Lehr- und Unterrichtsevaluationen durch Studierende und Schüler mittels Ratingskalen. Valide und nützlich oder verzerrt und schädlich?

ARNOLD HINZ

Zusammenfassung
Lehr- und Unterrichtsevaluationen mittels Ratingskalen zur Messung der Lehrqualität werden weltweit vermehrt eingesetzt. In diesem Beitrag werden Hauptlinien der Kontroverse über die Validität der Lehr- und Unterrichtsevaluationen verfolgt. Forschungsbefunde legen nahe, dass die Beziehung zwischen dem Lernerfolg und der durch Ratingskalen gemessenen Lehrqualität sehr gering ist. Forschungsbefunde und Theorien zu den möglichen Biasvariablen *erwartete Benotung* und *physische Attraktivität des Dozenten/der Dozentin* werden exemplarisch dargestellt und diskutiert. Beide Variablen scheinen einen bedeutenden Einfluss auf das Ergebnis von Lehrevaluationen zu haben. Forschungen zur Nützlichkeit und zu den Nebenwirkungen von Lehr-/Unterrichtsevaluationen mittels Ratingskalen werden mit einbezogen. Zuletzt wird die Frage aufgeworfen, ob im Sinne einer Dialektik der Aufklärung ein ursprünglich emanzipatorisch gedachtes Instrument zu einer Quelle von Anpassung und Unterordnung wurde.

Schüsselwörter: *Lehrevaluation, Lehrqualität, Lehreffektivität, Beurteilungsverzerrung, Notenerwartung, physische Attraktivität*

Summary
Students' evaluations of teaching effectiveness/quality (SETs) are increasingly being used throughout the world. This article traces the main lines of the contro-

versy about their validity. Research findings suggest that the relationship between pre/post measurements of learning and SETs is small to zero. In particular, the article reviews and discusses research findings and theories on the potential *grading leniency bias* and *physical attractiveness bias* in SETs. Both seem to have an important impact on students' evaluation of teaching. Research on the usefulness, effectiveness, and of the collateral damages SET's may cause is also taken into consideration. In this light the question is raised whether – in the sense of a dialectic of enlightenment – the emancipatory approach of SET's has turned into a source of subordination and conformity.

Keywords: *Student evaluations, teaching quality, teaching effectiveness, rating bias, expected grades, physical attractiveness*

1. Einleitung

Bei Sichtung der Literatur zu Lehr-/Unterrichtsevaluationen durch Studierende/Schüler mittels Ratingskalen fällt auf, dass die meisten Publikationen entweder vehemente Unterstützungen oder Kritiken sind. Obwohl es einen großen Konsens gibt, dass die Qualität der Lehre und des Unterrichts stärker zu gewichten wäre, sind Validität und Nützlichkeit der Lehr-/Unterrichtsevaluation nicht unumstritten. Der Ton der Vertreter der Validität und Nützlichkeit der *student evaluation of teaching quality/effectiveness* (im Folgenden abgekürzt als *SET*) ist zumeist scharf. So behaupten Marsh und Roche (2000), die Kritiken seien »unbegründet und dubios« und bezögen sich auf »atypische Anekdoten« oder »populäre Mythen« (S. 202, S. 205). Süffisant bemerken sie weiter, die Kritiken seien oft nur auf Internet-Pinwänden zu finden und SETs würden inzwischen weltweit vermehrt eingesetzt. So entsteht der Eindruck, dass sich bei der Debatte über Güte und Nutzen der Lehr-/Unterrichtsevaluation mittels Ratingskalen nicht nur Vertreter und Kritiker, sondern auch Gewinner und Verlierer gegenüberstehen. Tatsächlich erfolgte die Verbreitung der SET bis heute in einem rasanten Tempo, wobei die Entwicklung in Deutschland der in den USA, Kanada, Australien, Großbritannien, Hongkong und einigen anderen Ländern hinterherhinkt (Carr & Hagel, 2008; Centra, 2003; Chen & Hoshower, 2003).

Die Quellen zur Messung der Lehre beziehungsweise des Unterrichts können vielfältig sein. Neben studentischen Ratings sind Selbstbewertungen der Lehrenden

sowie Bewertungen durch Kollegen, externe Experten, Vorgesetzte oder ehemalige Absolventen durchführbar; andere Verfahren sind Interviews mit Studierenden, Videoaufnahmen, die Evaluation der Unterrichtsmaterialien, die Auswertung von Lerntagebüchern der Studierenden, die Auswertung eines Lehr-Portfolios oder die Messung der Lerneffekte bei den Studierenden (Berk 2006). Am weitesten verbreitet ist die Lehr-/Unterrichtsevaluation durch Studierende/Schüler mittels Ratingskalen.

Die erste Skala zur Lehrevaluation wurde 1927 von Hermann Henry Remmers entwickelt (Purdue Rating Scale for Instructors) und von ihm in den folgenden Jahren hinsichtlich ihrer Reliabilität, ihrer Übereinstimmung mit anderen Messungen sowie hinsichtlich ihrer Faktorenstruktur untersucht (Berk, 2006; Smalzried & Remmers, 1947). In den 1960er Jahren erschienen Seminarbewertungen vereinzelt in Studentenzeitungen. In den 1970er Jahren verwendeten etwa die Hälfte der Hochschulen in den USA Ratingskalen zur Evaluation, wobei es noch wenig Forschung hierzu gab. Gegenwärtig setzen 97 % aller amerikanischen Hochschulen Ratingskalen zur Lehrevaluation ein, wobei die Ergebnisse der Lehrevaluation in den USA ein wichtiges Kriterium für Gehalts-, Beschäftigungs- und Förderentscheidungen sind. Seit 1999 gibt es neben der offiziellen Lehr- und Unterrichtsevaluation die weit verbreitete und von überall einsehbare Internetevaluation RateMyProfessors.com (USA, Kanada, Großbritannien) sowie seit 2001 zusätzlich RateMyTeachers.com (USA, Kanada, Großbritannien, Irland, Australien, Neuseeland). Anfang der 1990er Jahre wurden an vielen deutschen Universitäten Lehrevaluationen erprobt, Mitte der 1990er Jahre erfolgte eine starke Ausbreitung. Bei Stellenbesetzungen werden vermehrt die Ergebnisse vergangener Lehrevaluationen eingefordert; 2008 forderte der Wissenschaftsrat eine lehrleistungsabhängige Mittelzuweisung. Ähnlich wie in den USA gibt es in Deutschland, Österreich und der Schweiz seit 2005 neben der offiziellen Lehrevaluation die Internetplattform MeinProf.de/at/ch sowie für die Schule seit 2007 Spickmich.de (siehe Hinz 2011).

2. Zielexplikation

Die Gütekriterien, die üblicherweise zur Prüfung der Qualität von Forschungen angelegt werden, sind auch an Lehr- und Unterrichtsevaluationen mittels Ratingskalen anzulegen. Zu prüfen ist, ob diese objektiv, reliabel, valide, nützlich, öko-

nomisch, vergleichbar, generalisierbar und ethisch vertretbar sind. In diesem Beitrag sollen zwei dieser Gütekriterien betrachtet werden: die Validität und die Nützlichkeit.

3 Repräsentativität, Objektivität und Reliabilität

Repräsentativität der Stichprobe, Objektivität und Reliabilität sind Voraussetzungen der Validität. Da eigentlich immer eine Totalerhebung angestrebt wird und es nur wenige Studierende gibt, die das Ausfüllen von Lehrevaluationsbögen verweigern, ist in solchen Untersuchungen zumeist eine hohe Repräsentativität der Stichprobe gewährleistet. Bei Nichtpflichtveranstaltungen kann jedoch bereits eine *Abstimmung mit den Füßen* erfolgt sein, sodass nur noch diejenigen anwesend sind, die eine positive Meinung zu der Lehrveranstaltung haben oder diese unbedingt absolvieren müssen (dropout bias). Bei internetgestützen Evaluationen (beispielsweise über stud.ip) wird die Repräsentativität bei einer hohen Verweigerungsquote eingeschränkt (nonresponse bias) (Meinefeld 2010). Die Durchführungs-, Auswertungs- und Interpretationsobjektivität der SET-Ratings kann als hoch angesehen werden. Es gibt nur leichte mögliche Beeinträchtigungen: Wenn Studierenden gesagt wird, dass ihr Rating über die Bezahlung, Förderung oder Weiterbeschäftigung des Dozenten/der Dozentin mitentscheidet, neigen sie zur Milde. Ebenfalls zu milderen Urteilen neigen sie bei Anwesenheit des Dozenten/der Dozentin (Cashin 1995; Ory 2001). Die Korrelation zwischen zwei Studierenden derselben Veranstaltung (single rater reliability) ist sehr gering, die Interraterreliabilität der Gesamtgruppe ist hingegen insbesondere bei vielen Ratern hoch. Bei fast allen eingesetzten Ratingskalen zur Lehrevaluation erwiesen sich die internen Konsistenzen und die Retestreliabilität als zufriedenstellend (Cashin 1995; Daniel 1998; Marsh 2007; Marsh/Ball 1989; Rindermann 1996).

4. Validität

Während Repräsentativität, Objektivität und Reliabilität der SET-Ratings somit als gegeben beziehungsweise als herstellbar angesehen werden können, liegen die Hauptprobleme im Bereich der Validität. Die Grundfrage der Validität lautet: *Wird wirklich das gemessen, was gemessen werden soll?* Mit den SET-Ratingskalen soll Lehrqualität gemessen werden. Wird jedoch wirklich Lehrqualität gemessen, also etwas, das zum Dozenten/zur Dozentin gehört? Oder messen »student ratings«

etwas, das zur Seite der Studierenden gehört, wie Wohlbefinden, Zufriedenheit, Interessiertheit, die Häufigkeit des Lachens? Bei der Einschätzung der Qualität eines Kinofilms gelten Zuschauerzahlen und -bewertungen eher nicht als ausreichend. Stattdessen werden hier, etwa anlässlich einer Preisvergabe, Experten herangezogen. Auch Preise für Literatur werden nicht nach Verkaufszahlen vergeben: Hohe Verkaufszahlen gelten hier manchmal sogar als Indiz für miserable Literatur. Verraten Lehr- und Unterrichtsevaluationen mehr über den Geschmack und das Niveau der Studierenden/Schüler als über die Qualität der Lehre/des Unterrichts? In diesem Fall wäre die SET nicht valide als Instrument zur Messung der Lehrqualität, sondern wäre so etwas wie eine Hitparade, wobei eine hohe Interraterreliabilität nur die Konformität des Geschmacks belegen würde. In seiner *Politik* führt Aristoteles an, dass an sich zwar nur Fachleute über Fachleute urteilen sollten und dass ein Laie schlechter urteilt als ein Fachmann, dass aber die Menge, wenn sie »nicht gar zu sklavenartig« sei, »besser oder doch nicht schlechter« (2005: 53-54) wie ein Fachmann urteile. Zudem gebe es Fälle, in denen der Laie besser urteile als der Fachmann: so könne der Hausherr das Haus besser beurteilen als der Baumeister und der Gast könne das Essen besser beurteilen als der Koch. Möglicherweise ist die studentische Urteilsfähigkeit doch eher mit der des Hausherrn oder Gastes als mit der des Kinobesuchers oder Romanlesers zu vergleichen.

Unter dem Aspekt der *Inhaltsvalidität* ist zu definieren, was Lehrqualität ist. Es gibt im angloamerikanischen und deutschen Sprachraum viele unterschiedliche Versuche einer Definition von Lehr- oder Unterrichtsqualität. Vorstellungen von guter und effektiver Lehre sind bei der Skalenkonstruktion Voraussetzung für die Auswahl von geeigneten Items. Die meisten Ratingskalen entstanden durch die Auswahl von Items durch Experten, aus der Analyse bereits vorhandener Messinstrumente sowie durch Befragungen von Studierenden und Lehrenden. Da Unterrichten eine vielschichtige Tätigkeit ist, wurden SET-Ratingskalen in der Regel multidimensional angelegt und die Dimensionen durch Faktorenanalysen überprüft. Bei dem häufig verwendeten SEEQ (Students' Evaluation of Educational Quality) werden neun Dimensionen berücksichtigt: Erkenntnisgewinn (Lernen), Enthusiasmus des Dozenten/der Dozentin, Organisation (Struktur), Gruppeninteraktion, Beziehung des Dozenten/der Dozentin zu den

Studierenden (Wertschätzung, Interesse), Breite des Inhalts, Fairness der Benotung, Qualität der Arbeitsaufträge und Anspruchsniveau (Marsh 2007).

In Frage steht, ob Studierende bei Lehrevaluationen wirklich zwischen den verschiedenen Dimensionen der Lehrqualität unterscheiden oder nicht vielmehr von einer Dimension auf alle anderen schließen. Von einem *Haloeffekt* spricht man, wenn ein Beurteiler aufgrund eines Merkmals einen globalen Eindruck einer anderen Person bildet und dann nicht mehr in der Lage ist, spezifische Charakteristiken zu beurteilen (Feeley 2002). Unter den angeführten Dimensionen des SEEQ kann der Enthusiasmus des Dozenten/der Dozentin als ein Faktor gelten, der im Sinne eines Haloeffektes auf die anderen Dimensionen abfärbt. So fanden beispielsweise Williams und Ceci (1997), dass bei einem enthusiastischeren Präsentationsstil (nach vorheriger Schulung) nicht nur der Vortragsstil, sondern auch die Fairness der Notengebung, die Organisation des Kurses, die Erreichbarkeit des Dozenten/der Dozentin, der eigene Erkenntnisgewinn und das identische Vorlesungsskript besser beurteilt wurden. Wie stark die Expressivität des Dozenten/der Dozentin auf die SET-Gesamtbewertung einwirkt, zeigt auch eine Studie von Ambady und Rosenthal (1993). Beim Vorspielen von videografierten Unterrichtsstunden vor fremden Ratern ergaben sich beeindruckende Korrelationen zwischen der Enthusiasmuseinschätzung der Rater (die nur 30 Sekunden ohne Ton gesehen hatten) und der SET-Bewertung der Studierenden am Semesterende. Angesichts dieser Befunde kann bezweifelt werden, dass Studierende bei der Evaluation von Lehrveranstaltungen unbeeinflusst von ihrem Gesamteindruck die einzelnen Dimensionen der Lehrqualität valide beurteilen. Eine hohe Expressivität des Vortragenden kann schließlich sogar dazu verleiten, inhaltlichen Unsinn zu übersehen, wie Naftulin, Ware und Donnelly (1973) mit ihrem berühmten Dr.-Fox-Experiment zeigen konnten (ein Charakterschauspieler wurde als Dr. Fox vorgestellt und hielt einen gesten- und mimikreichen, humorvollen und unterhaltsamen, aber inhaltlich bedeutungslosen Vortrag).

Indizien für *Kriteriumsvalidität* der SET-Messungen wären hohe Übereinstimmungen mit anderen Formen der Lehrevaluation. Verschiedene Studien zeigen eine geringe Übereinstimmung der SET mit der Selbstevaluation und eine mäßige bis mittelhohe Übereinstimmung mit der Expertenmeinung (Cashin 1995; Feldman 1989; Rindermann 1996). Eine recht hohe Übereinstimmung ergibt sich bei einem Vergleich der SET mit den Werten der RateMyProfes-

sors.com-Evaluation. Coladarci und Kornfield (2007) fanden zwischen dem SET-Wert der offiziellen Lehrevaluation und dem Overall-Quality-Wert der RateMyProfessors.com-Bewertung eine Korrelation von $r = .68$. Dieser Befund ist allerdings zunächst einmal nur ein Indiz für eine gewisse Güte der RateMyProfessors.com-Bewertungen, löst aber nicht das grundsätzliche Problem, ob mit den Ratings wirklich Lehrqualität oder studentische Zufriedenheit gemessen wird.

Mit Blick auf die *Konstruktvalidität* der student ratings ist zu erwarten, dass eine hohe Lehrqualität zu einem hohen Lernerfolg führt. Viele Autoren behaupten explizit, teilweise sogar in der Überschrift, dass mit den student ratings unmittelbar Lehreffektivität (teaching effectiveness) gemessen werde (so Aleamoni 1999; Berk 2005; Cashin 1995; Chen/Hoshover 2003; Kulik 2001; Marsh 2007; Marsh/Roche 1997; McKeachie 1997; Wachtel 1998). Hohe SET-Werte müssten dann mit einem hohen Lernerfolg einhergehen, wobei der Lernerfolg (wie in der Interventionsforschung üblich) durch Prä-Post-Messungen erhoben werden müsste. In einer älteren Metaanalyse kommt Cohen (1981) zu einer mittelhohen Korrelation zwischen SET und Lernerfolg, wobei allerdings keine Prä-Messungen durchgeführt wurden, so dass diese Studien nicht berücksichtigt werden können. In einer neueren Metaanalyse fand Clayson (2009) nur eine geringe Beziehung zwischen SET und dem Lernerfolg, wobei die Beziehung umso geringer war, je objektiver die Messung des Lernerfolgs war. So fanden Arthur, Tubré, Paul und Edens (2003) bei einer Stichprobe von 652 Studierenden unter Zugrundelegung der SET-Werte nur eine Varianzaufklärung von 1.5 % für den Prä-Posttest-Lernerfolg. Stark-Wroblewski, Ahlering und Brill (2007) ermittelten zwischen den SET-Werten und dem durch Prä-Post-Messungen erhobenen Lernerfolg nur eine geringe und nicht signifikante Korrelation ($r = .15$ bei $p = .06$). Delucchi und Pelowski (2000) fanden bei Regressionsberechnungen ($N = 1145$) einen signifikanten Effekt der Einschätzung der Liebenswürdigkeit des Dozenten/der Dozentin auf die Einschätzung der Lehrqualität (= .34), nicht aber auf den Lernerfolg der Studierenden (= -.04). Man könnte die Diskussion an dieser Stelle mit der Feststellung beenden, dass die Lehrqualität gemessen durch SET keinen bedeutenden Beitrag zur Varianzaufklärung des studentischen Lernerfolgs leistet. Möglicherweise ist aber der *Lernerfolg gar kein vernünftiges Kriterium für Lehrqualität*. Zur Erläuterung ein Beispiel: Extremer Leistungsdruck, eine hohe Durchfallquote sowie »teaching to the test« bei einem Seminar könnten zu einem hohen

Prä-Posttest-Lernerfolg führen, aber auf Kosten des studentischen Engagements bei anderen Seminaren, auf Kosten der Freude am Studieren sowie auf Kosten wichtiger, aber nicht testrelevanter Inhalte. Ein Beweis für hohe Lehrqualität ist der Lernerfolg in diesem Fall nicht (Kulik 2001).

4.1 Biasvariablen

Eine andere Annäherung an die Frage nach der Validität der SET ist die Untersuchung möglicher Biasvariablen. Hierbei handelt es sich um Variablen, die in Verdacht stehen, das studentische Urteil in unzulässiger Weise zu verfälschen. Während die Verteidiger der SET sich um den Nachweis bemühen, dass es kaum Verfälschungen des studentischen Urteils gibt, verweisen die Kritiker auf den Einfluss dieser Biasvariablen. Zudem gibt es noch eine dritte Gruppe, die die Auffassung vertritt, dass nachgewiesene Biasvariablen als Korrektiv bei der Interpretation von Lehrveranstaltungsevaluationen verwendet werden sollten (Cashin 1995). Der Vorteil der Diskussion über die verschiedenen Biasvariablen ist, dass die Debatte über die Validität der Lehrveranstaltungsevaluation hierdurch eine empirische Grundlage erhält. Die Liste der diskutierten Biasvariablen ist lang. Empirische Befunde gibt es unter anderem zur Teilnehmerzahl, zum Vorinteresse (Pflicht- versus Wahlveranstaltung), zum Niveau des Kurses bzw. der Teilnehmer, zum Fach, zur Veranstaltungszeit, zum Eindruck der letzten Sitzungen (Rezenzeffekt), zum Umfang der Hausaufgaben, zur wahrgenommenen Ähnlichkeit von Dozent/in und Studierendem sowie zu rassischer Herkunft, Alter, Geschlecht, Reputation, Persönlichkeit, Stimme oder politischer Meinung des Dozenten/der Dozentin. Im Folgenden werden exemplarisch die Befunde zu zwei möglichen Biasvariablen dargestellt.

4.1.1 Biasvariable: Erwartete Benotung

Es ist unbestritten, dass es eine signifikante Korrelation zwischen der SET und der erwarteten Benotung gibt (Mason/Edwards/Roach 2002; Marsh/Roche 1997). Umstritten ist jedoch, ob und welche Kausalität hinter dieser Korrelation steckt. Diskutiert werden vier Erklärungsmodelle.

Vertreter der SET (Centra 2003; Marsh/Roche 1997, 2000) engagieren sich für die *Validitätshypothese*. Hier wird behauptet, dass sich gute Lehrqualität sowohl auf die SET-Bewertung als auch auf den Lernerfolg und damit auch auf

die Noten auswirkt, da Studierende bei guter Lehrqualität auch gut und leicht lernen.

Die direkte Gegenthese ist die *Grading Leniency-Hypothese*, die in der neueren Forschung umfassender als *Reziprozitätshypothese* (Wolbring/Hellmann 2010) bezeichnet wird (»Wie du mir, so ich dir«-Hypothese). Hier wird die Auffassung vertreten, dass Studierende ihre Dozent(inn)en nach der Milde oder Strenge der Benotung beurteilen. Wenn der Dozent/die Dozentin milde bewertet, ist man bereit, ihn ebenfalls milde zu bewerten (Greenwald/Gillmore 1997). Bewertet der Dozent/die Dozentin hingegen streng, so bewerten auch die Studierenden streng. Bei der Grading Leniency-Hypothese liegt die Betonung auf dem Effekt der milden Benotung. Nach der Reziprozitätshypothese kann auch die Notengebung des Dozenten/der Dozentin durch SET-Bewertungen beeinflusst werden, vorausgesetzt, die SET-Bewertung findet etwa in der Mitte des Semesters statt und der Dozent/die Dozentin erhält frühzeitig die Ergebnisse. Schlecht bewertete Dozent(inn)en könnten sich über die Notengebung rächen, gut bewertete hingegen aus Dankbarkeit zu milden Noten neigen.

Eine nichtlineare Beziehung zwischen den Studierendennoten und der SET-Bewertung erwartet die *Attributionshypothese*. Hier wird angenommen, dass Studierende, die eine gute Benotung erwarten, ihr SET-Urteil unbeeinflusst abgeben, da sie die gute Benotung mit eigener Intelligenz und/oder Anstrengung attribuieren. Studierende hingegen, die eine schlechte Benotung erwarten, neigen nach der Attributionshypothese dazu, zur Erhaltung ihres Selbstwerts sowie zur Auflösung einer sonst bestehenden kognitiven Dissonanz ihre schlechte Leistung mit der schlechten Lehre oder dem schlechten Charakter des Dozenten/der Dozentin zu erklären (selbstwertdienlicher Attributionsfehler/self-serving bias).

Eine weitere Hypothese zur Erklärung der Korrelation zwischen Studierendennoten und SET ist die *Prior Characteristics-Hypothese*. Hier wird angenommen, dass sowohl die Studierendennoten als auch die SET-Noten abhängige Variablen einer oder mehrerer unabhängiger Variablen sind wie beispielsweise Vorinteresse, Motivation, äußere Seminarbedingungen etc. Wer mit einem hohen Vorinteresse ein Seminar besucht, wird nach dieser Hypothese unabhängig von der Lehrqualität das Seminar gut finden und auch eine gute Note erzielen, so dass sich eine Korrelation zwischen SET-Wert und erwarteter Benotung ohne Kausalzusammenhang ergibt.

Welche dieser Hypothesen zum Zusammenhang zwischen SET und Notenerwartung ist die richtige? Für alle angeführten Hypothesen gibt es sowohl empirische Befunde als auch gute theoretische Grundlagen. Eine Bias-Variable ist die erwartete Benotung nur dann, wenn entweder die Reziprozitätshypothese (Grading Leniency-Hypothese) oder die Attributionshypothese zutreffen. In einer Studie von Boysen (2008) gaben zwar nur 8 % der Studierenden zu, dass sie selbst aus Rache über eine schlechte Note schon einmal eine schlechte Bewertung abgaben; immerhin 30 % der Befragten hatten aber schon von anderen Studierenden gehört, dass sie aus Rache schlechtere Beurteilungen abgaben. Marsh und Roche (2000) berichten von experimentellen Untersuchungen mit willkürlich manipulierten Notenerwartungen unmittelbar vor Durchführung einer SET. Diese belegten vor allem einen Effekt schlechter Noten auf die SET. Auch Wolbring und Hellmann (2010) fanden in einem experimentellen Design (schwerer Leistungstest vor versus nach der SET-Bewertung) Belege für die Grading Leniency-Hypothese. Vor dem Hintergrund solcher Belege wurde der Vorschlag gemacht, als Korrektiv einen SET-Bewertungsabschlag bei Dozent(inn)en einzuführen, die überwiegend sehr gute Noten vergeben. Hiergegen wurde jedoch eingewendet, dass die sehr guten Noten im Sinne der Validitätshypothese auch durch eine gute Lehre entstanden sein können und dass dadurch Dozent(inn)en mit leistungsstarken Studierenden bestraft würden.

Unabhängig von den Belegen für die Grading Leniency Hypothese und/oder Attributionshypothese wirken beide auf die Notengebung. Nach einer Erhebung von Birnbaum (1998) glauben 2/3 der Hochschullehrenden, dass eine Erhöhung der Seminaranforderungen zu schlechteren Evaluationsergebnissen führen würde. Allein der Glaube an die Grading Leniency Hypothese oder die Attributionshypothese verursacht möglicherweise eine Noteninflation. Wenn die SET-Bewertungen Auswirkungen auf das Gehalt und die Karriere der Dozent(inn)en haben, so ist nach der Grading Leniency Hypothese eine Noteninflation zu erwarten, da dann die Dozent(inn)en lieber sehr gute Noten geben, um selbst gut bewertet zu werden. Eiszler (2002) belegte für eine amerikanische Universität, dass es dort in den 1990er Jahren eine deutliche Noteninflation gab, während die Noten in den 1980er Jahren relativ konstant geblieben waren. Nach Ausschluss alternativer Erklärungen kommt er zu dem Schluss, dass das Streben nach guten SET-Noten zu einer Senkung der Anforderungen führte.

Eine Schwäche der angeführten Studien ist, dass nur die erwartete Note, nicht aber die Beziehung zwischen Leistung und erwarteter Note erhoben wurde. Eine positive Notenerwartung kann damit zu tun haben, dass Noten *verschenkt* werden, kann aber auch Ergebnis von Anstrengung sein. Auf der Internetseite RateMyProfessors.com wird nicht nach der erwarteten Note gefragt, sondern unmittelbar danach, wie viel Arbeit man investieren muss, um eine gute Note zu erhalten (»easiness«). Die Milde der Bewertung wird also unmittelbar erhoben (»Is this class an easy A? How much work do you need to do in order to get a good grade?«). Felton, Mitchell und Stinson (2004) berechneten mit den Daten von RateMyProfessors.com zwischen »easiness« und »overall quality« eine hohe Korrelation von r = .62 (N = 6852). Somit werden 38 % der Varianz der SET durch niedrige oder hohe Ansprüche des Dozenten/der Dozentin aufgeklärt. Da die Korrelation zwischen den RateMyProfessors.com-Werten für »overall quality« und den offiziellen Hochschulevaluationswerten ebenfalls hoch ist (Coladarci/Kornfield 2007), erklärt sich ein erheblicher Teil der Varianz der SET-Bewertung mit der Höhe des Aufwands für gute Noten.

Auch wenn es für jede der hier angeführten Theorien zum Zusammenhang zwischen SET und der erwarteten Benotung Belege gibt, so fällt doch auf, dass es besonders gute Belege für die Grading Leniency Hypothese gibt (die Selbstaussagen der Studierenden, die experimentellen Befunde und die deutliche Korrelation zwischen »easiness« und »overall quality«).

4.1.2 Biasvariable: Physische Attraktivität des Dozenten/der Dozentin

Ein besonders wichtiger Haloeffekt ist die Wirkung der physischen Attraktivität auf die Wahrnehmung. In der inzwischen längeren Tradition der Attraktivitätsforschung ist gut belegt, dass physisch attraktive Menschen mehr Aufmerksamkeit erfahren, eher Hilfe erhalten, populärer sind, mehr Freundschaften, Verabredungen und sexuelle Erfahrungen haben, seltener verurteilt werden und geringere Strafen erhalten, ein höheres Gehalt beziehen und seltener entlassen werden, als motivierter gelten, besser überzeugen können, mehr Wählerstimmen erhalten und als intelligenter und kompetenter angesehen werden (Hamermesh/Biddle 1994; Hassebrauck/Niketta 1993; Henss 1998; Renz 2006). Vor dem Hintergrund dieser Befunde würde es verwundern, wenn die physische Attraktivität keinen Einfluss auf die SET-Beurteilungen hätte. Eine methodische Voraussetzung

für Studien zum Zusammenhang zwischen physischer Attraktivität und SET ist, dass die Einschätzung der physischen Attraktivität von anderen Personen erfolgt als von denen, die die SET-Bewertung abgeben. Sonst kann man argumentieren, dass eine Korrelation zwischen physischer Attraktivität und SET nur dadurch entsteht, dass die Zufriedenheit (oder Unzufriedenheit) mit einer Lehrveranstaltung dazu führt, dass man den Dozenten als schön (oder hässlich) ansieht. Eine solche Abfärbung ist besonders dann zu erwarten, wenn die Probanden unmittelbar nach der SET-Bewertung die physische Attraktivität des Dozenten/der Dozentin einschätzen sollen (empirischer Beleg hierzu in Feeley 2002). In vielen Studien (Bonds-Raacke/Raacke 2007; Felton/Koper/Mitchell/Stinson 2008; Felton/Mitchell/Stinson 2004; Freng/Webber 2009; Kindred/Mohammed 2005; Riniolo/Johnson/Sherman/Misso 2006) wurden die Angaben zur »hotness« in RateMyProfessors.com mit den Angaben zur »overall quality« in Beziehung gesetzt (die Korrelation liegt in den verschiedenen Studien zwischen $r = .31$ bis .64). Die Kausalrichtung ist hierbei unklar: Führt hohe Lehrqualität zur Wahrnehmung als »hot« oder führt »hotness« zur Wahrnehmung einer hohen Lehrqualität? Dies gilt auch für die Studien von Hultman und Oghazi (2008) sowie von Gurung und Vespia (2007), da in diesen die Einschätzung der physischen Attraktivität und der Lehrqualität ebenfalls durch dieselben Personen erfolgte.

Bei Ausschluss aller Studien mit identischen Schönheits- und SET-Ratern verbleiben sechs Untersuchungen (Bokek-Cohen/Davidowitz 2008a, 2008b; Hamermesh/Parker 2003; Klein/Rosar 2006; Rosar/Klein 2009, 2010; Süßmuth 2006; Wolbring 2010). Tabelle 1 gibt einen Überblick über die jeweilige Stichprobe, über die Anzahl der beurteilten Dozenten und Dozentinnen, über die Anzahl der Attraktivitätsrater und deren Beurteilerübereinstimmung, über die Korrelationswerte zwischen SET- und Attraktivitätsnote sowie über den maximalen Effekt der physischen Attraktivität auf den Lehrevaluationswert.

Mit Ausnahme der Studie von Rosar und Klein (2009, 2010), die sich auf MeinProf.de bezieht, verwenden alle Studien die vorhandenen Lehrevaluationsergebnisse ihrer Hochschulen. Die Attraktivitätseinschätzung erfolgte durch Rater unterschiedlichen Geschlechts, die die physische Attraktivität anhand eines Portraitfotos einschätzten. Nur in der Studie von Bokek-Cohen und Davidowitz (2008a, 2008b) erfolgte die Attraktivitätseinschätzung durch die Studierenden eines Semesterdurchgangs am Semesterende unmittelbar vor der SET-Erhebung.

	Stichprobe Anzahl Bewertungen Ort	*Dozenten N (♂,♀)*	*Rater_ für phys. Attrakt. N (♂,♀)*	*Beurteiler-übereinstimmung*	*Korrelation SET– Attr. Männl. Doz (weibl. Doz.)*	*Maximale SET-Noten-Verbesserung*
Bokek-Cohen & Davidowitz, 2008a, 2008b	1388 Stud., Ariel University, Israel	49 (31, 18)	1388		r = .77 (r = .37)	
Hamermesh & Parker, 2003	16957, University of Texas, Austin	94 (54, 40)	6 (3, 3)	.91	R^2 =.36 (R^2 = .16)	1.0
Klein & Rosar, 2006	1004 Seminare, Uni Köln	206 (174, 32)	25-36	.95		0.6
Rosar & Klein, 2009, 2010	56245, MeinProf.de	2745 (2466, 279)	24 (12, 12)	.95	r = .14	0.6
Süssmuth, 2006	111 Sem., LMU München	50	48 (33, 15)			0.5
Wolbring, 2010	12073, LMU München	110 (69, 41)	20 (11, 9)	.95	r = -.20 (r = .03)	0.8

Tabelle 1: Studien mit unterschiedlichen Ratern für Physische Attraktivität und SET

Verwendet wurde dann aber nicht diese SET-Bewertung, sondern eine frühere SET-Erhebung, damit die Attraktivitätseinschätzung und die SET-Bewertung durch unterschiedliche Personen erfolgt. Beim Vergleich der Studie von Bokek-Cohen und Davidowitz mit den anderen Studien ist anzunehmen, dass in dieser Studie der Attraktivitätseinfluss eher überschätzt, in den anderen Studien hingegen unterschätzt wird. In die Bewertung der Attraktivität gehen in der Studie von Bokek-Cohen und Davidowitz vermutlich nicht nur Gesichtsschönheit, Statur und Kleidung ein, sondern auch die Stimme, das nonverbale Verhalten und möglicherweise zusätzlich als Bias das Lehrverhalten. In den anderen Studien wird der Attraktivitätseinfluss eher unterschätzt, da erstens die Attraktivitätseinschätzung nur aufgrund von Internetportraitfotos erfolgte (wobei in den meisten Studien diese Fotos einheitlich in Schwarz-Weiß-Bilder übertragen wurden), da zweitens wenig attraktive Dozenten vermutlich seltener ihr Foto ins

Internet stellen, so dass auf diese Weise Varianz verloren geht (was problematisch ist, wenn sich der Attraktivitätseffekt vor allem im Bereich zwischen »unattraktiv« und »mittelattraktiv« zeigt) und da drittens damit zu rechnen ist, dass Internetfotos von Dozenten nicht immer aktuell sind (Hamermesh/Parker 2003).

Ein Befund aller Studien war, dass die SET-Bewertungen der männlichen Dozenten mit physischer Attraktivität korrelierten. Für die männlichen Dozenten ergab sich in der Studie von Bokek-Cohen und Davidowitz (2008a, 2008b) eine Korrelation von r =.77 zwischen der SET-Bewertung durch Studentinnen und der physischen Attraktivität, was bedeutet, dass nahezu 60 % der Varianz der SET-Bewertung durch die physische Attraktivität aufgeklärt wird, wobei allerdings in dieser Studie aus den dargelegten Gründen der Einfluss der physischen Attraktivität überschätzt wird. In der amerikanischen Studie von Hamermesh und Parker (2003) konnte bei Dozenten 36 % und bei Dozentinnen 16 % der Varianz des SET-Wertes durch die physische Attraktivität aufgeklärt werden. Wenn man sich auf der Schönheitsskala von einer Standardabweichung unterhalb des Mittelwertes hin zu einer Standardabweichung oberhalb des Mittelwertes bewegte, verbesserte sich bei den Dozenten die SET-Note um 0.46 Punkte (nahezu eine Standardabweichung). Süssmuth (2006) fand in einer deutschen Studie im Anschluss an Hamermesh und Parker nur einen halb so großen Einfluss der physischen Attraktivität. In seiner Erhebung sollten die Schönheitsrater allerdings vom Alter der Dozent(inn)en absehen; zudem zeigte sich, dass seine Rater (BWL-Studierende) stark auf attraktive Kleidung (»business look«) ansprachen. Klein und Rosar (2006) fanden als Attraktivitätseffekt eine Verbesserung des SET-Wertes um maximal 0.6 Punkte, was etwas mehr als die Standardabweichung war. Bei Einrechnung von Kontrollvariablen fanden sie, dass der Attraktivitätseffekt für Dozenten und Dozentinnen gleich groß war. In einer weiteren Studie fanden Rosar und Klein (2009) anhand der Daten in MeinProf.de ebenfalls einen Attraktivitätsbonus auf die SET-Bewertung im Umfang von 0.6 (für »Spaß am Kurs«), allerdings nur für Dozenten und nicht für Dozentinnen. Auch Wolbring (2010) fand für Dozenten bei einer Veränderung von 2 auf 7 beim Attraktivitätsindex (von 1 = sehr unattraktiv bis 10 = sehr attraktiv) eine Verbesserung der SET-Bewertung um fast eine ganze Note (0.8). Bei Einrechnung von Kontrollvariablen wie Vorinteresse, Veranstaltungsart, Teilnehmerzahl blieb dieser Attraktivitätseffekt erhalten und zeigte sich dann auch für Dozentinnen.

In den Studien von Hamermesh und Parker (2003), Süssmuth (2006), Klein und Rosar (2006) sowie Rosar und Klein (2009) konnten die SET-Bewertungen nicht auf das Geschlecht der SET-Rater bezogen werden. Dies war nur in der Studie von Bokek-Cohen und Davidowitz (2008a, 2008b) sowie in der Studie von Wolbring (2010) möglich. Der Attraktivitätseinfluss erwies sich in diesen Studien beim gegengeschlechtlichen Rating als stärker, was mit einer größeren Sensibilität gegenüber der Attraktivität des anderen Geschlechts zu tun haben könnte (Bokek-Cohen/Davidowitz 2008a). Der stärkste Attraktivitätseffekt zeigte sich, wenn Studentinnen männliche Dozenten beurteilten. Frauen geben zwar anders als Männer in Befragungen immer wieder an, dass physische Attraktivität für sie bei der Partnersuche weniger wichtig sei, experimentelle Studien in der Attraktivitätsforschung belegen aber das Gegenteil (Ha/Overbeek/Engels 2010; Hadjistavropoulos/Genest 1994; Luo/Zhang 2009).

Im Vergleich zu ihren männlichen Kollegen profitieren Dozentinnen weniger stark von physischer Attraktivität. Dies könnte damit zu tun haben, dass sich männliche Studenten eher für die Attraktivität jüngerer Frauen interessieren. Eine weitere Erklärung für den geringeren Beautyeffekt bei Dozentinnen ist, dass physische Attraktivität mit einer geschlechtstypischeren Wahrnehmung einhergeht. Physisch attraktive Männer werden eher als typisch »männlich«, physisch attraktive Frauen eher als typisch »weiblich« wahrgenommen, was beim traditionell eher als männlich wahrgenommenem Beruf des Hochschullehrers ein Nachteil ist (Bokek-Cohen/Davidowitz 2008a).

Die dargelegten Befunde zeigen, dass Dozent(inn)en mit hoher Attraktivität bessere SET-Noten erhalten. *Ist aber physische Attraktivität überhaupt ein Bias-Faktor oder sind die Veranstaltungen der schönen Dozent(inn)en einfach besser?* Zu unterscheiden ist zwischen Diskriminierungs- und Produktivitätseffekten der physischen Attraktivität. Ein Diskriminierungseffekt liegt vor, wenn eine geringe physische Attraktivität Ursache einer schlechten SET-Bewertung ist. Hierzu gehört auch die Diskriminierung in entgegengesetzter Richtung, wenn nämlich hohe Attraktivität zu hohen Erwartungen führt und dann aus Enttäuschung heraus attraktive Dozent(inn)en besonders abgestraft werden (»Beauty-is-Beastly-Effekt«). Neben diesen Diskriminierungseffekten gibt es auch Produktivitätseffekte der physischen Attraktivität. So werden physisch attraktive Menschen von Geburt an besser behandelt und gefördert, was zu einem höheren Selbstwertgefühl und

zu erhöhter Produktivität führen kann (lebensgeschichtlicher Produktivitätseffekt). Physische Attraktivität des Dozenten/der Dozentin führt aber auch zu einer größeren Aufmerksamkeit, zu mehr Mitarbeit und häufigerer Anwesenheit der Studierenden (Attractiveness Attention Boost), was eine positive Wirkung auf die Sicherheit und Zufriedenheit des Dozenten/der Dozentin hat, der/die dann besser motiviert ist, Energie in die Lehre zu stecken, so dass interaktionistisch ein Verstärkungskreislauf in Gang kommt, der die Lehrqualität und die SET-Bewertungen erhöht (interaktionistischer Produktivitätseffekt). Wolbring und Hellmann (2010) konnten in einer experimentellen Studie Produktivitätseffekte der physischen Attraktivität belegen. Obwohl jeweils derselbe 11-minütige Vortrag gehört wurde, lernten die Studierenden bei Vorlage der Fotographie eines/r attraktiven Dozenten/in besser als bei der Fotographie eines/r unattraktiven Dozenten/in.

5. Nützlichkeit

Die Einführung der SET erfolgte an den meisten Hochschulen vor dem Hintergrund der Erwartung, dass sich hierdurch die Lehrqualität verbessern ließe. Rindermann (2001) unterscheidet zwischen drei Hypothesen zur Wirksamkeit der SET: die Sensibilisierungshypothese, die Feedbackhypothese und das hochschuldidaktische Diskursmodell. Bei der Sensibilisierungshypothese wird angenommen, dass alleine die Durchführung der SET zu einer Sensibilisierung für Lehrqualität und damit zu einer Verbesserung der Lehre führt. Beim Feedbackmodell geht man davon aus, dass die individuelle Rückmeldung über die SET-Werte zu einer Verbesserung in den Bereichen motiviert, in denen man schlechter abgeschnitten hat. Beim hochschuldidaktischen Diskursmodell wird angenommen, dass ein Diskurs mit den Seminarteilnehmern über die SET-Ergebnisse (etwa in der Mitte des Semesters) dazu führt, dass sich die Lehrqualität erhöht. Bei Analysen mit verschiedenen Datensätzen berechnete Rindermann (2001) für alle drei Modelle Effektstärken, die sich um $d = 0$ bewegten. Vor dem Hintergrund, dass die Dozenten sich vermutlich kurz vor der SET-Wiederholungsmessung besonders anstrengten und sich die Studierenden durch die zweite Messung geschmeichelt fühlen konnten und somit Milde zu erwarten war, ist dies ein erstaunlicher Befund. Auch Marsh und Hocevar (1991) fanden bei einer Stichprobe von 195 Lehrern bei 13-jähriger SET-Erhebung keine signifikanten Effekte trotz ständigen

SET-Feedbacks. Ebenfalls keine signifikante Verbesserung der Lehrqualität fanden Kember, Leung und Kwan (2002) bei einer drei- bis vierjährigen Längsschnittstudie. In einer Studie über vier aufeinander folgende Semester fanden Lang und Kersting (2007) einen Anstieg der SET-Werte vom ersten zum zweiten Semester und dann einen allmählichen Abstieg bis zum vierten Semester.

Deutlich besser sind die Befunde für die Effekte ausführlicher *Beratungen*. Für Beratungen, die wiederholt durchgeführt wurden, sich auf Videofeedback stützten, Diskrepanzen zwischen Fremd- und Selbstwahrnehmung erfassten und Trainingsmöglichkeiten (Sprechtraining, Rollenspieltraining) anboten, ließen sich gute Effektstärken belegen (Aleamoni 1999; Dresel/Rindermann/Tinsner 2007; Penny/Coe 2004).

5.1 Nebenwirkungen

Zur Einschätzung der Nützlichkeit der SET gehört auch die Erörterung unerwünschter Nebenwirkungen. Auf Seiten der Studierenden gehört zu den Nebenwirkungen die Verstärkung einer Konsumhaltung (Anspruch, ohne Anstrengung gut unterhalten zu werden) sowie die Enttäuschung, dass sich trotz der SET-Erhebungen nichts ändert. Deutlich länger ist die Liste unerwünschter Nebenwirkungen bei Lehrkräften. Sie reicht von der Unterdrückung des Bedürfnisses nach Autonomie und Selbstbestimmung über die Verstärkung von Gefühlen wie Frustration, Entmutigung, Angst, Ärger, Demütigung, Spannung, Selbstzweifel, Scham und Depression bis hin zur dramatischsten Folge Suizid (Nichols/Berliner 2007; Rindermann/Kohler 2003). So kann man in der Washington Post vom 28.9.2010 (Strauss, 2010) einen Beitrag über den Suizid eines 39-jährigen Lehrers lesen, der zuvor auf einer publizierten Ratingskala für ihn enttäuschende Werte erhalten hatte.

Zu den unerwünschten Nebenwirkungen der SET gehören auch einige der Strategien, die Lehrkräfte einsetzen, um ihre Werte zu verbessern. In »How to improve your teaching evaluation scores without improving your teaching!« empfiehlt Trout (1997), dass man Studierende bei Unhöflichkeit, bei Fehlterminen, bei unvorbereiteten Referaten oder bei Arbeitsunwilligkeit niemals konfrontieren solle, dass man Ansprüche senken und schlechten Studierenden gute Noten geben solle, dass man niemals Notenerwartungen enttäuschen solle, dass man keine kontroversen Positionen vortragen solle, dass es besser sei, gute Nachrichten

zu verbreiten, dass man Studierende großzügig loben solle, dass man geschickt Mitleid provozieren könne und dass es weniger riskant sei, vor der SET-Bewertung Plätzchen oder Kuchen mitzubringen oder eine Feier zu organisieren, als die Evaluationsbögen zu manipulieren.

6. Fazit und Diskussion

Ein Fazit fällt angesichts der angeführten empirischen Studien und des historisch-politischen Hintergrunds nicht leicht. Die Durchführung von Lehrevaluationen ist eine alte studentische Forderung, und die Evaluation und Verbesserung der Lehre ist eine emanzipatorische Aufgabe im Sinne von mehr Transparenz, Selbstbestimmung und Demokratie. Die exemplarischen Darlegungen zu den Biasvariablen »erwartete Benotung« und »physische Attraktivität« legen jedoch eine starke Einschränkung der Validität der Ratings zur Lehrqualität nahe. Besonders gut aussehende und milde bewertende Dozent(inn)en erhalten häufiger unangemessen gute Evaluationsergebnisse. Dies ist problematisch, wenn Lehrevaluationswerte nicht nur der Rückmeldung dienen, sondern auch über Einstellungen, Weiterbeschäftigungen, Stipendien und Gehaltserhöhungen mitentscheiden. Lehr- und Unterrichtsevaluationen durch Schüler/Studierende mittels Ratingskalen messen recht valide die studentische Zufriedenheit (hierfür sprechen die Übereinstimmung zwischen den offiziellen SET- und den RateMyProfessors.com-Werten und die hohe Interraterreliabilität), aber nur eingeschränkt die Lehr- oder Unterrichtsqualität und nahezu gar nicht die Lehreffektivität. Die nahezu nicht vorhandene Beziehung zwischen den SET-Werten und dem durch Prä-Post-Messungen erhobenen Lernerfolg zeigt, dass das *Mögen* einer Lehrveranstaltung etwas anderes ist als das *Lernen* durch eine Lehrveranstaltung. Dabei hilft es wenig darauf zu insistieren, dass es nur auf das richtige Messinstrument ankomme und es von den jeweiligen Fragen abhänge, ob man studentische Zufriedenheit oder Lehreffektivität messe, da Haloeffekte eine separate Beurteilung erschweren. Die Kritik an der Validität der Lehrevaluation durch Ratingskalen ist offenbar weder »unbegründet« noch »dubios« und sie bezieht sich auch nicht auf »atypische Anekdoten« oder »populäre Mythen« (Marsh/Roche 2000).

Neben der zumindest eingeschränkten Validität ist die fehlende Nützlichkeit ein stark zu gewichtendes Argument. Im Kontext einer Einzelberatung oder eines Lehrertrainings kann eine SET-Messung ein sinnvolles Hilfsmittel sein. Eine

Rechtfertigung für den flächendeckenden Einsatz der SET ist dies jedoch nicht. Angesichts der teilweise erheblichen Nebenwirkungen der Lehrevaluation auf Dozentinnen und Dozenten stellt sich die Frage, ob man hier nicht einer Dialektik der Aufklärung gegenübersteht: Ein emanzipatorisch gedachtes Instrument wird auf der Seite der Lehrkräfte zu einer Quelle der Anpassung und Unterordnung. Wenn Lehrevaluationsergebnisse mitentscheiden über Einstellung, Bezahlung und Ansehen und wenn die Evaluationsergebnisse durch milde Benotungen und geschickte Anbiederungen verbessert werden können, ist zu befürchten, dass eher Dozent(inn)en gefördert werden, die nach allen Seiten konformistisch agieren. Verbesserungen der Lehrqualität sind durch Beratungs- und Trainingsprogramme zu erreichen, aber offenbar nicht durch ständig wiederholte Ratings.

Literatur

Aleamoni, Lawrence M. (1999): Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166.

Ambady, Nalini & Rosenthal, Robert (1993): Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431-441.

Aristoteles (2005): Politik. In: Weber-Fas, Rudolf (Hg.): Staatsdenker der Vormoderne. Tübingen (Mohr, Siebeck), S. 47-83 (original 335 v. Chr.).

Arthur, Winfred Jr.; Tubré, Travis; Paul, Don S. & Edens, Pamela S. (2003): Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology, 23*, 275-285.

Berk, Ronald A. (2005): Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17,* 48-62.

Berk, Ronald A. (2006): *Thirteen Strategies to Measure College Teaching. A Consumer's Guide to Rating Scale Construction, Assessment, and Decision Making for Faculty, Administrators, and Clinicians*. Sterlin, Virginia (Stylus).

Birnbaum, Michael H. (1998): A survey on faculty opinions concerning student evaluations of teaching. URL: http://psych.fullerton.edu/mbirnbaum/faculty3.htm (Stand: 09.10.2012).

Bokek-Cohen, Ya'arit & Davidowitz, Nitza (2008a): Beauty in the classroom: Are female students influenced by the physical appearance of their male

professors? *Journal of Education and Human Development, 2(1)*. UTR: http://www.scientificjournals.org/journals2008/articles/1371.pdf (Stand: 09.10..2012).

Bokek-Cohen, Ya'arit & Davidowitz, Nitza (2008b): Beauty in the classroom: Are students influenced by professors' appearance? *The Lookstein Center, 6(3)*. UTR: http://www.lookstein.org/online_journal_toc.php?id=1 (Stand: 09.10.22.05.2012).

Bonds-Raacke, Jennifer & Raacke, John D: (2007): The relationship between physical attractiveness of professors and students' ratings of professor quality. *Journal of Psychiatry, Psychology and Mental Health, 1(2)*, 1-7.

Boysen, Guy A. (2008): Revenge and student evaluation of teaching. *Teaching of Psychology, 35*, 218-222.

Carr, Rodney & Hagel, Pauline (2008): Students' evaluations of teaching quality and their unit online activity: An empirical investigation. In *Hello! Where are you in the landscape of educational technology? Proceedings ascilite Melbourne 2008*. URL: http://www.ascilite.org.au/conferences/melbourne08/procs/carr-r.pdf (Stand: 09.10.22.05.2012).

Cashin, William E. (1995): *Student ratings of teaching: The research revisited* (IDEA Paper Nr. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, John A. (2003): Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*, 495-518.

Chen, Yining & Hoshower, Leon B. (2003): Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*, 71-88.

Clayson, Dennis E. (2009): Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*, 16-30.

Cohen, Peter A. (1981): Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Coladarci, Theodore & Kornfield, Irv (2007): RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research & Evaluation, 12(6),* 1-15.

Daniel, Hans-Dieter (1998): Beiträge der empirischen Hochschulforschung zur Evaluierung von Forschung und Lehre: Hochschul-Ranking – Studentische Beurteilung von Lehrveranstaltungen – Selbststeuerung der Wissenschaft durch Peer-Review. In: Teichler, Ulrich; Daniel, Hans-Dieter & Enders, Jürgen (Hg.): *Brennpunkt Hochschule. Neuere Analysen zu Hochschule, Beruf und Gesellschaft*. Frankfurt am Main (Campus), S. 11-54.

Delucchi, Michael & Pelowski, Susan (2000): Liking or learning?: The effect of instructor likeability and student perceptions of learning on overall ratings of teaching ability. *Radical Pedagogy 2(2)*, 1-15. URL: http://radicalpedagogy.icaap.org/content/issue2_2/delpel.html (Stand: 09.10.22.05.2012).

Dresel, Markus; Rindermann, Heiner & Tinsner, Karen (2007): Beratung von Lehrenden auf der Grundlage studentischer Veranstaltungsbeurteilungen. In: Kluge, Annette & Schüler, Kerstin (Hg.), *Qualitätssicherung und –entwicklung an Hochschulen: Methoden und Ergebnisse*. Lengerich (Papst), S. 193-204.

Eiszler, Charles F. (2002): College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*, 483-501.

Feeley, Thomas H. (2002): Evidence of halo effects in student evaluations of communication instruction. *Communication Education 51*, 225-236.

Feldman, Kenneth A. (1989): Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education, 30*, 137-194.

Felton, James; Koper, Peter T.; Mitchell, John & Stinson, Michael (2008): Attractiveness, easiness, and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment and Evaluation in Higher Education, 33*, 45-61.

Felton, James; Mitchell, John & Stinson, Michael (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. *Assessment and Evaluation in Higher Education, 29*, 91–108.

Freng, Scott & Webber, David (2009): Turning up the heat on online teaching evaluations: Does »hotness« matter? *Teaching of Psychology, 36,* 189-193.

Greenwald, Anthony G. & Gillmore, Gerald M. (1997): No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-751.

Gurung, Regan A. R. & Vespia, Kristin M. (2007): Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology, 34*, 5-10.

Ha, Thao; Overbeek, Geertjan & Engels, Rutger C. M. E. (2010): Effects of attractiveness and social status on dating desire in heterosexual adolescents: An experimental study. *Archives of Sexual Behavior, 39*, 1063-1071.

Hadjistavropoulos, Thomas & Genest, Myles (1994): The underestimation of the role of physical attractiveness in dating preferences: Ignorance or Taboo? *Canadian Journal of Behavioural Science*, *26*, 298-318.

Hamermesh, D. S. & Biddle, Jeff E. (1994): Beauty and the labor market. *American Economic Review, 84*, 1174-1194.

Hamermesh, Daniel S. & Parker, Amy M. (2003): Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. National Bureau of Economic Research, Inc. *NBER Working Papers: 9853*. Economics of Education Review (Forthcoming). URL: http://www.math.wisc.edu/~miller/old/Teachingbeauty.pdf (Stand: 09.10.22.05.2012)

Hassebrauck, Manfred & Niketta, Reiner (1993): *Physische Attraktivität*. Göttingen (Hogrefe).

Henss, Ronald (1998): *Gesicht und Persönlichkeitseindruck*. Göttingen (Hogrefe).

Hinz, Arnold (2011): Attitudes of German Teachers and Students towards Public Online Ratings of Teaching Quality. *Electronic Journal of Research in Educational Psychology, 9(2),* 745-764.

Hultman, Magnus & Oghazi, Pejvak (2008): Good looks – good courses: The link between physical attractiveness and perceived performance in higher educational services. In: ANZMAC conference papers 2007. Dundedin, New Zeeland (IAHR), S. 2588-2597. URL: http://pure.ltu.se/portal/files/1884887/artikel.pdf (Stand 09.10.22.5.2012)

Kember, David; Leung, Doris Y. P. & Kwan, K. P. (2002): Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27*, 411-425.

Kindred, Jeannette & Mohammed, Shaheed N. (2005): «He will crush you like an academic Ninja!". Exploring teacher ratings on Ratemyprofessors.com. *Journal of Computer-Mediated Communication, 10(3)*, article 9. URL: http://jcmc.indiana.edu/vol10/issue3/kindred.html (Stand 09.10.2012).

Klein, Markus & Rosar, Ulrich (2006): Das Auge hört mit! Der Einfluss der physischen Attraktivität des Lehrpersonals auf die studentische Evaluation von Lehrveranstaltungen – eine empirische Analyse am Beispiel der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln. *Zeitschrift für Soziologie, 35,* 305-316.

Kulik, James A. (2001): Student ratings: Validity, utility and controversy. In: Theall, Michael; Abrami, Philip C. & Mets, Lisa A. (Hg.): *The student ratings debate: Are they valid? How can we best use them?* San Francisco (Jossey-Bass), S. 9-26.

Lang, Jonas W. B., & Kersting, Martin (2007): Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science, 35*, 187-205.

Luo, Shanhong & Zhang, Guangjian (2009): What leads to romantic attraction: Similarity, reciprocity, security, or beauty? Evidence from a speed-dating study. *Journal of Personality, 77*, 933-964.

Marsh, Herbert W. (2007): Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In: Perry, Raymond P. & Smart, John C. (Hg.): *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. New York (Springer), S. 319-383).

Marsh, Herbert W. & Ball, Samuel (1989): The peer review process uses to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education, 57*, 151-169.

Marsh, Herbert W. & Hocevar, Dennis (1991): Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*, 303-314.

Marsh, Herbert W. & Roche, Lawrence A. (1997): Making students' evaluations of teaching effectiveness effective. The critical issues of validity, bias and utility. *American Psychologist, 52*, 1187-1197.

Marsh, Herbert W. & Roche, Lawrence A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, of innocent bystanders? *Journal of Educational Psychology, 92,* 202-228.

Mason, Kevin H., Edwards, Robert R. & Roach, David W. (2002): Student evaluation of instructors: A measure of teaching effectiveness or of something else? *Journal of Business Administration Online, 1(2)*. URL: http://www.atu.edu/business/jbao/Fall2002/mason_edwards_roach.pdf (Stand 09.10.2012).

McKeachie, Wilbert J. (1997): Student ratings. The validity of use. *American Psychologist, 52*, 1218-1225.

Meinefeld, Werner (2010): Online-Befragungen im Kontext von Lehrevaluationen – praktisch und unzuverlässig. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*, 297-315.

Naftulin, Donald H.; Ware, John E. & Donnelly, Frank A. (1973): The Doctor Fox lecture: a paradigm of educational seduction. *Journal of Medical Education, 48,* 630-635.

Nichols, Sharon L. & Berliner, David C. (2007): *Collateral Damage. How high-stakes testing corrupts America's schools.* Cambridge (Harvard Education Press).

Ory, John C. (2001): Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 87*, 3-15.

Penny, Angela R. & Coe, Robert (2004): Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research, 74*, 215-253.

Renz, Ulrich (2006): *Schönheit. Eine Wissenschaft für sich*. Berlin (Berlin).

Rindermann, Heiner (1996): Zur Qualität studentischer Lehrveranstaltungsevaluationen: Eine Antwort auf Kritik an der Lehrevaluation. *Zeitschrift für Pädagogische Psychologie, 10*, 129-145.

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

Rindermann, Heiner & Kohler, Jürgen (2003): Lässt sich die Lehrqualität durch Evaluation und Beratung verbessern? *Psychologie in Erziehung und Unterricht, 50,* 71-85.

Riniolo, Todd C.; Johnson, Katherine C.; Sherman, Tracy R. & Misso, Julie A. (2006): Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology, 133*, 19-35.

Rosar, Ulrich & Klein, Markus (2009): Mein(schöner)Prof.de. Die physische Attraktivität des akademischen Lehrpersonals und ihr Einfluss auf die Ergebnisse studentischer Lehrevaluationen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 61*, 621-645.

Rosar, Ulrich & Klein, Markus (2010): Mein (nach-wie-vor-schöner)Prof.de. Einige klärende Bemerkungen zu einigen kritischen Anmerkungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*, 327-342.

Smalzried, Newell T. & Remmers, Hermann H. (1943): A factor analysis of the Purdue Rating Scale for Instructors. *Journal of Educational Psychology, 34*, 363-367.

Stark-Wroblewski, Kimberly; Ahlering, Robert F. & Brill, Flannery M. (2007): Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education, 32*, 403-415.

Strauss, Valerie (2010): About the suicide of an L.A. teacher. *The Washington Post*, September 28, 2010. URL: http://voices.washingtonpost.com/answer-sheet/teachers/about-the-suicice-of-an-la-tea.html (Stand: 09.10.2012).

Süßmuth, Bernd (2006): Beauty in the classroom: Are German students less blinded? Putative pedagogical productivity due to professors' pulchritude: Peculiar or pervasive. *Applied Economics, 38*, 231-238.

Trout, Paul (1997): How to improve your teaching evaluation scores without improving your teaching! *The Montana Professor, 7(3)*, 17-22. URL: http://mtprof.msun.edu/Fall1997/HOWTORAI.html (Stand: 09.10.2012)

Wachtel, Howard K. (1998): Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23,* 191-211.

Williams, Wendy M. & Ceci, Stephen J. (1997): How'm I doing? Problems with student ratings of instructors and courses. *Change, September/October*, 13-23.

Wissenschaftsrat (2008): Empfehlungen zur Qualitätsverbesserung von Lehre und Studium. Berlin. URL: http://www.wissenschaftsrat.de/download/archiv/8639-08.pdf (Stand: 09.10.2012)

Wolbring, Tobias (2010): Attraktivität, Geschlecht und Lehrveranstaltungsevaluation. Eine Replikationsstudie zu den Befunden von Hamermesh und Parker (2005) und Klein und Rosar (2006) mit Hilfe von Individualdaten. *Zeitschrift für Evaluation, 9*, 29-48.

Wolbring, Tobias & Hellmann, Anja (2010): Attraktivität, Reziprozität und Lehrveranstaltungsevaluation. Eine experimentelle Untersuchung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62,* 707-730.

Abbildungsverzeichnis

Über den Autor

Arnold Hinz

PD Dr., Institut für Pädagogische Psychologie und Soziologie
Pädagogische Hochschule Ludwigsburg
Reuteallee 46
D-71634 Ludwigsburg

E-Mail: hinz@ph-ludwigsburg.de

Web: http://www.ph-ludwigsburg.de/211.html