Lässt sich die Lehrqualität durch Evaluation und Beratung verbessern?*

Überprüfung eines Evaluations-Beratungs-Modells

Heiner Rindermann
Universität Magdeburg

Jürgen Kohler
Hochschule für Heilpädagogik Zürich

Does Evaluation and Consulting Improve Quality of Instruction? Test of an Evaluation-Consulting-Model

Summary: One important aim of the evaluation of instruction is the improvement of instruction. International as well as German studies, however, have provided little evidence that student evaluation of instruction alone would change the quality of teaching. It appears that mere feedback of evaluation results is not effective. The present study aimed at improving the effects of evaluation by combining evaluation feedback with intensive personal consultation of the teachers by experts. The study was conducted in a private school for logopedia (speech therapy) with 16 teachers in 35 classes with about 20 students per class. Evaluation was conducted twice (pre- and post intervention) within a period of six months using a standardized set of questionnaires (HILVE-II) developed for evaluation of university courses. The overall intervention effect on teaching quality as measured by *d* (effect size for dependent means) was more than half a standard deviation in most scales. Quality of teaching in the school not only improved quantitatively (measured in the scales) and qualitatively (verbal student evaluations) but also became more homogeneous in quality (reduction of standard deviation) although the relative rank order of teachers did not change appreciably.

Keywords: Student evaluation of instruction, consultation, quality of instruction, improvement of instruction, effect size *d* for dependent means

Zusammenfassung: Ein wichtiges Ziel der Lehrevaluation ist die Verbesserung der Lehre. In den meisten internationalen und deutschen Studien ließ sich ein Veränderungseffekt auf Grund von Lehrveranstaltungsevaluation aber nicht nachweisen. Reine Feedbackansätze zur Modifikation von Lehrerhandeln genügen nicht. In einer Studie an einem privaten Ausbildungsinstitut für Logopädie ($N = 16$ Dozenten in 35 Kursen mit jeweils etwa 20 Schülern) wurde deshalb ein Beratungs-Rückmelde-Verfahren entwickelt und seine Effektivität überprüft. Evaluationen wurden mittels des HILVE-II zweimal mit anschließender Beratung innerhalb von sechs Monaten durchgeführt. Es zeigten sich Verbesserungseffekte größer als eine halbe Standardabweichung (*d* für abhängige Stichproben) im studentischen Urteil. Verbesserte Lehrqualität war nicht nur quantitativ bestimmbar, sondern ließ sich auch qualitativ den studentischen Kommentaren entnehmen. Zudem wurden die Einschätzungen homogener, d. h. Lehrende erreichten ein einheitliches günstiges Niveau, auch wenn die relativen Positionen im studentischen Urteil weitgehend stabil blieben.

Schlüsselbegriffe: Lehrevaluation, Beratung, Lehrqualität, Verbesserung der Lehre, Effektgröße *d* bei abhängigen Stichproben

* **Danksagung:** Wir möchten allen Schülern und Ausbildern an der Schule für Logopädie für ihre Bereitschaft an der Teilnahme der Evaluation danken. Ebenso sind wir mehreren Mitarbeitern der vergangenen fünf Jahre in München und Magdeburg Dank schuldig, die Daten eingaben, auswerteten und Rückmeldungen für die Beratung erstellten. Für wertvolle Anregungen zur Lösung des Effektgrößenproblems sind wir Herrn Prof. Dr. Jürgen Bortz und Frau Dipl.-Math. Dipl.-Psych. Hella Klemmert (beide Berlin) verpflichtet.

Psychologie in Erziehung und Unterricht, 2003, *50,* 71–85
© Ernst Reinhardt Verlag München Basel

Die Einführung von Lehrbeurteilungen an Hochschulen setzt stillschweigend voraus, dass Evaluation einen Beitrag zur Qualitätssicherung in der universitären Ausbildung leistet. Evaluationen sollen sich durch praktische Nützlichkeit ausweisen. Es wurden aber bislang im deutschen Sprachraum – soweit uns bekannt – außer mit dem HILVE (Heidelberger Inventar zur Lehrveranstaltungsevaluation) keine Wirksamkeitsuntersuchungen zu Lehrevaluationen durchgeführt. Es liegen zwar eine Menge anekdotischer Hinweise vor (Übersicht in Rindermann, 2001), systematische Untersuchungen mit Messwiederholungen fehlen jedoch. Die Frage bleibt so offen, ob Evaluation den erwünschten Effekt bewirken kann, die Lehrqualität zu verbessern. Ein Beispiel für dieses Forschungsdefizit ist Teichmanns (1999) Arbeit zur „Wirksamkeit der Evaluation von Studium und Lehre": Außer einer Beschreibung des Vorgehens der Evaluation im Verbund norddeutscher Universitäten findet sich im Text die Behauptung, dass „nachweisliche Verbesserungen ... festzumachen sind" (S. 247) – eine Aussage, die empirisch aber nirgendwo belegt wird. Die Autorin fordert für zukünftige Evaluationen, dass „systematische Ergebnissicherungen eingeleitet werden ..., es ist noch nachdrücklicher zu fragen, welche Erfolge sich eingestellt haben" (S. 247) – eine Forderung, der unbedingt entsprechende Untersuchungen folgen sollten (s. a. Hackl & Sedlacek, 2001).

In verschiedenen Ansätzen von Lehrveranstaltungsevaluation wird eine Wirksamkeit ihrer Einführung angenommen: Lehrevaluation solle etwa durch den bloßen Einsatz als Intervention wirken (Will & Blickhan, 1987; Henninger, 1999), das Qualitätsbewusstsein fördern und somit Lehre verbessern *(Sensibilisierungs-Hypothese)*. Im *Feedbackmodell* wird dem Dozenten zusätzlich eine individuelle Rückmeldung seiner Ergebnisse angeboten. Dieses Feedback informiert über spezifische Schwächen und Stärken der Veranstaltung (Informationsfunktion) und motiviert (vor allem bei Ist-Soll-Diskrepanzen) zu Verbesserungen. Zum Ausgleich solcher Diskrepanzen würden Dozenten versuchen, ihre Lehre zu verbessern (vgl. Marsh, 1987, S. 342ff.; Balk, 2000). In einer Spezifikation der Feedback-Hypothese wird betont, dass das Feedback nur bei behavioralen und verhaltensnahen Informationen veränderungswirksam sei.

Angloamerikanische und australische Untersuchungen sowie HILVE-Studien an deutschsprachigen Universitäten, in denen nur Evaluationen durchgeführt (ohne Feedback) oder Ergebnisse ohne spezifische Diskussionsmethoden und Beratungs- und Weiterbildungsangebote rückgemeldet bzw. solche Varianten im Vergleich mit anderen Verfahren erprobt wurden, zeigen, dass entweder kein Effekt durch bloße Evaluationsdurchführung oder nur ein vergleichsweise geringer Verbesserungseffekt durch zusätzliches Feedback zu erreichen ist (ausf. Rindermann, 2001).

Auch Berichte von deutschen Lehrevaluationsprojekten unterstützen die Resultate, die unter kontrollierten Forschungsbedingungen bei Fehlen von Beratung und Weiterbildung gefunden wurden: Veränderungen sind meist nicht beobachtbar, Verbesserungen bleiben anekdotisch. Dozenten ignorieren häufig die Resultate, Feedback löst keine (günstige) Verhaltensänderung aus. Und negatives Feedback als solches stellt keine Motivation zur Verbesserung dar (McKeachie, 1997). Webler (1992) hält deshalb die Ergänzung durch eine veranstaltungsinterne Besprechung zwischen Studierenden und Lehrenden für notwendig. In einer solchen Besprechung sollten die Resultate berichtet, neue Informationen über das Feedback hinaus gewonnen sowie mögliche Veränderungen erörtert und angestoßen werden *(hochschuldidaktisches Diskursmodell)*. Doch auch die Ergebnisse hinsichtlich des (unmoderierten) hochschuldidaktischen Diskursmodells stimmen skeptisch. So spricht Webler (1996) von „Versandungsprozessen", ausgelöst durch großen Widerstand oder Gleichgültigkeit, meist würden keine Veränderungen auftreten und falls Änderungen beobachtbar seien, träten diese bei

den besseren Dozenten auf. Ein weiteres Problem dieses Ansatzes liegt darin, dass die Umsetzung der Besprechung der Resultate zwischen Dozent und Veranstaltungsteilnehmern nicht gewährleistet ist. Nach einem Vorschlag von Alean-Kirkpatrick, Hänni und Lutz (1997) sollte diese Besprechung deshalb durch Hochschuldidaktiker moderiert werden, um zu vermeiden, dass nur Ergebnisse ohne Diskussion präsentiert werden (s. a. Winter, 2000).

In Messwiederholungsstudien hat eine Osnabrücker Forschergruppe um Gediga (2000) das hochschuldidaktische Diskursmodell auf seine Wirksamkeit hin systematisch überprüft. Bei Messung Mitte und Ende des Semesters in sechs Veranstaltungen (95 Studierende) zeigten sich keine positiven Veränderungen in Lehrveranstaltungsevaluationen. „Massive Verbesserungen" traten in nur einer Veranstaltung auf, die zum ersten Messzeitpunkt als „sehr schlecht" beurteilt wurde und in der „ein Mitarbeiter des Projektes KIEL [Kommunikations-Instrument für die Evaluation von Lehrveranstaltungen] als Hochschuldidaktiker eingesprungen war, um bei der Problemlösung innerhalb der Veranstaltung mitzuwirken" (S. 62, 77).

Auch in drei Studien am Heidelberger Psychologischen Institut, an der Sozialwissenschaftlichen Fakultät sowie in den Fächern Romanistik und Medizin wurde in verschiedenen Varianten die Effektivität studentischer Lehrevaluation überprüft:

1. Evaluation Mitte und Ende der Vorlesungszeit in der gleichen Veranstaltung, Rückmeldung erst nach dem zweiten Messzeitpunkt – *Sensibilisierungshypothese.*
2. Evaluation Mitte und Ende der Vorlesungszeit in der gleichen Veranstaltung, Rückmeldung nach dem ersten und zweiten Messzeitpunkt – *Feedbackhypothese.*
3. Evaluation Mitte und Ende der Vorlesungszeit in der gleichen Veranstaltung, Rückmeldung nach dem ersten und zweiten Messzeitpunkt, Aufforderung zur Besprechung des ersten Feedbacks mit den Studierenden in der Veranstaltung ohne externe Unterstützung oder Beratung, Vergleich der Veranstaltungen mit und ohne Besprechung der Resultate (eigene Entscheidung der Dozenten) – *hochschuldidaktisches Diskursmodell.*
4. Schließlich noch ein Vergleich über die drei Studien hinweg: *mittelfristige Veränderungen* über verschiedene Semester bei gleichen Dozenten in gleichen Veranstaltungen (Rindermann, 1996).

Die Analysen mit verschiedenen Datensätzen, Urteilern und Maßen ergaben aber keine bedeutsamen Verbesserungen. Veränderungen waren nicht oder kaum feststellbar. Die durchschnittlichen Effektstärken ($d = \frac{\bar{x}_2 - \bar{x}_1}{s}$) bewegten sich um $d = 0$, z. T. waren auch negative Veränderungen beobachtbar (s. Rindermann, 2001). Auch die Differenzierung zwischen Veranstaltungen mit besprochenem Feedback (hochschuldidaktisches Diskursmodell) und nicht besprochenem Feedback führte nicht zu bedeutsamen Resultaten. Die Besprechung oder Nichtbesprechung der Evaluationsergebnisse wies dagegen eine *Indikatorfunktion* für Lehrkompetenz und -engagement auf: Dozenten, deren Lehre von Studierenden als kompetent und engagiert wahrgenommen wird (Mitte und Ende des Semesters), besprachen eher Evaluationsergebnisse und die Lehre mit ihren Studierenden als es solche Lehrende taten, die als weniger kompetent und engagiert wahrgenommen wurden. Dieser Effekt war entweder auf die Resultate selbst zurückzuführen – gute Resultate werden durch Lehrende eher besprochen – oder auf Dispositionen der Dozenten: In der Lehre engagierte Dozenten bieten bessere Lehre und sind auch an einer Besprechung von Evaluationsresultaten mit Studierenden stärker interessiert. Diese Resultate waren auch aus Fremdurteilersicht (die nicht wussten, ob Feedbacks besprochen wurden) zu bestätigen: Besprochene Veranstaltungen wurden zu beiden Messzeitpunkten besser beurteilt.

Die meisten Autoren befürworten eine Beratung des Dozenten (Erläuterung des Feedbacks, Anregungen, Motivierung) und ergänzend den Einsatz hochschuldidaktischer Weiterbildung als optimale Strategie zur Verbesserung von Lehrqualität. Dozenten müs-

sen nicht nur über ihre Stärken und Schwächen informiert werden, sondern sie benötigen darüber hinaus auch Hinweise, wie sie ihre Lehre anders gestalten können: In einem Beratungsgespräch sollten Lehrende informiert (Rückmelden), ihr Problembewusstsein gefördert (Amplifizieren, Konfrontieren), einzelne modifizierbare Bereiche ausgewählt (Vereinfachen, Akzentuieren), dysfunktionale Erklärungsmuster und Handlungsketten unterbrochen (z. B. Reattribuieren), konkrete Veränderungsvorschläge erarbeitet (Vorschlagen, Modellieren) und Lehrende sollten zu Veränderungen aktiviert und dabei emotional gestützt werden. Studien, die den Beratungsansatz mit Kontrollgruppen ohne Beratung und in Messwiederholungen überprüften, zeigten bedeutende Effekte (vgl. die Metaanalysen von P. A. Cohen, 1980, Feedback $d = 0.20$, Feedback und Beratung $d = 0.64$; und von Menges & Brinko, 1986, Feedback $d = 0.22$, Feedback und Beratung $d = 1.10$).

Evaluation kann anscheinend nicht ohne entsprechende Einbettung in ein Beratungsprogramm erfolgreich an der Universität durchgeführt werden. Die Evaluation bedarf einer Ergänzung durch Beratung, Training und Anreize. Die Effektivität eines solchen Ansatzes soll hier untersucht werden.

Reichenauer Studie zum Beratungsansatz

Ziel dieser Studie war es zu überprüfen, ob ein Lehrevaluationsverfahren, bei dem Ergebnisse von Veranstaltungskritik in einem persönlichen Beratungsgespräch an Dozenten rückgemeldet und Dozenten in der Gestaltung ihrer Lehre beraten werden, positive Veränderungen im Lehrverhalten der Dozenten – gemessen an studentischen Lehrveranstaltungseinschätzungen in Vorher-Nachher-Erhebungen – bewirken kann. Dieser Nachweis stand im deutschsprachigen Raum aus. Hierzu wurde eine in privater Trägerschaft stehende Institution gewählt – ein Ausbildungsinstitut für Logopädie mit fest und frei angestellten Lehrkräften.

Schule für Logopädie – Ausbildungsmerkmale

Die Schule für Logopädie auf der Insel Reichenau bildet innerhalb von drei Jahren zum staatlich anerkannten Logopäden aus. Die Ausbildung umfasst mindestens 1740 Schulstunden so genannten „theoretisch-praktischen Unterrichts" und mindestens 2100 Stunden „praktische Ausbildung". Die Veranstaltungen, die in der Regel in seminaristischer Form, aber meist ohne Referate gehalten werden, werden von einer Gruppe von 20 Auszubildenden besucht, die im Quasiklassenverband die dreijährige Ausbildung absolvieren. Die Kurse dienen der Vermittlung von Grundlagen- und Anwendungswissen auf den Gebieten Sprachtherapie, Medizin, Psychologie und Pädagogik. Neben den sozialwissenschaftlichen Fächern spielen die medizinischen Fächer wie Hals-Nasen-Ohren-Heilkunde, Anatomie, Neurologie oder Psychiatrie eine gewichtige Rolle. Schwerpunkt des theoretisch-praktischen Unterrichts bildet aber die Logopädie selbst. In dieser werden anwendungsorientierte Behandlungsmodelle der Sprachtherapie vorgestellt. Der Unterricht wird in den medizinischen und sozialwissenschaftlichen Fächern ausschließlich von Honorarkräften gegeben, die durch Ausbildung und Berufspraxis ein ausgewiesenes Fachwissen in den oben genannten Fächern erlangt haben. Der Unterricht im Fach Logopädie erfolgt durch Logopäden. Die evaluierten Seminare („theoretisch-praktischer Unterricht") ähneln Hochschulveranstaltungen.

Dieses Institut bot besonders positive Rahmenbedingungen, die sich in relevanten Merkmalen von den gegebenen Rahmenbedingungen an deutschen Universitäten unterscheiden, hierin aber denen internationaler Hochschulen ähneln: Lehre stellt ein zentrales Qualitätsmerkmal dieser Einrichtung dar. Das Institut ist privat, es gibt eine klare Führungsstruktur. „Schüler" haben auch die Stellung von Kunden, da sie für ihre Ausbildung bezahlen. Dozenten sind zwar fast unkündbar, entweder weil sie Lebenszeitstellen haben oder weil zu ihnen als freien Mitarbeitern im regionalen Umfeld kaum eine Alternative besteht, durchgehend negative Evaluationsergebnisse in ihrem zentralen Aufgabenbereich könnten aber zu einer Beendung des Arbeitsverhältnisses oder zu Veränderung des Aufgabenspektrums führen. Lehrqualität und Lehrevaluation stellen somit keine irrelevan-

ten behördlichen Verwaltungsvorgänge dar. Allerdings ist dieses Institut in Merkmalen der Lehre, in der Art der Veranstaltungsdurchführung und in der Klientel der Veranstaltungsteilnehmer nicht den deutschen Hochschulen unähnlich: Die „Schüler" sind junge Erwachsene, die meisten mit Abitur, Veranstaltungen variieren zwischen dem Typ der Vorlesung und dem des Seminars, nur die Größe bleibt fast immer mit etwa 15 – 20 Teilnehmern konstant. Das Institut stellte somit im Vergleich zur üblichen staatlichen Hochschule in Deutschland besonders günstige Rahmenbedingungen zur Verfügung, um die Effektivität eines Evaluations-Beratungs-Verfahrens zu bestimmen.

Rückmeldeverfahren

Auf Grund der oben geschilderten Erfahrungen wurde in einer Untersuchung an einem privaten Ausbildungsinstitut ein im Vergleich zu früheren HILVE-Studien aufwändigeres Rückmeldeverfahren gewählt: Dozenten erhielten wie gehabt ein statistisches Feedback, die Dimensionsbeschreibungen und eine Kopie der handschriftlichen Kommentare der Studierenden in – und das ist neu – einem *persönlichen Rückmeldegespräch* mit einem Diplom-Psychologen. Das standardisierte Feedback beinhaltete Mittelwerte (nicht-normiert und normiert), Streuungen, Minima, Maxima, Anzahl der Beobachtungen je Dimension und Item und eigener Veranstaltung und eine übersichtliche Grafik normierter Mittelwerte je Dimension und eigener Veranstaltung. Zusätzlich erhielten sie die studentischen Kommentare zu den Fragen „Was ist besonders gut an der Veranstaltung?", „Was ist schlecht?", „Verbesserungsvorschläge?" und Anmerkungen zum Fragebogen oder zur Erhebung.

In einem ca. einstündigen Einzelgespräch zwischen Schulleiter (dem Psychologen) und dem Dozenten wurden das Feedback erläutert, Gründe für positive oder negative Rückmeldungen eruiert, die Selbstsicht des Dozenten im Vergleich mit den studentischen Angaben diskutiert, Unterrichtsformen und einzelne didaktische Strategien besprochen und Möglichkeiten erörtert, die Lehre zu verbessern (ausführlich Kohler & Rindermann, 2000 a, 2000 b; Rindermann & Kohler, 2001). Das Feedback diente der Konfrontation mit der Fremdwahrnehmung eigenen Handelns – stabile Verhaltensstrukturen sollten außer Kraft gesetzt werden. Schwerpunkte der Beratung bildeten dann konkrete Vorschläge zur Veränderung des Handelns im Unterricht sowie dessen Vorbereitung, etwa Stoffmenge reduzieren oder neue didaktische Methoden einsetzen (z. B. zeitweise Kleingruppenarbeit statt durchgehend Vorlesungsstil). Zusätzlich wurde dazu angeregt, die Ergebnisse in der Klasse zu diskutieren. Abschließend wurde vom Berater (Schulleiter) sein Interesse an der weiteren Entwicklung des Dozenten hervorgehoben und jederzeitige Gesprächsbereitschaft signalisiert. Die Beratung diente der Stützung und dem Aufbau neuer Lehrtechniken und pädagogischer Handlungsweisen. Nicht zuletzt sollte sie auch die Relevanz der Evaluation erhöhen.

Die Bedeutung der Evaluation für das *Gesamtsystem* Schule insgesamt war ebenfalls wichtig. Das Evaluationsprojekt als offizielle Maßnahme zur Qualitätssicherung erfasste alle am Institut Tätigen: Der Austausch der Dozenten untereinander sollte verstärkt werden. Im günstigen Fall wurde erwartet, dass sie sich gegenseitig stützen und voneinander lernen. Die Möglichkeit einer Einflussnahme der Schülerschaft auf die Lehrqualität sollte die Identifikation mit der Schule steigern.

Dieser Evaluationsansatz unterschied sich hinsichtlich mehrerer Aspekte von den zuvor geschilderten und von den meisten an deutschsprachigen Hochschulen praktizierten Verfahren:

- kommentierte Rückmeldung und Beratung
- der Berater war Diplom-Psychologe und gleichzeitig Dozent und Leiter der Ausbildungsinstitution
- dadurch erhielten Evaluation und Rückmeldung ein besonderes Gewicht
- die Evaluation fand nicht an einer staatlichen

Hochschule statt, sondern an einem privaten Ausbildungsinstitut für Logopädie
- die Institution führte die Evaluation durch, um später mit der Qualität der Lehre werben zu können
- die Lehrkräfte waren keine Professoren oder verbeamtete Dozenten, sondern mit individuellen Lehr-Verträgen eingestellte Berufspraktiker (Logopäden, Ärzte, Psychologen; 52 % Frauen)
- Lehrkräfte könnten die Institution verlassen, prinzipiell wäre es auch möglich, die Verträge wegen konstant kritischer Beurteilungen nicht zu verlängern (kam nicht vor)
- die Veranstaltungsteilnehmer waren keine Universitätsstudenten, sondern Auszubildende des Faches Logopädie (65 % Abitur, 33 % Realschulabschluss, 2 % Hauptschulabschluss; Durchschnittsalter 31 Jahre, $s = 8$ Jahre, der jüngste war 19, der älteste 52; 82 % Frauen); es werden sowohl junge Abiturienten als auch ältere Umschüler mit mittlerem Bildungsabschluss oder abgeschlossener Hochschulausbildung zur Logopädenausbildung zugelassen
- Auszubildende zahlen für ihre Ausbildung zum Logopäden 700,– Euro pro Monat, die Einrichtung lebt zu einem großen Teil von diesen Gebühren (daneben noch eine staatliche Förderung).

Ablauf der Lehrevaluation

1. *Lehrevaluation* in einer Veranstaltung Mitte des Semesters, meist durch den Dozenten selbst ausgeteilt und eingesammelt. Dozenten hatten keine Vorerfahrung mit Evaluation.
2. *Sammlung der Bögen* verschiedener Dozenten beim Leiter der Schule und Verschickung an die Universität.
3. *Externe Auswertung* der Bögen (Dateneingabe, Erstellung des Feedbacks, Kopie der handschriftlichen Kommentare) und Rücksendung des Feedbacks an den Schulleiter. Zeitlicher Abstand zwischen Erhebung und Rückmeldung ca. drei Wochen bis drei Monate.
4. *Individuelle Besprechung* des Feedbacks zwischen Schulleiter und Dozent, Diskussion von Ursachen positiver oder negativer Rückmeldung, Erhebung der Selbstsicht des Dozenten, Erörterung von konkreten Verbesserungsmöglichkeiten, *Beratung* in didaktischen Problemen und Fragen der Dozenten-Studenten-Interaktion, die Beratung dient der Stützung und dem Aufbau neuer Lehrtechniken und didaktischer Handlungsweisen.
5. *Evaluation* einer weiteren Veranstaltung zu einem späteren Zeitpunkt (wenige Monate später bis zu einem Jahr, meist Kurs bei einer anderen Schülergruppe).
6. *Sammlung der Bögen* beim Leiter der Schule und Verschickung an die Universität.
7. *Externe Auswertung* der Bögen und Rücksendung des Feedbacks.
8. *Individuelle Besprechung* der Rückmeldung zwischen Schulleiter und Dozent und *Beratung* (in der Regel wurden somit zwei Veranstaltungen eines Dozenten erhoben).
9. Am Ende eines jeden Ausbildungsjahres eine *Dozenten- und Schülerkonferenz* mit dem Schulleiter, auf der die Gesamtergebnisse besprochen werden.

Dass es sich nicht um ein Hochschulstudium und nicht um Studenten handelt, dürfte zu den weniger ausschlaggebenden Rahmenbedingungen zählen. Wichtiger sind der private Charakter der Ausbildungseinrichtung, die Durchführung von Evaluation und Beratung durch den Leiter der Einrichtung und der andere Status der Dozenten. Die Qualität der Lehre und ihre Bewertung haben eine andere Bedeutung als an staatlichen Hochschulen, deren Mitglieder sich vornehmlich als Wissenschaftler verstehen und die nicht von den Studiengebühren ihrer Studierenden abhängig sind. Somit unterscheiden sich das Evaluationsverfahren (kommentierte Rückmeldung und Beratung) und die Rahmenbedingungen von den bislang durchgeführten deutschen Lehrevaluationsstudien.

Auf eine Kontrollgruppe wurde in der Reichenauer Studie verzichtet. Frühere Studien an Universitäten ließen es als unwahrscheinlich erscheinen, dass etwa nur durch Rückmeldung oder nur durch die Durchführung von Evaluation (ohne Rückmeldung oder Beratung) bedeutsame Verbesserungen erzielbar wären. Strenge methodische Designs waren nicht umsetzbar, etwa Isolierung der Wirkfaktoren: Durchführung von Kursevaluation, Rückmeldung, Beratung, Dozenten- und Schülerkonferenzen zur Evaluation, privates vs. staatliches Ausbildungsinstitut, Ausbildungsinstitut vs. Hochschule, Honorarkräfte vs. beamtete Dozenten und Professoren, Evaluation und Beratung durch Leiter vs. durch eine andere Person, Relevanz und Verpflichtungscharakter der Evaluation, etc. Ebenso wäre an eine Kontrollgruppe ohne Rückmeldung und Beratung in einem kleinen Ausbil-

dungsinstitut, das die Evaluation mit dem Ziel der Verbesserung der Lehrqualität und der Reputationssteigerung durchführt, nicht zu denken. Allerdings ist diese Evaluationsstudie gut mit früheren Studien des Autors vergleichbar, bei denen die aufgelisteten Wirkfaktoren in anderen Kombinationen vorhanden waren, was bei unterschiedlichen Ergebnissen Schlussfolgerungen auf die praktische Bedeutung dieser Faktoren zulässt.

Methode

Stichprobe

Es wurden 48 Veranstaltungen von 27 Dozenten durch meist 15–20 Schüler evaluiert. Schüler aus sechs Ausbildungsjahren waren zum Teil über verschiedene Dozenten und Veranstaltungen identisch. Von den Dozenten mit Messwiederholung wurde ein Dozent viermal evaluiert (drei Veranstaltungen zu einem Messzeitpunkt zusammengefasst, eine zu einem späteren) und ein Dozent dreimal (drei Veranstaltungen zu je einem Messzeitpunkt, hier wurden nur die ersten zwei berücksichtigt). Diese Dozenten wurden mehrfach evaluiert, weil mehrmalig intensive Beratung durchgeführt wurde und weil studentische Gruppenunterschiede geprüft werden sollten. Von 16 Dozenten liegen jeweils Kursbeurteilungen zu zwei verschiedenen Messzeitpunkten vor. Die Daten dieser 16 Dozenten bilden die Grundlage für die Auswertungen. Auswertungseinheit ist der Dozent. Die Veranstaltungen der Dozenten stammten jeweils aus dem gleichen Fachgebiet. Die genauen Fallzahlen sind in den Tabellen aufgeführt (in Referaten und Fähigkeiten geringere Fallzahlen). Die Untersuchung lief von 1997 bis 2002.

Instrument

Zur Messung der Lehrqualität wurde das HILVE-II in geringfügig an das Logopädie-Institut adaptierter Form eingesetzt. Veranstaltungen wurden nur durch die Teilnehmer eingeschätzt (keine Dozentenselbsteinschätzungen oder Fremdurteiler). Auszubildende zu verschiedenen Messzeitpunkten gehörten meist verschiedenen Jahrgängen an.
Die Dimensionen wurden nach theoretischen Kriterien in vier Gruppen zusammengefasst (s. Rindermann, 2001): *Dozentenskalen* (Struktur, Auseinandersetzung, Verarbeitung, Lehrkompetenz, Engagement, Klima, Betreuung, Interaktionsmanagement), *studentische Skalen* (Referate, Beteiligung, Disziplin, Fähigkeitsniveau), *Rahmenbedingungen* (Redundanz, Anforderung, Thema, veranstaltungsexterner Fleiß) und *Lehreffektivität* (Veranstaltungsinteressantheit, Lernen quantitativ und qualitativ, Interessenförderung, Allgemeinurteil). In allen Skalen bis auf Anforderung, Redundanz und Veranstaltungsbenotung sind große Werte optimal (Skala 1–7). In Anforderungen ist eine mittlere Ausprägung günstig, ebenso in Redundanz, in Veranstaltungsbenotung ein geringer Wert (1–6). Die Dozentenskalen sind in sich homogener als die studentischen und die Rahmenbedingungs-Skalen, die heterogenere Aspekte zusammenfassen.

Analysen, Berechnung von d

Die Analysen wurden in zwei Varianten berechnet: a) alle Dozenten ($n = 16$) und b) ohne die Dozenten, deren Verhalten zum ersten Messzeitpunkt auf der Lehrkompetenzskala mit besser als 6.5 (Skala 1 bis 7) und deren Veranstaltung besser als 1.5 benotet wurde (Notenskala von 1, sehr gut, bis 6, ungenügend; verbleibend $n = 13$ Dozenten). Bei drei äußerst günstig bewerteten Dozenten (Lehrkompetenzschnitt 6.71; Notenschnitt 1.23) war anhand des Instruments kaum eine messbare Verbesserung möglich, auch im Lehrverhalten dürften sich durch diese Maßnahme kaum noch Steigerungsmöglichkeiten ergeben. Diese (erfahrenen) Dozenten dienten z. T. bei Hospitationen als Ratgeber.
Zur Bestimmung der Größe des Effektes wurden Effektstärken herangezogen. Nur sie erlauben eine direkte Einschätzung des Treatmenteffektes und anders als Signifikanzangaben sind sie stichprobengrößenunabhängig und ermöglichen so den Vergleich zwischen verschiedenen Studien (J. Cohen, 1988; Wolf, 2001). Das Differenzmaß d wird konventionell durch Abziehen der ersten Messung von der zweiten und Relativierung der Differenz durch die Streuung der Kursmittelwerte gebildet ($d = \frac{\bar{x}_2 - \bar{x}_1}{s}$). Diese Variante hat den großen Vorteil der Einfachheit (man benötigt nur zwei Mittel und eine Streuung der Mittel, als Streuung in der Regel die Streuung der ersten Messung, seltener die Gesamtstreuung). Zudem ist sie die gängigste Variante und Ergebnisse sind somit gut zwischen verschiedenen Studien verschiedener Autoren vergleichbar. Allerdings ist diese Berechnungsvariante nur für unabhängige Stichproben optimal (etwa J. Cohen, 1992), bei abhängigen Stichproben und Korrelation zwischen den Messzeitpunkten unterschätzt sie den Veränderungseffekt.
Für abhängige Stichproben empfehlen deshalb Bortz und Döring (1995, S. 569; J. Cohen, 1988, S. 48) folgende Formel $d' = \frac{\bar{x}_2 - \bar{x}_1}{s_D} \cdot \sqrt{2}$. Hierbei ist s_D die Streuung der Differenzen $\bar{x}_2 - \bar{x}_1$, $\sqrt{2}$ dient der Anpassung an die Cohenschen Normen von $d = 0.2$ als kleinen, 0.5 (halbe Streuung) als mittleren und 0.8 als großen Effekt. Falls die Streuungen zu beiden Messzeitpunkten gleich sind, ließe sich auch

$d' = \frac{\bar{x}_2 - \bar{x}_1}{s \cdot \sqrt{1-r}}$ heranziehen. In diesem Falle ist s nicht die Streuung der Differenzen, sondern die der Veranstaltungsmittel (Erst- oder Zweitmessung).
Hohe (positive) Korrelationen zwischen Messzeitpunkten führen dazu, dass die konventionell berechneten d-Größen Veränderungen unterschätzen – je höher die Korrelation, desto mehr müsste der Effekt durch die Formel für abhängige Stichproben nach oben korrigiert werden. Hintergrund ist die Reduktion von Fehlervarianz durch mehrfache Heranziehung einer einzigen Stichprobe: Messfehler, die bei verschiedenen Stichproben auf Grund von irrelevanten Personenunterschieden innerhalb der Gruppen auftreten (hier Unterschiede zwischen Dozenten), treten bei Messwiederholungsstichproben nicht auf. Bei einem Vergleich von zwei Stichprobenmittelwerten aus zwei unabhängigen Gruppen werden zwei fehlerbehaftete Schätzwerte benötigt, während bei Messwiederholung nur ein Schätzwert nötig ist, die Fehlervarianz ist deshalb hier nur halb so groß (Klemmert, 2000). In der Gesamtvarianz nimmt die Fehlervarianz ab, die durch eine Maßnahme (hier Rückmeldung und Beratung) erklärbare Varianz nimmt somit zu. Demzufolge nimmt zusätzlich auch der Stichprobenumfang, der zur statistischen Absicherung eines Effektes benötigt wird, ab.

Ergebnisse

Im Gesamtdatensatz waren in Dozentenvariablen systematische Verbesserungen zu beobachten (s. Tabelle 1), in den Skalen Lehrkompetenz ($d = 0.94$), Engagement ($d = 0.78$) und Verarbeitung ($d = 0.70$) fielen sie am deutlichsten aus, am geringsten unter den Dozentenvariablen waren sie in der Betreuung (Feedback und Betreuung, $d = 0.38$) und im Klima (Freundlichkeit und Kooperativität des Dozenten, $d = 0.47$). Auch in den Lehreffektivitätsskalen waren günstige Veränderungen beobachtbar, hier traten die größten Verbesserungen in Lernen quantitativ ($d = 0.76$) und vor allem im Allgemeinurteil ($d = 0.85$) auf. Die Disziplin (keine Unruhe während der Sitzungen, geringe Fehlzeiten) der Teilnehmer hat erkennbar zugenommen ($d = 0.54$), die Beteiligung war höher, der veranstaltungsexterne Fleiß (Vor- und Nachbereitung, Arbeitsaufwand) blieb aber gleich, die Redundanz nahm zu. Letzteres ist nicht unbedingt als positiv oder negativ zu interpretieren: Kursinhalte waren nun bekannter, es gab inhaltliche Überschneidungen mit anderen Kursen; mittlere bis geringe, aber nicht bei Null liegende Redundanz ist optimal.
Auch in studentischen Fähigkeiten war eine Zunahme erkennbar (die Skala Fähigkeiten besteht aus den Items transformiertes Abitur/Schulabschlussnote und Selbsteinschätzung des Leistungsniveaus im Vergleich zu anderen Kursteilnehmern). Nur die Zunahme im reduzierten Datensatz ist bedeutend, d. h. es traten große „Verbesserungen" bei den mäßig beurteilten Dozenten auf. Worauf dies zurückzuführen ist, ist unklar: Unterschiedliche Beantwortungsquote (Abitur wurde bei der Zweitausfüllung häufig weggelassen, es wurden, um Verzerrungen zu vermeiden, für die Skala Fähigkeiten nur Veranstaltungen mit mindestens 10 in diesen beiden Items ausgefüllten Bögen genommen), zufällige Unterschiede der Jahrgänge? Auf jeden Fall sind Auswirkungen des Evaluations-Beratungs-Ansatzes unwahrscheinlich.
Insgesamt waren die Veränderungen von mittlerer Größe: Nach Cohen (1988) gelten Differenzen von $d = 0.2$ als klein, von 0.5 als mittel und von 0.8 als groß. Die meisten Veränderungen lagen hier bei $d = .30 - .80$. Bei sehr gut bewerteten Dozenten (s. u.) könnten Deckeneffekte aufgetreten sein. Zielgröße des Eingriffs war primär das Lehrhandeln (hier gab es mittlere bis große Effekte), nicht studentisches Verhalten und Rahmenbedingungen (hier geringe Effekte). Die Evaluierung verschiedener Veranstaltungen durch verschiedene Studierende könnte den Effekt noch unterschätzen. In gängigen Studien zur Analyse der Effektivität von Lehrevaluation und Feedback (vgl. Menges & Brinko, 1986) wurden Evaluation, Feedback und Beratung innerhalb einer Veranstaltung eines Semesters untersucht. Hier in Reichenau wechselten Teilnehmer (Urteilerstichproben) und Veranstaltung (aber gleiches Fachgebiet und selbstverständlich gleicher Dozent). Unsystematische Veränderungen durch diese Einflussvariablen verringern den Anteil messbarer Auswirkungen der Rückmelde-Beratungs-Maßnahme (geringere Präzision).

Bei schon zum ersten Messzeitpunkt sehr gut bewerteten Dozenten waren messbare Verbesserungen in Lehrevaluationen durch Evaluation und Beratung kaum möglich (Deckeneffekt). Deshalb wurden drei (erfahrene) Dozenten mit extrem guten Bewertungen in einer zweiten Analyse ausgeschlossen (s. Tabelle 2). Es wurde nun deutlich erkennbar, dass bei kritisch und mäßig bewerteten Dozenten Evaluation und Beratung ein sehr wirksames Mittel zur Verbesserung der Lehre darstellen. In Lehrerfolgsskalen lag die mittlere Verbesserung bei $d = 1.03$, in Dozentenskalen bei $d = 0.96$. In der zentralen Skala Lehrkompetenz ist eine Verbesserung von $d = 1.26$ beeindruckend. Bei diesen Dozenten waren auch in Disziplin, Redundanz/Anknüpfung an Vorwissen und Fleiß größere Zunahmen beobachtbar.

Tabelle 1: Veränderungen in den Evaluationsvariablen zwischen zwei Messzeitpunkten, erfasst im Effektstärkemaß *d*
$n = 16$ Dozenten wiederholt, 35 Veranstaltungen bzw. 592 Fragebögen zu beiden Messzeitpunkten zusammen, zwischen 12 und 41 Fragebögen je Dozent und Messzeitpunkt (Messung zweier Dozenten zu einem Messzeitpunkt in verschiedenen Kursen zusammengefasst), Analysegröße ist Dozent

Skala	**Dozenten-skalen**	Struktur	Auseinandersetz.	Verarbeitung	Lehrkompetenz	Engagement	Klima	Betreuung	Interaktionsma.
d	**0.65***	0.66^t	0.64^t	0.70^t	0.94*	0.78*	0.47	0.38	0.59
M_1	**5.08**	5.16	5.32	4.73	4.81	5.12	5.99	4.83	4.67
s_1	**1.00**	1.20	0.89	1.12	1.20	1.02	0.77	1.05	1.19
M_2	**5.38**	5.48	5.56	5.08	5.31	5.48	6.09	5.00	5.01
s_2	**0.62**	0.80	0.61	0.77	0.74	0.57	0.68	0.75	0.80

Skala	**studentische Skalen**	Referate ($n = 3$)	Beteiligung	Disziplin	Fähigkeiten ($n = 6$)
d	**0.43**	0.35	0.39	0.54	0.12
M_1	**4.84**	4.52	4.86	5.50	4.04
s_1	**0.45**	0.60	0.42	0.74	0.50
M_2	**5.13**	4.77	4.98	5.83	4.08
s_2	**0.34**	0.71	0.47	0.46	0.16

Skala	**Rahmenbedingungen**	Redundanz	Anforderung	Thema	Fleiß
d	**0.24**	0.47	0.23	0.21	0.05
M_1	**3.43**	2.13	4.06	5.07	2.48
s_1	**0.31**	0.32	0.56	0.42	0.71
M_2	**3.51**	2.27	4.12	5.15	2.51
s_2	**0.35**	0.41	0.32	0.54	0.56

Skala	**Lehrerfolg**	Interessantheit	Lernen quan.	Lernen qual.	Interessenförd.	Allgemein
d	**0.64^t**	0.45	0.76*	0.60	0.51	0.85*
M_1	**4.96**	4.80	4.84	5.40	4.67	5.07
s_1	**0.98**	1.34	0.94	0.69	1.01	1.11
M_2	**5.27**	5.06	5.21	5.66	4.94	5.48
s_2	**0.58**	0.87	0.54	0.41	0.64	0.69

Anmerkungen: t = tendenziell signifikant (10 %), * = signifikant (5 %), ** = sehr signifikant (1 %); in Referaten und in Fähigkeiten liegen nur von 3 bzw. 6 Dozenten ausreichend Daten vor (in Veranstaltungen keine Referate oder keine ausreichend großen Stichproben; nur Veranstaltungen mit mindestens 10 vollständigen Angaben zum eigenen Zeugnis und zur Selbsteinschätzung der eigenen Fähigkeiten), Mittel und Streuung in einer Antwortskala von 1–7

Tabelle 2: Veränderungen in den Evaluationsvariablen zwischen zwei Messzeitpunkten, erfasst im Effektstärkemaß *d*:
ohne sehr gute Dozenten (s_D von Tabelle 1 beibehalten)
$n = 13$ Dozenten wiederholt, 29 Veranstaltungen bzw. 485 Fragebögen zu beiden Messzeitpunkten zusammen, zwischen 12 und 41 Fragebögen je Dozent und Messzeitpunkt (Messung zweier Dozenten zu einem Messzeitpunkt in verschiedenen Kursen zusammengefasst), Analysegröße ist Dozent; Angabe von *d*

Skala	**Dozenten-skalen**	Struktur	Auseinandersetz.	Verarbeitung	Lehrkompetenz	Engagement	Klima	Betreuung	Interaktionsma.
d	**0.96****	0.91*	1.02*	1.06*	1.26**	1.14**	0.69	0.73t	0.87*

Skala	**studentische Skalen**	Referate ($n = 2$)	Beteiligung	Disziplin	Fähigkeiten ($n = 5$)
d	**0.63****	0.07	0.56	0.76t	0.69

Skala	**Rahmenbedingungen**	Redundanz	Anforderung	Thema	Fleiß
d	**0.33**	0.43	0.24	0.28	0.37

Skala	**Lehrerfolg**	Interessantheit	Lernen quan.	Lernen qual.	Interessenförd.	Allgemein
d	**1.03****	0.87*	1.14**	1.01**	0.93*	1.20**

Anmerkungen: t=tendenziell signifikant (10 %), * = signifikant (5 %), ** = sehr signifikant (1 %); in Referaten und in Fähigkeiten liegen nur von 2 bzw. 5 Dozenten ausreichend Daten vor (in Veranstaltungen keine Referate oder keine ausreichend großen Stichproben; nur Veranstaltungen mit mindestens 10 vollständigen Angaben zum eigenen Zeugnis und zur Selbsteinschätzung der eigenen Fähigkeiten); 0.63 in studentischen Skalen wird sehr signifikant, weil zur Signifikanzprüfung automatisch der reduzierte s_D-Wert herangezogen wurde, zur Effektberechnung aus Gründen der Vergleichbarkeit jedoch der Gesamt-s_D

Die Mittelwertsverbesserungen im Gesamtdatensatz waren vor allem auf Verbesserungen der zuerst kritisch bewerteten Dozenten zurückzuführen. Im mittleren und oberen Bereich waren insgesamt nur leichte Veränderungen, meist zum Positiven, zu verzeichnen. Demzufolge gingen die Standardabweichungen auch stark zurück. Das Kollegium wurde hinsichtlich seiner didaktischen Fähigkeiten geschlossener. Dies lässt sich exemplarisch an den Ergebnissen in der Veranstaltungsbenotung verdeutlichen (s. Abbildung 2).
Die Mehrheit der Dozenten verbesserte sich, allerdings fielen die Veränderungen bei schwächer beurteilten Dozenten deutlicher aus. Die Korrelation zwischen Ausgangswert und Differenz beträgt $r = -.71$, was für einen Regressionseffekt spricht. Regression zur Mitte wäre nur bei perfekter Korrelation zwischen erster und zweiter Messung ausschließbar (Cohen & Cohen, 1983, S. 45). Die Regression ist für die Interpretation der Veränderungen als Verbesserung jedoch unproblematisch, da hier ($N = 16$, Tabelle 1) nicht eine auf Grund eines Vortestergebnisses ausgewählte Stichprobe untersucht wurde und da in der Gesamtgruppe eine bedeutsame Verbesserung feststellbar war. Und Verbesserungen in der Lehre ließen sich sowohl den günstigeren mittleren Einschätzungen als auch den studentischen Kommentaren entnehmen: So waren bei einem Dozenten bei der Zweitmessung auf die Frage „Was ist besonders gut an der Veranstaltung?" folgende Antworten zu finden: „Dozent ist seit neuestem wie verwandelt: sehr gut vorbereitet, wirkt kompetent und ist sehr freundlich. Weiter so!" Oder: „Dozent ist in den letzten Wochen sehr gut vorbereitet. Gute Planung."
Trotz der Homogenisierung der Stichprobe

Messzeitpunktveränderung in HILVE-II-Skalen

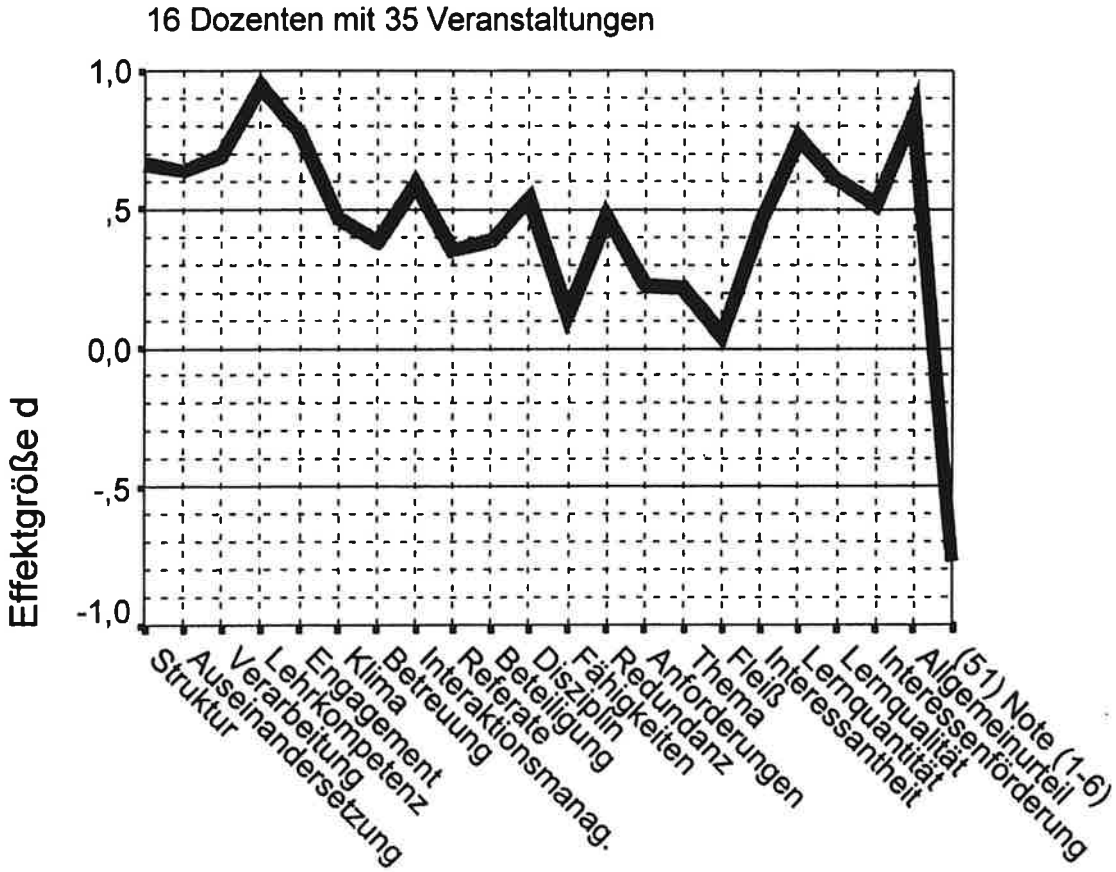

Abbildung 1: *d*-Effektgrößen für alle Skalen und die Veranstaltungsbenotung (eine numerisch negative Veränderung in der Veranstaltungsbenotung Item 51 ist inhaltlich positiv)

Messzeitpunktveränderung in der Note

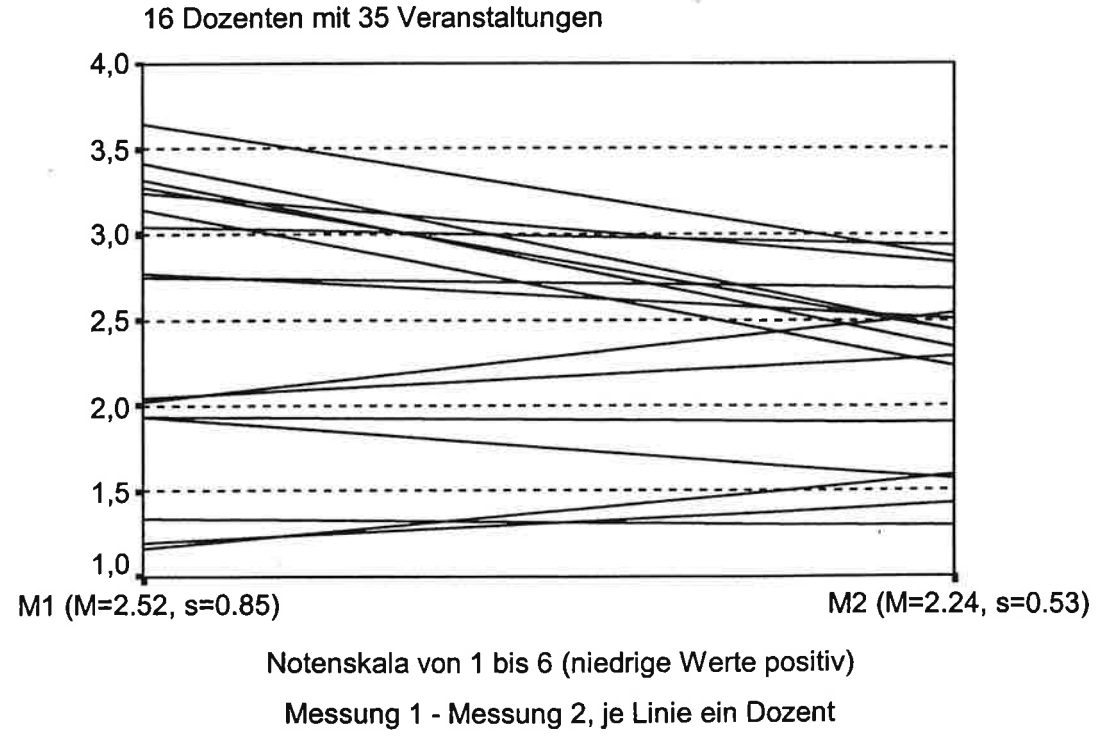

Abbildung 2: Mittelwertsverbesserung und Streuungsreduktion in der Veranstaltungsbenotung

durch den verwirklichten Evaluations-Beratungs-Ansatz bleibt unübersehbar, dass kritisch bewertete Dozenten ungeachtet deutlicher Verbesserungen meist schwächer blieben als günstig bewertete Dozenten. Die Korrelation zwischen erster und zweiter Messung in den Dozentenskalen lag bei $r_{tt}=.87$, in den Effektivitätsskalen betrug sie $r_{tt}=.79$, in Rahmenbedingungen $r_{tt}=.57$ und in studentischen Skalen $r_{tt}=.19$. Eine Grenze des Beratungsansatzes für das Dozentenverhalten ist hiermit erkennbar.

Diskussion – Bedeutung für die Evaluationspraxis

Durch das praktizierte Beratungs-Rückmelde-Verfahren ließ sich die Lehre am Reichenauer Ausbildungsinstitut für Logopädie aus Sicht der Studenten bedeutend verbessern. Vor allem in der zentralen Zieldimension Lehrkompetenz ist ein starker Veränderungseffekt in der Höhe von etwa einer Standardabweichung (der Differenzen) dokumentierbar. Besonders große Veränderungen sind bei vor der Maßnahme nur mäßig gut bewerteten Dozenten zu beobachten. Allerdings erreichen auch nach der Beratung die zuvor kritisch beurteilten Dozenten noch nicht die Werte der zuvor überdurchschnittlich bewerteten Lehrenden. Weitere Potenziale wären vermutlich erst durch Trainingsansätze erschließbar (vgl. Leitner, 2001). Der didaktische Optimismus sollte aber nicht überstrapaziert werden: Nicht aus jedem kann ein begnadeter Rhetoriker und Didaktiker werden! Die aus der Literatur berichteten Effekte bewegen sich sowohl im Beratungs- wie auch im Trainingsansatz auf mittlerem Niveau. Hauptziel sollte sein, negative Ausreißer zu vermeiden und allen Dozenten Möglichkeiten zu eröffnen, an ihrer Lehre erfolgreich zu arbeiten. Dies ist durch ein Beratungsmodell, wie in Reichenau praktiziert, auf jeden Fall zu realisieren.

Inwieweit die spezifisch nicht-universitären Bedingungen eine Rolle spielen (Honorarkräfte, keine Beamten, Lehr- statt Forschungsorientierung, private Institution, klare Führung, institutionelle Identität und Geschlossenheit etc.), ist schwierig zu beantworten. Das Untersuchungsdesign genügt nicht den harten Standards einer versuchsplanorientierten Evaluation (Hager, Patry & Brezing, 2000): Nach dieser Form der Evaluationsforschung werden die Wirksamkeit (ob und wie stark?) und Wirkfaktoren (was wirkt?) von lokal begrenzten Interventionen bei Beachtung strenger methodischer Standards bestimmt. Versuchsplanerische Gütekriterien wie interne und externe Validität und Präzision haben bei diesem Modell Priorität (vgl. Campbell & Stanley, 1963). Bei der Evaluation der Lehre am Reichenauer Institut wurde dagegen der Schwerpunkt auf die Veränderung, die Wirksamkeitsmessung und den praktischen Erfolg gelegt – den Nutzen der Evaluation (vgl. Patton, 1997) –, ohne methodische Maßstäbe (etwa Vorher-Nachher-Vergleich, Heranziehung eines wissenschaftlich überprüften Instrumentes) zu vernachlässigen. Evaluationsforschung ist primär angewandte Forschung und als solche auf Praxis gerichtet und auf Bedingungen der Praxis angewiesen.

Ein Vergleich ist mit früheren HILVE-Studien an Universitäten möglich. Effekte durch die Art des Eingriffes (Beratung) und institutioneller Hintergrundvariablen, in denen sich diese Studie von den deutschen universitären Studien unterscheidet, sind aber nicht trennbar. Positiv auffällig ist zudem, dass neben Veränderungen aufseiten des *Dozentenverhaltens* auch *systemische Auswirkungen* auf studentisches Verhalten und Rahmenbedingungen erkennbar sind: Auszubildende fehlen seltener, stören seltener im Unterricht, sie sind engagierter, zwischen Kursen wird auf größere inhaltliche Anknüpfung geachtet. Auch die Institution hat sich verändert, die Dozenten- und Studentenkonferenzen mögen hier eine Rolle gespielt haben (Kohler & Rindermann, 2000 b). Evaluation hatte hier ein hohes institutionelles Gewicht. Sie stellte keinen unverbindlichen Verwaltungsakt dar, der en passant zwischen Studentensprechstunde und Abzeichnungsmappe erledigt werden

konnte. Eine Konsequenz für die Hochschulevaluation muss deshalb sein, neben der Bereitstellung von Beratungs- und Fortbildungsangeboten der Evaluation ein höheres Gewicht zu verleihen. Bestimmte Rahmenbedingungen sind Voraussetzung für ein Gelingen von Evaluation und dürfen nicht ausgeblendet werden! Der Effekt von Evaluation und Beratung an deutschen Hochschulen und die Übertragbarkeit der Effekte unter gegebenen Bedingungen sollte aber durch eine zusätzliche analoge Universitätsstudie überprüft werden.[1]

Nicht zuletzt muss darauf verwiesen werden, dass sich die Lehre als Ganzes am Reichenauer Institut auch durch den nicht zufälligen Weggang von Dozenten verbessert hat. In den oben vorgestellten Analysen wurden zwar nur die Dozenten berücksichtigt, bei denen mindestens zwei Veranstaltungen evaluiert wurden. Analysegröße ist der individuelle Dozent, nicht die Lehre insgesamt. Da vor allem aber kritisch beurteilte Dozenten nach der ersten oder einer späteren Messung sich beruflich veränderten und die gut bewerteten (erfolgreich lehrenden) eher blieben, steigt die durchschnittliche Veranstaltungsbeurteilung neben individuellen Verbesserungseffekten durch die Beratung weiter institutionell durch berufliche Veränderung von Dozenten an. Dies ist für kommende Generationen von Auszubildenden an dieser Institution von Bedeutung. Als Nutzen oder Auswirkung der Evaluation kann nicht nur die individuelle, dozentenbezogene Verbesserung der Lehre, sondern auch die institutionelle Verbesserung der Lehre insgesamt betrachtet werden.

Festhalten lässt sich auf Grund verschiedener internationaler Studien und mehrerer HILVE-Studien an verschiedenen Institutionen, dass Lehrevaluation mit und ohne Feedback sowie mit und ohne veranstaltungsinterne Besprechung keine oder nur geringfügige Verbesserungen erzielt, dass jedoch bei Ergänzung des Feedbacks durch Beratung oder Weiterbildung bedeutsame Optimierungseffekte erreichbar sind.

Evaluation und Evaluationsforschung müssen sich neben binnenwissenschaftlichen Kriterien, wie der Anschlussfähigkeit an den Theorienkanon und der Passung zur Forschungstradition, vor allem dem *Kriterium der Praxis* stellen. Das Destillat nationaler und internationaler Studien zeigt für die Praxis, dass Evaluation, Rückmeldung und Beratung in einem gratifizierenden Umfeld die Lehre verbessern können. Dozenten greifen solche Beratungs- und Weiterbildungsangebote vor allem dann auf, wenn sich Engagement in der Lehre für sie lohnt (vgl. Spiel & Fischer, 1998).

Dass reine Rückmeldung von Verhalten – Information – wenig oder keine Auswirkung auf Verhalten zeigt, ist nicht überraschend. Messung impliziert nicht Verbesserung (vgl. Gray, 1991). Moderne Lerntheorien unterstreichen, dass der Erwerb prozeduralen Wissens und von Kompetenzen einer angeleiteten Instruktion in praxisnahen Lernumwelten bedarf (vgl. Renkl, 2002; Rindermann, 2002). Bloße Information (hier Rückmeldung von Evaluationsergebnissen), ohne *Beispiele* zu geben, wie man es besser machen kann, ohne *Einübung,* ohne *Reflexion* über eigenes Handeln und ohne *beratende Rückmeldung* durch Experten, führt nicht zum Aufbau neuen Lehrerhandelns. Und Handlungen werden bei motivationalen oder volitionalen Defiziten nicht initiiert. Reine Beratung ohne Evaluation dürfte auch nicht effektiv sein: Es fehlt eine Informationsgrundlage über Schwächen und Stärken der Lehre. Selbsteinschätzungsstudien, die Dozentenurteile mit denjenigen von Studierenden und Fremdeinschätzern vergleichen, zeigen, dass Lehrende mit den Wahrnehmungen von Hörern und Fremdurteilern (Kollegen, Didaktiker, geschulte ehemalige Studenten) kaum übereinstimmen ($r = .24$ bzw. $r = .06$; Rindermann, 2001), Lehrende also Qualitätsmerkmale wie Struktur oder Anforderungen kaum angemessen einschätzen können. Es bedarf somit der studentischen und ex-

[1] Hierzu läuft zur Zeit eine Studie in Kooperation mit Dr. Markus Dresel (Ulm).

ternen Perspektive. Zusätzlich dient Evaluation der Erfolgskontrolle.

Beratung ist neben funktionalen auch aus ethischen Gründen angezeigt: Evaluation – die bloße Messung der Lehrqualität – ohne Personen und Institutionen zu ermöglichen, Qualität zu verbessern, könnte zu Frustration und Entmutigung bei kritisch bewerteten Dozenten führen, zu Überforderung oder Rückzug. Cranton und Knoop (1991) sprechen sogar von „professorial melancholia" (S. 100). Doyle würde einen solchen Ansatz für inhuman und unökonomisch halten: „I do not think it is humane to open someone to the possibility of a negative evaluation without at the same time providing some meaningful help toward improvement. Business firms know that they waste money if they discourage or dismiss potentially productive employees and so they spend large sums of money on intensive training programs" (Doyle, 1991, S. 126). Ähnlich Marsh und Roche (1999, S. 517): „There is an ethically dubious but widespread custom of giving potentially negative feedback to teachers without providing access to cost-effective interventions to assist them to improve their teaching effectiveness. This, perhaps, is the most serious indictment of the current practice."

Wenn man an der Verbesserung der Qualität der Lehre interessiert ist, sollte man die Lehre im Rahmen eines Beratungs-Rückmelde-Verfahrens evaluieren. Dagegen ist ein Evaluationsansatz ohne Möglichkeit für die Beteiligten, die Qualität zu verbessern, ohne Beratung und fördernde Rahmenbedingungen weder effektiv noch vertretbar. Universität, Lehrende und Studierende gewinnen bei Evaluationen, die Beratung und adäquate institutionelle Gewichtung guter Lehre einschließen, am meisten.

Literatur

Alean-Kirkpatrick, P., Hänni, H. & Lutz, L. (1997). Internal quality monitoring of the teaching at the ETH, Zürich: Model design and initial impacts. *Quality in Higher Education, 3*(1), 63–71.

Balk, M. (2000). *Die Evaluation von Lehrveranstaltungen. Die Wirkung von Evaluationsrückmeldung.* Frankfurt a. M.: Peter Lang.

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

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Hrsg.), *Handbook of research on teaching* (S. 171–246). Chicago: Rand McNally.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

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

Cranton, P. & Knoop, R. (1991). Incorporating job satisfaction into a model of instructional effectiveness. In M. Theall & J. Franklin (Hrsg.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (S. 99–109). San Francisco: Jossey-Bass.

Doyle, K. O. (1991). Report on a trip downtown. In M. Theall & J. Franklin (Hrsg.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (S. 123–126). San Francisco: Jossey-Bass.

Gediga, G., Kannen, K. v., Schnieder, F., Köhne, S., Luck, H. & Schneider, B. (2000). *Kiel: Ein Kommunikations-Instrument für die Evaluation von Lehrveranstaltungen. Bericht über die Entwicklung und Anwendungsmöglichkeiten einer formativen Evaluationsprozedur im universitären Alltag.* Bangor: Methodos.

Gray, P. J. (1991). Using assessment data to improve teaching. In M. Theall & J. Franklin (Hrsg.), *Effective practices for improving teaching (New directions for teaching and learning 48)* (S. 53–63). San Francisco: Jossey-Bass.

Hackl, P. & Sedlacek, G. (2001). Evaluierung als Chance zur kontinuierlichen Verbesserung der Lehre: Das Beispiel der Wirtschaftsuniversität Wien. In C. Spiel (Hrsg.), *Evaluation universitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck* (S. 111–129). Münster: Waxmann.

Hager, W., Patry, J.-L. & Brezing, H. (Hrsg.). (2000). *Evaluation psychologischer Interventionsmaßnahmen.* Bern: Huber.

Henninger, M. (1999). *Evaluation: Diagnose oder Therapie.* München: Forschungsberichte, LMU, Empirische Pädagogik, No. 102.

Klemmert, H. (2000). *Warum werden beim Vergleich zweier unabhängiger Stichproben-Mittelwerte größere Effekte gefordert als beim Vergleich eines Stichproben-Mittelwertes mit einem Populationsparameter bzw. wo kommt der Faktor Wurzel 2 bei abhängigen Stichproben her?* Berlin: Papier aus der Abteilung Methodenlehre.

Kohler, J. & Rindermann, H. (2000 a). Evaluation der Lehre als Maßnahme zur Qualitätssicherung in der Logopädieausbildung – Vorstellung des Reichenauer Modells und seine Effektivitätsprüfung. *L.O.G.O.S. interdisziplinär, 8*(3), 164–171.

Kohler, J. & Rindermann, H. (2000 b). Evaluation der Logopädieausbildung: Beratung der Dozenten und institutionelle Verankerung des Evaluierungsprozesses als Bedingungen für effektive Evaluationen. *L.O.G.O.S. interdisziplinär, 8*(4), 244–252.

Leitner, E. (2001). Die hochschuldidaktische Qualifikation der Lehrenden. *BUKO-Info, 1,* 44–47.

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

Marsh, H. W. & Roche, L. A. (1999). Reply upon SET research. *American Psychologist, 54*(7), 517–518.

McKeachie, W. J. (1997). Student ratings. *American Psychologist, 52*(11), 1218–1225.

Menges, R. J. & Brinko, K. T. (1986). *Effects of student evaluation feedback: A meta-analysis of Higher Education research.* San Francisco: Paper presented at the meeting of the American Educational Research Association.

Patton, M. Q. (1997). *Utilization-focused evaluation.* Thousand Oaks: Sage.

Renkl, A. (2002). Lehren und Lernen. In R. Tippelt (Hrsg.), *Handbuch Bildungsforschung* (S. 589–602). Opladen: Leske + Budrich.

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

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

Rindermann, H. (2002). Evaluation perspectives: An overview of evaluation of communication technologies for education and teaching. In H. H. Adelsberger, B. Collis & J. M. Pawlowski (Eds.), *Handbook on Information Technologies for Education & Training* (S. 309–329). Berlin: Springer.

Rindermann, H. & Kohler, J. (2001). Was hilft, die Ausbildung zu verbessern: Evaluation oder Beratung in einem systemischen Eingriff? *L.O.G.O.S. interdisziplinär, 9*(2), 123.

Spiel, C. & Fischer, U. (1998). Evaluierung eines Weiterbildungsangebots für Hochschullehrende. *Zeitschrift für Hochschuldidaktik, 22*(1), 83–99.

Teichmann, S. (1999). Wirksamkeit der Evaluation von Studium und Lehre. Evaluation im Verbund norddeutscher Universitäten – Follow-up-Elemente des Verfahrens. In J.-H. Olbertz & P. Pasternack (Hrsg.), *Profilbildung, Standards, Selbststeuerung* (S. 241–248). Weinheim: Deutscher Studienverlag.

Webler, W.-D. (1992). Evaluation der Lehre: Praxiserfahrungen und Methodenhinweise. In D. Grühn & H. Gattwinkel (Hrsg.), *Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes* (S. 143–161). Berlin: FU-Dokumentationsreihe.

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

Will, H. & Blickhan, C. (1987). Evaluation als Intervention. In H. Will, A. Winteler & A. Krapp (Hrsg.), *Evaluation in der beruflichen Aus- und Weiterbildung* (S. 43–59). Heidelberg: Sauer.

Winter, M. (2000). Quantitative und qualitative Methoden der Lehrveranstaltungsevaluation. *Handbuch Hochschullehre,* 1–20 (D 2.4). Bonn: Raabe.

Wolf, B. (2001). Effektstärkenmaße. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 96–102). Weinheim: PVU.

Dr. Heiner Rindermann
Institut für Psychologie
Otto-von-Guericke-Universität Magdeburg
Postfach 41 20
D-39016 Magdeburg
E-Mail: heiner.rindermann@gse-w.uni-magdeburg.de

Dipl.-Psych. Jürgen Kohler
Hochschule für Heilpädagogik
Schaffhauserstr. 239
CH-8057 Zürich
E-Mail: juergen.kohler@hfh.ch

Psychologie in Erziehung und Unterricht

Zeitschrift für Forschung und Praxis

Herausgeber
Hellgard Rauh
Ernst Hany
Andreas Krapp
Sabine Walper

PEU: 50 Jahre im Dienst von Wissenschaft und Praxis

Themenheft

Evaluation in Hochschule und Schule

Gast-Mitherausgeber: Willi Hager

- Begriffe, Modelle, Methoden
- Normierung von studentischen Evaluationsfragebögen
- Lehrveranstaltungszufriedenheit und Leistung
- Erste Prüfungen – weiterer Studienerfolg
- Verbesserung der Lehrqualität durch Evaluation und Beratung?
- Evaluation eines schulischen Modellprogramms (QuiSS)

Forum: McKinsey Manifest zur Bildung

50. Jahrgang 1. Quartal

1/2003

50. Jahrgang

ERV reinhardt