vor, die Gefahr laufen, zu Datenfriedhöfen zu werden. Daher ist zu fragen, inwieweit diese Daten für die Qualitätssicherung des Lehrevaluationsinstrumentariums an den einzelnen Hochschulen genutzt werden können, oder auch hochschulübergreifend, sofern vergleichbare Instrumentarien, z. B. das Heidelberger Inventar zur Lehrveranstaltungsevaluation (HILVE, *Rindermann/Amelang 1994)*, eingesetzt werden (Benchmarking). Diese Form der Qualitätssicherung des eingesetzten methodischen Instrumentariums ist dabei nicht Selbstzweck, sondern wird beispielsweise im Rahmen von Systemakkreditierungen oder „Quality Audits" auch von den Universitäten erwartet *(Kultusministerkonferenz 2007)*.

Aus diesen Gründen ist die zentrale Zielsetzung dieser Arbeit, am Beispiel der Universität Zürich Möglichkeiten aufzuzeigen, wie Daten einer breit gefächerten Lehrevaluationserhebung genutzt werden können, um die Qualität des eingesetzten Instrumentariums (Reliabilität, Validität, Fairness) zu prüfen. Im Folgenden wird zuerst auf die methodischen Grundlagen von Lehrveranstaltungsevaluationen eingegangen, dann kurz die Datenlage des Pilotprojekts der Universität Zürich erläutert und danach werden ausgewählte Ergebnisse zu den von Studierenden beurteilten Vorlesungen dargestellt.

2 Methodische Grundlagen

Die studentische Lehrevaluation kann als eine Art psychologisches Testverfahren aufgefasst werden *(Daniel 1998; Loewenthal 2001; Mutz 2003; Wirtz/Caspar 2002)*. Psychologische Tests sind nach *Lienert (1969, S. 7)* „... wissenschaftliche Routineverfahren zur Untersuchung eines oder mehrerer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung". Die Güte eines Testverfahrens wird anhand der Testgütekriterien der Reliabilität, Validität und Fairness beurteilt. Unter Reliabilität versteht man, wie genau ein Test misst, unter Validität, ob ein Test inhaltlich auch das erfasst, was er zu messen beabsichtigt, und unter Fairness, ob die Testergebnisse unabhängig sind von Studierendenmerkmalen oder anderen Variablen wie studentischem Interesse am Thema der Lehrveranstaltung.

Während die üblichen psychologischen Testverfahren (z. B. Intelligenztests) im Rahmen der Klassischen Testtheorie (KTT) den wahren Messwert der jeweils getesteten Person schätzen, sollen in studentischen Lehrevaluationen die wahren Werte von Lehrveranstaltungen bzw. Lehrpersonen bestimmt werden *(Cranton/Smith 1990)*. Die Frage der *Reliabilität* wird zur Frage, wie genau das wahre Veranstaltungsmittel durch ein Lehrevaluationsinventar erfasst werden kann. Die Berechnung der üblichen Reliabilitätskoeffizienten auf der Basis der Einzelurteile allein reicht aus zwei Gründen nicht aus, die Reliabilität eines

Lehrevaluationsinventars zu sichern *(Marsh 1987)*: Erstens sind Einzelurteile stark durch implizite subjektive Theorien „guter Lehre" der Studierenden geprägt, die bei aggregierten Urteilen eliminiert werden. Zweitens können Studierende verschiedene Lehrpersonen sehr unterschiedlich bewerten, dies aber mit hoher Urteilerübereinstimmung, d. h. mit vergleichbarem Skalenwert. Dies hat zur Folge, dass die Reliabilität bezüglich der Einzelurteile innerhalb der Lehrveranstaltungen infolge geringer Streuung (Urteilshomogenität) auf Null sinkt, die Reliabilität auf Veranstaltungsebene infolge hoher Urteilerübereinstimmung jedoch auf 1.0 ansteigt. Dieser Sachverhalt kommt in der sogenannten Urteilerübereinstimmung, beispielsweise gemessen über eine Intraklassenkorrelation (ICC), zum Ausdruck. Dies setzt aber voraus, dass die Lehrveranstaltungen tatsächlich recht heterogen in ihrem Qualitätsniveau sind. Liegt eine Stichprobe von Lehrveranstaltungen mit geringen Unterschieden im Qualitätsniveau vor, ist eine beobachtete geringe Interraterreliabilität nicht unbedingt auf die mangelnde Urteilsfähigkeit der Studierenden, sondern auf die Homogenität der Stichprobe von Lehrveranstaltungen hinsichtlich des Qualitätsniveaus zurückzuführen. Daher sind zu den Intraklassenkorrelationen auch immer die Standardabweichungen der Lehrveranstaltungsmittelwerte und der Gesamtmittelwert anzugeben. In Untersuchungen zur Lehrveranstaltungsevaluation sind Intraklassenkorrelationen in der Regel relativ gering ausgeprägt mit Werten von .20/.30 *(Feldman 1977; Rindermann 2003)*. Neben den klassischen Koeffizienten der internen Konsistenz (Cronbach Alpha) auf Studierenden- und Veranstaltungsebene kann mittels einer Split-Half-Reliabilität geprüft werden, inwieweit die durchschnittliche Beurteilung bei der einen Hälfte zufällig ausgewählter Studierender einer Lehrveranstaltung mit der durchschnittlichen Beurteilung bei der anderen Hälfte der Studierenden derselben Lehrveranstaltung über alle Lehrveranstaltungen zusammenhängt. Je höher dieser Zusammenhang, desto höher ist die Split-Half-Reliabilität und damit die Genauigkeit des Instrumentariums.

Zur Frage der Reliabilität gehört auch die Frage des angemessenen *Antwortformats* der verwendeten Antwortskala (z. B. dreistufige oder sechsstufige Ratingskala) zur Beurteilung des Lehrgeschehens durch die Studierenden. Ratingskalen sind ordinal skaliert, was die Berechnung von Mittelwerten nicht ausschließt, da der Erwartungswert einer diskreten Zufallsvariablen der Summe aus den Produkten der Skalenwerte (z. B. 1, 2) mit den Antworthäufigkeiten je Skalenwert entspricht *(Ross 2007)*. Zur Prüfung des Antwortformates hat die Item-Response-Theorie (IRT) als eine probabilistische Theorie psychologischer Tests mit der Berechnung von Schwellenparametern für Ratingskalen einen wichtigen Beitrag geleistet *(Rost 2004)*: In IRT wird die Beantwortung von Ratingskalen als Überschreiten von Schwellen mit einer bestimmten Wahrscheinlichkeit verstanden. Ist beispielsweise die Schwellenwahrscheinlichkeit von Stufe 1 zu Stufe 2 einer sechsstufigen

Ratingskala gering, so fällt es diesen Personen offenbar leicht, die erste Stufe zu überschreiten. Unterscheiden sich zwei Schwellenwahrscheinlichkeiten eines Items nicht, so geben die zugehörigen Antwortstufen keine zusätzliche Information. Die zwei Stufen einer sechsstufigen Ratingskala ließen sich in eine fünfstufige umkodieren, ohne wesentliche Informationen zu verlieren. Mit diesem Ansatz lässt sich die Angemessenheit des gewählten Antwortformats prüfen und gegebenenfalls lassen sich spezifische Antworttendenzen, z. B. Tendenz zu extremen Antworten, ermitteln.

Bezüglich der *Validität* wird zwischen Kriterienvalidität, inwieweit der Fragebogen, die Skalen und Items mit anderen, z. B. durch Beobachtung des Lehrgeschehens oder Bewertungen der Lehrenden selbst gewonnenen, externen Kriterien übereinstimmen, und Konstruktvalidität unterschieden. Letzteres fragt, ob ein Fragebogen inhaltlich die Dimensionen, sprich Konstrukte, erfasst, für die das Verfahren entwickelt wurde *(Cronbach/ Meehl 1955)*. So zeigen Untersuchungen, dass Studierende das Lehrgeschehen nicht eindimensional (z. B. gut-schlecht), sondern auf mehreren Dimensionen beurteilen *(Marsh 1987)*. Ein Konstrukt wären beispielsweise die didaktischen Fertigkeiten der Lehrperson, ein anderes der Schwierigkeitsgrad einer Lehrveranstaltung. Diese Mehrdimensionalität kann mittels konfirmatorischer Faktorenanalyse, eines statistischen Verfahrens zur Reduktion der Zusammenhänge zwischen den Items des Evaluationsbogens auf wenige zugrunde liegende Dimensionen, überprüft werden, wobei explizit der hierarchische Aufbau der Daten (studentische Urteile als Ebene 1, Lehrveranstaltungen als Ebene 2) in der Analyse zu berücksichtigen ist. Statt der klassischen Faktorenanalyse, die intervallskalierte Variablen erfordert, bieten heute einschlägige Softwareprogramme auch Faktorenanalyseverfahren an, die explizit den Rangskalencharakter der Antwortskalen und die hierarchische Struktur der Daten berücksichtigen *(z. B. MPLUS 4.1, Muthén/ Muthén 2006)*. Eine weitere Möglichkeit, die Konstruktvalidität zu überprüfen, bietet der Multitrait-Multimethod-Ansatz (MTMM): Studierende füllen zusätzlich zu einem Evaluationsbogen mit unbekannter Konstruktvalidität einen oder mehrere weitere Evaluationsbögen aus, deren Dimensionalität und Testgütekriterien allgemein bekannt sind. So müssen inhaltlich vergleichbare Konstrukte in den verschiedenen Evaluationsbögen miteinander höher korrelieren als inhaltlich unterschiedliche Konstrukte, um die Konstruktvalidität des neu entwickelten Instrumentariums zu gewährleisten *(Marsh 1982)*.

Haben Faktoren auf die Lehrevaluation einen bedeutsamen Einfluss, die weder einen Zusammenhang mit dem eigentlichen Lehr- und Lernprozess aufweisen, noch unter der Kontrolle der Lehrpersonen stehen, z. B. das Geschlecht der Lehrperson, Pflicht- vs. Wahlveranstaltung, allgemeine Rahmenbedingungen (z. B. Überfüllung), dann ist die *Fairness*

der studentischen Evaluation gefährdet *(Greenwald 1997; Marsh 1987; Spiel 2001)*. Die Evaluationsergebnisse wären gegebenenfalls durch Bias-Variablen verzerrt. In einer Übersichtsarbeit nennt *Marsh (1987)* fünf zentrale Bias-Variablen: Veranstaltungsgröße, das studentische Interesse am Thema der Lehrveranstaltung, das Arbeitspensum, das Anforderungsniveau der Lehrveranstaltung und die erwartete Note. Mittels einer Korrelationsanalyse kann geprüft werden, inwieweit die Skalen des Fragebogens mit den potenziellen Bias-Variablen zusammenhängen. Fehlende Zusammenhänge geben Hinweise für die Fairness des Verfahrens. Ein weiterer Faktor, der die Fairness des Verfahrens gefährden kann, sind systematische Antworttendenzen: Bei Befragungen wird häufig beobachtet, dass Personen Sachverhalte generell zu positiv oder zu negativ beurteilen, was als Milde- oder Strenge-Effekt bezeichnet wird. Solche Antworttendenzen können aber nur dann ermittelt werden, wenn eine Kohorte von Studierenden mehrere unterschiedliche Lehrveranstaltungen mit demselben Evaluationsbogen bewertet. Daher ist es in der Organisation von Lehrevaluationen wichtig, die ausgefüllten Evaluationsbögen einer Person mittels eines persönlichen Codes oder der Matrikelnummer zusammenführen zu können. Letzteres würde zusätzlich die Möglichkeit eröffnen zu prüfen, ob die in einer Prüfung erwartete oder erzielte Note einen Einfluss auf die Lehrveranstaltungsbeurteilung hat. Weitere Antworttendenzen, wie die Tendenz, mittlere Kategorien einer Ratingskala oder Extremkategorien anzukreuzen, lassen sich mittels der oben genannten Item-Response-Theorie feststellen.

Die Skalen zur Lehrveranstaltungs- oder Dozierendenzufriedenheit werden häufig als Indikator der globalen Bewertung einer Lehrveranstaltung interpretiert. Unabhängig von den Testgütekriterien kann es daher von Interesse sein zu ermitteln, welche Faktoren mit der Zufriedenheit mit der Lehrveranstaltung bzw. mit dem jeweiligen Lehrenden korrelieren. Mittels multipler Regression kann ermittelt werden, welche Kombination von Skalen der jeweiligen Fragebögen am besten die Zufriedenheit vorhersagen kann, und in welchem Ausmaße. Da es sich bei den Zufriedenheitsurteilen um Ratingskalen, d.h. Rangskalen handelt, ist die multiple ordinale logistische Regression anstelle der klassischen multiplen Regression zu verwenden, die kontinuierlich gemessene Variablen voraussetzt.

3 Datenmaterial und Methoden

3.1 Stichprobe

Die Untersuchung an der Universität Zürich wurde im Sommersemester 2006 in den letzten Wochen vor Ende der Vorlesungszeit durchgeführt. Sie hatte das Ziel, ein neu ent-

wickeltes Lehrevaluationsinstrument im Rahmen einer Pilotstudie vor dem flächendeckenden Einsatz im Hinblick auf ihre Qualität und Güte an einer Stichprobe zu prüfen. Es wurden insgesamt 1.345 Studierende aus 48 Lehrveranstaltungen befragt. In die Untersuchung einbezogen wurden Lehrveranstaltungen aus der Mathematisch-naturwissenschaftlichen Fakultät, der Philosophischen Fakultät, der Rechtswissenschaftlichen Fakultät, der Wirtschaftswissenschaftlichen Fakultät und der Veterinärmedizinischen Fakultät. Die Lehrveranstaltungen waren unterschiedlichen Typs: 14 Lehrveranstaltungen waren Vorlesungen, 22 Seminare, neun Übungen und drei sogenannte Blended-Learning-Lehrveranstaltungen mit Präsenzveranstaltungen in Kombination mit E-Learning-Angeboten.

Ein Teil der Lehrveranstaltungen wurde mit der jeweiligen Papierform des Evaluationsbogens bewertet. Den Studierenden wurde im Rahmen der Lehrveranstaltung Zeit gegeben, den jeweiligen Bogen auszufüllen. Der andere Teil der Lehrveranstaltungsbeurteilung erfolgte online. Über den Internet-Zugang wurden die Studierenden in den jeweiligen Lehrveranstaltungen informiert. Der Online-Bogen wurde aber außerhalb der Lehrveranstaltung ausgefüllt. Die Auswahl der Lehrveranstaltungen und die Zuweisung (Online versus Papier) erfolgte nicht zufällig.

3.2 Fragebögen

Es handelt sich hier um eine im Wesentlichen vollstandardisierte Befragung mit 27 Items und sechs- bzw. dreistufigen Antwortskalen (Likertskalen) und offenen Fragen zur Lehrveranstaltung. Eine „nicht beantwortbar"-Kategorie soll den Studierenden ermöglichen, bestimmte Sachverhalte, die für sie nicht aufgetreten oder beantwortbar sind, zu kodieren.

Es wurde im Vorfeld für jeden Lehrveranstaltungstyp jeweils ein Evaluationsbogen entwickelt. Die Inhalte der Bögen orientieren sich am Heidelberger Inventar zur Lehrevaluation (HILVE) *(Rindermann/Amelang 1994; Rindermann 1995; 2004)*, wobei der HILVE-Bogen keinen Unterschied macht zwischen verschiedenen Lehrveranstaltungstypen. Die Items des Evaluationsbogens für Vorlesungen (siehe Itemliste in Anhang I) lassen sich zu fünf Konstruktbereichen gruppieren: Items zur Beurteilung (1) der *„Lehraktivitäten"* (Organisation, Auseinandersetzung, Interaktion, ...), (2) der *„Lehrperson"* (Fachkompetenz, Klima, Gesamtbeurteilung des Lehrenden, ...), (3) der *„Lehrveranstaltung"* (Interessantheit, Lernen, Allgemeinbeurteilung der Lehrveranstaltung, ...), (4) des *„Workload"* (die von den Studierenden subjektiv wahrgenommenen Leistungsanforderungen, z. B. Stoffumfang) und (5) des *„Studierenden"* (Fleiß, Eigenaktivität, ...). *Rindermann* und *Amelang (1994)* kommen in ihren Untersuchungen zum HILVE mittels exploratorischer Faktorenanalyse

auf sieben Dimensionen, wobei die obigen Konstruktbereiche 1, 2, 3 und 5 den HILVE-Dimensionen 1, 2, 4 und 7 weitgehend entsprechen. Folgende Einzelkonstrukte des Lehrgeschehens sollen durch den Fragebogen abgebildet werden, die im Anhang II näher erläutert werden: *Struktur*, *Auseinandersetzung*, *Verarbeitung*, *Lehrkompetenz*, *Lehrengagement*, *Klima*, *Interessantheit*, *Thema*, *Anforderungen*, *Lernen*, *Interaktion*, *Eigenaktivität*, *Fleiß*, *Hilfsmittel*, *Allgemeinbeurteilung*. Die Items zu diesen Konstrukten werden zu Skalen durch Summierung der Werte je Person zusammengefasst.

Zu Beginn des Fragebogens wird eine Auswahl potenzieller Bias-Faktoren wie „Geschlecht", „Hauptfach oder Nebenfach", „Pflicht-, Wahlpflicht- oder Wahlveranstaltung" und „Bachelor- oder Masterstudium" erhoben. Am Schluss wird gefragt, was den Studierenden besonders gut bzw. weniger gut gefallen hat und um Kommentare/Verbesserungsvorschläge gebeten, jeweils als offene Frage formuliert. Um einen Vergleich über alle Lehrveranstaltungen zu erlauben, wurde ein Set von Items gebildet, das die Bias-Variablen einschließt, die in allen vier Fragebögen vorkommen.

Ergänzt wurden die Bögen um eine Frage zum Verfahren selbst „Mit diesem Fragebogen habe ich meine Beurteilung dieser Lehrveranstaltung angemessen zum Ausdruck bringen können", um die Akzeptanz der Befragung zu erfassen. Zusätzlich wurde um eine Verschlüsselung des Fragebogens (Erster Buchstabe des Vornamens der Mutter, ...) gebeten, um später Fragebögen, die vom selben Studierenden ausgefüllt wurden, zusammenführen zu können. Eine Identifikation des Studierenden selbst (z. B. Name) ist auf diese Weise ausgeschlossen. Um die Online-Version zu erstellen und die automatisierte Dateneingabe und die Feedbackerstellung an die Dozierenden zu gewährleisten, wurden die Evaluationsbögen mittels der Lehrevaluations-Software EVA-SYS *(www.electricpaper.de)* elektronisch implementiert. Aufgrund zeitlicher und finanzieller Restriktionen waren eine umfangreiche Validierung mit externen Kriterien (z. B. Lehrveranstaltungsbewertung durch einen Didaktikexperten), Wiederholungserhebungen oder der zusätzliche Einsatz bekannter Lehrevaluationsinventare für eine MTMM-Analyse nicht möglich.

4 Ausgewählte Ergebnisse

Die folgenden Ergebnisse beziehen sich nur auf die Vorlesungen, da es in dieser Darstellung weniger um die inhaltlichen Ergebnisse geht als um die Aufarbeitung und Nutzung von Lehrevaluationsdaten für die Optimierung der Evaluationsinstrumente.

4.1 Reliabilität

Tabelle 1 enthält die wesentlichen Informationen und Koeffizienten, die eine Bewertung der Reliabilität des Evaluationsbogens für Vorlesungen erlauben. In Fettdruck sind jeweils die Skalen als Summe der genannten Items aufgeführt. Die Mittelwerte sagen zwar über die Reliabilität nicht unmittelbar etwas aus, geben jedoch Hinweise, inwieweit der Wertebereich nach oben oder nach unten schon ausgeschöpft ist, wie symmetrisch die Verteilung ist. Eine starke Abweichung des Mittelwerts von der mittleren Skalenposition z. B. 3.5 bei einer 6-stufigen Antwortskala (1 = *trifft gar nicht zu*, 6 = *trifft voll und ganz zu* bzw. 1 = *sehr unzufrieden*, 6 = *sehr zufrieden*) hat tendenziell auch eine Reduktion der Reliabilität zur Folge. In dieser Untersuchung liegen die Mittelwerte bei den 6-stufigen Ratingskalen zwischen 4.0 und 5.0, sodass der Einfluss auf die Reliabilität noch als gering zu werten ist. Die Nichtbeantwortungsquote (NB %) ist mit Ausnahme der „Teilnehmerzahl" (7 %) recht gering, d. h., die Items erfassen Aspekte des Lehrgeschehens, die in jeder Vorlesung auch auftreten.

Als zentrales Maß der Reliabilität wird häufig Cronbach Alpha als Maß interner Konsistenz angegeben *(Schulz/Greve/Koch/Koops/Wilmers 2006)*, hier berechnet sowohl auf der Ebene der Rohwerte (CRBAR), als auch der Ebene der aggregierten Veranstaltungen (CRBAM) für die gebildeten Subskalen. Während auf der Rohdatenebene (Einzelurteile) die Reliabilitäten mit .67 bis .80 nur bedingt zufriedenstellen, sind die Werte auf der Lehrveranstaltungsebene mit Werten von .72 bis .93 recht hoch. Werden alle Items ohne die Subskalen zu einer Skala summiert, so liegt die interne Konsistenz dieser Summenskala auf Rohdatenebene sehr hoch bei .91, auf Lehrveranstaltungsebene bei .94.

Tabelle 1: Ergebnisse der Itemanalyse für den Evaluationsbogen „Vorlesung"

Item-Kürzel	NB %	M	S	ICC	STD	SHR	CRBAR	CRBAM
Struktur		**4.56**	**0.63**	**.18**	**0.89**	**.93**	**.72**	**.91**
LV gut organisiert	0.6	4.47	0.70	.23	1.06	.72	.57	.84
inhaltlicher Aufbau logisch	0.3	4.65	0.60	.17	1.15	.40	.57	.84
Auseinandersetzung		**4.62**	**0.69**	**.34**	**0.88**	**.94**	**.72**	**.91**
stellt Theorie-Praxis-Bezug her	2.7	4.33	0.80	.35	1.00	.94	.57	.85
erläutert mit Beispielen	1.1	4.87	0.67	.34	0.98	.80	.57	.85
Verarbeitung								
regt zum Mitdenken an	1.4	4.37	0.53	.13	1.03	.90		
Lehrkompetenz		**4.81**	**0.44**	**.20**	**0.68**	**.91**	**.79**	**.87**
ist gut vorbereitet	0.3	5.23	0.48	.22	0.83	.87	.55	.62
macht kompetenten Eindruck	0.5	5.59	0.30	.12	0.62	.68	.40	.47
erklärt Kompliziertes verständlich	1.1	4.49	0.67	.21	0.95	.81	.66	.73
fasst den Stoff zusammen	1.8	4.21	0.56	.13	1.12	.74	.65	.81
gliedert den Stoff übersichtlich	0.8	4.49	0.61	.15	1.06	.82	.66	.93
Lehrengagement								
ist lebendig und engagiert	0.0	4.69	0.65	.26	1.01	.88		
Klima		**5.20**	**0.35**	**.19**	**0.78**	**.45**	**.69**	**.93**
sorgt für angenehme Atmosphäre	0.2	4.99	0.39	.11	0.96	.31	.55	.89
ist freundlich und aufgeschlossen	0.3	5.41	0.31	.13	0.78	.61	.55	.89
Interessantheit								
vermag mich zu interessieren	0.0	4.70	0.42	.11	0.99	.70		
Thema								
Thema ist interessant	0.2	4.70	0.30	.02	1.03	.60		
Anforderungen		**2.13**	**0.18**	**.18**	**0.30**	**.85**	**.67**	**.79**
Stoffumfang*	1.7	2.16	0.21	.19	0.40	.56	.44	.42
Schwierigkeitsgrad*	1.8	2.14	0.26	.39	0.38	.76	.49	.80
Tempo*	1.2	2.10	0.18	.07	0.42	.61	.52	.77
Lernen								
lerne viel in der Veranstaltung	0.2	4.25	0.44	.09	1.08	.14		
Interaktion								
fördert Fragen und Mitarbeit	1.8	4.26	0.86	.40	1.02	.95		
Eigenaktivität								
besuche die LV regelmäßig	0.2	5.37	0.30	.04	0.93	.63		
Fleiß								
bereite die LV vor bzw. nach	0.2	3.50	0.52	.12	1.39	.35		

Tabelle 1, Fortsetzung

Item-Kürzel	NB %	M	S	ICC	STD	SHR	CRBAR	CRBAM
Hilfsmittel		**4.21**	**1.00**	**.43**	**0.99**	**.86**	**.71**	**.72**
setzt Hilfsmittel hilfreich ein	3.0	4.16	1.35	.59	0.97	.95	.55	.60
Lernhilfen sind nützlich	3.2	4.14	0.94	.37	1.25	.71	.55	.60
Sonstiges								
Teilnehmerzahl*	7.0	2.08	0.28	.47	0.38	.77		
Allgemeinbeurteilung		**4.75**	**0.45**	**.18**	**0.82**	**.82**	**.80**	**.83**
Zufriedenheit mit der LV		4.52	0.56	.21	0.99	.60	.67	.74
Zufriedenheit mit der Lehrperson		4.98	0.40	.16	0.89	.49	.67	.74

Anmerkungen: N = 660 Studierende aus 14 Vorlesungen (LV), NB % = Anteil „nicht beantwortbar (N. B.)", M = Mittelwert, S = Standardabweichung der Veranstaltungsmittelwerte, ICC = Intraklassenkorrelation, STD = Mittelwert der Innerhalbstreuungen, SHR = Split-Half-Reliabilität, CRBAR = Cronbach Alpha der Skalen (Fettdruck) und Trennschärfen bezogen auf die Rohdaten, CRBAM = Cronbach Alpha der Skalen (Fettdruck) und Trennschärfen bezogen auf Veranstaltungsmittelwerte.

* Antwortskala (1 = *zu klein/zu tief*, 2 = *gerade richtig*, 3 = *zu groß/zu hoch*)

Rindermann (2003) gibt, bezogen auf die Ebene der Einzelurteile, ein Cronbach Alpha von durchschnittlich .76, bezogen auf die Ebene der Lehrveranstaltungen, eine interne Konsistenz von .86 für das HILVE-Inventar an.

Da letztlich die Veranstaltungsmittel zuverlässig erfasst werden sollen, sind die Maße auf Veranstaltungsebene für die Bewertung der Reliabilität entscheidend. Hierfür ist auch die Intraklassenkorrelation (ICC) interessant. Für die Einzelitems wurde dabei das Ordinalskalenniveau in der Berechnung des ICC beachtet (SAS-Software, proc nlmixed). Auffallend ist, dass mit Ausnahme der Items der Skalen „Auseinandersetzung", „Interaktion", „Hilfsmittel" und „Sonstiges" (Teilnehmerzahl) die Intraklassenkorrelationen gering ausgeprägt sind mit Werten um .10/.20. *Rindermann (2003)* gibt aus seinen Studien zum HILVE eine Interraterreliabilität von .30, aus internationalen Studien eine Interraterreliabilität von .24 an *(*siehe auch *Feldman 1977).* Die Ursache liegt in der hohen Variabilität innerhalb der Lehrveranstaltungen, worauf die relativ hohen Mittelwerte der Innerhalbstreuungen (STD) hinweisen, die bei den meisten Items einen Wert bis zu einer Skalenstufe (1.0) annehmen. Die Unterschiede zwischen den Lehrveranstaltungen, gemessen über die Standardabweichung der Mittelwerte der Lehrveranstaltungen (S), liegen für einzelne Skalen wie „Struktur", „Auseinandersetzung" und „Lehrengagement" bei fast zwei Dritteln einer Skalenstufe (0.66), was darauf hinweist, dass Studierende die einzelnen Vorlesungen in ihrem Urteil sehr wohl unterscheiden. Dies gilt nicht für alle Skalen.

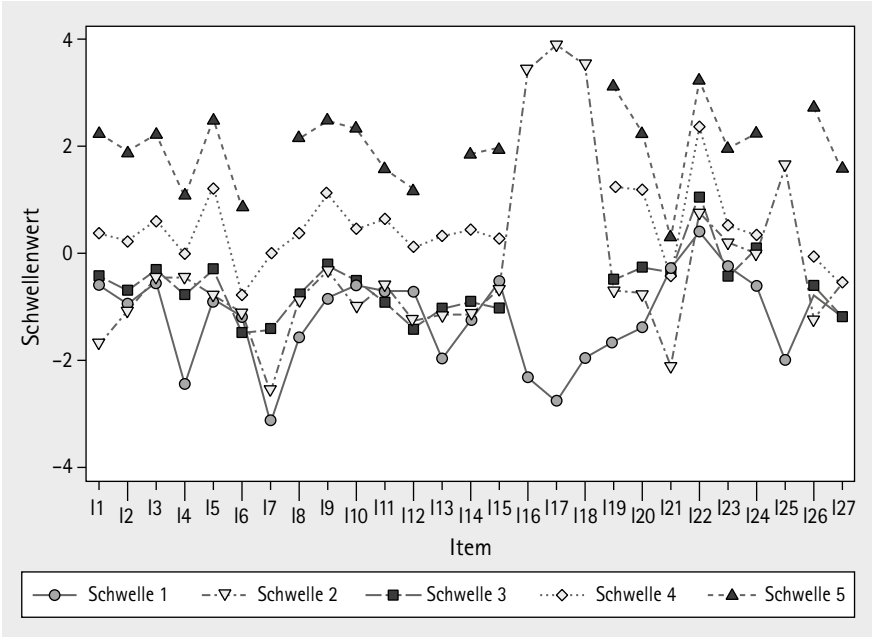
So ist für die Skala „Klima" die Variabilität zwischen den Lehrveranstaltungen (S) deutlich geringer, was auch eine geringe Intraklassenkorrelation (ICC) und interne Konsistenz (CRBAR, CRBAM) zur Folge hat.

Überraschend ist, dass die Split-Half-Reliabilität (SHR) für einzelne Items oder Skalen recht hohe Werte annehmen kann, z. B. Struktur, Auseinandersetzung, Lehrkompetenz, d. h., die mittleren Bewertungen innerhalb einer Lehrveranstaltung recht eng zusammenhängen mit Ausnahme der Skala „Lernen" mit einem SHR von .14. Die berechneten Reliabilitätskoeffizienten müssen jedoch mit Vorsicht interpretiert werden, da die Anzahl der Lehrveranstaltungen mit n = 14 nicht hoch ist. Die Ergebnisse können jedoch genutzt werden, Items zu eliminieren, gegebenenfalls zu verändern oder umzuformulieren.

4.2 Antwortformat

Zur Prüfung des Antwortformats wird ein ordinales Rasch-Modell verwendet, das es erlaubt, für jedes Item Schwellenparameter für den Übergang von einer Stufe der 6-stufigen Ratingskala zur nächsten Stufe zu schätzen. Die Anzahl der Schwellenparameter ergibt sich aus der Anzahl Stufen der Ratingskala minus 1, d. h., bei einer 6-stufigen Ratingskala sind es 5 Schwellenparameter, bei einer 3-stufigen Skala 2. Niedrige Schwellenparameter geben an, dass es den Personen einfach fällt, eine Stufe höher zu gehen in der Beantwortung, bei hohen Schwellenparametern fällt es den Personen schwer, die jeweilige Stufe zu verlassen. In Abbildung 1 sind die Schwellenparameter für jedes Item des Vorlesungsbogens dargestellt. Der Inhalt des Items lässt sich entsprechend dem Itemcode (z. B. I1) Anhang I entnehmen.

Abbildung 1: Schwellenparameter für die ordinalskalierten Items des Evaluationsbogens „Vorlesung"



Zwei Sachverhalte werden in der Abbildung deutlich: Zum einen besteht zwischen den Schwellenwerten 1, 2 und 3 bei den meisten Items mit 6-stufigen Ratingskalen kein großer Unterschied, was daraufhin weist, dass Studierende im unteren Bereich der Skala (negativer Bereich der Bewertungen) nicht sehr differenzieren. Eine Zusammenfassung der Werte würde zu keinem Informationsverlust führen. Die meiste Trennkraft besteht zwischen den Skalenwerten 4 und 5 und vor allem zwischen den Werten 5 und 6. Zum anderen ist bei den dreistufigen Items (I16, I17, I18) der erste Schwellenparameter sehr niedrig, der zweite sehr hoch, d. h., es ist auf der einen Seite sehr leicht, die erste Stufe der Ratingskala zu übersteigen, auf der anderen Seite aber sehr schwer, die zweite Stufe zu verlassen. Dies führt zu einer Tendenz zur Mitte und zu einer geringen Differenzierung der Lehrveranstaltungen im Hinblick auf die Anforderungen. Bei dem Item „Teilnehmerzahl" ist dieses Phänomen nicht so stark ausgeprägt.

4.3 Validität

Aufgrund der geringen Anzahl von Lehrveranstaltungen wurde die Dimensionalität des Fragebogens bezüglich der Rohdaten geprüft unter Einsatz einer konfirmatorischen Faktorenanalyse für ordinale Variablen *(Muthén/Muthén 2006)* ohne Berücksichtigung der Mehrebenenstruktur (Studierende, Lehrveranstaltung). Die oben genannten fünf zentralen Konstruktbereiche des Fragebogens (*Lehraktivitäten*, *Lehrperson*, *Lehrveranstaltung*, *Workload*, *Studierende*) wurden als Faktoren innerhalb eines Strukturgleichungsmodells definiert. Der Modelltest im Rahmen der konfirmatorischen Faktorenanalyse zeigt eine akzeptable Passung des Modells auf die Daten: Der Comparative Fit Index (CFI) liegt bei .93, der Root Mean Square Error of Approximation (RMSEA) bei .079, was nach *Hair, Anderson, Tatham* und *Black (1998)* noch als angemessen eingestuft wird (RMSEA < .80). Der Modell-χ^2-Test (H_0: Modell gilt) ist zwar signifikant ($\chi^2(97) = 397.5$ $p < .05$), d. h. Ablehnung des Modells, aber infolge des großen Stichprobenumfangs nur bedingt geeignet zur Modellbeurteilung. Die Einzelskalen mit Ausnahme des Workload können mit der Faktorenanalyse nicht repliziert werden.

4.4 Fairness

Bias-Variablen sind Größen, die zwar Einfluss auf das Lehrgeschehen haben, aber nicht unter der Kontrolle der Lehrenden stehen, sie können die Fairness des Verfahrens beeinträchtigen. So können beispielsweise Studierende, die sich unabhängig von der Präsentation des Lehrstoffes für das Thema der Lehrveranstaltung sehr interessieren, eine Lehrveranstaltung anders beurteilen als Studierende, die sich weniger für das Thema interessieren. Als potenzielle Bias-Variablen wurden, neben dem Interesse am Thema, Geschlecht, Haupt-/Nebenfach, Pflichtfach oder Wahl-/Wahlpflichtfach, Veranstaltungsgröße, Papier/Online-Befragung in die Analyse einbezogen.

Um den Einfluss von Bias-Variablen zu prüfen, wurden in Anlehnung an *Rindermann (1995)* Rangkorrelationen (Kendall Tau-b) von potenziellen Bias-Variablen mit den Skalen des Fragebogens berechnet (Tabelle 2). Die Bias-Variable „Interesse am Thema unabhängig von der Art der Vermittlung" (D1) zeigt die höchsten – fast durchgehend positiven – Korrelationen mit den Skalen des Evaluationsbogens „Vorlesung" auf. Insgesamt ist der Einfluss der Bias-Variablen mit Korrelationen unter .30 oder gar .10 als gering einzuschätzen trotz statistischer Signifikanz, die aber bei einer Stichprobengröße von N = 660 auch nicht überrascht.

Tabelle 2: Zusammenhang der Skalen des Evaluationsbogens „Vorlesung" mit den Bias-Variablen (Kendallsche Tau-b Korrelation)

Skalen	Thema	Gender	Haupt-/ Nebenfach	Pflicht-/ Wahlfach	Veranstal-tungsgröße	Papier/ Online
Struktur	.28*	-.11*	.03	-.03	.14*	-.07*
Auseinandersetzung	.20*	-.12*	-.07*	-.12*	.23*	-.04
Verarbeitung	.25*	-.09*	.01	.14*	-.03	.03
Lehrkompetenz	.27*	-.11*	-.01	-.04	.18*	-.05
Engagement	.15*	-.12*	-.11*	-.08*	.23*	-.01
Klima	.16*	-.06	-.16*	-.01	.20*	-.06
Interessantheit	.38*	-.13*	.01	.01	.12*	-.05
Anforderungen	-.14*	-.07	-.01	.11*	-.15*	-.02
Lernen	.36*	-.07	.00	.01	.03	-.05
Interaktion	.18*	-.17*	-.01	-.04	.20*	-.02
Eigenaktivität	.18*	-.04	-.09*	-.07	.11*	-.02
Fleiß	.19*	-.04	-.05	-.06	-.03	.09*
Hilfsmittel	.17*	-.12*	-.02	-.10*	.25*	-.07*
Teilnehmerzahl	.02	-.07	-.22*	-.28*	.49*	-.27*
Allgemeinbeurteilung	.31*	-.09	.01	-.01	.17*	-.09*

Anmerkungen: Gender (0 = Frau, 1 = Mann), Haupt-(= 0)/Nebenfach und anderes (= 1), Pflicht (= 0)/Wahlpflicht (= 1), Papier (= 1)/Online(= 0) (N = 660).

* $p < .05$

Überraschend ist auch nicht, dass die objektive Veranstaltungsgröße mit der subjektiv eingeschätzten Größe (Teilnehmerzahl) mäßig korreliert (r = .49). Interessant ist auch, dass die Veranstaltungsgröße gering positiv mit den Skalen „Lehrkompetenz", „Lehrengagement", „Klima" korreliert. Mit steigender Veranstaltungsgröße werden diese Aspekte tendenziell positiver gesehen. Die Modalität des Fragebogens (Papierform oder Online) hat keinen verzerrenden Einfluss auf das Antwortverhalten der Studierenden.

4.5 Prädiktoren der Allgemeinbeurteilung

Da die Allgemeinbeurteilungen von Lehrveranstaltung und Lehrperson häufig für Vergleiche von Lehrveranstaltungen herangezogen werden und Lehrpersonen sich zuerst an diesen globalen Urteilen orientieren (werde ich gut oder weniger gut beurteilt?), ist es

von besonderem Interesse zu wissen, welche Skalen mit den globalen Urteilen korrelieren. Dies erfolgt mit der multiplen Regression, spezifisch der multiplen ordinal logistischen Regression, da es sich bei den Allgemeinbeurteilungen um Ratingskalen handelt.

Die Ergebnisse zeigen (Likelihood-Ratio-χ^2 (7) = 520.9* $p < .05$), dass in der Reihenfolge der Wichtigkeit (Grad der erklärten Varianz der Veranstaltungszufriedenheit) „Lehrkompetenz", „Lernen", „Interessantheit", „Struktur", „Thema" und „Hilfsmittel" einen positiven Einfluss, Teilnehmerzahl dagegen einen negativen Einfluss auf die Zufriedenheit mit der Lehrveranstaltung insgesamt haben (65.7 % Varianzerklärung, wobei 55.5 % allein die Lehrkompetenz erklärt). Wird die Teilnehmerzahl als zu hoch eingeschätzt, sinkt die Zufriedenheit mit der Lehrveranstaltung. Je positiver die Lehrkompetenz bewertet wird, je mehr die Studierenden meinen, etwas in der Vorlesung gelernt zu haben, je mehr die Vorlesung Interesse weckt, strukturiert ist, das Thema interessant und die Teilnehmerzahl nicht zu hoch ist, desto besser fällt die Gesamtbeurteilung der Lehrveranstaltung aus.

Für die Zufriedenheit mit den Dozierenden ergab sich Folgendes (Likelihood-Ratio-χ^2 (7) = 520.9* $p < .05$): Je positiver die Lehrkompetenz und das Lehrengagement der Lehrenden eingeschätzt werden, je mehr die Lehrenden imstande sind, eine angenehme Atmosphäre zu schaffen und je mehr Studierende meinen, in der Vorlesung etwas gelernt zu haben, desto besser fällt die Gesamtbeurteilung der Lehrenden aus (59.4 % Varianzerklärung, wobei die Lehrkompetenz 49.1 % erklärt). Diese Informationen dürfen nicht kausal interpretiert werden, da die Gesamtbewertung einer Lehrveranstaltung umgekehrt die Einzelbeurteilungen beeinflussen kann.

5 Zusammenfassung und Schlussfolgerung

Um die Qualität der Lehre zu sichern, ist es nicht ausreichend, Lehrveranstaltungen durch Studierende beurteilen zu lassen. So hat beispielsweise die *Kultusministerkonferenz (2007)* beschlossen, ab 2008 eine Überprüfung und Zertifizierung hochschulinterner Qualitätssicherungssysteme durchführen zu lassen. Dies ist sehr sinnvoll, da Hochschulen durch die heute fast flächendeckende Umsetzung von Lehrevaluationen, meist zentralisiert und Web-basiert, über ein ständig wachsendes Reservoir von Daten zur Lehrveranstaltungsevaluation verfügen, die nicht nur für die Rückmeldung an die Lehrpersonen, sondern auch zur Sicherung der Qualitätssicherungsinstrumente selbst eingesetzt werden können.

Welche Schlussfolgerungen für die Sicherung der Qualität von Lehrevaluationsinventaren lassen sich aus Ergebnissen von testtheoretischen Analysen beispielhaft für die Universität Zürich ziehen?

- *Reliabilität*: Die interne Konsistenz konnte nur für einzelne Subskalen überprüft werden, da für andere Bereiche, z. B. Dozentenengagement, nur jeweils ein Item vorlag. Sie ist auf der Einzelurteilerebene weniger zufriedenstellend als auf der Lehrveranstaltungsebene. Hieraus folgt, für die ausgewählten Lehrevaluationsdimensionen Skalen mit mindestens drei bis vier Einzelitems zu konstruieren. Bezüglich Items mit Intraklassenkorrelationen unter .20, wie z. B. für die Skala „Klima", müsste nochmals überprüft werden, ob dies an den Items oder an der geringen Streuung der untersuchten Veranstaltungen liegt. Das Item „lerne viel" ist als Maß der wahrgenommenen Effektivität einer Lehrveranstaltung zwar sehr wichtig, trennt aber mit einem ICC = .09 wenig zwischen den Lehrveranstaltungen und könnte entfernt werden, ebenso die Variable „Fleiß" (ICC = .04).

- *Konstruktvalidität*: Die Konstruktbereiche können zwar durch die konfirmatorische Faktorenanalyse vergleichbar den Ergebnissen von *Rindermann* und *Amelang (1994)* repliziert werden, nicht jedoch die Einzelkonstrukte, sprich Subskalen mit Ausnahme des Workload, was aber auch in *Rindermann* und *Amelang (1994)* nicht gelang. Hieraus ergibt sich die Notwendigkeit, Subskalen mit einer größeren Zahl von Items zu konstruieren oder die Konstruktbereiche zu Subskalen zu machen.

- *Gesamtindikator*: Für die Steuerung einer Hochschule sind quantitative Indikatoren oder Zielgrößen wichtig. So könnte die Skala „Allgemeinbeurteilung" mit einer internen Konsistenz von .80 zufriedenstellend als Gesamtindikator für die Akzeptanz der Lehre verwendet werden. Darüber hinaus hängen einzelne wichtige Subskalen und Items sehr stark mit der Allgemeinbeurteilung zusammen.

- *Einfluss von potenziellen Biasfaktoren*: Es wurde ein geringer bis mäßiger Einfluss der Veranstaltungsgröße und des Interesses am Thema („Thema") auf die Items und Subskalen des Fragebogens nachgewiesen. Daher ist dringend zu empfehlen, zusätzlich zu den Lehrevaluationsitems potenzielle Bias-Variablen wie Interesse am Thema, Geschlecht, Studiengang zu erheben. Zum einen soll damit der Einfluss dieser Variablen im Rahmen eines Monitorings kontinuierlich überprüft werden. Zum anderen soll die Möglichkeit eröffnet werden, Datenkorrekturen mit statistischen Verfahren vorzunehmen, um die Fairness der studentischen Beurteilung überprüfen zu können.

- *Antwortformat*: Es zeigte sich, dass die dreistufigen Ratingskalen wenig geeignet sind, das Lehrgeschehen trennscharf zu bewerten. Hier ist zu empfehlen, die dreistufigen Antwortskalen auf mindestens sechsstufige zu erweitern, um Differenzierungsmöglichkeiten für die Studierenden zu ermöglichen. Die sechsstufigen Antwortskalen sind im unteren negativen Bereich wenig trennscharf. Hier ist zu empfehlen, z. B. mit optischen Mitteln (Smilies, Gesichter mit unterschiedlichen Emotionen) den vollen Wertebereich der Beurteilung deutlich zu machen und/oder die Skalenpolarität (unipolar, bipolar) zu variieren *(Sedlmeier 2006)*. Die sechsstufige Antwortskala könnte auch auf sieben oder acht Skalenpunkte erweitert werden.
- *Ableitung von statistischen Normen*: Für ein Feedback an die Lehrenden, wie auch zur externen Prüfung der Qualität der Lehre können statistische Normen hilfreich sein, die sich aus umfangreichen Datenbeständen durchgeführter Lehrevaluationen oder aus einer Normstichprobe ableiten. Sie geben an, wie viel Prozent der Lehrveranstaltungen mit welchem Skalenwert (z. B. Allgemeinbeurteilung) bewertet werden. Wenn beispielsweise lediglich 5 Prozent aller Lehrveranstaltungen mit einem durchschnittlichen Skalenwert unter 3.0 auf einer sechsstufigen Antwortskala (z. B. 1 = *sehr unzufrieden*, 6 = *sehr zufrieden*) bewertet werden, so ist eine Lehrveranstaltung mit einem mittleren Skalenwert von 2.8 vergleichsweise als sehr schlecht beurteilt.

Eine Auswahl methodischer Möglichkeiten zur Prüfung der Testgütekriterien wie Reliabilität, Validität und Fairness wurde aufgezeigt, die prinzipiell Gegenstand von Sekundäranalysen von bestehenden Datenbeständen bereits durchgeführter Lehrevaluationen sein können. Aufgrund der eher geringen Zahl von Lehrveranstaltungen und des Zeitpunkts der Befragung am Ende des Semesters, der eine selektive Auswahl von Studierenden zur Folge haben kann, lassen sich die konkreten Ergebnisse nicht unbedingt auf andere Hochschulen übertragen.

Insgesamt wäre es wünschenswert, Standards der Güte von Lehrevaluationsinstrumenten zu entwickeln, um die Qualität der Lehrevaluationen sowohl innerhalb als auch über die Universitäten hinweg zu garantieren. Würden sich mehrere Universitäten bereit erklären, denselben Fragebogen einzusetzen und die erhobenen Daten für statistische Zwecke einer zentralen Datenbank zur Verfügung zu stellen, könnte eine breite empirische Datengrundlage geschaffen werden für die Beantwortung von Fragen zur Güte des studentischen Urteils im deutschsprachigen Raum.

Anhang I Originalformulierung der Items des Evaluationsbogens für Vorlesungen

Item-Kürzel	Originalformulierung des Items mit 6-stufiger Antwortskala (1 = trifft gar nicht zu, ..., 6 = trifft voll und ganz zu)
I1 Die LV ist gut organisiert	Die Veranstaltung ist gut organisiert (Information, Kommunikation, Durchführung)
I2 inhaltlicher Aufbau logisch	Der inhaltliche Aufbau der Veranstaltung ist logisch/ nachvollziehbar
I3 stellt Theorie-Praxis-Bezug her	Stellt einen Bezug zwischen Theorie und Praxis her
I4 erläutert mit Beispielen	Erläutert den Stoff anhand von Beispielen
I5 regt zum Mitdenken an	Regt mich zu kritischem Mit- und Selberdenken an
I6 ist gut vorbereitet	Wirkt bei den einzelnen Lektionen gut vorbereitet
I7 macht kompetenten Eindruck	Macht fachlich einen kompetenten Eindruck
I8 erklärt Kompliz. verständlich	Erklärt komplizierte Sachverhalte verständlich
I9 fasst den Stoff zusammen	Fasst den behandelten Stoff der Veranstaltung übersichtlich zusammen
I10 gliedert den Stoff übersichtlich	Gliedert den Stoff übersichtlich
I11 ist lebendig und engagiert	Gestaltet die Veranstaltung lebendig und engagiert
I12 sorgt für angenehme Atmosphäre	Sorgt für eine angenehme Atmosphäre in der Veranstaltung
I13 ist freundlich und aufgeschlossen	Ist im Umgang mit Studierenden freundlich und aufgeschlossen
I14 vermag mich zu interessieren	Vermag mich für den Stoff der Veranstaltung zu interessieren
I15 Thema ist interessant	Am Thema der Veranstaltung bin – unabhängig von der Art der Vermittlung durch die Lehrperson – sehr interessiert
I16 Stoffumfang*	Der Stoffumfang, der in der Veranstaltung behandelt wird, ist
I17 Schwierigkeitsgrad*	Der Schwierigkeitsgrad der Veranstaltung ist
I18 Tempo*	Das Tempo der Veranstaltung ist für mich
I19 lerne viel in der Veranstaltung	Ich lerne viel in der Veranstaltung
I20 fördert Fragen und Mitarbeit	Fördert Fragen und aktive Mitarbeit
I21 besuche die LV regelmäßig	Ich besuche die Veranstaltung regelmäßig
I22 bereite die LV vor bzw. nach	Ich bereite die Veranstaltung regelmäßig vor bzw. nach
I23 setzt Hilfsmittel hilfreich ein	Setzt die Hilfsmittel (z. B. Folien, Beamer, Internet, Computer) hilfreich ein
I24 Lernhilfen sind nützlich	Die Lernhilfen (z. B. Skript, Internet, CD) sind für das Verständnis/ Lernen nützlich
I25 Teilnehmerzahl*	Die Zahl der Teilnehmerinnen und Teilnehmer in der Veranstaltung ist
I26 Zufriedenheit mit der Vorlesung‡	Wie zufrieden sind Sie mit der Veranstaltung insgesamt?
I27 Zufriedenheit mit der Lehrperson‡	Wie zufrieden sind Sie mit der Dozentin/dem Dozenten insgesamt?

* 3-stufige Antwortskala (1 = *zu klein/zu tief*, 2 = *gerade richtig*, 3 = *zu groß / zu hoch*)

‡ 6-stufige Antwortskala (1 = *sehr unzufrieden*, ..., 6 = *sehr zufrieden*)

Anhang II Beschreibung der Konstrukte des Evaluationsbogens für Vorlesungen
(in Anlehnung an *Rindermann und Amelang (1994)*)

Struktur: beschreibt Aufbau und Organisation der Veranstaltung.

Auseinandersetzung: thematisiert die erläuternde Behandlung des Stoffes. Beispiele und ein Theorie-Praxis-Bezug vertiefen das Thema und zeigen die Relevanz des Stoffes.

Lehrkompetenz fragt danach, ob die Dozentin oder der Dozent didaktisch überzeugt. Wird Kompliziertes verdeutlicht, der Stoff regelmäßig zusammengefasst und sind die Veranstaltungen gut vorbereitet?

Lehrengagement: Motiviert die Lehrkraft, nimmt sie den Erfolg der Lehre wichtig?

Mit **Klima** wird die Atmosphäre in der Veranstaltung erhoben. Gibt es einen guten Umgang zwischen den Dozierenden und den Studierenden? Das Klima wird sowohl von Lehrenden als auch Studierenden beeinflusst.

Interessantheit erhebt im Gegensatz zur Interessantheit des Themas die der Veranstaltung. Ist die Vorlesung interessant? Hier soll die Gestaltung gemessen werden, nicht das Thema.

Thema: Haben sich die Studierenden schon vor Beginn des Kurses für das Thema interessiert? Thema stellt weitgehend eine „Biasvariable" dar, d. h. nicht die Veranstaltung oder der Dozent werden bewertet.

Anforderungen sollen messen, ob die Teilnehmerinnen und Teilnehmer stark in der Stoffschwierigkeit, der Stoffmenge und der Geschwindigkeit gefordert werden. Im Gegensatz zu anderen Skalen sind hier mittlere oder nur leicht höhere Werte optimal.

Effektivität des Lernens: Diese Dimension erfasst die Selbsteinschätzung, ob die Studierenden etwas in der Veranstaltung lernen können.

Interaktionsmanagement beurteilt die Förderung und Moderation von Interaktionen unter den Studierenden.

Mit **Eigenaktivität** wird erfasst, inwieweit der Studierende die Vorlesung regelmäßig besucht.

Mit **Fleiß** soll die Mitarbeit der Studierenden erhoben werden. Werden gestellte Aufgaben wie Literaturlesen oder Vorbereitung der Stunden gemacht?

Hilfsmittel misst, inwieweit der Lehrende fähig ist, mit unterschiedlichen Mitteln zur Unterstützung der Lehre umzugehen und inwieweit die abgegebenen Unterlagen hilfreich sind für das Lernen.

Mit der **Allgemeineinschätzung** wird die allgemeine Bewertung sowohl der Veranstaltung als auch der Dozierenden erhoben, orientiert an der schweizerischen Notenskala.

Literatur

Cranton, P.; Smith, R. A. (1990): Reconsidering the unit of analysis: a model of student rating of instruction. In: Journal of Educational Psychology, 82 (2), S. 207–212

Cronbach, L. J.; Meehl, P. E. (1955): Construct validity in psychological tests. In: Psychological Bulletin, 52, S. 281–302

Daniel, H.-D. (1996): Evaluierung der universitären Lehre durch Studenten und Absolventen. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 16 (2), S. 149–164

Daniel, H.-D. (1998): Beiträge der empirischen Hochschulforschung zur Evaluierung von Forschung und Lehre. In Teichler, U.; Daniel, H.-D.; Enders, J. (Hrsg.): Brennpunkt Hochschule. Neuere Analysen zu Hochschule, Beruf und Gesellschaft. Frankfurt, S. 11–53

Daniel, H.-D. (2000): Die Bewertung der Lehre durch Studierende. Ein Beispiel aus Baden-Württemberg. In: Beiträge zur Hochschulforschung, 3, S. 275–296

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

Greenwald, A. G. (1997): Validity concerns and usefulness of student ratings of instruction. In: American Psychologist, 52 (11), S. 1182–1186

Greimel-Fuhrmann, B.; Geyer, A. (2005): Die Wirkung von Interesse und Sympathie auf die Gesamtbeurteilung in der Lehrevaluation. Direkte und indirekte Effekte unter Berücksichtigung des Lehrverhaltens. In: Empirische Pädagogik, 19 (2), S. 103–120

Hair, J. F.; Anderson, R. E.; Tatham, R. L.; Black, W. (1998): Multivariate data analysis (5th ed.). New Jersey

Hufen, F. (1995): Rechtsfragen der Lehrevaluation an wissenschaftlichen Hochschulen, Rechtsgutachten. Bonn

Klein, M.; Rosar, U. (2006): Das Auge hört mit! Der Einfluss physischer Attraktivität des Lehrpersonals auf die studentische Evaluation von Lehrveranstaltungen – eine empirische Analyse am Beispiel der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln. In: Zeitschrift für Soziologie, 4, S. 305–316

Kultusministerkonferenz (2007): Ergebnisse der 318. Plenarsitzung der Kultusministerkonferenz am 14.6.2007. Bonn www.kmk.org (Zugriff am 12.7.2007)

Kromrey, H. (1994): Wie erkennt man „gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen. In: Empirische Pädagogik, 8 (2), S. 153–168

Lienert, G. A. (1969): Testaufbau und Testanalyse. Weinheim

Loewenthal, K. M. (2001): An introduction to psychological tests and scales. Church Road

Marsh, H. W. (1982): Validity of students' evaluation of college teaching: A multitrait-multimethod analysis. In: Journal of Educational Psychology, 74 (2), S. 264–279

Marsh, H. W. (1984): Students' evaluation of teaching: dimensionality, reliability, validity, potential bias, and utility. In: Journal of Educational Psychology, 76 (5), S. 707–754

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

Marsh, H. W.; Roche, L. A. (1997): Making student evaluation of teaching effective. The critical issues of validity, bias, and utility. In: American Psychologist, 52 (11), S. 1187–1197

Mußgnug, R. (1992): Gefährden Lehrevaluationen die Freiheit der Wissenschaft? In: Mitteilungen des Hochschulverbandes, 40 (4), S. 253–256

Muthén, L. K. ; Muthén, B. O. (2006): Mplus User's Guide. Fourth Edition. Los Angeles, CA

Mutz, R. (2000): Studienreform als Programm. Programmevaluation zur Akzeptanz des reformierten Studiengangs Forstwissenschaften bei Lehrenden und Studierenden. Landau

Mutz, R. (2003): Multivariate Reliabilitäts- und Generalisierbarkeitstheorie in der Lehrevaluationsforschung. In: Zeitschrift für Pädagogische Psychologie, 17 (3–4), S. 245–254

Rindermann, H. (1995): Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen (Psychologie Bd. 6). Landau

Rindermann, H. (2003): Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. In: Zeitschrift für Evaluation, 2, S. 233–256

Rindermann, H. (2004): HILVE-II in einer computerbasierten Form mit Normen, individualisierter, ergebnisabhängiger Rückmeldung, mit Interpretationshilfen und Beratungsvorschlägen zur Verbesserung der Lehre und mit automatisierter Auswertung und Ergebniszustellung über EvaSys. Lüneburg

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

Rost, J. (2004): Lehrbuch Testtheorie – Testkonstruktion. Bern

Ross, S. M. (2007): Introduction to probability models (9th ed.). New York

Schulz, N.; Greve, W.; Koch, U.; Koops, T.; Wilmers, N. (2006): Wie gut erfassen Fragebögen die Qualität der Lehre? In: Krampen, G.; Zayer, H. (Hrsg.): Didaktik und Evaluation in der Psychologie. Göttingen, S. 75–89

Sedlmeier, P. (2006): The role of scales in student ratings. In: Learning and Instruction 16 (5), S. 401–415

Spiel, C. (2001): Der differentielle Einfluss von Bias-Variablen auf studentische Lehrveranstaltungsbewertungen. In: Engel, U. (Hrsg.): Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre. Frankfurt a. M., S. 61–82

Süllwold, F. (1992): Welche Realität wird bei der Beurteilung von Hochschullehrern durch Studierende erfasst? In: Mitteilungen des Hochschulverbandes 40 (1), S. 34–35

Wirtz, M. ; Caspar, F. (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen

Anschrift der Verfasser:
Prof. Dr. Hans-Dieter Daniel
Dr. Rüdiger Mutz
Kontaktadresse:
Dr. Rüdiger Mutz
Professur für Sozialpsychologie und Hochschulforschung
ETH Zürich
Zähringerstrasse 24
CH-8091 Zürich
E-Mail: mutz@gess.ethz.ch

Prof. Dr. Hans-Dieter Daniel ist wissenschaftlicher Leiter der Evaluationsstelle der Universität Zürich und Professor für Sozialpsychologie und Hochschulforschung an der ETH Zürich.
Dr. Rüdiger Mutz ist wissenschaftlicher Mitarbeiter an der Professur für Sozialpsychologie und Hochschulforschung der ETH Zürich.