

Welche Faktoren beeinflussen die Evaluation von Lehrkräften?

Which factors affect the evaluation of teachers?

Bettina Greimel und Alois Geyer

Wirtschaftsuniversität Wien

Zusammenfassung: In der vorliegenden Untersuchung wird analysiert, welche Faktoren die Gesamtbeurteilung von Lehrenden durch Lernende in welchem Ausmaß beeinflussen. Der gewählte multivariate Ansatz berücksichtigt potenziell verzerrende Faktoren im Kontext von logo- und paidotropen Lehrverhaltensfaktoren. Neben den üblicherweise überprüften Biasvariablen wird eine Reihe weiterer Faktoren wie beispielsweise die Leistungsangst und – motivation der Lernenden, ihre Attribuierung von Erfolg im Fach, die Attraktivität der Lehrkraft und der Zweifel der Lernenden bezüglich ihrer Urteilsfähigkeit betrachtet. Außerdem berücksichtigen wir bei den Signifikanztests die Heteroskedastizität in den Daten. Die Ergebnisse zeigen, dass die Gesamtbeurteilung der Lehrkräfte zwar hauptsächlich von den Lehrverhaltensvariablen abhängt, dass jedoch zusätzlich auch das Alter der Lehrkraft, der Zweifel der Lernenden und die Erfolgsattribution wirksam sind.

Schlüsselwörter: Evaluation, Biasvariablen, Bewertung durch Lernende, Einflussfaktoren

Summary: This study aims to examine which factors affect students' global ratings of teachers. The multivariate approach takes potential biasing variables into account in the context of «logo- and paidotropic» factors of teaching behaviour. In addition to the usually analyzed biasing variables we also considered factors such as the students' motivation to learn and their fear of failure, their attribution of success in the subject, the teachers' attractiveness and students' own doubts whether they are able to adequately evaluate their teachers. Our significance tests account for heteroscedasticity in the data. The results show that the global ratings of teachers mainly depend on their teaching behaviour. However, we also find additional effects introduced by the teacher's age as well as the students' doubt and attribution of success.

Keywords: evaluation, biasing variables, student's ratings, influences on SETs

1 Einleitung

Zur Lehrerevaluation werden häufig die Lernenden gebeten, ihre Lehrenden und deren Lehrverhalten zu beurteilen sowie ihre Zufriedenheit mit dem Unterricht der Lehrkraft anzugeben. Dementsprechend stehen Untersuchungen zur Aussagekraft solcher Beurteilungen im Mittelpunkt der Forschung zur Unterrichtsevaluation. Die Fragestellung, ob die Beurteilungen der Lernenden valide Rückschlüsse auf die Qualität der Lehrkräfte (und damit auch auf die Güte des Unterrichts) zulassen, hat einen nahezu unüberschaubaren Korpus an Forschungsarbeiten zur Untersuchung der Validität von Urteilen der Lernenden zur Beurteilung der Lehrenden produziert. Im Wesentlichen ist jedoch Weinert (2001) zuzustimmen, der von einem «inflationär anwachsenden, zugleich aber hoch redundanten Schrifttum zur Evaluationsproblematik» spricht.

Zur Untersuchung und Beantwortung der Validitätsfrage wurden hauptsächlich drei Ansätze verfolgt: Der erste Ansatz beruht auf der Annahme, dass gute Lehrkräfte die Lehrziele in hohem Ausmaß erreichen. Diese Lehrziele bestehen in der Regel darin, dass die Lernenden Lernerfolge wie Wissenszuwachs und/oder Kompetenzerwerb erzielen. Daher wurde der Lernerfolg der Lernenden als Außenkriterium für die Güte des Unterrichts herangezogen und mit der Beurteilung der Lernenden korreliert (Kriteriumsvalidität). Ein zweiter Ansatz kritisiert die Fixierung auf den Lernerfolg als einziges Unterrichtsprodukt und fordert die Berücksichtigung weiterer – auch affektiver – Unterrichtsziele (Konstruktvalidität). Der dritte Ansatz untersucht die Effekte von potenziell urteilsverzerrenden Einflüssen (Biasvariablen) auf die Beurteilungen, die mit der Qualität des Unterrichts in keinem systematischen Zusammenhang stehen. Wenn Validität vorliegt, sollten diese Biasvariablen keinen Einfluss auf die Beurteilungen der Lehrenden durch die Lernenden haben.

Wir greifen den dritten Ansatz auf, unterscheiden uns in unserem Forschungskonzept jedoch in fünf wesentlichen Punkten von den meisten bisherigen Forschungsarbeiten, die in Abschnitt 2 überblicksmäßig dargestellt und kurz analysiert werden:

Erstens werden in der vorliegenden Untersuchung wesentlich mehr potenziell bedeutsame Biasvariablen berücksichtigt als in bisherigen Studien. Neben Variablen wie dem Interesse der Lernenden für das Fach oder der Note, die in diesem Zusammenhang schon häufig empirisch untersucht worden sind, beziehen wir weitere Variablen in die Analyse mit ein. Dazu zählen beispielsweise die Einschätzung des Nutzens des Faches für das spätere

Berufsleben, Zweifel der Lernenden bezüglich ihrer Fähigkeit, ihre Lehrkräfte beurteilen zu können, die Einschätzung der Attraktivität der Lehrkraft sowie die Attribuierung von Erfolg im Fach. Die Auswahl der Variablen beruht nicht nur auf theoretischen Überlegungen zu Faktoren, deren Wirkung auf die Personenbeurteilung im Allgemeinen gezeigt wurde (vgl. Kanning, 1999), sondern es wurden auch Variablen ausgewählt, die sich bereits in einer Vorstudie zum vorliegenden Beitrag als bedeutend für den Beurteilungsprozess erwiesen haben (vgl. Greimel, 2001).

Zweitens beschränken wir uns nicht auf bivariate Analysen, in denen die einfache Korrelation zwischen jeweils einer potenziellen Biasvariable und der Beurteilung berechnet wird, sondern führen multiple Regressionsanalysen durch. Regressionsanalysen werden zum Teil auch in anderen Untersuchungen durchgeführt, diese berücksichtigten jedoch bis auf wenige Ausnahmen (vgl. Rindermann, 1999) ausschließlich eine bestimmte Anzahl von potenziellen Biasvariablen, nicht jedoch jene Aspekte des Lehrverhaltens, die typischerweise in einem Beurteilungsbogen enthalten sind, wie zum Beispiel die Verständlichkeit der Erklärungen der Lehrkraft. Entscheidend für die Bewertung eines potenziellen Bias ist jedoch nicht nur, ob er einen signifikanten Einfluss auf die Beurteilung zeigt, sondern auch wie groß dieser Einfluss im direkten Vergleich zu den Variablen ist, die einen Einfluss auf die Gesamtbeurteilung haben sollten. Im Falle der Lehrerbeurteilungen durch die Lernenden sind dies die Variablen, die das Lehrverhalten abbilden. Bei hoher Aussagekraft der Urteile der Lernenden über ihre Lehrenden sollten Lehrverhaltensvariablen einen starken Zusammenhang haben und gleichzeitig potenzielle Biasvariablen, die im Hinblick auf die Qualität der Lehre irrelevant sind, keinen Zusammenhang mit der Beurteilung aufweisen. Wir untersuchen also eine Reihe von potenziell urteilsverzerrenden Variablen im Kontext der Lehrverhaltensvariablen.

Drittens untersuchen wir, ob die Streuung der Urteile über die Stichprobe konstant ist oder ob Heteroskedastizität vorliegt, sodass für bestimmte Ausprägungen von Variablen eine höhere Streuung vorliegt als für andere. Man könnte zum Beispiel vermuten, dass Zweifel bezüglich einer gerechten Beurteilung auch dazu führen könnten, dass mit zunehmendem Zweifel die Präzision der beobachteten Gesamtbeurteilung abnimmt (d.h. die zufällige Streuung um den 'tatsächlichen' Wert zunimmt). Dies kann dazu führen, dass die Varianz der Residuen mit dem Ausmaß des Zweifels zunimmt. Neben dem Zweifel könnten das Interesse der Beurteilenden am Fach und/oder der eingeschätzte Nutzen des Faches ebenfalls einen derartigen Effekt verursachen. Hohes Interesse am Fach könnte ebenso wie eine hohe Nutzeneinschätzung dazu

führen, dass die Beurteilenden entweder besonders gut beurteilen, weil sie auf Grund ihrer hohen intrinsischen Motivation sowieso sehr positiv eingestellt sind, möglicherweise gegenüber fachdidaktischen und/oder menschlichen Schwächen der Lehrkraft toleranter werden, und diese starke positive Einstellung auf die Beurteilung der Lehrkraft abfärbt. Andererseits wäre auch denkbar, dass die Beurteilenden gerade wegen ihrer hohen intrinsischen Motivation besondere Erwartungen und Ansprüche an den Unterricht in diesem Fach haben und die Beurteilung ihrer Lehrkräfte sehr streng vornehmen. Ein Einfluss auf die Heteroskedastizität der Fehler erscheint bei diesen Variablen daher plausibel. Für den Fall, dass Heteroskedastizität vorliegt, könnten die Signifikanztests der einzelnen Biasvariablen beeinträchtigt sein. Wir versuchen die damit verbundenen Effekte bei den Signifikanztests der Biasvariablen zu berücksichtigen und zu korrigieren.

Viertens verbinden wir in unserem Fragebogen theoretische Überlegungen zur Gestaltung des Unterrichts im Unterrichtsgegenstand Rechnungswesen, die auf dem Modell der Direkten Instruktion von Weinert (1996) basieren, mit einer Reihe von empirischen Ergebnissen zur Unterrichtsqualität in betriebswirtschaftlichen Fächern (vgl. Greimel, 1996; Schneider, 1995). Ein Defizit vieler bereits vorliegender Untersuchungen besteht nämlich darin, dass kein oder nur ein mangelhafter Bezug zu einer zugrundeliegenden Theorie von gutem Unterricht in einem bestimmten Unterrichtsgegenstand festzustellen ist. Damit sind die in den jeweiligen Untersuchungen verwendeten Fragebögen und Beurteilungskriterien ebenfalls derart zu hinterfragen, ob sie hinsichtlich des Faches und der Lehrveranstaltungsform, für die sie verwendet wurden, inhaltssvalide sind. Dieser Prüfung wurde der in der vorliegenden Untersuchung eingesetzte Evaluierungbogen sowohl in theoretischer als auch in empirischer Hinsicht unterzogen.

Schließlich konzentrieren wir unsere Untersuchung nicht auf Lehrende an Hochschulen – wie in einem Großteil der US-amerikanischen und auch der europäischen Forschung –, sondern auf die Lehrerbeurteilung durch Lernende in Schulen der Sekundarstufe, konkret auf Lernende einer Handelsakademie kurz nach ihrem Schulabschluss.

Das Ziel der vorliegenden Studie ist also die empirische Überprüfung der Frage, welche Faktoren die Beurteilung von Lehrkräften in welchem Ausmaß beeinflussen. Untersucht werden sowohl relevante als auch irrelevante Faktoren im Sinn der oben erwähnten Unterscheidung. Es wurde besonderer Wert darauf gelegt, eine möglichst umfassende

Auswahl von Faktoren aus beiden Gruppen zu verwenden. Zur Liste der als relevant angenommenen Kriterien zählen Faktoren, die das Lehrverhalten beschreiben, wobei wir entsprechend der Typologie von Caselmann (1967) zwischen logo- und paidotropen Lehrverhaltensvariablen unterscheiden. Zur Liste der als irrelevant vermuteten Kriterien zählen unter anderem Spezifika des Faches aus der Sicht der Absolvent/inn/en, für das Lehrverhalten irrelevante Merkmale der Lehrkraft sowie besonders auch Merkmale der Schülerindividualebene, unter anderem die Einstellung der Schüler zur Beurteilung von Lehrkräften. Dieser zuletzt genannte Aspekt wurde unseres Wissens bisher in diesem Zusammenhang noch nicht überprüft. In Kapitel 3 werden die einzelnen Variablen näher erläutert.

Zunächst stellt Kapitel 2 dieses Beitrags einen Überblick über die wichtigsten Einflussfaktoren dar, die bisher im Rahmen von empirischen Untersuchungen eingehend analysiert wurden. In Kapitel 3 wird über die Methode, in Kapitel 4 über die Ergebnisse der Untersuchung berichtet. Kapitel 5 schließt mit der Diskussion der Ergebnisse und deren Implikationen für die zukünftige Gestaltung und den Einsatz von Lehrerbeurteilungen einerseits, sowie andererseits für weitere empirische Untersuchungen im Bereich der Unterrichtsevaluationsforschung.

2 Bislang vorliegende Ergebnisse

Die Ergebnisse der bislang vorliegenden Forschungsarbeiten stellen sich im Wesentlichen folgendermaßen dar: Biografische Merkmale der Lehrenden wie auch der Lernenden zeigen keinen systematischen Zusammenhang mit den Bewertungen, ebenso wenig wie die Klassengröße oder Raumverhältnisse. Vergleichsweise stärkere und über viele Untersuchungen konsistente Beziehungen zwischen den Beurteilungen und potenziellen Biasvariablen sind beim Interesse der Lernenden am Unterrichtsgegenstand sowie bei den (erwarteten) Noten und der damit möglicherweise zusammenhängenden Milde der Lehrperson bei der Notengebung festgestellt worden (vgl. Marsh, 1987).

Das Interesse für den Unterrichtsgegenstand ist meist der stärkste von allen untersuchten Einflussfaktoren. In einem Regressionsmodell von Spiel und Gössler (2000) zur Lehrbeurteilung durch Studierende in verschiedenen Fakultäten werden für das Interesse Betagewichte von etwa .5 bis .6 ermittelt. Das Interesse war die einzige Variable, die über alle Substichproben hinweg systematisch mit den Urteilen der Studierenden kovariiert.

Die Korrelation zwischen Beurteilungen und der (erwarteten) Note beträgt in der Regel etwa .2 bis maximal .3, mindestens jedoch .1. Vergleichsweise hohe Korrelationen ermittelten Greenwald und Gillmore (1997) an der University of Washington. Drei unterschiedliche Stichproben ergaben standardisierte Pfadkoeffizienten von .48, .38 und .50.

Diese Ergebnisse lassen zumindest drei Interpretationen zu, die jedoch ein ganz unterschiedliches Licht auf die Aussagekraft von Lehrerevaluationen durch die Lernenden werfen (vgl. Marsh & Roche, 1997; Marsh, 1987; Greenwald & Gillmore 1997):

Lehrpersonen, die milde benoten und gute Noten erwarten lassen, werden von den Lernenden durch (ungerechtfertigt) gute Beurteilungen bei den Evaluationen «belohnt». Diese Interpretation wird in der Diskussion meist als «*grading leniency hypothesis*» bezeichnet und bezieht sich auf den in der Sozialpsychologie bekannten Effekt, dass Lob Sympathie für den Lobenden hervorrufen kann, insbesondere wenn das Lob über das erwartete Maß hinausgeht (vgl. Greenwald & Gillmore, 1997). In diesem Fall könnte zweifellos berechtigterweise von einem Bias gesprochen werden.

Zu diesem Punkt gehört auch die Überlegung, dass die Studierenden aus guten Noten Rückschlüsse auf ihre eigene Leistung(sfähigkeit), ihre (intellektuellen) Fähigkeiten und auch auf die Unterrichtsqualität und auf die Fähigkeiten der Lehrperson ziehen könnten. Diese Rückschlüsse könnten die Studierenden geneigt machen, den Unterricht und die Lehrperson gut zu evaluieren.

Die nächsten beiden Interpretationsmöglichkeiten gehen davon aus, dass eine dritte Variable wirksam ist, und zwar entweder die Unterrichtsqualität selbst oder Merkmale der evaluierenden Lernenden:

Die besseren (erwarteten) Noten könnten auch besonders gute Schülerleistungen widerspiegeln, die wiederum das Ergebnis guten Unterrichts durch eine gute Lehrperson sein könnten. Diese Interpretationsmöglichkeit würde die Validitätshypothese unterstützen. Ergebnisse von Untersuchungen in Parallelveranstaltungen – so genannten «*multisection validity studies*», in denen die gleiche Lehrveranstaltung an einem College oder an einer Universität von unterschiedlichen Lehrer/inne/n unterrichtet wird und der Lernerfolg der

Studierenden in einer gemeinsamen standardisierten Abschlussprüfung ermittelt wird – zeigen einen Zusammenhang zwischen den Evaluationsergebnissen und den Lernleistungen, wobei die Effekte durch Charakteristika der Studierenden und der Milde der Lehrperson bei der Notengebung berücksichtigt wurden. Diese Befunde unterstützen die Validitätshypothese.

Eine dritte mögliche Erklärung stützt sich auf Charakteristika der beurteilenden Lernenden wie beispielsweise das bereits vorhandene Eingangsinteresse, das sowohl ihr Lernen und damit ihre (erwartete) Note beeinflusst und darüber hinaus auch die Lehrperson dazu veranlassen könnte, den Unterricht dem höheren Interesse der Lernenden anzupassen, sodass der Effekt durch die erwartete Note an sich hinfällig, ein Scheineffekt ist (vgl. Marsh, 1987). Weitere derartige potenzielle Einflussfaktoren wären die (Leistungs-)Motivation der Studierenden, die sich entweder auf alle Unterrichtsgegenstände beziehen kann oder auch nur auf einzelne spezifische Unterrichtsgegenstände (vgl. Greenwald & Gillmore, 1997).

Allerdings gehen diese beiden letzten Interpretationsmöglichkeiten von der Analyseebene der Klasse aus. Sie sind daher nicht geeignet, die Korrelation zwischen guter erwarteter Note und Lehrerbeurteilungen durch die Lernenden innerhalb der Klasse zu erklären. Die Korrelation innerhalb der Klasse ließe eine weitere Interpretationsmöglichkeit zu: Möglicherweise unterrichtet die Lehrkraft nicht für alle Lernenden in gleicher Weise, sie könnte sich bestimmten Lernenden – beispielsweise jenen, die bereits besonders gute Leistungen erbringen – besonders zuwenden oder auch die Art und Qualität der Rückmeldung bestimmten Merkmalen der Lernenden anpassen.

Die Problematik bei jedem Interpretationsversuch liegt darin, dass jede Note sowohl die individuellen Beurteilungsprinzipien der Lehrkraft als auch die Prüfungsleistung wiedergibt, die wiederum natürlich von einer Reihe von Merkmalen der Lernenden wie beispielsweise dem Interesse an dem Unterrichtsgegenstand beeinflusst wird. Daher ist es auch durchaus plausibel anzunehmen, dass alle drei bzw. vier Interpretationsmöglichkeiten zusammen den oftmals ermittelten Zusammenhang von guten erwarteten Noten und guter Lehrerevaluation durch die Lernenden erklären. Deshalb wäre es entscheidend, mehr über die Art und das Ausmaß des Zusammenwirkens der einzelnen Faktoren herauszufinden.

Marsh (1987) untersuchte in einer Pfadanalyse die Beziehungen zwischen der erwarteten Note, dem Eingangsinteresse und den Lehrerevaluationen durch die Lernenden und fand

heraus, dass sich in etwa ein Drittel des Effekts der erwarteten Note durch das Eingangsinteresse erklären ließ. Dieses Ergebnis unterstützt primär die dritte Interpretationsmöglichkeit, indirekt jedoch auch die Validitätshypothese (vgl. auch Howard & Maxwell, 1982).

Die Frage, ob die *grading leniency hypothesis* nun zutrifft oder nicht, bleibt auf Grund verschiedener Studien, die unterschiedliche Interpretationsmöglichkeiten des Zusammenhangs zwischen guten Noten und guten Lehrerevaluationsergebnissen unterstützen, unbeantwortet und dementsprechend umstritten. Während manche Wissenschaftler wie Greenwald und Gillmore (1997) weiterhin die Auffassung vertreten, dass «Yes, I can get higher ratings by giving higher grades», widersprechen andere (vgl. Marsh & Roche, 1997; d'Apollonia & Abrami, 1997).

3 Methode

Die vorliegende Untersuchung kann als Pre-Test eines Fragebogens betrachtet werden, dessen Einsatz für eine umfassende Erhebung an österreichischen Handelsakademien im Unterrichtsgegenstand Rechnungswesen geplant ist (vgl. Greimel, 2001). Für die Entwicklung des Fragebogens wurden 40 teilstrukturierte Interviews mit Schülerinnen und Schülern von zweiten bis fünften Jahrgängen von Handelsakademien in Niederösterreich und im Burgenland durchgeführt. Dabei wurden zu den Themenfeldern «guter Unterricht», «gute Lehrkraft» und «Lehrerbeurteilungen durch Schüler/innen» besonders relevante Faktoren zur positiven Beurteilung von Unterricht und Lehrkräften, sowie die Einstellung der Schüler/innen zur Lehrerevaluation durch ihre Beurteilungen inhaltsanalytisch ermittelt. Etwa 35% der befragten Schülerinnen und Schüler stehen der Lehrerbeurteilung durch die Schüler/innen an Handelsakademien durchaus ablehnend gegenüber und bezweifeln selbst die Objektivität und Gerechtigkeit der Beurteilungen (vgl. Greimel, 2001). Fast die Hälfte der Schüler/innen, welche die Lehrerbeurteilung durch die Schüler/innen an Handelsakademien zwar prinzipiell befürworten, haben ebenso Bedenken, ob ihre Mitschüler/innen ihre Lehrkräfte tatsächlich gerecht beurteilen können oder sich eher von Sympathie für bestimmte Lehrpersonen und guten Noten in ihrem Urteil beeinflussen ließen.

Die Ergebnisse dieser Interviews sowie einer qualitativen, mehrstufigen, schriftlichen Befragung von Schüler/innen an einer Wiener Handelsakademie und Handelsschule (vgl. Greimel, 1996) wurden für die Entwicklung eines strukturierten Fragebogens herangezogen.

Die darin enthaltenen Lehrverhaltensvariablen orientieren sich darüber hinaus auch am Unterrichtsmodell der Direkten Instruktion, wie es vor allem Weinert (1995) als erwiesenermaßen äußerst lernwirksames Modell für systematischen Wissenserwerb beschrieben hat. Deshalb ist es ein adäquates Unterrichtsmodell für den Rechnungswesenunterricht an Handelsakademien, das auch in den Phasen der Unterrichtsplanung für den betriebswirtschaftlichen Unterricht von Posch, Schneider & Mann (1989) umgesetzt wird.

Für alle Items im Fragebogen wird eine fünfstufige Skala verwendet, die von «trifft überhaupt nicht zu» bis «trifft vollkommen zu» reicht.

Zur Messung der Gesamtbeurteilung der Lehrkraft wurden fünf Items verwendet (siehe Tabelle 1), die anhand einer Faktoranalyse auf einen einzigen Faktor-Score reduziert wurden. Die Verwendung der Summe über die fünf Items führt zu keinen wesentlichen Änderungen in den unten präsentierten Ergebnissen.

Tabelle 1: Statistische Kennzahlen der Skala «Gesamtbeurteilung der Lehrkraft».

Item	Mittelwert	Standard- abweichung	Trennschärfe
Ich hätte die Lehrkraft gerne auch in anderen Fächern gehabt	2.9	1.5	.80
Insgesamt war die Lehrkraft ein/e gute/r Lehrer/in	2.3	1.3	.85
Ich habe bei dieser Lehrkraft etwas Wichtiges gelernt	2.3	1.1	.74
Mein Interesse am Fach ist durch den Unterricht gestiegen	2.8	1.3	.75
Im Vergleich zählt er/sie zu den besten Lehrer/innen	2.7	1.5	.86
Skala	13.1	5.8	$\alpha = .92$

Zur Abbildung des logotropen Lehrverhaltens wurden insgesamt 18 Items verwendet. Sie umfassen gemäß der Typologie von Caselmann (1949) hauptsächlich fachdidaktische Lehrverhaltenskomponenten, wie die folgenden drei Beispiele zeigen:

- Die Lehrerin/der Lehrer hat neue Sachverhalte so erklärt, dass ich sie gut verstehen kann.
- Die Lehrerin/der Lehrer hat uns gezeigt, wofür wir das Gelernte verwenden können.
- Die Lehrerin/der Lehrer konnte die Fragen der Schüler/innen verständlich beantworten.

Die paidotrope Einstellung der Lehrkraft, ihre Zugewandtheit zu den Schüler/inne/n, ihr Interesse an der persönlichen und schulischen Situation der Schüler/innen, wurde in 14 verschiedenen Items abgebildet, wie zum Beispiel:

- Die Lehrerin/der Lehrer war zu Schüler/inne/n freundlich.
- Die Lehrerin/der Lehrer war allen Schüler/inne/n gegenüber gerecht.
- Die Lehrerin/der Lehrer hat Schüler/innen in den Unterricht miteinbezogen.

Zusätzlich wurden drei Items abgefragt, die sowohl dem logo- als auch dem paidotropen Verhalten zugeordnet werden könnten. Zur Dimensionsreduktion wurden die insgesamt 35 Items zum Lehrverhalten mit einer Faktorenanalyse auf zwei Faktoren reduziert.¹ Die inhaltsorientierten, fachdidaktischen Verhaltensitems (Beispiele: «Die Lehrerin/der Lehrer hat neuen Stoff logisch aufgebaut präsentiert»; «Die Lehrerin/der Lehrer hat neue Sachverhalte so erklärt, dass ich sie gut verstehen konnte») zeigten hauptsächlich hohe Ladungen auf einem Faktor. Die personenbezogenen Variablen, die sich vor allem die Beziehung zu den Schüler/innen bezogen (Beispiele: «Die Lehrerin/der Lehrer kann mit Schüler/innen/n gut umgehen»; «Die Lehrerin/der Lehrer ist geduldig»), zeigten mehrheitlich hohe Ladungen auf dem anderen Faktor. Diese Trennung in eher fachorientierte und eher schüler-personenbezogenen Variablen hatte sich bereits in anderen Untersuchung zum Lehrverhalten aus Schülersicht in wirtschaftlichen Fächern gezeigt (vgl. Greimel, 1996; Schneider, 1995).

Entsprechend der Zielsetzung der vorliegenden Untersuchung sind im Fragebogen neben den Items zum Lehrverhalten zahlreiche potenzielle Einflussfaktoren enthalten, die apriori irrelevant für die Beurteilung sein sollten, wenn Evaluationen durch Lernende ein valides Instrument zur Beurteilung der Unterrichtsqualität einer Lehrkraft sind. Die Einflussfaktoren können wie folgt gruppiert werden:

- kognitive und affektive Merkmale der Schülerindividualebene, wie
 - ihre Note in diesem Unterrichtsgegenstand im fünften Jahrgang der HAK oder HLA sowie ihre Note bei der Matura,
 - das (auch daraus resultierende) Fähigkeitsselbstkonzept,
 - die Art ihrer Attribuierung von Erfolg und Misserfolg,
 - ihre Leistungsängstlichkeit und –motivation,

¹ Eine Analyse auf Basis der ursprünglichen Skalen (Itemsumme) führt zu keinen wesentlichen Änderungen in den unten präsentierten Ergebnissen. Die Reliabilität der logotropen (paidotropen) Skala beträgt .94 (.91).

- ihre Einschätzung des Unterrichtsgegenstands Rechnungswesen hinsichtlich verschiedener Kriterien (Interesse, Aufwand, Nutzen, Schwierigkeitsgrad, ...)
- ihre Einstellung zur Lehrerbeurteilung durch Schüler/innen,
- ihr subjektiv empfundener Anreizwert von Erfolg
- sowie Alter und Geschlecht.

Bezüglich der Einschätzung des Unterrichtsgegenstands Rechnungswesen ist zu bemerken, dass acht Items zur Beschreibung des Fachs Rechnungswesen mit Hilfe einer Faktoranalyse auf zwei Faktoren reduziert wurden. Es zeigte sich, dass aufgrund der Struktur der Ladungen der erste der beiden Faktoren als «Interesse für das Fach» interpretiert werden kann (Beispiel: «Ich habe gerne für Rechnungswesen gelernt»). Der zweite Faktor ist vor allem durch ein Item repräsentiert («Im Vergleich zu anderen Fächern ist der Aufwand für das Fach Rechnungswesen hoch») und wird daher als «Aufwand für das Fach» bezeichnet.

- Faktoren auf der Klassenebene, wie die Klassengröße und die Klassenzusammensetzung (Mädchenanteil, Anteil der deutschsprachigen Schüler/innen)
- für die Unterrichtsqualität irrelevante Merkmale der Lehrkraft, wie Alter und Geschlecht der Lehrkraft, sowie deren Attraktivität² (in den Augen der beurteilenden Schüler/innen).

Im Dezember 2000 wurden alle Studierenden in der Vorlesung «Einführung in die Wirtschaftspädagogik» an der Wirtschaftsuniversität Wien gebeten, den gesamten Fragebogen auszufüllen und damit jene Lehrkraft im Unterrichtsgegenstand Rechnungswesen zu beurteilen, die sie zuletzt in der Schule gehabt hatten. Von insgesamt 120 befragten Studierenden hatten 96 eine Schule besucht, in der Rechnungswesen fünf Jahre lang unterrichtet worden war (davon 67 eine Handelsakademie und 29 eine Höhere Bundeslehranstalt für wirtschaftliche Berufe). Die restlichen Befragten hatten eine allgemein bildende höhere Schule besucht, in der der Unterrichtsgegenstand Rechnungswesen nicht vorgesehen ist. Sie wurden gebeten, ihre Lehrkräfte aus Mathematik zu beurteilen, da Mathematik und Rechnungswesen von der Art der Inhalte und Zielsetzungen vergleichbar

² Zusätzlich zu den direkten Urteilen der Attraktivität wurde für die Auswertung eine zweite Variable zur Attraktivität definiert, sodass nur gegengeschlechtliche Beurteilungen berücksichtigt wurden (d.h. Urteile wurden nur berücksichtigt, wenn Schüler weibliche Lehrkräfte und Schülerinnen männliche Lehrkräfte beurteilten).

sind. In die nachfolgend beschriebene Auswertung wurden diese Studierenden jedoch nicht einbezogen.

Die Befragung von Absolvent/inn/en anstelle von Schüler/inne/n hat Vor- und Nachteile. Aufgrund der zeitlichen Distanz zwischen der Schulzeit, auf die sich die Fragen beziehen und dem Zeitpunkt der Erhebung, könnte die Präzision der Aussagen leiden, weil die Lehrsituation den Befragten eventuell nicht mehr so präsent ist. Allerdings könnten gerade wegen dieser zeitlichen Distanz die Urteile nüchterner und realistischer ausfallen, als in der – eventuell emotional belasteten – Schulzeit. Darüber hinaus zeigt unsere Analyse, dass 73% der Befragten zwischen 19 und 21 Jahre alt sind und daher offenbar sofort nach dem Schulabschluss oder nur mit geringer zeitlicher Verzögerung begonnen haben zu studieren. Ein für die Beurteilung interessanter Aspekt besteht darin, dass die Befragten durch ihre Studienwahl bereits teilweise erfahren konnten, inwieweit ihr in der Schule erworbenes Wissen im Fach Rechnungswesen im Studium anwend- und verwertbar ist.

Als deskriptives Instrument werden einfache Korrelationen zwischen der Gesamtbeurteilung und allen Einflussgrößen ermittelt, vor allem auch deshalb, um Vergleiche mit anderen Studien zu ermöglichen, die nur auf Korrelationsanalysen beruhen. Der Nachteil einer Analyse der Zusammenhänge auf Basis von Korrelationen liegt darin, dass der Einfluss dritter (oder weiterer) Variablen bei diesen paarweisen Vergleichen nicht kontrolliert wird. Um die verzerrenden Effekte vernachlässigter Variablen auszuschließen, führen wir eine multiple Regressionsanalyse durch, wobei wir zunächst die beiden Lehrverhaltensvariablen sowie jene Variablen in die Regressionsgleichung aufnehmen, die in der Literatur häufig als (signifikante) Biasvariablen diskutiert werden. Anschließend fügen wir jene Variablen hinzu, die signifikante Korrelationen mit der Gesamtbeurteilung aufweisen. Dann überprüfen wir die verbleibenden potenziellen Biasvariablen auf Signifikanz. Nach Elimination der nicht signifikanten Variablen gelangen wir schließlich zu einem Regressionsmodell, dessen Residuen auf Heteroskedastizität überprüft werden. Wir berücksichtigen die Heteroskedastizität auf Basis einer gewichteten Kleinst-Quadrat Schätzung des Regressionsmodells. Dabei wird den Fehlern umso weniger Gewicht beigemessen, je größer die Streuung der Residuen ist.

4 Ergebnisse

Tabelle 2 enthält die Korrelationen aller betrachteten Variablen mit der Gesamtbeurteilung der Lehrkraft. Es zeigt sich, dass die beiden Verhaltensvariablen hoch signifikant mit der Beurteilung korrelieren, das logotrope Lehrverhalten mit .73 jedoch deutlich stärker als das paidotrope Verhalten mit .47. Dies bestätigt ihre oben erfolgte Zuordnung zur Gruppe der relevanten Variablen, weil nur diese die Beurteilung der Lehrkraft beeinflussen sollten.

Weitere fachbezogene Aspekte (Nutzen des Fachs, Erfolgsattribution, Anreiz und Motivation, Leistungsangst), jedoch nicht der Aufwand für das Fach weisen eine signifikant positive Korrelation mit der Gesamtbeurteilung auf.

Tabelle 2: Korrelation einzelner Variablen mit der Gesamtbeurteilung

logotropes Lehrverhalten	.73**	Geschlecht der/des Befragten	.12
paidotropes Lehrverhalten	.47**	Alter der/des Befragten	-.12
Interesse am Fach	.37**	Klassengröße	.10
Aufwand für das Fach	-.11	Mädchenanteil in %	-.03
Nutzen des Fachs	.25*	Nicht Deutsch als Muttersprache in %	.16
Anreiz und Motivation	.36**		
Erfolgsattribution	.27*	Geschlecht der Lehrkraft	.10
Prüfungsangst	.29**	Attraktivität der Lehrkraft	.41**
Note bei der Matura	.21	Attraktivität (nur gg.geschl.)	.15
Note im 5. Jahrgang	.28**	Alter der Lehrkraft	.05
Fähigkeitsselbstkonzept	.17	wie viele Jahre Unterricht bei dieser Lehrkraft	-.05
Evaluation durch Schüler ist sinnvoll	-.06	Lehrkraft sollte zuerst über Unzufriedenheit des Schülers erfahren	.30**
Zweifel, ob Schüler gerecht beurteilen	.19		

** kennzeichnet signifikante Korrelationen bei einem Niveau von 1%.

* kennzeichnet signifikante Korrelationen bei einem Niveau von 5%.

Tabelle 2 zeigt außerdem, dass von den weiteren in Betracht gezogenen Variablen, nur drei signifikant mit der Gesamtbeurteilung korrelieren: die Noten im fünften Jahrgang (d.h. im letzten Schuljahr), die Attraktivität der Lehrkraft und ein Item zur Frage, ob die Lehrkraft zuerst darüber erfahren sollte, wenn ein Schüler mit dem Unterricht unzufrieden ist. Das positive Vorzeichen der Korrelation mit der Note bedeutet, dass die Beurteilung umso schlechter ausfällt, je schlechter die Note ist. Dies entspricht den in Kapitel 2 dargestellten, bisher vorliegenden Befunden zum Einfluss der Note auf die Beurteilung.

Wir formulieren und schätzen nun ein multiples Regressionsmodell, das jene Variablen enthält, die in der Literatur häufig als signifikante Biasvariablen³ diskutiert werden. Dieses Modell wird um jene Variablen erweitert, die aufgrund der Korrelationsanalyse als weitere signifikante Biasvariablen in Frage kommen. Tabelle 3 zeigt allerdings, dass nur wenige erklärende Faktoren einen signifikanten Einfluss aufweisen. Die beiden Lehrverhaltensvariablen treten als eindeutig statistisch signifikante Einflussgrößen hervor: das logotrope Lehrverhalten mit einem Betagewicht von etwa .73 und das paidotrope Lehrverhalten mit einem Betagewicht von etwa .48 (in beiden Fällen ist der P-Wert gleich null). Es gibt jedoch noch zwei Variablen, die einen P-Wert aufweisen, der kleiner als 5% ist: das Interesse am Fach und das Alter der Lehrkraft. Wenn das Regressionsmodell um jene Variablen erweitert wird, die eine signifikante Korrelation aufweisen, zeigt sich, dass keine der hinzugefügten Variablen einen signifikanten Einfluss auf die Gesamtbeurteilung hat.

Tabelle 3: Standardisierte Regressionskoeffizienten und P-Werte von multiplen Regressionsmodellen (abhängige Variable ist die Gesamtbeurteilung).

	stand. Koeff.	P-Wert	stand. Koeff.	P-Wert
logotropes Lehrverhalten	.738	.000	.724	.000
paidotropes Lehrverhalten	.484	.000	.480	.000
Interesse am Fach	.141	.035	.145	.085
Note bei der Matura	-.075	.246	-.071	.304
Geschlecht der/des Befragten	.036	.543	.036	.587
Alter der/des Befragten	.005	.920	-.003	.956
Klassengröße	-.009	.872	-.013	.829
Mädchenanteil in %	.017	.780	.015	.813
Alter der Lehrkraft	-.116	.044	-.116	.057
Geschlecht der Lehrkraft	.020	.727	.012	.844
Nutzen des Fachs			.013	.833
Anreiz und Motivation			-.008	.918
Erfolgsattribuierung			.049	.435
Prüfungsangst			.046	.467
Attraktivität der Lehrkraft			-.040	.551
Lehrkraft sollte zuerst über Unzufriedenheit erfahren			.022	.704
R ²	.80		.81	
adj. R ²	.78		.77	

³ In den Regressionsmodellen wurden sowohl die Note bei der Matura als auch die Note im letzten Schuljahr verwendet. Dabei zeigte sich, dass keine der beiden Variablen einen signifikanten Einfluss aufweist. Da der Zusammenhang mit der Gesamtbeurteilung für die Maturanote jedoch etwas stärker ausfällt, wurde diese Variable im Modell belassen.

Auf Basis dieses Regressionsmodells wurde für die verbleibenden Variablen überprüft, ob deren Aufnahme in die Regressionsgleichung eine signifikante Verbesserung der Erklärungskraft bringt (jeweils einzeln auf Basis eines F-Tests). Dabei stellte sich vor allem ein Item mit dem Wortlaut: «Ich bezweifle, ob HAK-Schüler/innen in der Lage sind, ihre Lehrer/innen gerecht zu beurteilen» als signifikant heraus. Ein Item, das ebenfalls die Einschätzung von Evaluationen zum Ausdruck bringt («Ich halte es für sinnvoll, wenn Schüler/innen Lehrkräfte anonym beurteilen dürfen») weist ebenfalls einen signifikanten Effekt auf. Bei gleichzeitiger Aufnahme dieser beiden Variablen in die Regressionsgleichung stellt sich jedoch nur die Variable Zweifel als signifikant heraus. Nach Hinzufügen der Variable Zweifel sowie sukzessiver Elimination der nicht signifikanten Variablen (unter Verwendung eines Signifikanzniveaus von 20%) erhält man die Ergebnisse in Tabelle 4.

Tabelle 4: Standardisierte Regressionskoeffizienten und P-Werte von multiplen Regressionsmodellen (abhängige Variable ist die Gesamtbeurteilung).

	stand. Koeff.	P-Wert	P-Wert ¹⁾	P-Wert ²⁾
logotropes Verhalten	.719	.000	.000	.000
paidotropes Verhalten	.484	.000	.000	.000
Interesse am Fach	.113	.065	.070	.123
Note bei der Matura	-.084	.140	.115	.127
Alter der Lehrkraft	-.125	.017	.011	.001
Erfolgsattribuierung	.076	.154	.166	.047
Zweifel	.125	.016	.018	.001
R ²	.84			
adj. R ²	.77			

¹⁾ P-Wert bei Verwendung von heteroskedastizitäts-konsistenten Standardfehlern für die geschätzten Parameter.

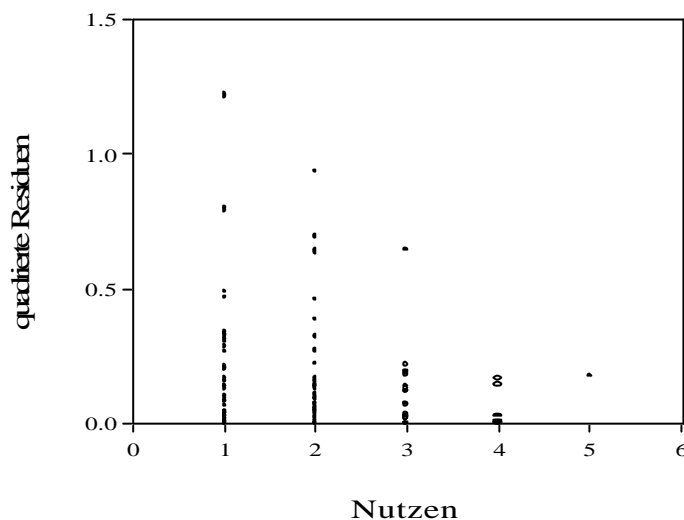
²⁾ P-Wert bei Verwendung einer gewichteten Kleinst-Quadrat Schätzung, unter Verwendung der Variable Nutzen des Fachs als Gewicht (Details siehe Text).

Abschließend wenden wir uns der Frage zu, ob die Varianz der Residuen im Regressionsmodell konstant ist. Umstände, die dazu beitragen können, dass Heteroskedastizität in den Daten vorliegt, wurden in der Einleitung bereits diskutiert. Weiters ist anzumerken, dass prinzipiell alle Variablen dafür in Frage kommen und nicht nur jene, die einen signifikanten Einfluss auf die Gesamtbeurteilung haben. Bei der Berücksichtigung der Heteroskedastizität versucht man, Unterschiede in der Streuung der Daten und nicht Unterschiede in den Niveaus zu erklären (vgl. Miles & Shevlin, 2001).

Für den Fall, dass Heteroskedastizität der Fehler vorliegt, könnten die Signifikanztests der einzelnen Variablen beeinträchtigt sein. Zunächst berücksichtigen wir diesen Umstand durch Verwendung von heteroskedastizitäts-konsistenten Standardfehlern nach White (1980). Diese

Korrektur ist allerdings unspezifisch in dem Sinn, dass keine konkrete Ursache für die Heteroskedastizität festgelegt wird. Aus einer Korrelationsanalyse zwischen den quadrierten Residuen des Modells und einigen in Frage kommenden Variablen schließen wir, dass vor allem die Einschätzung des Nutzens des Fachs (und in geringerem Ausmaß das Alter der Befragten) als potenzielle Quelle von Heteroskedastizität betrachtet werden kann. Je höher der Nutzen eingeschätzt wird, desto größer ist die Varianz der Residuen (siehe Abbildung 1).

Abbildung 1: Quadrierte Residuen des multiplen Regressionsmodells in Abhängigkeit vom eingeschätzten Nutzen des Fachs.



Eine Berücksichtigung des Umstands der Heteroskedastizität kann auf Basis einer gewichteten Kleinst-Quadrat Schätzung des Regressionsmodells erfolgen. Dabei wird den Fehlern umso weniger Gewicht beigemessen, je größer die Streuung der Residuen ist. Konkret verwenden wir als Gewicht die beobachteten Werte der Variable Nutzen, die durch deren Stichprobenmittelwert dividiert werden. Die Merkmalsausprägungen dieser Variablen («Rechnungswesen kann man später sicher gut gebrauchen») reichen von 1 («trifft vollkommen zu») bis 5 («trifft gar nicht zu»). Wenn diese Form der Modellschätzung durchgeführt wird, stellt sich heraus, dass das Interesse für das Fach und die Note bei der Matura keinen signifikanten Einfluss mehr auf die Gesamtbeurteilung haben (siehe letzte Spalte in Tabelle 4). Die Signifikanz des Alters der Lehrkraft und des Zweifels aus Tabelle 3 wird bestätigt und sogar erhöht. Die Erfolgsattribution stellt sich nunmehr als signifikant heraus (P-Wert=.047).

Das Alter der Lehrkraft hat einen signifikanten Einfluss auf die Gesamtbeurteilung. Aus der Analyse von Interviews mit Schüler/innen ging hervor, dass manche Befragte den routinierten

Umgang mit Schüler/inne/n und den Erfahrungsschatz der Lehrkraft mit deren Alter in Verbindung bringen (vgl. Greimel, 2001). Das ist eine mögliche Erklärung dafür, dass Lehrkräfte umso besser beurteilt werden je älter sie sind, und zwar unabhängig davon, wie sich die mit dem Alter verbundenen Aspekte in konkreten Lehrverhaltenskomponenten manifestieren.

Besonders interessant ist der Einfluss des Zweifels, den die Befragten bezüglich ihrer Beurteilung haben. Wer Zweifel an einer gerechten Beurteilung hat, beurteilt die Lehrer/innen *ceteris paribus* besser als Befragte, die diese Zweifel nicht hegen. Wie lässt sich dieses Ergebnis erklären? Bei der Evaluation besteht die Herausforderung für die Lernenden zweifellos darin, zu erkennen, ob zum Beispiel ihre Schwierigkeiten in einem Fach rein vom Unterricht der Lehrkraft herrühren oder doch mit mangelndem Interesse oder der Schwierigkeit des Stoffes zu tun haben. Wenn man etwas nicht versteht, kann es einerseits daran liegen, dass es sehr schwer zu verstehen ist, oder daran, dass es schlecht erklärt wurde. Wenn Vergleichsmöglichkeiten fehlen, entstehen beim Schüler Zweifel, woran es wirklich liegt und er oder sie gibt im Zweifelsfall – *«in dubio pro reo»* – eine bessere Beurteilung für den Lehrer ab. Zweifel können jedoch auch ohne Verständnisschwierigkeiten im Unterrichtsgegenstand auftreten. Aus den oben erwähnten Interviews ging hervor, dass selbst Befragte, die der Lehrerbeurteilung durch Schüler/innen positiv gegenüberstehen, die Gerechtigkeit der Beurteilungen bezweifeln, da sie Verzerrungen durch gute Noten und durch die Sympathie für eine Lehrkraft vermuten.

Dieser Erklärungsansatz korrespondiert auch mit der Erfolgsattribution: Je mehr der Schüler seinen Erfolg im Fach Rechnungswesen eindeutig in der Person der Lehrkraft begründet sieht, desto besser fällt seine Beurteilung für die Lehrkraft aus: Wer stärker attribuiert, hat weniger Zweifel hinsichtlich der Gerechtigkeit seiner Beurteilung. Diese beiden Variablen weisen auch eine signifikant negative Korrelation von -0.24 auf.

5 Diskussion und Ausblick

Das Ziel der vorliegenden empirischen Untersuchung war festzustellen, in welchem Ausmaß die Gesamtbeurteilung der Lehrkraft bei gleichzeitiger Betrachtung der beurteilten Lehrverhaltensvariablen von verschiedenen potenziellen Biasvariablen abhängt. Die Ergebnisse unseres Regressionsmodells zeigen, dass das von den beurteilenden Absolventinnen und Absolventen eingeschätzte Lehrverhalten am stärksten deren

Gesamtbeurteilung beeinflusst (vgl. dazu auch die Ergebnisse von Rindermann 1999). Dabei hat das logotrope Lehrverhalten einen etwas stärkeren Einfluss auf die Beurteilung als das paidotrope Lehrverhalten. Neben dem Alter der Lehrkraft fließt noch ein weiterer Faktor in statistisch signifikanter Weise in die Beurteilung mit ein: Wenn Schüler/innen bezweifeln, ob ihre Lehrkräfte überhaupt gerecht beurteilt werden können, führt dies zu einer etwas besseren Gesamtbeurteilung der Lehrkraft. Außerdem stellen wir fest, dass die Einschätzung des praktischen Nutzens des Faches die Präzision der Aussagen über die Gesamtbeurteilung beeinflusst. Die Beurteilung selbst wird davon nicht berührt, sondern nur das Ausmaß der Streuung der Urteile um den erwarteten Wert. Dies kann zur Folge haben, dass die statistischen Tests über die Signifikanz von Einflussfaktoren beeinträchtigt sind. Eine Berücksichtigung der Heteroskedastizität zeigt jedoch, dass dies nicht der Fall ist. Die (ursprüngliche) Beurteilung der Signifikanz wird im Wesentlichen bestätigt. Dennoch kann man ableiten, dass in zukünftigen Untersuchungen der Aspekt der Heteroskedastizität unbedingt berücksichtigt werden sollte.

Obwohl in der vorliegenden Untersuchung eine relativ kleine Stichprobe verwendet wurde, sind die Ergebnisse sowohl hinsichtlich der Praxis des Einsatzes und der Durchführung von Lehrerbeurteilungen als auch hinsichtlich der Evaluationsforschung relevant und richtungweisend: Für die Praxis bedeuten die Ergebnisse, dass die Urteile der Lernenden in hohem Ausmaß vom Lehrverhalten abhängen und von einer Reihe von verzerrenden Einflussfaktoren weitgehend unberührt bleiben. Das bedeutet jedoch nicht, dass Lehrerevaluation durch die Lernenden in allen Schul- und Universitätsbereichen, in allen Fächern in gleicher Weise standardisiert und unreflektiert eingesetzt werden sollen. Vielmehr bedarf es bei jedem Einsatz einer gut überlegten Auswahl des Instrumentes, das für das jeweilige Fach und die jeweilige Unterrichtsform passende Items enthält. Weiters sollten nicht nur der beurteilten Lehrkraft, sondern auch den beurteilenden Lernenden die Zielsetzung und der Stellenwert der Evaluierung deutlich gemacht werden. Das würde nicht nur die Akzeptanz von Lehrerevaluationen durch Lernende erhöhen, sondern vielleicht auch dazu beitragen, den Zweifel der Lernenden, welche die Beurteilungen vornehmen sollen, zu reduzieren.

Für die Evaluationsforschung bedeutet die vorliegende Studie, dass ein differenzierter Ansatz, der der Komplexität des Unterrichts- und Beurteilungsgeschehens gerecht wird, nicht nur frühere Forschungsergebnisse relativiert, sondern auch vollkommen neue Erkenntnisse hervorbringt. Frühere Forschungsergebnisse werden insofern relativiert, als unsere Ergebnisse

die außerordentliche Bedeutung der Lehrverhaltensvariablen zeigen, die in viele Untersuchungen gar keinen Eingang gefunden hat. Bei den untersuchten Biasvariablen ist wiederum bemerkenswert, dass wir für Faktoren wie dem Interesse und den Noten, die bislang in der Forschung breiten Raum eingenommen haben, bei der Betrachtung von einfachen Korrelationen mit bisherigen Studien übereinstimmende Ergebnisse erhalten. Allerdings besitzen sie gemeinsam mit anderen Biasvariablen und den Lehrverhaltensvariablen keine Relevanz mehr. Im Modell verbleiben als signifikante Faktoren neben den Lehrverhaltensvariablen das Alter der Lehrkraft, die Erfolgsattribution und ein bisher nicht beachteter Störfaktor bei der Beurteilung von Lehrkräften durch Lernende: die eigene selbstkritische Haltung gegenüber ihren Beurteilungen.

Literatur

- Caselmann, C. (1967). Differentielle Psychologie des Lehrers und Erziehers. In Strunz, K. (Hrsg.), *Pädagogische Psychologie für Höhere Schulen* (453–465). Ernst Reinhardt Verlag München Basel.
- d'Apollonia, S. & Abrami, P. (1997). Navigating Student Ratings of Instruction. *American Psychologist*, 52, 1198–1208.
- Greenwald, A. und Gillmore, G. (1997). Grading Leniency Is a Removable Contaminant of Student Ratings. *American Psychologist*, 52, 1209–1217.
- Greimel, B. (1996). Ein Anforderungsprofil an Lehrer. In: Fortmüller, Richard / Aff, Josef: *Wissenschaftsorientierung und Praxisbezug in der Didaktik der Ökonomie*. Festschrift Wilfried Schneider, Manz Verlag, Wien, 229–254.
- Greimel, B. (2001). Können Schüler/innen Unterrichtsqualität beurteilen? Zwischenbericht zum Projekt für das Bundesministerium für Bildung, Wissenschaft und Kultur, Wien.
- Howard, G. & Maxwell, S. (1982). Do Grades Contaminate Student Evaluations of Instruction. *Research in Higher Education*, 16, 175–188.
- Kanning, U. (1999). *Die Psychologie der Personenbeurteilung*. Hogrefe Verlag für Psychologie Göttingen.
- Marsh, H. (1987). Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Educational Research*, 11, 253–387.

- Marsh, H. & Roche, L. (1997). Making Students' Evaluations of Teaching Effectiveness Effective. The Critical Issues of Validity, Bias, and Utility. *American Psychologist*, 52, 1187–1197.
- Miles, J. & Shevlin, M. (2001). *Applying Regression & Correlation*. Sage Verlag London.
- Posch, P., Schneider, W. & Mann, W. (1989). *Unterrichtsplanung mit Beispielen für den betriebswirtschaftlichen Unterricht*. Manz Verlag Wien
- Rindermann, H. (1999). Bedingungs- und Effektivvariablen in der Lehrevaluationsforschung: Konzeption und Prüfung des Münchner multifaktoriellen Modells der Lehrveranstaltungsqualität. *Unterrichtswissenschaft*, 4, 357–380.
- Schneider, W. (1995). Was motiviert Schüler an Wirtschaftsschulen wirklich? – Affektive Lehrer-Schüler-Beziehung oder kognitive Unterrichtsqualität? In: Metzger, Christoph / Seitz, Hans (Hrsg.): *Wirtschaftliche Bildung, Träger, Inhalte Prozesse* (357–373). Verlag des Schweizerischen Kaufmännischen Verbandes, Zürich.
- Spiel, C. & Gössler, M. (2000). Zum Einfluss von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. *Zeitschrift für Pädagogische Psychologie*, 14, 38–47.
- Weinert, F. (1996). Für und wider die «neuen Lerntheorien» als Grundlagen pädagogisch-psychologischer Forschung. *Zeitschrift für Pädagogische Psychologie*, 10, 1–12.
- Weinert, F. (2001). Die evaluierte Universität. Manuskript zur Heidelberger Universitätsrede an der Ruprecht-Karls-Universität in Heidelberg am 25. Jänner 2001.
- White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48, 817–838.