BETTINA GREIMEL

Lehrerevaluation durch Beurteilungen der Lernenden – eine Analyse des Standes der Evaluationsforschung

KURZFASSUNG: Die Fragestellung, ob die Unterrichts- und Lehrerevaluationen durch die Lernenden einen Rückschluss auf die tatsächliche Qualität der Unterrichtsarbeit der Lehrenden zulassen, hat einen nahezu unüberschaubaren Korpus an Forschungarbeiten hervorgebracht. Die meisten Veröffentlichungen in diesem Bereich lassen sich jedoch zu drei wesentlichen Forschungsansätzen zusammenfassen, deren Ergebnisse in diesem Beitrag analysiert und diskutiert werden. Insgesamt sprechen die empirischen Befunde für die Durchführung von Lehrerevaluation durch die Lernenden, zeigen aber auch typische Problemfelder der Evaluierung.

1 Zielsetzung

Lehrerevaluationen durch die Lernenden stellen eine seit vielen Jahren zunehmend – vielleicht überhaupt die am häufigsten – eingesetzte Lehrevaluationsform dar. Die Aussagekraft der Urteile der Lernenden zur Lehrerevaluation ist jedoch nicht zuletzt deswegen seit langem umstritten (vgl. z. B. GREENWALD 1997). Meinungsäußerungen wie „Student ratings of instruction are easily biased by the instructors popularity. (…) Faculty can manipulate student ratings by giving lots of high grades. Student ratings are notoriously unreliable and completely invalid" (WAGENAAR 1995, S. 64) sind m.E. typisch für die seit vielen Jahrzehnten andauernde Diskussion, ob die Lernenden beurteilen können, welche Lehrkräfte gut oder schlecht sind. „It is no secret", meint FELDMAN (1997, S. 368), dass einige Lehrkäfte der Beurteilung keine große Bedeutung beimessen, dass „(…) teachers have little regard for them". Dementsprechend prägen unzählige Untersuchungen zur Aussagekraft (zur so genannten „Validität") der Beurteilungen der Lernenden zur Lehrerevaluation die Unterrichtsevaluationsforschung, die sich jedoch auf Grund vergleichbarer Forschungsfragen und -designs zu einigen Forschungsansätzen zusammenfassen lassen. In diesem Beitrag sollen Forschungsergebnisse zu drei wesentlichen Forschungsansätzen[1] zusammengefasst, analysiert und ihre Bedeutung für die Durchführung von Lehrerevaluationen diskutiert werden. Die Analyse des Forschungsstandes basiert auf Grund des beträchtlichen Korpus an Forschungsarbeiten sowohl auf einer Reihe von Meta-analysen und Reviews von einzelnen Untersuchungen als auch auf ausgewählten Einzelstudien[2].

1 Thematisch zu den jeweiligen Forschungsansätzen passende Experimentalstudien werden im Rahmen dieser Ansätze erläutert und nicht als eigener Forschungsansatz dargestellt.

2 Ein Anspruch auf Vollständigkeit wird nicht erhoben, denn „literally thousands of papers have been written" (Marsh 1984, S. 708). Allerdings spricht auch Weinerts (2001, S. 1) Auffassung von einem „inflationär anwachsenden, zugleich aber hoch redundanten Schrifttum zur Evaluationsproblematik" für eine Zusammenfassung der Forschungsarbeiten.

Zeitschrift für Berufs- und Wirtschaftspädagogik, 98. Band, Heft 2 (2002)

2 Wesentliche Forschungsansätze zur Validität von Lehrerevaluationen durch die Lernenden

Ein erster Ansatz beruht auf der theoretischen Annahme, dass gute Lehrkräfte die Lehrziele des Unterrichts 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. Dies entspricht jener Vorgangsweise, mit der die Kriteriumsvalidität eines Testverfahrens bestimmt wird (vgl. BORTZ/DÖRING 1995, MUMMENDEY 1995). Deshalb ist auch der Begriff der „validity of student ratings" in den entsprechenden Publikationen weit verbreitet, auch im deutschsprachigen Raum wird von „Validität" gesprochen (vgl. RINDERMANN 2001, ders. 1999, sowie SPIEL 2001, dies. 2001a, SPIEL/GÖSSLER 2000). Dieser Ansatz wurde insbesondere in den siebziger und achtziger Jahren in den USA verfolgt. Charakteristisch dafür ist die Aussage von CENTRA (1977, S. 17): „Studies of what student ratings of instruction really measure have frequently employed student achievement as a validity criterion. The hope is that ratings of teaching or course quality would be at least moderately related to how much students learn in a course". Ein verwandter Zugang zur Untersuchung der Validitätsproblematik besteht in der Korrelation der Beurteilungen der Lernenden mit jenen anderer Beurteiler. Im Sinne einer Expertenvalidierung gelten die Evaluationen der Lernenden dann als gültig, wenn sie sehr hoch mit den Evaluationsergebnissen anderer Beurteiler, die als Experten gelten, korrelieren, d. h. weitgehend mit diesen übereinstimmen (vgl. RINDERMANN/AMELANG 1994, RINDERMANN 2001).

Ein zweiter Ansatz kritisiert die Fixierung auf den Lernerfolg als einziges Unterrichtsprodukt und fordert die Berücksichtigung weiterer – auch affektiver – Unterrichtsziele. Eine Reihe von Evaluierungsdaten soll mit mehreren Unterrichtsprodukten in Beziehung gesetzt werden. Damit weist dieser Ansatz Parallelen zur Bestimmung der Konstruktvalidität auf, bei der ein ganzes System von Zusammenhängen und Beziehungen, von Hypothesen, geprüft wird (vgl. GRUBITZSCH/REXILIUS 1985). So erläutern LIENERT und RAATZ (1998, S. 226), dass man bei einer Konstruktvalidierung von einem bestimmten Konstrukt und der entsprechenden Theorie ausgeht und daraus Hypothesen ableitet, die unter Einsatz des zu validierenden Tests empirisch überprüft werden. Tests können zu diesem Zweck zum Beispiel mit mehreren Außenkriterien korreliert werden wie auch mit Kriterien, die ein anderes Konstrukt messen, um zu zeigen, dass es keinen Zusammenhang mit diesen Größen gibt. Vor allem seit den (späten) achtziger Jahren wandte sich das Augenmerk der Bestimmung der Konstruktvalidität zu, da Unterrichtsqualität als Konstrukt und der Lernerfolg als wichtiges, jedoch nicht ausreichendes Kriterium für die Güte des Unterrichts angesehen wurden (vgl. MARSH 1987, MARSH/ROCHE 1997).

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. Biasvariablen könnten also zum Beispiel Verzerrungen der Beurteilungen von spezifischen Lehrverhaltensdimensionen wie auch der Gesamtbeurteilungen bewirken, die nicht mit dem Unterrichtsgeschehen und der Unterrichtsarbeit der Lehrkräfte zusammenhän-

gen. So könnte man etwa vermuten, dass die äußere Attraktivität einer Lehrkraft die Evaluationsergebnisse beeinflusst, ohne jedoch eine (positive oder negative) Auswirkung auf das tatsächliche Lehrverhalten und die Qualität der Unterrichtsarbeit der beurteilten Lehrkraft zu haben. Wenn die Beurteilungen der Lehrenden durch die Lernenden aussagekräftig sind, sollten solche Variablen keinen Einfluss auf die Beurteilungen der Lehrenden durch die Lernenden haben. Fehlende Biaseffekte sprechen daher „für die Validität von Evaluationen" (RINDERMANN/AMELANG 1994, S. 32).

Bevor die wesentlichen Ergebnisse der oben erläuterten Forschungsansätze analysiert und diskutiert werden können, ist zunächst die allen Ansätzen zugrundeliegende Frage nach den Beurteilungskriterien für Unterrichtsqualität zumindest in Kürze zu diskutieren.

3 Was ist guter Unterricht und anhand welcher Kriterien soll er beurteilt werden?

Bevor die Güte von Unterricht und der Unterrichtsarbeit von Lehrkräften beurteilt werden kann, müssen relevante Beurteilungskriterien für Unterrichtsqualität ermittelt werden (vgl. KREMPKOW 1998). Damit stellt sich die grundsätzliche Frage, was guter Unterricht überhaupt ist. Genau darin liegt ein wesentliches Problem der Evaluation, dass „diese Frage nicht eindeutig beantwortbar ist, weil es *die* (eine) gute Lehre nicht gibt, gar nicht geben kann" (KROMREY 1994, S. 92). Wie noch näher ausgeführt werden wird, sollten Evaluationen und die dafür eingesetzten Instrumente auf die Ziele, Inhalte und die Zielgruppe des Unterrichts sowie auf die eingesetzten Lehrmethoden Rücksicht nehmen.

In der angloamerikanischen Literatur wird die Qualität der Unterrichtsarbeit häufig als „teaching effectiveness" und eine gute Lehrkraft als ein „effective teacher" bezeichnet (vgl. MCKEACHIE 1979, COHEN 1981, MARSH 1987). Da Effektivität in der Regel das Ausmaß der Zielerreichung meint, bedeutet Effektivität des Unterrichts einer Lehrkraft das Ausmaß, in dem sie bei den Lernenden die angestrebten Unterrichtsziele tatsächlich erreicht. Denn „although teaching effectiveness is difficult to define, it is generally thought of as the degree to which an instructor facilitates student achievement" (COHEN 1981, S. 281, ebenso MCKEACHIE 1979, S. 385). Die Effektivität bezieht sich auf die Produktqualität, die Ergebnisse des Unterrichts, ebenso bedeutend sollte jedoch auch die Prozessqualität sein, wie vor allem die Unterrichtsarbeit der Lehrkräfte (vgl. DUBS 1998, ABRAMI et al. 1997). Daher müssen jene Dimensionen guten Lehrverhaltens identifiziert werden, die unter Berücksichtigung von Lehrzielen, -inhalten, -methoden und der Zielgruppe des Unterrichts tatsächlich guten Unterricht ausmachen und von den Lernenden beurteilt werden können, so schwierig dies auch sein mag. Denn „there is ample research evidence that good teachers come in many styles" (MCKEACHIE 1979, S. 392, ders. 1997a). Andererseits konnten verschiedene Komponenten guter Unterrichtsarbeit von Lehrkräften in einer Reihe von – meist empirischen – Studien auf verschiedenen Ebenen des Bildungswesens in unterschiedlichen Ländern ermittelt werden. Wenn Unterrichtsqualität auch auf Grund der Komplexität von Unterrichtsgeschehen (vgl. KREMER-HAYON 1993) und der Vielzahl der angestrebten Unterrichtsziele ein nicht leicht fassbarer Begriff ist, sodass BERLINER (1987, S. 106) sogar von einem „most elusive concept of „quality" instruction" spricht, so

können doch eine Reihe von Lehrverhaltensfaktoren wie Verständlichkeit der Erklärungen, Struktur des Unterrichts, Interaktion mit den Lernenden über viele Untersuchungen hinweg als lernwirksam und auch bedeutend aus Sicht der Lernenden identifiziert werden (vgl. BROPHY/GOOD 1986, BROPHY 2000, FELDMAN 1997, für Unterricht in kaufmännischen Fächern an österreichischen Schulen vgl. SCHNEIDER 1995 und GREIMEL 1996, dies. 2002). Wenn sich guter Unterricht auch auf sehr unterschiedliche Art und Weise verwirklichen lässt (vgl. WEINERT/HELMKE 1996), so trotzdem – und das ist die notwendige Einschränkung – keineswegs auf beliebige Weise (vgl. WEINERT 1998).

Die meisten Evaluationsinstrumente bestehen aus Items, die Gesamtbeurteilungen darstellen (z. B. „… ist insgesamt eine gute Lehrkraft") und die Produktqualitäten abbilden (z. B. „insgesamt bin ich zufrieden"), und aus Items, die sich auf einzelne Aspekte des Lehrverhaltens beziehen und sich damit vorwiegend auf die Prozessqualität des Unterrichts beziehen. Vor allem wenn die Evaluationsergebnisse diagnostisches Feedback für die Lehrkraft sein sollen, die sie für die Weiterentwicklung des Unterrichts verwenden können soll, sollten eine Reihe spezifischer Lehrverhaltensvariablen beurteilt werden und nicht nur Gesamtbeurteilungen abgegeben werden (vgl. MCKEACHIE 1997a).

Ein grundlegendes Problem für die Durchführung von Evaluationen und noch mehr für die Evaluationsforschung liegt darin, dass die Auswahl der Evaluationskriterien nicht nachvollziehbar ist und die Formulierung der Items im Evaluierungsinstrument nicht immer den Kriterien für Itemformulierungen in Fragebögen (vgl. MUMMENDEY 1995) entsprechen. Schlecht formulierte Items und fehlerhaft konstruierte Evaluierungsinstrumente beeinträchtigen jedoch die Aussagekraft der Evaluierungsergebnisse (vgl. TAGOMORI/BISHOP 1995) ebenso wie Evaluierungsinstrumente, die inhaltlich für bestimmte Formen der Lehre nicht geeignet sind. So werden z. B. für fragend-entwickelnden Unterricht zumindest teilweise andere Kriterien relevant sein als für Projektunterricht. Wesentlich erscheint mir daher, dass vor der Durchführung von Evaluationen bei der Wahl des Evaluierungsinstrumentes darauf geachtet wird, dass es für den zu evaluierenden Unterricht inhaltlich passend ist und außerdem eindeutig und zielgruppenadäquat formulierte Items enthält. Um Monotonie und mechanistisches Ausfüllen von Standardbögen wie bei einem Ritual zu vermeiden (vgl. DUBS 2000), sollten auch nicht immer dieselben Instrumente verwendet werden. Inwieweit solche Überlegungen bei den Instrumenten, die bei den Untersuchungen zur Validitätsproblematik zum Einsatz kamen, überhaupt berücksichtigt worden sind, geht aus der Mehrzahl der Publikationen nicht hervor.

4 Untersuchungen zur „Validität" von Lehrerevaluationen durch die Lernenden

Experimentalstudien wie die Untersuchungen von NAFTULIN et al. (1973) zum DR. FOX – Effekt erschütterten das Vertrauen in die Lehrerevaluationen der Lernenden, weil sie den Eindruck vermitteln, dass die Einschätzungen der Lernenden maßgeblich von persönlichen, für das Lernen irrelevanten Merkmalen des Lehrenden und nicht vom tatsächlich Gelernten beeinflusst würden. Ein charismatischer, witziger Vortragender erhielt nämlich nach einer inhaltlich schwachen Präsentation von seinem fachkundigem Publikum sehr gute Evaluationen. „Because the lecturer – actually a professional actor – was introduced as Dr. Myron L. Fox, the

phenomenon became known as the „Dr. Fox“ effect. (They) concluded that a lecturer's authority, wit, and personality can „seduce“ students into the illusion of having learned, even when the educational content of the lecture was missing“ (ABRAMI et al. 1982, S. 446). Die Metaanalysen von ABRAMI et al. (1982) kamen zu folgendem Ergebnis: „Instructor expressiveness“, das Charisma des oder der Vortragenden sowie dessen/deren Freundlichkeit, Witz und Begeisterung für das Fach, haben einen großen, meist statistisch signifikanten Einfluss auf die Beurteilungen durch die Lernenden, wesentlich geringer ist jedoch der Einfluss dieser „expressiveness“ auf den Lernerfolg. Dieser wird maßgeblich durch „lecture content“, den Inhalt des Unterrichts, beeinflusst, der jedoch nur geringen Einfluss auf die Beurteilungen durch die Lernenden hat.

Natürlich muss eingeschränkt werden, dass die – auch methodisch umstrittenen – Studien zum „Dr. Fox Effect“ in der Regel unter Laborbedingungen durchgeführt wurden, die die interessierenden Effekte besonders deutlich herausarbeiten können. Allerdings lassen sich in natürlichen Untersuchungssettings nicht annähernd so große Effekte festgestellen (vgl. RINDERMANN 1999). Eine Reanalyse der Daten durch MARSH und WARE (1982) ergab außerdem, dass sich bei Studierenden, denen ein Anreiz gegeben wurde zu lernen („incentive to learn“), die expressiveness (naheliegenderweise!) nur auf das Evaluationskriterium Enthusiasmus auswirkte, nicht jedoch auf die anderen Dimensionen. „The other four rating dimensions and the total rating score were more affected by content coverage, though primarily when expressiveness was low“ (MARSH/WARE 1982, S. 132).

Lange unbeachtet blieb auch die Möglichkeit, dass eine gute Lehrkraft sowohl inhaltlich kompetent sein sollte als auch mit Engagement und Enthusiasmus vortragen können sollte. Ein inhaltlich vollkommen richtiger Lehrvortrag, der jedoch ohne Begeisterung und persönliches Engagement gehalten wird, entspricht m. E. nicht den Vorstellungen guter Lehre. „Expressiveness“ könnte also ebenso ein Merkmal guter Lehre wie „content“ sein. Da es für die Lernenden schwieriger sein dürfte, die inhaltliche Richtigkeit und die Angemessenheit der Lerninhalte zu beurteilen als die Begeisterung und das Engagement der vortragenden Person, schlagen sich Defizite in diesem Bereich stärker in den Beurteilungen nieder als inhaltliche Mängel. „To sum up, students know when they are learning, but they do not know whether what they are learning is current, biased, or appropriate for course goals. They do rate lower an instructor who provides less content, but their ratings of effectiveness are probably less affected by amount of content than by other characteristics of teaching“ (MCKEACHIE 1979, S. 388).

Jedenfalls kann aus einer hohen Korrelation zwischen „expressiveness“ und den Evaluationen (z. B. 0,7) und einer niedrigen Korrelation zwischen „expressiveness“ und dem Lernerfolg (z. B. 0,2) noch wenig über die Korrelation zwischen den Evaluationen und dem Lernerfolg ausgesagt werden. Diese könnte bei den angegebenen Korrelationen Werte zwischen + 0,84 und – 0,56 annehmen (vgl. ABRAMI et al. 1982)[3]. Das bedeutet, dass die Ergebnisse der Dr. Fox-Studien die Frage der „Validität“ von Beurteilungen der Lehre durch die Lernenden (im Sinne eines

3 „expressiveness“ = Variable 1, Evaluationen = Variable 2 und Lernerfolg = Variable 3. Annahmen: $r_{12} = 0{,}7$ und $r_{13} = 0{,}2$. Mögliche Werte für r_{23} (vgl. McNemar 1962, S. 167):

limits for $r_{23} = r_{12} \cdot r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 \cdot r_{13}^2}$

Zusammenhangs mit dem Lernerfolg) nicht beantworten, sondern diese Frage erst recht stellen. Zur Größe dieses Zusammenhangs liegen mittlerweile zahlreiche empirische Ergebnisse vor, die in der Folge erläutert werden.

4.1 Korrelation mit dem Außenkriterium Lernleistung

COHEN fasste 1981 in einer Meta-Analyse die Daten und Ergebnisse von 41 voneinander unabhängigen Korrelationssstudien zusammen, „what appears to have been the most influential meta-analysis of multisection studies“ (MARSH/DUNKIN 1997, S. 266). „Multisection study“ bedeutet, dass die Untersuchung zur Erhebung des Zusammenhangs zwischen Evaluationen und Lernerfolg bei Studierenden in Parallelveranstaltungen mit demselben Inhalt (und derselben Prüfung) jedoch unterschiedlichen Lehrpersonen durchgeführt wurde. Experimentalstudien wie insbesondere sämtliche Dr. Fox-Untersuchungen wurden nicht in die Analyse miteinbezogen. Weitere Auswahlkriterien waren die Klasse als Analyseebene (nicht der oder die einzelne Studierende) und ein standardisierter Leistungstest zur Erhebung der Lernleistung, also nicht die von der unterrichtenden Lehrerin oder dem unterrichtenden Lehrer gegebene Note. Diese beiden Merkmale sind für die Qualität der Untersuchung entscheidend[4].

Die Meta-Analyse bezog sich sowohl auf global ratings, also Gesamtbeurteilungen der Lehrperson und des Kurses, wie etwa „Der/die Lehrveranstaltungsleiter/in ist ein/e hervorragende/r Lehrer/in“ („The instructor is an excellent teacher“) und „Dieser Kurs ist hervorragend“ („This is an excellent course“), als auch auf Beurteilungen spezifischer Lehrverhaltensdimensionen. Diese können den Überbegriffen Skill, Rapport, Structure, Difficulty, Interaction und Feedback zugeordnet werden.

Als durchschnittliche Korrelation zwischen der Gesamtbeurteilung der Lehrperson und dem Lernerfolg (bzw. der Lernleistung) der Beurteilenden wurde $r = 0{,}43$ ermittelt, die Beurteilung des Kurses korrelierte mit den Lernleistungen im Durchschnitt mit $r = 0{,}47$. Beide Verteilungen weisen lt. COHEN (1981) keine hohe Streuung auf, das heißt, die meisten Korrelationswerte liegen nahe beim ermittelten Mittelwert.

Die höchste durchschnittliche Korrelation wurde bei der Dimension Skill mit $r = 0{,}50$ erreicht, gefolgt von Structure mit 0,47. Rapport und Feedback wiesen jeweils durchschnittliche Korrelationen in der Höhe von $r = 0{,}31$ auf, Interaction erreichte 0,22 und die durchschnittliche Korrelation mit Difficulty war ein leicht negativer Wert mit $r = -0{,}02$. COHENS (1981) Schlussfolgerungen aus den Ergebnissen

4 Die Analyseebene für Validitätsuntersuchungen sollte die Lehrperson bzw. die Klasse (der Kurs) sein. Einige Untersuchungen hatten die einzelnen beurteilenden Lernenden als Analyseebene gewählt, wodurch sie jedoch eine Antwort auf eine andere Forschungsfrage erhalten hatten, nämlich ob Lernende, die höhere Lernleistungen erzielen, auch bessere Beurteilungen der Lehrperson abgeben. Die Frage muss jedoch sein, ob Lehrpersonen, die bessere Beurteilungen von ihren Lernenden erhalten, auch wirklich besseren Unterricht machen. Diese Frage kann nur auf der Analyseebene der Klasse bzw. der Lehrperson beantwortet werden. Zur Erhebung des Lernerfolgs muss ein standardisierter Leistungstest in gleichartigen Kursen mit denselben Lernzielen, die von unterschiedlichen Lehrkräften unterrichtet werden, eingesetzt werden, der nicht von der unterrichtenden Lehrerin / vom unterrichtenden Lehrer korrigiert und benotet wird, sondern von einer externen Person (vgl. Cohen 1981).

dieser Meta-Analyse sind, dass „students do a pretty good job of distinguishing among teachers on the basis of how much they have learned“ (S.□305) und dass die Analyse „provides strong support for the validity of student ratings as measures of teaching effectiveness“ (S.□300). Seine Schlussfolgerungen werden von anderen Wissenschaftlern nicht geteilt, die die Validität nicht als eindeutig gegeben sehen (vgl. ABRAMI et al. 1988). Wie in diesem Fall eine Korrelation von durchschnittlich etwa 0,4 tatsächlich zu beurteilen ist, soll in der Analyse der empirischen Befunde näher beleuchtet werden.

Eine weitere Schlussfolgerung COHENS bezog sich darauf, dass bestimmte Lehrverhaltensdimensionen offensichtlich stärker mit dem Lernerfolg korrelieren als andere. In COHENS Analyse waren dies vor allem Skill, also die fachdidaktischen Kompetenzen der Lehrperson, und Structure, also die Nutzung der zur Verfügung stehenden Unterrichtszeit und die Organisation des Unterrichts an sich. Faktoren, die sich mehr auf das Verhältnis zwischen Lehrenden und Lernenden bezogen, wie Rapport und Interaction korrelierten im Durchschnitt schwächer mit dem Lernerfolg, jedoch auch durchwegs positiv. Möglicherweise korrelieren sie stärker mit anderem Unterrichtsoutput wie beispielsweise mit nicht ausschließlich fachbezogenem Lernerfolg, der Motivation der Lernenden oder deren emotionaler Befindlichkeit. Darüber hinaus wäre auch denkbar, dass es zwar einen nicht nur mäßigen Zusammenhang gibt, dieser jedoch nicht linear ist.

Interessanterweise korrelierte die Dimension Difficulty gar nicht mit dem Lernerfolg der Lernenden. Die Einschätzung des Anspruchsniveaus und der Arbeitsbelastung durch den Kurs zeigt damit keinen Zusammenhang mit den Lernleistungen der Studierenden.

Die Korrelation war in jenen Fällen höher, in denen externe Beurteiler die Leitungstests korrigierten und nicht die unterrichtende Lehrerin oder der unterrichtende Lehrer selbst. Ob die Studierenden ihren Kurs selbst auswählten oder auf Zufallsbasis einem Kurs zugewiesen wurden, beeinflusste die Höhe der Korrelation nicht, ebenso wenig, ob die Untersuchungen die Eingangsvoraussetzungen und die allgemeinen Fähigkeiten („abilities“) der Lernenden berücksichtigten oder nicht (vgl. dazu auch FELDMAN 1997). Dies entspricht den Ergebnissen von MARSH und OVERALL (1980, S.□473), die in ihrer Untersuchung partielle Korrelationen zwischen Lernleistungen bei der Endprüfung eines Computerkurses und den Lehrergesamtevaluationen berechneten, in denen sie die Effekte von Eingangsinteresse und Vorwissen statistisch kontrollierten. Sie ermittelten einen statistisch signifikanten Korrelationskoeffizienten von (nur) r = 0,38. Die Lernleistung korrelierte (ebenfalls partiell) mit der Gesamtbeurteilung des Kurses nur 0,26 (keine statistische Signifikanz).

1987 baute COHEN seine Analyse noch aus, er erzielte bei seiner aktualisierten Analyse durchwegs Ergebnisse, die mit jenen von 1981 vergleichbar waren. FELDMAN (1989) führte die Meta-Analysen von COHEN fort, indem er dieselbe Datenbasis verwendete wie COHEN (1981 und 1987), jedoch mehr Evaluationskriterien zu spezifischen Lehrverhaltens- und Unterrichtsdimensionen analysierte als COHEN. Er stellte fest, dass die hohe Korrelation, die sich in COHENS Analysen zwischen Skill und den Lernerfolgen zeigte, hauptsächlich auf die in Skill enthaltenen Faktoren Verständlichkeit und Klarheit zurückzuführen war und weniger auf die auch von COHEN der Dimension Skill zugeordnete reine Fachkompetenz der Lehrperson oder deren Anpassung an das Niveau der Klasse. Ein vergleichbarer

Schluss konnte für den Faktor Planung und Organisation der Lehrveranstaltung innerhalb des Überbegriffs Structure, der auch vergleichweise hoch mit dem Lernerfolg korrelierte, gezogen werden. Innerhalb der COHENschen Dimension Rapport, die mäßig mit dem Lernerfolg korrelierte, schien auf Grund der Datenanalyse von FELDMAN der Faktor Hilfsbereitschaft und Zugänglichkeit vorherrschend zu sein (vgl. FELDMAN 1989). Insgesamt lässt sich feststellen, dass die ermittelten Korrelationen bei allen drei Analysen durchwegs vergleichbare Werte aufweisen und die Bedeutung von Klarheit und Verständlichkeit besonders unterstreichen.

Ein vergleichender Überblick von RINDERMANN (2001) zeigt, dass mehrere Metaanalysen aus den Jahren 1973 bis 1997 (darunter auch jene von COHEN 1981 und 1987 sowie FELDMAN 1989) im Schnitt eine Korrelation von r = 0,52 zwischen den Evaluierungen und den Lernleistungen der Lernenden ermittelten. Dieser Durchschnitt entspricht auch den Ergebnissen einer Reihe von Einzelstudien in RINDERMANNs Analyse, die in der Mehrzahl der Fälle zu einem Ergebnis im Bereich von 0,3 bis 0,5 kommen. In einzelnen Fällen werden jedoch auch höhere Korrelationen erreicht. Auch OBIEKWE (1999, S. 11) ermittelte in seiner Untersuchung, dass in ausgewählten Lehrveranstaltungen die durchschnittliche Korrelationen auch Werte bis über r = 0,9 erreichen. Insgesamt bedeutet das für die gesamte durchschnittliche Korrelation von z. B. 0,4 oder 0,5, dass in manchen Lehrveranstaltungen nur sehr schwache oder sogar negative Beziehungen zwischen den Evaluationen und den Lernergebnissen gefunden werden können. Einzelne Korrelationskoeffizienten können also höchst unterschiedliche Werte aufweisen, was in den Durchschnittswerten von Meta-analysen und Literaturreviews meistens nicht zum Ausdruck kommt (vgl. dazu auch ABRAMI et al. 1988 sowie ABRAMI et al. 1997).

4.2 Korrelation mit den Evaluationen weiterer Beurteiler

RINDERMANN und AMELANG (1994) sowie RINDERMANN (1999, ders. 2001) untersuchten die Validität von studentischen Beurteilungen der Lehre an Universitäten in Deutschland. Sie korrelierten die Beurteilungen der Studierenden mit jenen von geschulten Fremdbeurteilern, den „Experten“. Die Korrelationen zwischen den Evaluationsergebnissen von Studierenden mit jenen von Fremdurteilern erzielten bei einzelnen Dimensionen Werte bis über r = 0,8. Die Mehrzahl der Korrelationskoeffizienten liegt jedoch zwischen 0,5 und 0,7 und entspricht damit in etwa dem über mehrere Studien von RINDERMANN (2001, S. 166) ermittelten Durchschnitt von r = 0,54 und dem Ergebnis einer Metaanalyse von FELDMAN (1989a) in der Höhe von r = 0,53.

Ein Mittelwertvergleich zwischen den Urteilen der Studierenden und den Selbstbeurteilungen der Vortragenden zeigt, dass auch hier eine Beziehung zwischen den Urteilen besteht. Bei Abweichungen neigen die Vortragenden „nicht grundsätzlich zu besseren Einschätzungen, sie unterschätzen sich (relativ zu studentischen und Fremdurteilern) in der Lehrkompetenz“ (RINDERMANN/AMELANG 1994, S. 29). RINDERMANN (1999, S. 11) schließt daraus, dass „Beurteilungen durch Studierende somit kein unrealistisches Bild der Lehre zeichnen und sensibel für Lerngewinn sind“. In diesem Zusammenhang betont er jedoch zu Recht, dass nicht die Ergebnisse einzelner Veranstaltungen oder gar einzelner Studierender

herangezogen werden sollte, sondern jeweils von mehreren (mindestens vier) Veranstaltungen, die von zumindest zehn bis 15 Studierenden besucht werden. Fremdbeurteilungen durch Fachgutachter sollten zusätzlich herangezogen werden.

Vergleichbare Untersuchungen liegen auch aus den USA vor. Dort untersuchten beispielsweise HOWARD et al. (1985) den Zusammenhang zwischen den Beurteilungen von Studierenden mit jenen von ehemaligen Studierenden, von Kollegen und acht geschulten Unterrichtsbeobachtern sowie mit den Ergebnissen der Selbstbeurteilung der 43 an der Untersuchung teilnehmenden Lehrpersonen an einer amerikanischen Universität. Am höchsten korrelierten die Beurteilungen von Studierenden mit denen von ehemaligen Studierenden ($r = 0,74$). Die Selbstevaluationen korrelierten mit $r = 0,31$ mit den Evaluationsergebnissen von ehemaligen Studierenden und mit $r = 0,34$ mit jenen der (aktuell) Studierenden. Alle weiteren Evaluationsergebnisse korrelieren nur schwach miteinander. Die Analysen umfassten unter anderem auch die Durchführung einer konfirmatorischen Faktorenanalyse, bei der die ermittelten Koeffizienten zwar etwas höher sind als die zuvor dargestellten Korrelationskoeffizienten; Die höchsten Werte weisen allerdings wiederum die Beurteilungen der ehemaligen Studierenden (0,88) und der (aktuell) Studierenden auf.

4.3 Ansätze zur Konstruktvalidierung

Seit Ende der 80er Jahre mehrten sich die Stimmen, dass der Lernerfolg oder der Lernzuwachs der Studierenden als Indikator für Unterrichtsqualität nicht ausreiche, dass die Untersuchung der kriteriumsbezogenen Validität als Ansatz zu kurz greife und ein Konstruktvalidierungansatz gewählt werden solle, der mehr Indikatoren für Unterrichtsqualität als nur den kognitiven Lernerfolg der Studierenden in Betracht ziehen soll. Der Mehrdimensionalität von Unterrichtsprozessen und -ergebnissen müsse bei der Evaluation von Unterricht Rechnung getragen werden. Lehrverhalten ist mehrdimensional und kann nicht nur mit einem Item gemessen werden, auch ein globales Rating kann Lehrverhalten nicht adäquat wiedergeben. Ebenso kann das Unterrichtsergebnis nicht nur an der Produktqualität Lernerfolg festgemacht werden (vgl. MARSH 1984, COHEN 1981, MARSH/ROCHE 1997). Beim Konstruktvalidierungsansatz („construct validity approach“) werden verschiedene Beurteilungsdimensionen zur Lehrerevaluation mit einer Reihe von Produktqualitäten von Unterricht korreliert.

Die zentrale Produktqualität ist weiterhin der Lernerfolg der Schülerinnen und Schüler, entscheidend ist jedoch auch, wie gerne und motiviert gelernt wird, die emotionale Befindlichkeit der Lernenden, die Entwicklung ihres Selbstkonzepts, die Stimulierung ihres Interesses, viele weitere Faktoren sind prinzipiell denkbar (vgl. MARSH/ROCHE 1997, 1190).

Das Konzept der Konstruktvalidierung, in dem die Beurteilungen einer Reihe von einzelnen Beurteilungsdimensionen mit jenen einer Reihe von Produktqualitäten von Unterricht in Zusammenhang gebracht werden, ist statistisch mit einem Strukturgleichungsmodell zu untersuchen. Solche Strukturgleichunsgsmodelle liegen auf Grund ihrer Komplexität in wesentlich geringerem Umfang vor als zum Beispiel Korrelationsstudien. Auch MARSH, der zu den vehementesten Befürwortern des Konstruktvalidierungsansatzes zählt, hat selbst erst in seinen jüngsten

Publikationen die empirischen Ergebnisse von Strukturgleichungsmodellen dargestellt, in denen jedoch nur der von den Lernenden subjektiv eingeschätzte Lernerfolg berücksichtigt wurde, kein im Rahmen einer für alle Lernenden standardisierten Prüfung erhobenes Lernergebnis. Schwerpunkt dieser Untersuchungen war die Prüfung der Wirkung der Arbeitsbelastung – er differenziert zwischen aus Sicht der Studierenden nützlichen Lehrveranstaltungsstunden („good workload“) und nicht nützlichen Stunden („bad workload“) sowie des Eingangsinteresses und der Motivation auf den Zusammenhang zwischen Noten und Evaluationen. Zu den wesentlichen Erkenntnissen zählt MARSH (2001) das Ergebnis, dass von Studierenden als „good workload“ bezeichnete Lern- und Arbeitszeit eine positive Wirkung auf die Lehrerevaluation habe, während „bad workload“ negative Auswirkungen habe (vgl. MARSH 2001).

Die Strukturgleichungsmodelle von RINDERMANN (2001, S.□224) zeigen, dass die Lehrkraft und deren Lehrverhalten „die zentrale Größe für den wahrgenommenen Lehrerfolg darstellt“. Lehrerfolg ist dabei eine der latenten Variablen im Modell von RINDERMANN, die durch die manifesten Variablen Interessantheit der Veranstaltung, Lernen-quantitativ, Lernen-qualitativ, Interessenförderung und Allgemeinurteil gebildet wird. Beide Dimensionen des Lernens beruhen auf Selbsteinschätzungen der Lernenden, einerseits zum Umfang des Gelernten (quantitativ), andererseits zur Sinnhaftigkeit des Inhalts (qualitativ). Darüber hinaus wäre es m. E. interessant, Prüfungsergebnisse als zusätzliche „Lernen-Variable“ in das Modell einzuführen. Die latente Variable Dozent-Lehrkompetenz bildet wesentliche Komponenten des Lehrverhaltens ab und bestimmt am stärksten den Lehrerfolg, während die übrigen latenten Variablen Interaktion und Anforderungen-Arbeitslast nur geringe Wirkung zeigen. RINDERMANN (2001) schränkt jedoch selbst ein, dass die Untersuchung der Bedeutung der Rahmenbedingungen und der studentischen Merkmale weiterer Analysen bedarf. Sie sind in seinen Modellen bislang nicht berücksichtigt.

4.4□Zusammenfassende Analyse der Korrelationsstudien und des Konstruktvalidierungsansatzes

Da die methodische Vorgangsweise in den skizzierten Untersuchungen jener bei der Validierung von Testverfahren vergleichbar ist, wird in der Literatur weitverbreitet der Begriff von der „Validität“ der Urteilen der Lernenden zur Lehrerbeurteilung gesprochen. Interessant erscheint die Überlegung, dass in der gesamten Literatur eigentlich nicht diskutiert wird, ob der der klassischen Testtheorie entlehnte Begriff in diesem Zusammenhang angemessen verwendet wird. In der klassischen Testtheorie bezieht sich die Validität als Testgütekriterium darauf, ob mit dem Testverfahren tatsächlich das gemessen wird, was zu messen vorgegeben wird. Damit zielt der Begriff Validität auf die Qualität eines Tests ab, der dann als valide bezeichnet wird, wenn er „in der Lage ist, genau das zu messen, was er zu messen vorgibt“ (BORTZ/DÖRING 1995, S.□185). So wird die Validität eines Intelligenztests oder eines Interessenstests unter anderem damit bestimmt, wie hoch die ermittelten Intelligenztestwerte oder Interessenswerte mit jenen Werten korrelieren, die die gleichen Probanden bei anderen Intelligenz- oder Interessenstestverfahren erzielen. Je nach dem Testzweck werden verschieden hohe Korrelationskoeffizienten erwartet, die im Idealfall über 0,7, zumindest jedoch über 0,5

liegen sollten (vgl. LIENERT/RAATZ 1998, GRUBITZSCH/REXILIUS 1985). Der Fokus der Validitätsbestimmung liegt also beim eingesetzten Verfahren, nicht bei den Personen, die mit diesem Verfahren befragt werden, und deren Urteilsfähigkeit. In den Untersuchungen zur Lehrerevaluation durch die Lernenden geht es jedoch primär um die Glaubwürdigkeit der Angaben der befragten Personen und erst in zweiter Linie um das Instrument. Inwiefern das eingesetzte Evaluationsinstrument für den zu beurteilenden Unterricht adäquat ist, ob es der Art der Lehrziele und Lerninhalte, den eingesetzten Unterrichtsmethoden und der Zielgruppe des Unterrichts angepasst ist, bleibt in vielen Fällen unberücksichtigt oder wird zumindest in den Veröffentlichungen unzureichend diskutiert. Eine Ausnahme stellt beispielsweise RINDERMANN (2001) dar, der in seine Definitionen von Validität der Lehrerevaluation durch die Lernenden auch dezidiert das verwendete Evaluationsinstrumentarium miteinbezieht.

Abgesehen von der Bedeutung des Begriffs Validität in der klassischen Testtheorie lassen sich trotz der teilweise inkonsistenten Ergebnisse der zahlreichen Untersuchungen zu den Zusammenhängen zwischen Evaluationsergebnissen von Lernenden und Lernerfolgen einerseits sowie den Evaluationen von Fremdbeurteilern andererseits in der Regel positive, jedoch im Durchschnitt nicht allzu hohe Korrelationen feststellen (vgl. COHEN 1981, ders. 1987, MARSH 1987, FELDMAN 1989, ders. 1989a). Das bedeutet, dass in der Regel, nicht jedoch unbedingt in jeder Klasse oder Lehrveranstaltung eine gewisse Beziehung zwischen der Beurteilung der Lehrkraft durch die Lernenden und dem Lernerfolg besteht, der auch als „moderately valid“ oder „mittelvalide“ bezeichnet wird (RINDERMANN 2001, S.□166). Es ist jedoch in Anbetracht der oben ausgeführten Überlegungen zum Begriff „Validität“ fraglich, ob diese Korrelation für die angestrebte „Validierung“ der Schülerbeurteilung hinsichtlich ihres Lernerfolges als Außenkriterium ausreichend ist. ABRAMI et al. (1982, S.□459) bezeichnen diese Frage wörtlich als „matter of judgement“, weil eine durchschnittliche Korrelation von 0,4 oder 0,5 für manche bereits ausreicht, um die Validität zu begründen (vgl. CASHIN 1988, COHEN 1981), für andere hingegen ist sie zu gering. Eine Korrelation von z. B. 0,4 bedeutet, dass weniger als ein Fünftel der Varianz in den Beurteilungen durch den Lernerfolg erklärt wird, die Validität von Schülerurteilen zur Lehrerbeurteilung wird hinsichtlich des Lernerfolges daher verneint oder zumindest angezweifelt. So meinte auch FRANZ E. WEINERT, dass die Lehrerbeurteilungen der Lernenden zwar valide hinsichtlich ihrer Zufriedenheit mit der Lehrkraft, nicht jedoch hinsichtlich ihres Lernerfolges sind (persönliche Kommunikation am 1.2.2001).

Beim Ansatz der Expertenvalidierung besteht darüber hinaus das Problem, dass „die zum Teil hohen Korrelationen und die geringen Mittelwertsunterschiede (…) nicht das Problem lösen, ob die gefundenen Urteilsübereinstimmungen als Kennzeichen der Validität studentischer Werte interpretiert werden können oder ob sie nur Urteilskonkordanzen entsprechen. – eine Konkordanz, die auch im unzutreffenden Urteil bestehen könnte“ (RINDERMANN 2001, S.□172).

Diese Feststellungen bedeuten jedoch aus zahlreichen Gründen nicht, dass die Lehrerbeurteilungen der Lernenden nicht aussagekräftige Urteile über den Unterricht darstellen. Die folgenden Überlegungen sollten unbedingt berücksichtigt werden, bevor voreilige Schlüsse gezogen werden:

Die Validierung mit einem Außenkriterium kann nur so gut sein wie das Außenkriterium selbst. Wie valide sind jedoch die Validitätskriterien selbst? Kognitive

Lernziele sind für gewöhnlich nicht die einzigen Lernziele, die eine gute Lehrkraft anstrebt. „Teachers effective in teaching a good deal of knowledge are not necessarily effective in teaching critical thinking“ (McKeachie 1979, S. 384). Als affektive Ziele nennt McKeachie (1979) beispielsweise die Förderung von Interesse am Unterrichtsgegenstand und der Motivation, für das Fach zu lernen und sich auch nach dem Unterricht, ja sogar nach Abschluss der Ausbildung mit der Materie zu beschäftigen. Je komplexer jedoch das Spektrum an Lehr-/Lernzielen ist – nicht nur Wissenserwerb und -anwendung, sondern auch Entwicklung fachübergreifender Fähigkeiten, Einstellungs- und Verhaltensänderungen –, umso schwieriger können sie gemessen und in einer empirischen Untersuchung berücksichtigt werden.

Manche Lehrziele wie etwa selbständige Auseinandersetzung mit Inhalten und Problemen ohne Unterstützung durch die Lehrkraft oder Konfliktbewältigung in der Gruppe können schwierig zu erreichen sein und/oder den Interessen der Lernenden (zumindest anfangs) zuwiderlaufen, sodass sie im Lernprozess eine negative Einstellung zur Lehre und zur Lehrkraft entwickeln könnten, obwohl sie Lernerfolge erzielen. Unterricht kann daher je nach Unterrichtsgegenstand und -zielen, Art der Lerninhalte und Zielgruppe des Unterrichts mit unterschiedlichen Methoden realisiert werden. In vielen Studien wird nicht zwischen verschiedenen Unterrichtsgegenständen differenziert und für alle Studienrichtungen und Lehrveranstaltungstypen (die gleiche Unterrichtsform angenommen und) der gleiche Evaluierungsbogen eingesetzt. Die Frage, ob er für den zu evaluierenden Unterricht überhaupt ein taugliches Instrument war, wird damit ausgeblendet.

Weiters sind die erhobenen Prüfungsergebnisse vermutlich kein zufriedenstellendes Maß für den tatsächlichen Lernerfolg der Lernenden. Das Untersuchungsdesign verlangt eine große Anzahl von Prüfungsteilnehmern, die dieselbe Lehrveranstaltung mit denselben Lehrinhalten bei verschiedenen Vortragenden besucht haben und nun dieselbe standardisierte Prüfung ablegen. Diese Bedingungen sind meistens nur für einführende Lehrveranstaltungen typisch. Die Aufgabenstellungen, die für gewöhnlich in derart großen Prüfungen eingesetzt werden, erfordern zur objektiven Korrektur eindeutig richtige Antworten und prüfen deshalb oft nur auf einem niedrigen Lehrzielniveau, auf dem Wissensreproduktion mehr zählt als Wissensanwendung, Problemlösung oder andere kognitive Ziele (vgl. McKeachie 1997, Feldman 1997). Werden multiple-choice- und/oder „true/false“-Aufgaben eingesetzt, besteht darüber hinaus das Problem der Ratemöglichkeiten (vgl. Wagenaar 1995). Dies könnte auch erklären, warum Feldman (1997) hohe Korrelationen der Faktoren „intellectual challenge“ und „encouragement of students' independent thought“ mit den Gesamtbeurteilungen des Unterrichts durch die Studierenden, jedoch nur schwache Korrelationen dieser Faktoren mit den bei den Prüfungen festgestellten Lernleistungen festgestellt hat.

Schließlich müssen zahlreiche Faktoren berücksichtigt werden, welche die Beziehung zwischen Unterrichtsqualität und Lernerfolg moderieren. Dazu zählen insbesondere das Vorwissen der Schülerinnen und Schüler, ihr Interesse am Fach, ihre Eigenarbeit beim Lernen zu Hause, ihre Motivation, für das Fach zu lernen und die Prüfung zu bestehen, sowie deren Intelligenz (vgl. Helmke & Weinert 1997). Dadurch könnte der fachbezogene Lernerfolg wesentlich unempfindlicher gegenüber unterschiedlicher Unterrichtsqualität sein als die Beurteilungen der Lernenden. Wenn es den Lernenden nämlich wichtig ist, diese Prüfung zu bestehen,

werden sie schlechten Unterricht möglicherweise durch zusätzliches Lernen zu kompensieren versuchen (vgl. HOWARD et al. 1985). „Thus performance of students on a final examination is not an ideal measure of teaching effectiveness. Nevertheless, this is the best we have in most of our validity studies“ (MCKEACHIE 1979, S. 385).

Bivariate Analysen wie Korrelationen zwischen Evaluationen und fachbezogenem Lernerfolg müssen angesichts der Komplexität von Unterricht und des Zustandekommens von Lernerfolgen zu kurz greifen. Strukturgleichungsmodelle wie jene von MARSH (2001) und RINDERMANN (2001) sind daher ein bedeutender Beitrag zur Erforschung der Zusammenhänge im Unterrichts- und Beurteilungsprozess. Sie müssten allerdings noch mehr als die bislang berücksichtigen Variablen beinhalten[5], um die Forschungsfragen umfassend zu beantworten.

Zuletzt – und dies ist vermutlich der wichtigste Punkt – sind die Lernenden immer die Hauptzielgruppe des Unterrichts. Sie sind jene Personengruppe, für die der Unterricht gehalten wird und die den Unterricht am längsten miterleben. Sie sind eine größere Zahl von Beurteiler/inne/n als jede andere Beurteilergruppe (vgl. RINDERMANN/AMELANG 1994). Wie sie den Unterricht erleben und beurteilen, muss daher für jede Lehrkraft relevant sein. Selbst wenn ihre Beurteilung hinsichtlich des in Prüfungen ermittelten Lernerfolgs nicht valide oder nur „mittelvalide“ sein sollten, so stellen ihre Einschätzungen dennoch für jede Lehrkraft wertvolles Feedback zum Unterricht dar. Eine wichtige Voraussetzung für sinnvolles Evaluieren ist natürlich die Auswahl eines für die jeweilige Unterrichtsform passenden und für die Zielgruppe verständlichen Instrumentes.

4.5 Untersuchung von Einflussfaktoren auf die Lehrerbeurteilung durch die Lernenden

„One need not to talk with faculty very long to be aware of their concern about possible biases in student ratings“ (CASHIN 1995, S. 6). Denn selbst wenn die Korrelationen zwischen Evaluierungsergebnissen und den in Prüfungen festgestellten Lernleistungen für (die Validität von) Lehrerevaluation durch die Lernenden sprechen, „this does not mean that they are impervious to influences by other factors“ (MCKEACHIE 1979, S. 389). Insbesondere der Einfluss der (erwarteten) Noten – und der damit möglicherweise in Zusammenhang stehenden Milde der Lehrkraft bei der Beurteilung der Leistungen – sowie des Interesses der Lernenden am Unterrichtsgegenstand auf die Evaluierungsergebnisse wurden eingehend in zahlreichen Studien empirisch untersucht. Ebenfalls als Biasvariablen vermutet wurden biografische Merkmale der Lernenden wie auch der Lehrenden, die Motivation der Lernenden, ihre Sympathie für die Lehrkraft, die Klassengröße, Raumverhältnisse sowie der Unterrichtsgegenstand selbst und der Schwierigkeitsgrad des Lehrstoffes. Die vermuteten Biasvariablen werden in der Regel durch Berechnung der Interkorrelationen mit den entsprechenden Beurteilungsskalen untersucht (vgl. RINDERMANN 2001), der Einfluss der Noten(erwartungen) auch durch Experimentalstudien.

5 so z. B. auch die in Abschnitt 4.5 diskutierten potenziellen Einflussfaktoren in Form von manifesten Variablen im Strukturgleichungsmodell

Ein grundlegendes Problem besteht lt. MARSH (1987) bereits darin, dass die Unterscheidung von Einflussfaktoren und Lehrverhaltensvariablen nicht trennscharf durchgeführt worden sei. Dass bestimmte Lehrverhaltensvariablen die Beurteilung der Lehrkraft und/oder des Unterrichts beeinflussen, dürfe nicht überraschen und sei auch nicht als Einfluss im ursprünglichen und eigentlichen Sinn zu werten. MARSH und ROCHE (1997) kritisieren weiters, dass ein Großteil der Forschungsarbeiten in diesem Bereich einer theoretischen Grundlage entbehren und deshalb nicht einmal definieren, was in diesem Zusammenhang als Bias verstanden wird. Einflussfaktoren auf die Beurteilungen durch die Lernenden sind nicht automatisch einem Bias gleichzusetzen, da sie aus sachlichen Gründen mit einer oder mehreren Beurteilungsdimensionen zusammenhängen können (vgl. dazu auch MARSH/DUNKIN 1997).

Ein stark simplifizierender Ansatz („simplistic bias hypothesis") besteht darin, nur in den folgenden drei Faktoren mögliche Biasvariablen zu sehen: in der Milde der Benotung der Leistungen der Lernenden, in dem geringen Anspruchsniveau des Unterrichts der Lehrkraft und darin, dass sich die Lehrkraft nur in kleinen Klassen evaluieren lassen will (unter der Annahme, dass Evaluationen in kleinen Klassen tendenziell besser ausfallen als in großen). Alle drei Faktoren wurden in einer Reihe von Studien hinsichtlich ihrer Wirkung auf die Evaluationen untersucht, die Ergebnisse werden in den folgenden Passagen näher diskutiert. Dieser simplifizierende Ansatz klammert jedenfalls eine Reihe von relevanten potenziellen Faktoren aus und impliziert isolierte, singuläre Betrachtungsweisen, die der Komplexität des Unterrichts und der Lehrerevaluation nicht gerecht werden können (vgl. MARSH 1987).

Andere Wissenschaftler bezeichnen alle jene Einflussvariablen auf die Evaluationen als Bias, welche die Lehrperson nicht steuern kann. Doch auch diese Definition greift zu kurz, da sie beispielsweise die Milde der Lehrperson bei der Benotung der Studierenden ausklammert, obwohl diese wahrscheinlich der meist diskutierte und untersuchte Einflussfaktor auf die Beurteilungen war und ist. Da dieser zentrale Faktor jedoch eindeutig von der Lehrperson gesteuert werden kann und mit der Qualität der Lehre nicht systematisch zusammenhängt, wurde auch dieser Ansatz als inadäquat zurückgewiesen.

Ein Bias beeinflusst lt. MARSH (1987) die Beurteilungen, obwohl er mit ihnen und auch mit den Indikatoren für Unterrichtsqualität wie beispielsweise dem Lernerfolg in keinem sachlich begründbaren Zusammenhang steht. In diesem Fall kann von einer Verzerrung der Ergebnisse durch einen Bias gesprochen werden (vgl. auch MARSH/ROCHE 1997). Die Schwierigkeit besteht nun darin, dass zwar relativ einfach gezeigt werden kann, dass eine Variable mit den Lehrerevaluationen durch die Lernenden zusammenhängt, gleichzeitig aber auch gezeigt werden muss, dass diese Variable nicht mit dem Lernerfolg oder mit anderen Indikatoren von Unterrichtsqualität oder Unterrichtseffektivität logisch zusammenhängt. Dieses Problem stellt sich bereits bei der Betrachtung der Wirkung des Interesses der Evaluierenden.

Interesse

Insgesamt zeigen die Ergebnisse verschiedener Untersuchungen im Überblick, dass das Interesse der Lernenden an der Materie ihre Lehrerevaluationen zu den stärksten von allen potenziellen Einflussfaktoren gehört (vgl. RINDERMANN 2001).

In einem Regressionsmodell von SPIEL und GÖSSLER (2000) zur Lehrbeurteilung durch Studierende in verschiedenen Fakultäten der Wiener Universität werden für das Interesse an den Inhalten der Lehrveranstaltung zum Beispiel Betagewichte von etwa 0,5 bis 0,6 ermittelt. Keinen statistisch signifikanten Einfluss hatten die Raumverhältnisse und biografische Merkmale der Lernenden und Lehrenden. Das Interesse war damit die einzige Variable, die über alle Substichproben hinweg systematisch mit den Urteilen der Studierenden kovariierte (vgl. dazu auch SPIEL 2001a).

Ein Vergleich mit empirischen Untersuchungen an amerikanischen Universitäten ergibt ein konsistentes Bild. Auch hier zählten das (bereits eingangs bestehende) Interesse am Unterrichtsgegenstand und Interesse als Besuchsgrund für die Lehrveranstaltung zu den einzigen Faktoren, die einen Zusammenhang mit den Lehrevaluationen der Studierenden zeigten (vgl. NASSER/GLASSMANN 1997).

Eine Zusammenfassung mehrerer Untersuchungen bei RINDERMANN (2001) zeigt, dass höhere Zusammenhänge bei den Faktoren Vorinteresse, Thema und Besuchsgrund (einer Lehrveranstaltung) festzustellen sind. Alle diese Faktoren sind Indikatoren für das Interesse am Lehrstoff. Auch bei eigenen Untersuchungen mit dem Heidelberger Inventar zur Lehrveranstaltungsevaluation findet RINDERMANN (2001, S.□190) entsprechende Ergebnisse: „Studierende, die Interesse für ein Veranstaltungsthema empfinden und aus Interesse oder wegen des Dozenten (vs. Pflicht) eine Veranstaltung besuchen, beurteilen diese, den Dozenten und den Lehrerfolg günstiger als Studierende, die kein Interesse entwickeln können und aus extrinsischen Gründen die Veranstaltung belegen“.

Die Wirkungsrichtung des Zusammenhangs bleibt unklar, insbesondere wenn die Faktoren zur Abbildung des Interesses nicht zu Beginn des Unterrichts bzw. der Lehrveranstaltung erhoben werden, sondern zu einem späteren Zeitpunkt (z. B. gemeinsam mit der Beurteilung der Lehrkraft): Guter Unterricht könnte auch das Interesse am Thema fördern und in diesem Fall wäre Interesse nicht mehr als Bias zu interpretieren, sondern als Unterrichtsprodukt, weil Konfundierungen mit dem durch die Lehrkraft im Laufe des Unterrichts erweckten Interesse vorliegen könnten (vgl. SPIEL 2001a). Nur eine Korrelation zwischen –/+0,2 und 0 könnte eindeutig interpretiert werden, eine Verzerrung könnte hier auf Grund des geringen Zusammenhangs eher ausgeschlossen werden. Tatsächlich sinken die Korrelationen, wenn nach einer retrospektiven Einschätzung der Interessantheit des Themas vor Besuch der Lehrveranstaltung gefragt wird (vgl. RINDERMANN 2001). MARSH (1987) berichtet, dass in Hinblick auf die Konfundierungsproblematik durchgeführte Untersuchungen sehr hohe Korrelationen zwischen dem zu Beginn eines Kurses eingeschätzten Eingangsinteresse und dem zu Ende des Kurses eingeschätzten Eingangsinteresse fanden. Eine retrospektive Erhebung des Eingangsinteresses erscheint daher vertretbar.

Eine dem Interesse vergleichbare Konfundierungsproblematik mit Unterrichtsqualität liegt bei den Faktoren Sympathie für die Lehrkraft und Lernmotivation für den Unterrichtsgegenstand vor. Auch diese Faktoren können Unterrichtsprodukt sein und daher nicht Biasvariablen, die in keinem sachlich begründbaren Zusammenhang mit dem Unterricht stehen. Schließlich ist auch nicht eindeutig, wie der Einfluss der Noten oder der Notenerwartungen auf die Evaluierungsergebnisse zu interpretieren ist.

Noten und Notenerwartungen

„Grading perhaps generates the most suspicion about the validity of student evaluations“ (GIGLIOTTI/BUCHTEL 1990, 342). Deshalb zählen die tatsächlichen Noten, die von den Lernenden erwarteten Noten und die damit zusammenhängenede Milde der Lehrkraft bei der Beurteilung der Lernleistungen zu den am häufigsten untersuchten Einflussfaktoren auf die Evaluationen durch die Lernenden. Dabei wurde angenommen, dass bessere Noten generell oder Noten, die besser waren als die von den Lernenden erwarteten Noten, zu besseren Evaluationen führen würden.

Zunächst hatten Experimentalstudien wie jene von WORTHINGTON und WONG (1979) oder von SNYDER und CLAIR (1976) signifikante Auswirkungen von (manipulierten) Notenerwartungen und (wiederum manipulierten) tatsächlichen Noten auf die Evaluierungsergebnisse gezeigt. Diese und ähnliche Experimente wurden allerdings wiederholt kritisiert. Es ist plausibel anzunehmen, dass Lernende, die hohe Leistungen erbringen und sich deshalb eine sehr gute Note erwarten, jedoch eine schlechte bekommen sollen, sich anders verhalten als andere Personen, denen als zu erwartende Note eine Beurteilung genannt wird, die der eigenen Einschätzung entspricht oder sogar deutlich besser ist als die eigene Einschätzung. In der Untersuchung von SNYDER und CLAIR (1976) beispielsweise sahen die Probanden einen zehnminütigen Lehrvortrag, der auf Kassette aufgenommen worden war, mussten danach einen Test schreiben, ihre Testnoten wurden jedoch manipuliert. Nach Bekanntgabe der Noten mussten sie die Lehrkraft beurteilen. Es erstaunt m. E. nicht, dass solche Untersuchungen nicht nur vom ethischen Standpunkt, sondern auch auf der Grundlage methodischer Überlegungen als zweifelhaft angesehen worden sind. Weiters muss berücksichtigt werden, dass Ergebnisse aus Experimentalstudien nicht 1:1 auf nicht-experimentelle Settings wie Unterrichtssituationen und deren Beurteilung durch die Lernenden übertragen werden können (vgl. MARSH/ROCHE 1997).

In Korrelationsstudien wurden in der Regel Korrelationskoeffizienten im Bereich von r = 0,1 bis 0,3 festgestellt – die meisten davon statistisch signifikant –, wobei die Daten für gewöhnlich über mehrere Klassen oder Kurse hinweg zusammengefasst analysiert wurden (vgl. HOWARD/MAXWELL 1982, MARSH 1984, ders. 2001). Eine vergleichsweise hohe Korrelation von r = 0,44 fanden CENTRA und LINN in ihrer Studie (1973), der Großteil der anderen Untersuchungen berichtet von Korrelationen von etwa r = 0,2, Effekte, die FELDMAN (1976) als „small but not unimportant“ bezeichnet (S.□69, vgl. dazu auchFELDMAN 1997). Einige Studien verglichen die Korrelationen in den einzelnen Kursen bzw. Klassen und stellten große Unterschiede zwischen den Klassen fest, sodass in einzelnen Klassen auch noch etwas höhere Korrelationen (bis r = 0,57), dafür jedoch in anderen Klassen wesentliche geringere, bis 0 gehende Korrelationen ermittelt wurden.

GIGLIOTTI (1987) berichtet ebenso wie FELDMAN (1976) von geringen Effekten (wenn auch teilweise statistisch signifikant). GIGLIOTTI (1987) erhob in der ersten Einheit der Lehrveranstaltung die Erwartungen der Studierenden (u. a. hinsichtlich der Relevanz des Fachs, wie interessant der Stoff sein würde, des Lehrverhaltens der vortragenden Person und der Note) und verglich die Ergebnisse mit jenen einer Befragung derselben Studierenden am Ende des Kurses. In einer multiplen Regressionsanalyse konnte dadurch nicht nur der relative Einfluss der erwarteten

und der tatsächlichen Werte, sondern insbesondere auch der Einfluss der Diskrepanz zwischen den beiden untersucht werden. Die Ergebnisse von GIGLIOTTIS Regressionsmodell zeigen, dass die Erwartungen der Studierenden ihre Beurteilungen der Lehrkraft oder des Kurses praktisch nicht beeinflussen, „they account for none of the variance in intentions to take the professor or field again“ (S.□409f). Die tatsächliche Note erklärte etwa 2% der Varianz, ob die Studierenden einen weiteren Kurs in diesem Fach (Soziologie) belegen würden, die Diskrepanz zwischen tatsächlicher und erwarteter Note erklärte etwa 2% der Varianz, ob die Studierenden bei dieser Lehrkraft wieder einen Kurs belegen würden, und etwa 1% der Varianz im Interesse am Kurs (gemessen am Ende des Kurses). Die höchsten Betagewichte zur Erklärung der Varianz der Gesamtbeurteilung der Lehrkraft und der Lehrveranstaltung – z. B. ob die Studierenden wieder eine Lehrveranstaltung bei dieser Lehrkraft besuchen würden – wurden bei den am Ende der Lehrveranstaltung von den Studierenden eingeschätzten Lehrverhaltensfaktoren ermittelt.

Sehr hohe Korrelationen zwischen der erwarteten Note und den Lehrerevaluationen durch die Studierenden ermittelten GREENWALD (1996) bzw. GREENWALD und GILLMORE (1997) an der University of Washington. Drei unterschiedliche Stichproben ergaben standardisierte Pfadkoeffizienten von 0,48, 0,38 und 0,50.

Der Zusammenhang zwischen Note und Evaluationsergebnissen lässt prinzipiell zumindest drei Interpretationen zu, die jedoch ein ganz unterschiedliches Licht auf die Lehrerevaluationen durch die Lernenden und damit auch auf deren Aussagekraft 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, S.□1211). In diesem Fall könnte zweifellos berechtigterweise von einem Bias gesprochen werden.

Die nächsten beiden Interpretationsmöglichkeiten gehen davon aus, dass eine dritte Variable wirksam ist, und zwar entweder die Unterrichtsqualität selbst oder Merkmale der evaluierenden Lernenden (vgl. MARSH 1987, ders. 2001):

Die besseren erwarteten Noten könnten jedoch auch besonders gute Lernleistungen widerspiegeln, die wiederum das Ergebnis guten Unterrichts durch eine gute Lehrkraft sein könnten: „… strong instructors teach courses in which students both learn much (therefore, they earn and deserve high grades) and give appropriately high ratings to the course and to the instructor“ (GREENWALD/GILLMORE 1997, S.□1210). Diese Interpretationsmöglichkeit würde die so genannte Validitätshypothese („validity hypothesis“) unterstützen, dass gute Beurteilungen von Lehrkräften durch die Lernenden mit guten Lernleistungen zusammenhängen.

Eine dritte mögliche Erklärung stützt sich auf Charakteristika der beurteilenden Lernenden („student characteristics hypothesis“) 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, so-

dass der Effekt durch die erwartete Note ein Scheineffekt ist („the expected grade effect is spurious", MARSH 1987, S. 317). Weitere derartige potenzielle Einflussfaktoren wären die (Leistungs-)Motivation der Studierenden, die sich entweder auf alle Unterrichtsgegenstände (der Schullaufbahn oder des Studiums) beziehen kann oder auch nur auf einzelne spezifische Unterrichtsgegenstände (vgl. GREENWALD/GILLMORE 1997).

Die beiden zuletzt genannten Interpretationsmöglichkeiten gehen von der Analyseebene der Klasse aus. Sie sind nicht geeignet, die Korrelation zwischen guter erwarteter Note und Lehrerbeurteilungen durch die Lernenden innerhalb der Klasse zu erklären. GREENWALD und GILLMORE (1997) betonen dies in ihren Analysen, die den Zusammenhang auch innerhalb des Kurses bzw. der Klasse zeigen. Allerdings ließe die Korrelation innerhalb der Klasse noch 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, diese mit anspruchsvollem Unterricht besonders fördern, gleichzeitig aber schwächere Lernende damit überfordern, oder auch die Art und Qualität der Rückmeldung bestimmten Merkmalen der Lernenden anpassen.

Ein weiteres Argument, das GREENWALD und GILLMORE (1997) gegen die dritte o. a. Erklärung vorbringen, ist die negative Beziehung, die sie in ihren Stichproben zwischen den erwarteten Noten und dem Arbeitsaufwand („workload") festgestellt haben. Großes Interesse für ein Fach und hohe Motivation, für dieses Fach zu lernen, müssten jedoch in einer positiven Beziehung zwischen erwarteten Noten und Arbeitsaufwand resultieren. Dieses Argument ist jedoch m. E. nicht stichhaltig. Die „workload" wurde u. a. durch die Anzahl der Stunden abgebildet, die die Lernenden für das Fach aufwenden. Diese rein quantitative Größe sagt jedoch noch nichts über die Qualität des Lernens aus. Eine höhere Aufmerksamkeit und lernwirksamerer Unterricht im Kurs bedingen möglicherweise sogar, dass die Lernenden außerhalb des Kurses weniger Zeit mit Lernen verbringen müssen oder aber sich mit dem Fach beschäftigen, ohne dass das als „Lernzeit" wahrgenommen wird. Es erscheint durchaus plausibel, dass die meiste Lernzeit für Fächer aufgewendet wird oder werden muss, in denen man Schwierigkeiten hat, weniger begabt ist oder für die man sich weniger interessiert, damit man trotzdem positive Prüfungsleistungen erbringen kann (vgl. dazu auch MCKEACHIE 1997a, MARSH 2001).

Auch die Feststellung, dass die Korrelation zwischen erwarteter Note und Evaluation höher ist, wenn man nicht die erwartete Note als absoluten Wert ($r = 0{,}37$), sondern als relativen Wert – als erwartete Note im Verhältnis zu den in den anderen Fächern erwarteten Noten ($r = 0{,}53$) – heranzieht, ist kein ausreichender Hinweis auf grading leniency. Es bedeutet lediglich, dass die erwartete Note in einem Fach besser ist als in anderen Fächern, wofür die Milde der Lehrkraft bei der Beurteilung ein Grund von vielen möglichen Ursachen ist. Hier sind ebenso Konfundierungen mit Interesse, Motivation und selbst Unterrichtsqualität denkbar. Es ist daher zu überlegen, wie die Milde bei der Notengebung überhaupt erhoben werden kann. Kann die Milde direkt gemessen werden oder ist es möglich, auf bestimmte Weise aus der Note den Faktor Milde zu ermitteln (vgl. MARSH 2001)?

Die Problematik bei jedem Interpretationsversuch liegt darin, dass jede Note sowohl die individuellen Beurteilungsprinzipien der Lehrperson als auch die Schü-

lerleistung wiedergibt, die wiederum natürlich von einer Reihe von Schülermerkmalen wie beispielsweise dem Interesse an dem Unterrichtsgegenstand beeinflusst wird. Daher kann auch angenommen werden, dass alle drei bzw. vier Interpretationsmöglichkeiten zusammen den oftmals ermittelten Zusammenhang von guten erwarteten Noten und guter Lehrerevaluation durch die Lernenden erklären könnten. 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 in etwa ein Drittel des Effekts der erwarteten Note sich durch das Eingangsinteresse erklären ließ (vgl. auch HOWARD/MAXWELL 1982). Ein derartiges Zusammenwirken der einzelnen Faktoren hatte bereits FELDMAN (1976) vermutet. Die Wirkung von Interesse und Motivation auf die Beziehung zwischen Noten und Evaluationen wurde in weiteren empirischen Studien gezeigt, sie wird deutlicher kleiner (vgl. MARSH/ROCHE 2000) oder verschwindet ganz (vgl. MARSH 2001).

Die Fragen, ob die grading leniency hypothesis nun zutreffe oder nicht, sowie ob die Evaluierungen durch die Lernenden hauptsächlich von der Note beeinflusst werden, bleibt auf Grund der dargestellten 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). Die Ergebnisse der nachstehend noch näher erläuterten empirischen Untersuchung von GIGLIOTTI und BUCHTEL (1990) unterstützen die letztere Auffassung. Sie resümieren, dass „there is very little evidence to support the popular beliefs that easy or hard grading affect evaluations of instructors" (S. 350). Denn weder gaben Studierende, die eine bessere Note als erwartet bekommen hatten, bessere Beurteilungen ab, noch beurteilten jene Studierenden schlechter, die eine schlechtere Note bekommen hatten als erwartet. Insgesamt scheint noch immer zuzutreffen, was FELDMAN bereits 1976 festgestellt hat, dass nur auf Grund der positiven Beziehung zwischen Noten und Evaluationen „it cannot be said that grades tend to bias evaluation. But neither can it be concluded that they do not" (1976, S. 100).

Attribuierung von Erfolg und Misserfolg

Die Attributionsforschung hat gezeigt, dass die Art und Weise, wie sich die Menschen Ereignisse in ihrem Leben erklären, einen Einfluss auf deren Beurteilungen von Personen und Situationen haben. GIGLIOTTI und BUCHTEL (1990) untersuchten deshalb bei Undergraduate students, inwieweit die Attribuierung von Erfolg und Misserfolg die Lehrerevaluation beeinflussen könnte. Sie gingen dabei von der Annahme aus, dass viele Menschen dazu neigen, sich Erfolge selbst zuzuschreiben, Misserfolge jedoch external zu attribuieren, d. h. die Ursachen dafür außerhalb ihrer Person zu suchen. Ihre Ergebnisse zeigten, dass die Studierenden ihre Noten sowohl sich selbst als auch der Lehrperson und dem Kurs zuschrieben, und zwar unabhängig davon, ob die Note über oder unter dem

Klassendurchschnitt war. Die Lernenden nahmen also eine kombinierte Attribuierung vor und schrieben nicht – wie angenommen – die Erfolge gänzlich sich selbst zu und die Misserfolge gänzlich der Lehrperson. Darüber hinaus untersuchten sie, ob die Annahme zutraf, dass Abweichungen von der erwarteten Note (bzw. Leistung) external attribuiert wurden und im Falle von positiven Abweichungen – die Note war besser als die erwartete Leistung – zu besseren Evaluationen, im Falle von negativen Abweichungen zu schlechteren Evaluationen führten. Die Erwartungen wurden dazu zu Beginn des Kurses in der ersten Einheit erhoben, die tatsächlichen Noten, die Evaluierung der Lehrkräfte und weitere Daten am Ende des Kurses. Die Daten unterstützten diese Annahme jedoch nicht. Während die Attribuierungen kaum einen Effekt auf die Lehrerevaluationen hatten, ließ sich bei den Noten allerdings ein Einfluss auf die Evaluationen feststellen. Schlechte Noten resultierten auch in schlechteren Evaluationen. Doch auch als vergleichsweise stärkster Prädiktor vermochten die Noten nur etwa 3% der Varianz der Evaluationen zu erklären (vgl. die Ausführungen zum Einfluss der Note).

Klassengröße

Die überwiegende Mehrheit von Studien zur Untersuchung des Zusammenhangs zwischen der Klassengröße und den Lehrerbeurteilungen durch die Lernenden zeigten schwach negative Korrelationen. In kleineren Klassen wurden die Lehrenden demnach tendenziell etwas besser beurteilt als in größeren (vgl. FELDMAN 1997, OBIEKWE 1999). Etwas stärkere negative Effekte (in etwa $r = –0,3$) ließen sich bei Korrelationen mit spezifischen Lehrverhaltensitems feststellen, die sich auf die Interaktion der Lehrperson mit den Studierenden im Unterricht und die individuelle Unterstützung der Studierenden durch die Lehrperson bezogen. Dieser Zusammenhang ist jedoch sachlich begründbar und daher nicht automatisch als Bias zu betrachten. Je größer eine Klasse ist, umso schwieriger wird es für die Lehrperson, alle Lernenden in vergleichbarer Weise in den Unterricht miteinzubeziehen und sich individuell den einzelnen Lernenden zu widmen, sodass Lehrkräfte in größeren Klassen durchaus ein anderes Lehrverhalten zeigen dürften als in kleinen Klassen. Deshalb meint MARSH (1987), dass dieses Ergebnis nicht so sehr aus der Perspektive der Erforschung der Einflussfaktoren interessant ist, sondern vielmehr seine Validitätshypothese stärkt, da sich die Unterrichtsbedingungen in adäquater und logischer Weise in den Beurteilungen der Studierenden niederschlagen (vgl. dazu auch FELDMAN 1984, ders. 1979 und 1997). FELDMAN vermutete bereits, dass der Zusammenhang zwischen den Evaluationen und der Klassengröße nicht linear sei, was CENTRA (1979) mit seinen Forschungsergebnissen bestätigen konnte. Er fand heraus, dass jene Klassen bzw. Kurse mit 35 bis 100 Studierenden die vergleichsweise schlechteren Beurteilungen bekamen, jene mit weniger als 35 oder mehr als 100 Studierenden vergleichsweise bessere. Insgesamt spielt die Klassengröße jedoch für die Gesamtbeurteilung und für die meisten spezifischen Beurteilungsdimensionen, die nicht unmittelbar und direkt mit der Klassengröße zusammenhängen, nur eine sehr untergeordnete Rolle. Beim Vergleich von Evaluationsergebnissen verschiedener Kurse oder Klassen oder auch verschiedener Lehrkräfte ist jedenfalls darauf zu achten, dass ausschließlich vergleichbare Kurse dafür herangezogen werden, denn „while it may be somewhat „unfair“ to compare teachers in classes of widely different sizes, the unfairness lies

in the difference in teaching conditions, not in a rating bias“ (FELDMAN 1997, S.□372).

Schwierigkeitsgrad des Unterrichts

Entgegen den Erwartungen, dass die Lernenden geneigt wären, vor allem leichte, nicht so aufwendige Kurse besonders gut zu evaluieren, weisen die empirischen Befunde darauf hin, dass schwere, arbeitsaufwendige Lehrveranstaltungen besser beurteilt werden als leichte „Mickey Mouse courses“. Mehrere Untersuchungen zu dieser Fragestellung zeigen konsistente Ergebnisse (vgl. MARSH 1987). Auch RINDERMANN (2001, S.□196) bestätigt, dass der häufig befürchtete (negative) Zusammenhang zwischen (niedrigen) Anforderungen und (guten) Lehrevaluationen nicht durch die Empirie zu belegen ist. Das Herabsetzen des Anforderungsniveau trägt nicht dazu bei, sich gute Evaluierungsergebnisse zu „erschleichen“. Vielmehr ist es natürlich auch ein Merkmal guten Unterrichts, dass das Anforderungsniveau den Zielen, Inhalten und der Zielgruppe des Unterrichts angepasst ist.

Einfluss durch weitere Faktoren?

Höchst inkonsistente und meist auch sehr schwache Effekte sind bei den folgenden potenziellen Einflussfaktoren festgestellt worden (vgl. MARSH 1987, MARSH/DUNKIN 1997):

Der Rang der Lehrperson scheint keinen wesentlichen Einfluss auf die Beurteilungen zu haben. Zwar werden teaching assistants tendenziell schlechter beurteilt als die Angehörigen der faculty, innerhalb der Lehrenden einer faculty gibt es jedoch praktisch keine Effekte eines Zusammenhangs zwischen den Beurteilungen und dem Rang der Lehrperson.

Hinsichtlich des Alters der Lehrpersonen orteten manche Studien tendenziell bessere Beurteilungen für jüngere Lehrpersonen, ebenso viele Studien zeigen jedoch keinen Einfluss des Alters auf die Beurteilungen (vgl. auch MCKEACHIE 1979, RINDERMANN 2001).

Auch das Geschlecht der Beurteilenden sowie das der Beurteilten scheinen für gewöhnlich keinen Einfluss auf die Beurteilungen zu haben (vgl. RINDERMANN/AMELANG 1994, RINDERMANN 2001). In manchen Studien zeigten die Ergebnisse, dass Frauen geringfügig bessere Beurteilungen gaben als Männer. Dieser Effekt konnte jedoch auch in anderen Studien nicht repliziert werden. Ähnlich sind die Ergebnisse der Untersuchungen zum Einfluss des Geschlechts der Lehrkraft zu beurteilen. Wenn statistisch signifikante Ergebnisse gefunden wurden, waren sie so gering ($r = 0{,}02$), dass sie keine praktische Bedeutsamkeit besitzen (vgl. FELDMAN 1992, 1993 und 1997). Auch zu Wechselwirkungen zwischen dem Geschlecht der Beurteilenden und dem Geschlecht der Lehrperson konnten keine konsistenten Ergebnisse gefunden werden.

Der Unterrichtsgegenstand dürfte ebenfalls nur geringfügige Auswirkungen auf die Evaluationen der Studierenden haben. Vor allem naturwissenschaftliche Fächer, in denen hauptsächlich so genannte „hard facts“ vermittelt werden wie zum Beispiel in Mathematik, werden etwas schlechter beurteilt als andere Fächer wie

beispielsweise Fremdsprachen und Kunst (vgl. dazu auch FELDMAN 1997). Obwohl in manchen Untersuchungen signifikante Unterschiede ermittelt werden konnten, erklärten diese Unterschiede nicht einmal 1% der Varianz der Lehrerevaluationen durch die Studierenden.

Die Persönlichkeit der Lehrperson war Gegenstand von vergleichsweise wenig Untersuchungen. Von einer Reihe von Persönlichkeitsmerkmalen, welche die Lehrpersonen selbst einschätzen sollten, zeigten nur zwei Variablen eine auch praktisch bedeutsame signifikante Korrelation mit den Evaluationen: „positive self-regard, self-esteem" mit r = 0,3 und „energy and enthusiasm" mit r = 0,27. Anders war das Bild, wenn nicht die Lehrpersonen selbst die Persönlichkeitsmerkmale einschätzten, sondern die Studierenden oder Kolleg/inn/en. In diesem Fall stiegen die Korrelationskoeffizienten auf 0,3 bis 0,6 für nahezu alle eingeschätzten Persönlichkeitsmerkmale. Die Problematik dieser Ergebnisse liegt in der Konfundierung von Lehrverhaltenseinschätzungen, dem Rückschluss vom Lehrverhalten auf Persönlichkeitsmerkmale und der Einschätzung dieser Persönlichkeitsmerkmale. Darüber hinaus verlassen sich Kolleg/inn/en häufig auch auf das Urteil der Studierenden. Während die Lehrperson selbst alle Lebensbereiche in die Einschätzung der Persönlichkeitsmerkmale einbeziehen kann, stützen sich die Einschätzungen von Studierenden und Kolleg/inn/en hauptsächlich auf das Verhalten im beruflichen Leben (vgl. dazu auch FELDMAN 1986).

Zusammenfassende Analyse der Untersuchung von Biasvariablen

Insgesamt lässt sich für die Untersuchung von potenziell verzerrenden Variablen festhalten, dass bei den meisten Faktoren keine oder nur schwache Effekte ermittelt werden konnten (Geschlecht, Alter, Rang der Lehrkraft). Teilweise widersprechen die empirischen Ergebnisse den alltäglichen Vermutungen (schwere aufwendige Kurse werden eher besser beurteilt als leichte Kurse). Der Einfluss der Noten und Notenerwartungen ist schwer interpretierbar. Höhere Effekte werden bei Interesse und Motivation (Besuchsgrund) festgestellt. Bei diesen Faktoren ist nicht auszuschließen, dass Konfundierungen mit der Wirkung der Unterrichtsqualität vorliegen. Selbst wenn das Interesse vor Beginn des Unterrichts erhoben worden ist, könnten dennoch interessiertere Lernende der Lehrkraft die Unterrichtsgestaltung durch ihr Interesse erleichtern und deshalb mehr zur Unterrichtsqualität beitragen als weniger interessierte (vgl. WOLF et al. 2001).

Aus methodischer Sicht erscheint mir wichtig, dass verzerrende Variablen nicht nur isoliert, sondern mit einem multivariaten Verfahren in einer Gesamtschau untersucht werden. Dabei ist vor allem die Frage interessant, wie stark potenzielle Biasvariablen (noch) auf die Gesamtbeurteilung einer Lehrkraft und ihres Unterrichts wirken, wenn die Faktoren in das multivariate Verfahren miteinbezogen werden, die diese Gesamtbeurteilung ausmachen sollten, nämlich jene Faktoren, die das Lehrverhalten der Lehrkraft im Unterricht abbilden (vgl. GREIMEL/GEYER 2001, DIES. 2002).

Auch wenn es nicht möglich ist, alle eine Beurteilung möglicherweise beeinflussenden Faktoren (vgl. KANNING 1999) in einer empirischen Untersuchung zu berücksichtigen, stellt sich dennoch vor allem hinsichtlich der Skepsis vieler Lehrkräfte gegenüber der Evaluation durch die Lernenden die Frage, ob nicht doch noch (naheliegende) entscheidende Faktoren bislang in der Forschung

unberücksichtigt geblieben sind (vgl. dazu auch WACHTEL 1998). Die Haltung der Lehrkräfte gegenüber der Lehrerevaluation und deren Wirkung auf die Beurteilungen durch die beurteilenden Lernenden wurden bislang kaum untersucht. Dabei erscheint es durchaus plausibel anzunehmen, dass die Einstellung der Lehrkräfte zur Evaluation von den Lernenden wahrgenommen wird und sich auf den Beurteilungsprozess auswirken könnte. Manche Lehrkräfte „display their own skepticism or lack of interest in student evaluations" (WAGENAAR 1995, S.□66). Diese Haltung könnte die Beurteilenden einerseits einschüchtern, verunsichern oder auch verärgern. Es ist jedoch für den Evaluierungsprozess wichtig, dass sowohl die Evaluierenden als auch die Evaluierten die Evaluierung ernstnehmen, Feedback geben bzw. nehmen wollen und „both need to have confidence in the methods used" (MCKEACHIE 1979, S.□397). Noch weniger wurde die Einstellung der Evaluierenden berücksichtigt. Dabei zeigte sich in qualitativen und quantitativen empirischen Untersuchungen bei österreichischen Schüler/innen an Handelsakademien, dass etwa ein Drittel der Befragten Zweifel hat, ob Lehrerbeurteilungen durch die Schüler/innen sinnvoll sind. Die Befragten begründen ihren Zweifel einerseits mit der Vermutung, dass die Lehrkräfte ihr Lehrverhalten nicht auf Grund der Evaluationen ändern würden. Andererseits bezweifeln vor allem die jüngeren Befragten, dass Schüler/innen in der Lage sind, ihre Lehrkräfte gerecht zu beurteilen (vgl. GREIMEL 2002). In diesem Zusammenhang stellt sich die Frage, ob Einstellungen der Beurteilenden wie beispielsweise ihr Zweifel einen Einfluss auf die Beurteilungen haben könnten (vgl. GREIMEL/GEYER 2001, DIES. 2002).

5□Abschließende Betrachtung und Ausblick

Die Fülle an empirischen Befunden spricht im Wesentlichen für die Durchführung von Lehrerevaluationen durch die Lernenden. Trotz mancher widersprüchlicher Ergebnisse lassen sich in der Regel positive Zusammenhänge zwischen Lehrerevaluation und Lernerfolgen der Lernenden feststellen. Außerdem sind die Effekte der meisten vermuteten Biasvariablen schwach. Interesse, Motivation und Sympathie, bei denen stärkere Zusammenhänge gefunden werden, könnten Unterrichtsprodukte sein und würden dann keine Verzerrung der Beurteilungen bedeuten. Doch selbst wenn die Beurteilungen möglicherweise durch manche – auch verzerrende – Faktoren beeinflusst sein könnten, ist ihr Nutzen als Feedback an die Lehrkraft zur Weiterentwicklung von Unterricht sowie ihre Aussagekraft über die Unterrichtskraft deswegen noch nicht grundsätzlich in Frage zu stellen. Denn „die Besorgtheit über Nebenwirkungen ist (…) kein wissenschaftlich legitimes Immunisierungsargument gegen eine sorgfältige Nutzung der Evaluation" (WEINERT 2001, S.□4).

RINDERMANN (2001) ist m. E. recht zu geben, wenn er meint, dass die Verwendung von Evaluationsergebnissen in vielen Fällen problematischer ist als die Evaluation selbst. Dem entspricht die Feststellung WEINERTS (2001, S.□4), dass „der Nutzen von Evaluation (…) von der sachgerechten Nutzung der Evaluationsdaten abhängt". Ein Vergleich von Evaluationsergebnissen verschiedener Lehrkräfte verschiedener Disziplinen, die unter verschiedenen Unterrichtsbedingungen unterschiedliche Studierende lehren, – möglicherweise in Form von Rankings – ist tatsächlich fragwürdig (vgl. dazu auch KRIZ 1994, SIMPSON 1995). MCKEACHIE (1997, S.□402) stellt fest, dass die Tatsache, dass die meisten Evaluationsergeb-

nisse in Zahlen vorliegen, diese Form der Verwendung fördert: „Not that qualities of teaching are not quantifiable. Numbers are often useful. The fault is not in the numbers, but rather in their use. Once numbers are assigned, faculty (…) begin to make comparisons between teachers and assume that if one number is larger than another, there is a real difference between the teachers to whom the numbers have been assigned". Die Interpretation und die Verwendung von Evaluationsergebnissen muss demnach reflektiert und wohl überlegt passieren.

Die Wahl eines für die jeweilige Unterrichtsform passenden Evaluationsinstrumentes ist für die ausführliche Vorbereitung der Evaluation ebenso bedeutend wie die eingehende Information der Lernenden und deren Motivierung, an der Evaluation teilzunehmen und die Beurteilungen überlegt und gewissenhaft durchzuführen. Zweifellos sollten jedoch für ein umfassendes Bild zur Beurteilung einer Lehrkraft auch nach andere Beurteilungsquellen herangezogen werden, und zwar nicht, weil die Beurteilungen der Lernenden eine schlechte Quelle wären, sondern weil sie für ein allumfassendes Bild der Leistungen einer Lehrkraft eine wertvolle, aber nicht ausreichende, weil „notwendigerweise einseitige" (WEINERT 2001, S. 9) Quelle darstellen. Denn „no single source of data, including student rating data, provides sufficient information to make a valid judgment about teaching effectiveness" (CASHIN 1988, S. 1).

Für die empirische Unterrichtsevaluationsforschung ist die Weiterentwicklung von Strukturgleichungsmodellen bedeutend, die die komplexen Zusammenhänge zwischen einer Reihe von Lehrverhaltensvariablen, Gesamtbeurteilungen und potenziellen Biasvariablen adäquat prüfen und darstellen können.

Das wichtigste Argument für die Lehrerevaluation durch die Lernenden ist m. E. jedoch darin zu sehen, dass alle Lernenden sich ein Bild von ihren Lehrkräften machen. Und jede Lehrkraft sollte im Sinne ihrer Professionalität daran interessiert sein, dieses Bild zu kennen und bei der Weiterentwicklung ihres Unterrichts darüber zu reflektieren und zu berücksichtigen.

Literaturverzeichnis

Abrami, Philip C. / d'Apollonia, Sylvia / Rosenfield, Steven (1997): The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not. In: Perry, Raymond P. / Smart, John C. (Hrsg.): Effective Teaching in Higher Education: Research and Practice. Agathon Press, New York, Seite 321–367

Abrami, Philip / Cohen, Peter / d'Apollonia, Sylvia (1988): Implementation Problems in Meta-Analysis. In: Review of Educational Research, Vol. 58, Seite 151–179

Abrami, Philip / Leventhal, Les / Perry, Raymond (1982): Educational Seduction. In: Review of Educational Research, Vol. 52, Seite 446–464

d'Apollonia, Sylvia / Abrami, Philip (1997): Navigating Student Ratings of Instruction. In: American Psychologist, Vol. 52, Seite 1198–1208

Berliner, David (1987): Simple Views of Effective Teaching and a Simple Theory of Classroom Instruction. In: Berliner, David / Rosenshine, Barak: Talks to Teachers. Random House, New York, Seite 93–110

Bortz, Jürgen / Döring, Nicola (1995): Forschungsmethoden und Evaluation. Springer Verlag, Berlin u. a.

Brophy, Jere (2000): Teaching. Educational Practices Series-1, International Academy of Education & International Bureau of Education, Brüssel

Brophy, Jere / Good, Thomas (1986): Teacher Behavior and Student Achievement. In

Wittrock, M.: Handbook of Research on Teaching. Macmillan Publishing Company, New York, Seite 328–375
Cashin, William E. (1988): Student Ratings of Teaching: A Summary of the Research. IDEA Paper No. 20, Center for Faculty Evaluation & Development, Division of Continuing Education, Kansas State University
Cashin, William E. (1995): Student Ratings of Teaching: the Research Revisited. Center for Faculty Evaluation and Development, IDEA No. 32, Manhattan
Centra, John A. (1977): Student Ratings of Instruction and Their Relationship to Student Learning. In: American Educational Research Journal, Winter 1977, Vol. 14, Seite 17–24
Centra, John A. (1979): Determining Faculty Effectiveness. Jossey Bass, San Francisco
Centra, John A. / Linn, R. L. (1973): Student Points of View in Ratings of College Instruction. Research Bulletin RB-73-60, Educational Testing Service, Princeton N.J.
Cohen, Peter A. (1981): Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. In: Review of Educational Research, Fall 1981, Vol. 51, Seite 281–309
Cohen, Peter A. (1987): A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis. ERIC paper ED283876, Paperpräsentation im Rahmen des Annual Meeting der American Educational Research Association, Washington DC
Dubs, Rolf (1998): Qualitätsmanagement für Schulen. Herausgegeben vom Institut für Wirtschaftspädagogik der Universität St. Gallen
Dubs, Rolf (2000): Lehrerbeurteilung und Lehrerqualifikationen. Ein vielschichtiges Postulat und sein Königsweg. In: Neue Zürcher Zeitung vom 19. September 2000
Feldman, Kenneth A. (1976): Grades and College Students' Evaluations of their Courses and Teachers. In: Research in Higher Education, Vol. 4, Seite 69–111
Feldman, Kenneth A. (1979): The Significance of Circumstances for College Students' Ratings of Their Teachers and Courses. In: Research in Higher Education, Vol. 10, Seite 149–172
Feldman, Kenneth A. (1984): Class Size and Student Evaluations of College Teacher and Courses: A Closer Look. In: Research in Higher Education, Vol. 21, Seite 45–116
Feldman, Kenneth A. (1986): The Perceived Instructional Effectiveness of College Teachers as Related to Their Personality and Attitudinal Characteristics: A Review and Synthesis. In: Research in Higher Education, Vol. 24, Seite 139–213
Feldman, Kenneth A. (1989): The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data From Multisection Validity Studies. In: Research in Higher Education, Vol. 30, Seite 583–645
Feldman, Kenneth A. (1989a): Instructional Effectiveness of College Teachers as Judged by Teachers themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers. In: Research in Higher Education, Vol. 30, Seite 137–194
Feldman, Kenneth A. (1992): College Students' Views of Male and Female College Teachers: Part I Evidence from the Social Laboratory and Experiments. In: Research in Higher Education, Vol. 33, Seite 317–375
Feldman, Kenneth A. (1993): College Students' Views of Male and Female College Teachers: Part II Evidence from Students' Evaluations of Their Classroom Teachers. In: Research in Higher Education, Vol. 34, Seite 151–211
Feldman, Kenneth A. (1997): Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings. In: Perry, Raymond P. / Smart, John C. (Hrsg.): Effective Teaching in Higher Education: Research and Practice. Agathon Press, New York, Seite 368–395
Gigliotti, Richard J. (1987): Expectations, Observations, and Violations: Comparing their Effects on Course Ratings. In: Research in Higher Education, Vol. 26, Seite 401–415

Gigliotti, Richard J. / Buchtel, Foster S.□(1990): Attributional Bias and Course Evaluation. In: Journal of Educational Psychology, Vol.□82, Seite 341–351

Greenwald, Anthony G. (1996): Applying Social Psychology to Reveal a Major (But Correctable) Flaw in Student Evaluations of Teaching. Paperpräsentation im Rahmen der Annual Meeting of the American Psychological Association (EDRS ED 400 754), New York 1995

Greenwald, Anthony G. / Gillmore, Gerald (1997): Grading Leniency Is a Removable Contaminant of Student Ratings. In: American Psychologist, Vol.□52, Seite 1209–1217

Greenwald, Anthony G. (1997): Validity concerns and Usefulness of Student Ratings of Instruction. In: American Psychologist, Vol.□52, Seite 1182–1186

Greimel, Bettina (1996): Ein Anforderungsprofil an Lehrer. In: Fortmüller, Richard / Aff, Josef (Hrsg.): Wissenschaftsorientierung und Praxisbezug in der Didaktik der Ökonomie. Festschrift Wilfried Schneider. Manz Verlag, Wien, Seite 229–254

Greimel, Bettina (2002): Evaluation von Lehrkräften aus Schülersicht – Relevante Kriterien und Einstellung zum Evaluationsprozess. In: Erziehung und Unterricht, im Erscheinen

Greimel, Bettina / Geyer, Alois (2001): Zum Einfluss des Zweifels der Lernenden auf die Lehrerevaluation. In Moosbrugger, Helfried et at. (Hrsg.): Methoden und Evaluation. Tagungsband zur 5. Fachgruppentagung an der Johann Wolfgang Goethe-Universität Frankfurt am Main 2001, Seite 34

Greimel, Bettina / Geyer, Alois (2002): Does a Student's Attitude Towards Teacher Evaluation Influence Global Ratings? In: Klein, Hans E. (Hrsg.): Creative Teaching ACT5. Selected Papers of the Fifth International Conference on Creative Teaching. Madison, USA 2002, im Erscheinen

Grubitzsch, Siegfried / Rexilius, Günter (1985): Testtheorie – Testpraxis. Voraussetzungen, Verfahren, Formen und Anwendungsmöglichkeiten psychologischer Tests im kritischen Überblick. Rowohlt, Reinbek bei Hamburg

Helmke, Andreas / Weinert, Franz (1997): Bedingungsfaktoren schulischer Leistungen. In: Psychologie des Unterrichts und der Schule. Enzyklopädie der Psychologie, 3, Hogrefe Verlag für Psychologie, Göttingen u. a., Seite 71–176

Howard, George S.□/ Maxwell, Scott E. (1982): Do Grades Contaminate Student Evaluations of Instruction. In: Research in Higher Education, Vol.□16, Seite 175–188

Howard, George S.□/ Conway, Christine G. / Maxwell, Scott E. (1985): Construct Validity of Measures of College Teaching Effectiveness. In: Journal of Educational Psychology, Vol.□77, Seite 187–196

Kanning, Uwe Peter (1999): Die Psychologie der Personenbeurteilung. Hogrefe Verlag für Psychologie, Göttingen u. a.

Kremer-Hayon, Lya (1993): Teacher Self-Evaluation. Teachers in Their Own Mirrors. Kluwer Academic Publishers, Boston u. a.

Krempkow, René (1998): Ist „gute Lehre" messbar? In: Das Hochschulwesen 4/98, Seite 195–199

Kriz, Jürgen (1994): Die Wirklichkeit von (Vor-)Urteilen. In: In: Mohler, Peter (Hrsg.): Universität und Lehre: ihre Evaluation als Herausforderung an die empirische Sozialforschung. Waxmann, Münster, New York, Seite 11–28

Kromrey, Helmut (1994): Evaluation der Lehre durch Umfrageforschung? In: Mohler, Peter (Hrsg.): Universität und Lehre: ihre Evaluation als Herausforderung an die empirische Sozialforschung. Waxmann, Münster, New York, Seite 91–114

Lienert, Gustav A. / Raatz, U. (1998): Testaufbau und Testanalyse. Beltz Psychologie VerlagsUnion, Weinheim

Marsh, Herbert W. (1984): Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. In: Journal of Educational Psychology, Vol.□76, Seite 707–754

Marsh, Herbert W. (1987): Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. International Journal of Educational Research, Vol.11, Seite 253–387

Marsh, Herbert W. (2001): Distinguishing Between Good (Useful) and Bad Workloads on Students' Evaluations of Teaching. In: American Educational Research Journal, Vol. 38, Seite 183–212

Marsh, Herbert W. / Dunkin, Michael J. (1997): Students' Evaluations of University Teaching: A Multidimensional Perspective. In: Perry, Raymond P. / Smart, John C. (Hrsg.): Effective Teaching in Higher Education: Research and Practice. Agathon Press, New York, Seite 241–319

Marsh, Herbert W. / Overall, J. U. (1980): Validity of Students' Evaluations of Teaching Effectiveness: Cognitive and Affective Criteria. In: Journal of Educational Psychology, Vol. 72, Seite 468–475

Marsh, Herbert W. / Roche, Lawrence (1997): Making Students' Evaluations of Teaching Effectiveness Effective. The Critical Issues of Validity, Bias, and Utility. In: American Psychologist, Vol. 52, No. 11, Seite 1187–1197

Marsh, Herbret W. / Roche, Lawrence (2000): Effects of Grading Leniency and Low Workloads on Students' Evaluations of Teaching: Popular Myth, Bias, Validity or Innocent Bystanders? In: Journal of Educational Psychology, Vol. 92, Seite 202–208

Marsh, Herbert W. / Ware, John E. (1982): Effects of Expressiveness, Content Coverage, and Incentive on Multidimensional Student Rating Scales: New Interpretations of the Dr. Fox Effect. In: Journal of Educational Psychology, Vol. 74, Seite 126–134

McKeachie, Wilbert J. (1979): Student Ratings of Faculty: A Reprise. In: Academe, October 1979, Seite 384–397

McKeachie, Wilbert J. (1997): Good Teaching Makes a Difference – And We Know What It Is. In: Perry, Raymond P. / Smart, John C. (Hrsg.): Effective Teaching in Higher Education: Research and Practice. Agathon Press, New York, Seite, 396–407

McKeachie, Wilbert J. (1997a): Student Ratings – The Validity of Use. In: American Psychologist, Vol. 52, Seite 1218–1225

Mummendey, Hans Dieter (1995): Die Fragebogen – Methode. Hogrefe Verlag für Psychologie, Göttingen u. a.

Naftulin, Donald H. / Ware, John E. / Donnelly, Frank A. (1973): The Doctor Fox Lecture: A Paradigm of Educational Seduction. In: Journal of Medical Education, Vol. 48, Seite 630–635

Nasser, Fadia / Glassman, David (1997): Student Evaluation of University Teaching: Structure and Relationship with Student Characteristics. Paperpräsentation im Rahmen des Annual Meeting der American Educational Research Association in Chicago

Obiekwe, Jerry C. (1999): The Multidimensional Character of Teaching Effectiveness: A Comparative Analysis of Student Evaluation Responses of Full and Part-time Faculty. ERIC paper ED 443 336, Paperpräsentation im Rahmen der Annual Conference of the Mid-Western Educational Research Association, Chicago

Rindermann, Heiner / Amelang, Manfred (1994): Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE). Roland Asanger Verlag Heidelberg

Rindermann, Heiner (1999): Die studentische Beurteilung von Lehrveranstaltungen – Forschungsstand und Implikationen. Vortragspapier auf dem Symposium Evaluierung an der Universität – Zwischen Qualitätsmanagement und Selbstzweck am 2.12.1999

Rindermann, Heiner (2001): Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. Verlag Empirische Pädagogik, Landau

Schneider, Wilfried (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. Verlag des Schweizerischen Kaufmännischen Verbandes, Zürich, Seite 357–373

Simpson, Ronald D. (1995): Uses and Misuses of Student Evaluations of Teaching Effectiveness. In: Innovative Higher Education, Vol.□20, Seite 3–5

Snyder, C.R. / Clair, Mark (1976): Effects of Expected and Obtained Grades on Teacher Evaluation and Attribution of Performance. In Journal of Educational Psychology, Vol.□68, Seite 75–82

Spiel, Christiane (Hrsg., 2001): Evaluation universitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck. Waxmann Münster u. a.

Spiel, Christiane (2001a): Der differentielle Einfluss von Biasvariablen auf studentische Lehrveranstaltungsbewertungen. In: Engel, Uwe (Hrsg.): Hochschulranking. Zur Qualitätsbewertung von Studium und Lehre, Campus, Frankfurt/Main, Seite 61–82

Spiel, Christiane / Gössler, P. Martin (2000): Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. In: Zeitschrift für Pädagogische Psychologie, Vol.□14, Seite 38–47

Tagomori, Harry T. / Bishop, Laurence A. (1995): Student Evaluation of Teaching: Flaws In the Instruments. In: Thought and Action, The NEA Higher Education Journal, Spring 1995, Seite 63–78

Wachtel, Howard (1998): Student Evaluation of College Teaching Effectiveness: A Brief Review. In: Assessment & Evaluation in Higher Education, Vol.□23, Seite 191–211

Wagenaar, Theodore C. (1995): Student Evaluation of Teaching: Some Cautions and Suggestions. In: Teaching Sociology, Vol.□23, Seite 64–68

Weinert, Franz E. (1998): Guter Unterricht ist ein Unterricht, in dem mehr gelernt als gelehrt wird. In: Freund, Josef / Gruber, Heinz / Weidinger, Walter (Hrsg. 1998): Guter Unterricht – Was ist das? Aspekte von Unterrichtsqualität. ÖBV Pädagogischer Verlag, Wien, Seite 7–18

Weinert, Franz E. (2001): Die evaluierte Universität. Manuskript zur Heidelberger Universitätsrede an der Ruprecht-Karls-Universität in Heidelberg am 25. Jänner 2001.

Weinert, Franz E. / Helmke, Andreas (1996): Der gute Lehrer: Person, Funktion oder Fiktion? In: Leschinsky, Achim (Hrsg.): Die Institutionalisierung von Lehren und Lernen. Beiträge zu einer Theorie der Schule. Zeitschrift für Pädagogik, 34. Beiheft, Beltz Verlag, Weinheim und Basel, Seite 223–233

Wolf, Patrick / Spiel, Christiane / Pellert, Ada (2001): Entwicklung eines Fragebogens zur globalen Lehrveranstaltungsevaluation – ein Balanceakt zwischen theoretischem Anspruch, Praktikabilität und Akzeptanz. In: Spiel (Hrsg.): Evaluation unversitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck, Waxmann, Münster u. a., Seite 89–109

Worthington, Alan G. / Wong, Paul T.P. (1979): Effects of Earned and Assigned Grades on Students Evaluations of an Instructor. In Journal of Educational Psychology, Vol.□71, Seite 764–775

Anschrift der Autorin: Dr. Bettina Greimel, Augasse 2–6, A-1090 Wien (Wirtschaftsuniversität Wien, Abteilung für Wirtschaftspädagogik)